Modeling and Estimation of Motion Over Manifolds with Motion Capture Data

Nathan R Powell

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mechanical Engineering

Andrew J. Kurdila, Chair
Steve C. Southward
Alexander Leonessa
Craig A. Woolsey

September 22, 2022
Blacksburg, Virginia

Keywords: Learning Theory, Estimation, Motion Capture, Dynamical Systems

Copyright 2022, Nathan R Powell
Modeling and Estimation of Motion Over Manifolds with Motion Capture Data

Nathan R Powell

(ABSTRACT)

Modeling the dynamics of complex multibody systems, such as those representing the motion of animals, can be accomplished through well-established geometric methods. In these formulations, motions take values in certain types of smooth manifolds which are coordinate-free and intrinsic. However, the dimension of the full configuration manifold can be large. The first study in this dissertation aims to build low-dimensional models from motion capture data. This study also expands on the so-called learning problem from statistical learning theory over Euclidean spaces to estimating functions over manifolds. Experimental results are presented for estimating reptilian motion using motion capture data. The second study in this dissertation utilizes reproducing kernel Hilbert space (RKHS) formulations and Koopman theory, to achieve some of the advantages of learning theory for IID discrete systems to estimates generated over dynamical systems. Specifically, rates of convergence are determined for estimates generated via extended dynamic mode decomposition (EDMD) by relating them to estimates generated by distribution-free learning theory. Some analytical examples illustrate the qualitative behavior of the estimates. Additionally, a examination of the numerical stability of the estimates is also provided in this study. The approximation methods are then implemented to estimate forward kinematics using motion capture data of a human running along a treadmill. The final study of this dissertation contains an examination of the continuous time regression problem over subsets and manifolds. Rates of convergence are determined using a new notion of Persistency of Excitation over flows of
manifolds. For practical considerations, two approximation methods of the exact solution to the continuous regression problem are introduced. Characteristics of these approximation methods are analyzed using numerical simulations. Implementations of the approximation schemes are also performed on experimentally collected motion capture data.
Modeling and Estimation of Motion Over Manifolds with Motion Capture Data

Nathan R Powell

(GENERAL AUDIENCE ABSTRACT)

Modeling the dynamics of complex multibody systems, such as those representing the motion of animals, can be accomplished through well-established geometric methods. However, many real-world systems, including those representing animal motion, are difficult to model from first principles. Machine learning, on the other hand, has proven to be extremely powerful in its ability to leverage ”big data” to generate estimates from typically independent and identically distributed (IID) data. This dissertation expands on the so-called learning problem from statistical learning theory over Euclidean spaces to those over manifolds. This dissertation consists of three studies, the first of which aims to build low-dimensional models from motion capture data. Using the distribution-free learning theory, estimates discussed in this dissertation minimize a proxy of the expected error, which cannot be calculated in closed form. This dissertation also includes a study into approximations of the so-called Koopman operator. This study determined that the rate of convergence of the estimate to the true operator depends on the reduced dimensionality of the embedded submanifold in the high-dimensional ambient input space. While most of the current work on machine learning focuses on cases where the samples used for learning or regression are generated from an IID, stochastic, discrete measurement process, this dissertation also contains a study of the regression problem in continuous time over subsets and manifolds. Additionally, two approximation methods of the exact solution to the continuous regression problem are introduced. Each of the aforementioned studies also includes several analytical
results to illustrate the qualitative behavior of the approximations and, in each study, im-
plemetations of the estimation schemes are performed on experimentally collected motion
capture data.
Dedication

To my parents, Richard and Ellen.
Acknowledgments

The completion of my degree would not have been possible without the help of Dr. Andrew Kurdila. His instrumental guidance and insight shaped both my current work as well as my career plans moving forward. I would also like to extend my deepest thanks to my doctoral committee members, Dr. Alexander Leonessa, Dr. Steve Southward, and Dr. Craig Woolsey. Their valued critiques greatly improved my current work including this dissertation, and their positive feedback provided great motivation for my future research plans. I would also like to thank my former lab mates Dr. Sai Tej Paruchuri, and Dr. Jia Guo for their constant advice, especially early on when I needed direction. Additionally, I want to acknowledge my former lab mate, Roger Liu, for his substantial assistance in collecting and processing the data used throughout my research. Lastly and most importantly, I want to express my most sincere gratitude to my wife, Macy Powell. I could not have done this without her continuous support throughout my graduate career.
Contents

List of Figures xiii

List of Tables xx

1 Introduction 1

1.1 Background and Review of Literature 2

1.1.1 Animal Motion Modeling and Bio-Inspired Robotics 2

1.1.2 Regression and Distribution-Free Learning in Euclidean Spaces 3

1.1.3 Manifold Learning 5

1.1.4 Estimation Methods over Manifolds 7

1.1.5 Koopman Theory and Dynamical Systems 8

1.2 Overview of this dissertation 11

1.2.1 Learning Theory to Estimate Animal Motion Submanifolds 11

1.2.2 Koopman Theory to Estimate Functions over Manifolds of Animal Motion 14

1.2.3 Kernel Methods for Regression in Continuous Time over Subsets and Manifolds 16

Bibliography 19
# Distribution-Free Learning Theory for Approximating Submanifolds from Reptile Motion Capture Data

## 2.1 Introduction

## 2.2 Problem Formulation

- 2.2.1 Kinematics and Inherent Geometry
- 2.2.2 The Distribution-Free Learning Problem on Manifolds

## 2.3 Function Spaces on Manifolds

- 2.3.1 Approximants in $L^2_{\mu}(S)$
- 2.3.2 Approximants in $\mathcal{H}^S$

## 2.4 Empirical Risk Minimization over Manifolds

## 2.5 Experiments and Numerical Examples

- 2.5.1 Experimental Setup
- 2.5.2 Data collection procedure
- 2.5.3 Numerical Results
- 2.5.4 Statistical Estimates of Error in Stereo-Triangulation
- 2.5.5 Variations in Bone Lengths and Skin Artifacts
- 2.5.6 Computational Cost

## 2.6 Future Work

## 2.7 Appendix

## 2.8 Lebesgue and Sobolev Function Spaces
3 Koopman Methods for Estimation of Animal Motions over Unknown, Regularly Embedded Submanifolds

3.1 Introduction

3.1.1 Example 1

3.1.2 Example 2

3.1.3 Related Research

3.1.4 Deterministic Koopman Operators

3.1.5 Direct Approximation Theorems

3.1.6 Approximations of Koopman Operators

3.1.7 Overview of New Results

3.2 Background Theory

3.2.1 Symbols and Nomenclature

3.2.2 Reproducing Kernel Hilbert Spaces

3.2.3 Koopman Theory and EDMD Algorithm

3.3 Problem Formulation

3.3.1 Distribution-Free Learning in Euclidean Spaces
3.3.2 Distribution-Free Learning over an Unknown Manifold .......... 106
3.3.3 Projection for Regression in the Pullback Space .................. 108
3.3.4 Data-Driven Approximations ........................................ 109

3.4 Examples and Numerical Results ..................................... 112
3.4.1 Example: The Pendulum ............................................. 112
3.4.2 Example: Human Kinematics Study ................................. 123

3.5 Conclusion ................................................................. 124

Bibliography ................................................................. 127

4 Kernel Methods for Regression in Continuous Time over Subsets and Manifolds ................................................................. 134

4.1 Introduction ........................................................................ 135
4.1.1 Motivation ..................................................................... 135
4.1.2 Problem Description ...................................................... 136
4.1.3 Summary of New Results .............................................. 139
4.1.4 Notation, Symbols, and Background ............................... 141

4.2 Regression in Continuous Time .......................................... 142
4.2.1 The Offline Optimal Regression Estimate in an RKHS ........ 142
4.2.2 Galerkin Approximations of the Regression Estimate ........ 148

4.3 Persistency of Excitation (PE) in Native Spaces ................. 149
4.3.1 A New Persistence of Excitation Condition .................... 150
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Abstract illustration of the relationship between the known manifold, $S$, and the unknown submanifold, $Q$, where the observed motion occurs.</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Illustrating our uniform refinement of the manifold $S$. As the refinement level $\ell$ increases, which is illustrated as moving down the figure, the dimension of the space, $N$, which is represented by characteristic functions defined over the partitions represented by the connected black lines, increases by a factor of 2.</td>
<td>40</td>
</tr>
<tr>
<td>2.3</td>
<td>The experimental setup of the motion capture recordings. Three Photron FASTCAM cameras face a central container housing an inclined wooden plank that the specimen crawls up. Because high shutter speeds diminish the brightness of the recording, floodlights were used to illuminate the scene.</td>
<td>48</td>
</tr>
<tr>
<td>2.4</td>
<td>Coordinates of the trajectories in the inertial frame of the markers placed on a specimen. Note that while the body coordinates move smoothly, the markers on the shoulder, elbow, and wrist were momentarily occluded from either one or both of the cameras leading to small sharp edges in the shoulder, elbow, and wrist trajectories.</td>
<td>50</td>
</tr>
<tr>
<td>2.5</td>
<td>The body frame coordinate system defined on the specimen. Note that $b_1$ points in the anterior direction towards the head, $b_2$ points medially left, and $b_3$ points ventrally towards the stomach of the specimen. Note that all units are in mm.</td>
<td>51</td>
</tr>
</tbody>
</table>
2.6  The articulated skeleton built from reconstructing the marker coordinates into body-fixed coordinates. The frame is parallel to anatomically defined planes that bisect the body. ................................................................. 53

2.7  The elbow motion trajectory over several gaits. Note that the motion of the elbow and wrist follow a periodic trajectory in the body fixed frame. Note that all units are in mm. ................................................................. 54

2.8  Comparison of two methods of approximation of the unknown map \( \gamma : S \to Q \) for a fixed number of samples, \( M = 366 \) and a dimension of the approximant space, \( N = 32 \). This corresponds to a uniform partitioning depth \( \ell = 4 \) The blue dashed line represents the approximation of the mapping of the wrist coordinates \( \gamma^{(4)} = \zeta^1_w \) corresponding to the \( b_1 \) vector using the characteristic functions \( 1_{S_{\ell,k}} \) in the partition \( S_{\ell} \), while the solid red line represents regression approximation using a kernel function basis in \( H^S_{N} \). The known low dimensional manifold \( S \) is depicted as a black solid line and the centers of each of the approximations are represented by the crosses on \( S \). .......................... 56

2.9  The approximations \( \gamma_{N,M}^{(j)} \) for various dimension, \( N \), of the approximant space. \( N = 8 \) is used for (a-b) \( N = 32 \) for (c-d), and \( N = 64 \) for (e-f). The estimates from \( H^S_N \) are on the left (a,c,e) while estimates from \( H^S_N \) are on the right (b,d,f). The estimates are mapped over the corresponding manifold given by the black circle and the observations are plotted as the red dots. It is evident the global behavior is consistent given the space dimension, \( N \). Note that larger error due to sample variance is evident with the larger dimension, \( N = 64 \) ................................................................. 59
2.10 The approximations $\gamma_{N,M}^{(j)}$ for $j = 5$ corresponding to the wrist coordinate along the $b_2$ vector using $M = 254$ and $N = 16$ for various values of the kernel hyperparameter, $\beta$. The dimension of the approximant space $N$ is demonstrated by the centers marked as crosses on the manifold labeled $S$. For larger values of $\beta$, the evaluation of the kernel basis decays rapidly away from the kernel centers resulting in highly oscillatory approximations. The empirical risk minimization over $H_N^S$ has no corresponding hyperparameter.

2.11 The approximates of the observation of the elbow coordinate of the $b_1$ vector using 16 centers, $\gamma_{N,M}^{(j)}$, from four male specimens using a basis of (a) approximation partitions in $H_N^S$ and (b) exponential kernels in $H_N^S$. Each approximation generated from an individual specimen is designated by the dashed lines while the average is designated by the solid line in both graphs. Note that the parameter $M$ differs per specimen, as well as the average approximation, because each approximation uses a different number of samples.

2.12 The estimates of the observation of the elbow coordinate of the $b_1$ vector using 16 centers $\gamma_{N,M}^{(j)}$ over a number of specimens of different sexes. The solid red line represents the female specimens while the dashed blue line represents the males. Using a basis of approximation partitions in $H_N^S$. It would appear that the average male specimen has a larger, albeit slight, range of motion in the $b_1$ direction.

2.13 Coordinates of the approximations of the motion corresponding to (a) the elbow coordinates, and (b) the wrist coordinates in the body frame using $N = 16$ basis functions. For each graph, one can see that the mappings act as an interpolating closed curve over the sample data. Note that all units are in mm.
2.14 The approximations of the elbow trajectory using (a) $\ell = 3$ (b) $\ell = 4$, (c) $\ell = 5$, and (d) $\ell = 6$ refinement depths of the space $S$. Note that the dimension of the space increases with increased refinement (a) $N = 8$ (b) $N = 16$, (c) $N = 32$, and (d) $N = 64$. The samples measured in the body frame are also depicted. Note the oscillatory behavior due to the increase in the variance for a larger number of basis functions. All units are in mm.

2.15 The approximations of the elbow motion using (a) $N = 8$, (b) $N = 16$, (c) $N = 32$, and (d) $N = 64$ centers with an exponential kernel function basis. The samples measured in the body frame are also depicted. Notice that the behavior does not change drastically while increasing the basis functions. This seems to indicate that the submanifold is well-approximated given a small number of basis functions. Note like the characteristic function basis, the increase in variation due to the increase in the second term of our bound, which we have labeled as the variance. All units are in mm.

2.16 Measured limb lengths compared to an ideal rigid limb length over a recording session for one specimen in the inertial frame for (a) the upper arm, (b) the forearm, (c) the thigh and (d) the shin. In an ideal case, the distances between markers on these limbs should remain fixed.

3.1 An example of the $U_fG$ approximation in red of $G$, represented by the light green surface. Using $M$ samples collected over the composition $U_fG$ the finite dimensional approximation $U_f^M G$ is generated from the RKHS represented by the mesh blue surface. Minimizing the empirical risk over the samples corresponds to interpolating the function over the samples as represented by the red dashed line.
3.2 The estimate $U^M_jG$ using the Matern-Sobolev kernel (a) with $\beta = 1$ and the estimate using the compactly supported kernel $R_2^{CS}$ (b). Both estimates decay in magnitude away from the manifold. However, the estimate (a) is still nonzero everywhere while the estimate (b) itself quickly decays to zero away from the manifold as a consequence of its bases being compactly supported.

3.3 The approximations $U^M_jG$ for various dimension, $\beta$, of the Matern kernel with $\beta = 5$ (a), $\beta = 1$ (b), $\beta = 0.5$ (c), $\beta = 0.2$ (d). Regardless of the hyperparameter selected, it is clear that each estimate is still determined by an algorithm which fits the estimate to the empirical samples collected along the red line over the manifold $Q$.

3.4 Three cases of the condition number of the Kernel Matrix $K$ given the minimum separation distance of the collected centers from our simulated data using kernels belonging to specific smoothness spaces. In this example, the samples are roughly uniformly distributed and the separation distance $q_X$ is approximately decrease along with the fill distance $h_{\Xi_M, Q}$. Note that as the separation distance decreases the condition number increase exponentially. Additionally, the condition number of the Kernel matrix associated with higher order kernels is larger for equivalent separation distances.

3.5 This figures demonstrates the relationship between the fill distance and the condition number for the Matern-Sobolev kernel $R^M S_2$ for various values of the hyperparameter $\beta$. The smaller spread of the kernel function corresponding to smaller $\beta$ lowers the condition number for estimates of similar fill distances. However, the rate of increase given by the slope is determined by the kernel type as is, therefore, relatively the same for each curve in this example.
3.6 A figure illustrating the behavior of the minimum eigenvalue of the Kernel Matrix \( K \) given the minimum separation distance of the collected centers from the simulate data. For each curve, the separation, \( q_X \) is decreased between only two centers while keeping a constant fill distance \( h_{\Xi,M,Q} \). Note that, just like the experimental data the condition number increases exponentially as the separation distance decreases. 

3.7 The rate of convergence of the estimate of the Koopman approximation. The rate of convergence is bounded by the fill distance represented by the dashed line above the curve.

3.8 An illustration of the inputs \( \theta^{(1)} \) and \( \theta^{(2)} \), which roughly correspond to hip and knee flexion. These input variables can be mapped to measured marker coordinates placed on joints such as the knee or ankle.

3.9 The estimates of the Koopman operator acting on observable functions \( G^{(1)} \) and \( G^{(2)} \) corresponding to the \( b^{(1)} \) coordinate (a) and \( b^{(2)} \) coordinate (b) of the ankle over the submanifold \( Q \) of the input space given by the leg coordinates \( \theta^{(1)} \) and \( \theta^{(2)} \).

4.1 (Top) An illustration of regression in continuous time. The orbit \( \Gamma(\phi_0) = \bigcup_{\tau \geq 0} \phi(\tau) \) is a subset of the state space \( X \). The output is determined by a function \( G \) represented by the green surface over \( X \). (Bottom) An illustration of the estimate \( \Pi_{\nu} G \) represented by the blue mesh is generated by a subspace \( V \) of \( \mathcal{H} \). Here the subspace consists of kernel functions illustrated by the “bumps” with centers at various points along the orbit. In this figure, one center is represented by \( \phi(\tau) \).
4.2 The dynamics of the Lorenz system for several initial conditions (a) in three-dimensions (b) projected onto the $x - y$ plane.

4.3 An illustration of the kernel estimate $\hat{g}_N(t, \cdot)$ of $G$ for $t = 200$ seconds defined over the orbit $\phi(t)$ starting at some initial conditions $\phi_0 = \{1, 1, 1\}^T$. The orbit is generated by projecting the Lorenz system dynamics onto the x-y plane. The output $G(\phi(t))$ for $t \geq 0$ is represented by the red curve hovering over the dynamics of the input orbit $\phi(t)$. The estimate $\hat{g}_N(t, \cdot)$ represented by the colored mesh minimizes its error from the projection of the true function $G$ (the curved surface) over the orbit $\phi(t)$.

4.4 An illustration of the various approximations from Equation 4.8 over an orbit after spanning different amounts of time from the same system dynamics and initial conditions. It is evident that the error $\|\Pi_N G - \hat{g}_N(t, \cdot)\|$ converges to zero over the domain of attraction as more time is passed and more samples are collected.

4.5 Estimates $\hat{g}_N(t, \cdot)$ of the function $G$ for different values of $\gamma$ using the same time-span, initial condition, and underlying dynamics. Notice that for larger $\gamma = 20$, the estimate has a slower convergence over the domain of attraction. While estimates using smaller $\gamma$, may have reduced error over the orbit, the estimate can potentially over-fit noisy training data.

4.6 An illustration of the evolution approximation from Equation 4.10 over an orbit as time $t \to \infty$. The error $\|\Pi_N G - \hat{g}_N(t, \cdot)\|$ converges over the domain of attraction as more time is passed and more samples are collected.
4.7 Demonstration of estimates $\hat{g}_N(t, \cdot)$ of the function $G$ for different values of $\gamma$ using the same time-span, initial condition, and underlying dynamics. Notice that the larger $\gamma = 20$ has a slower convergence over the domain of attraction. For smaller $\gamma$, such the estimate generated using $\gamma = 0.001$, the estimate suffers from less penalty on the regularization of the estimate. However, the estimate can over-fit noisy training data.

4.8 An illustration of the inputs $\theta^{(1)}$ and $\theta^{(2)}$, which roughly correspond to hip and knee flexion respectively. These input variables can be mapped to measured marker coordinates placed on joints such as the knee or ankle.

4.9 The estimates $\hat{g}_N(t, \cdot)$ of the ankle coordinate in the $b^{(2)}$ direction over the input space given by the angles $\theta^{(1)}$ and $\theta^{(2)}$ associated with hip and knee flexion respectively using (a) Approximation Method (1), and (b) Approximation Method (2). It must be noted that the first Approximation method shows relatively small error for the given sample size compared to Method (2).
List of Tables

2.1 Measured Limb Lengths ....................................................... 67
Chapter 1

Introduction

Modeling the dynamics of complex multibody systems, such as those representing the motion of animals, can be accomplished through well-established geometric methods such as [1, 2]. In these formulations, motions take values in certain types of smooth manifolds, and forward kinematic maps are defined over these manifolds. One attractive feature of these formulations is that such formulations are coordinate-free and intrinsic. [1] For many animal motion studies, however, the dimension of the full configuration manifold can be large. The study of human motion, bat motion, or the flight of birds are just a few such cases [3, 4]. For these animals, as well as others, it is often important to be able to identify submanifolds that provide accurate representations of particular motion regimes. In most animal motion studies, many assumptions are made about the underlying constraints and generalized coordinates of the observed motion. Ideal methods would work in as general a case as possible and minimize the assumptions about the geometry of the system. Furthermore, these methods should give some rigorous degree of confidence in the low-dimensional representation. A fundamental question that is both motivated and directly addressed in this dissertation is how to come up with a principled, low-dimensional model of a complex animal motion.

Additionally, this dissertation expands on the so-called learning problem from statistical learning theory in order to build estimates of functions over manifolds. Many real-world systems, including those representing animal motion, are difficult to model from first principles. Machine learning, on the other hand, has proven to be extremely powerful in its ability
to leverage "big data" to generate estimates from typically independent and identically distributed (IID) data. As described in [5, 6, 7], it is standard that the learning problem in Euclidean spaces is often used to motivate and to study adaptive estimation and control theory in Euclidean spaces. With the motivation from classic adaptive control texts such as [5, 6, 7], this dissertation investigates reproducing kernel Hilbert space (RKHS) formulations in the estimation problem to achieve some of the advantages of learning theory for IID discrete systems. At its core, this defense is concerned with learning kinematic models from motion capture data, and building estimates of functions defined over these models, which has been the focus of the recent works in [8], [9] and [10].

1.1 Background and Review of Literature

1.1.1 Animal Motion Modeling and Bio-Inspired Robotics

In many studies of animal motion, assumptions are made about the configuration space in order to simplify the kinematics [3, 11, 12, 13, 14], which can be limiting if the underlying generalized coordinates are difficult to determine. The goal of identifying the configuration manifold is critical in the field of bio-inspired robotics where researchers seek to create robots that emulate some observed animal motion. Many motion studies involve mechanisms that roughly imitate a biological counterpart such as a joint [15]. Many of these mechanisms are designed after trial-and-error attempts from experienced designers [16, 17], and typically involve sophisticated control algorithms based on the complex geometry of the designs themselves [18, 19, 20]. Many studies assume the configuration space as a collection of rigid bodies defined by bones that are interconnected by ideal joints [21, 22, 23, 24, 25, 26, 27, 28, 29]. The main difficulty stems from determining the relationship between the relative motion of
the joints and appendages (the shape of the body) and the global position of the body. Several studies have attempted to identify kinematic models through data-driven linear regression on discrete samples of the motion itself [30, 31]. As mentioned previously, the primary advantage of data-driven regression for motion modeling is the ability to use large samples of empirical data to model systems rather than building them from first-principles that are often difficult to determine. However, one well known difficulty with linear regression is that, regardless of the dimensionality of the system or the sample size collected, these techniques will not converge to the desired function if the model or function we want to estimate is nonlinear. In other words, these techniques will only work if the underlying function lies in the space of linear functions of the inputs. This provides the primary motivation for the non-parametric regression technique of distribution-free learning used throughout this dissertation; this method allows for estimates which better represent a much more general class of functions. Additionally, the theoretical results in this dissertation examine the rates of convergence of the estimated models, which have not been done before in the context of animal motion studies.

1.1.2 Regression and Distribution-Free Learning in Euclidean Spaces

The distribution-free learning problem is a general problem that has been studied in a wide variety of contexts, most commonly for mappings between Euclidean spaces [32]. This dissertation lifts learning over some known compact subset such as \( \Omega := [a, b]^d \subset \mathbb{R}^d \), \( a, b \in \mathbb{R} \) to an unknown manifold \( Q \). There are several features of this problem that make it challenging in comparison to the classical problem over \( \mathbb{R}^d \). When the distribution-free learning problem is applied to Euclidean space, we are faced with building approximations of some unknown function over \( \Omega \) that takes values in \( Y = \mathbb{R}^n \). We are given samples \( \{(x_i, y_i)\}_{i=1}^M \subset X \times Y \).
Here $X$ is the input space and $Y$ is the output space, and we can view the samples as perhaps noisy measurements of the functional relationship $y = G(x)$ where $G : X \to Y$ is unknown. It is usually assumed that the samples are IID, with $\mu$ the probability distribution over $X \times Y$ that generates the samples. In the distribution-free learning problem the probability distribution $\mu$ is unknown. The aim of learning theory is to build an approximation of the unknown function using the samples that makes some error measure small, and here we use the familiar quadratic error for functions $g : X \to Y$,

$$E_\mu(g) := \int_X \int_Y \|y - g(x)\|^2_\mu \mu(dx, dy),$$

(1.1)

where again we interpret $y = G(x)$ in the ideal case. A great deal is known about the structure of this particular problem in the space $L^2_\mu(\mathbb{R}^d)$, or even in $L^2_\mu(\Omega)$ with $\Omega$ a compact subset of $X$ like $\Omega = [a, b]^d$. The minimization of the risk $E_\mu$ above is equivalent to minimizing the expression

$$E_\mu(g) := \int_X \|g(x) - G_\mu(x)\|^2_\mu \mu_X(x) + E_\mu(g_\mu)$$

(1.2)

where the measure $\mu$ is written as $\mu(dx, dy) = \mu_Y(dy|x)\mu(dx)$, $\mu_Y(dy|x)$ is the conditional probability over $Y$ given $x \in X$, $\mu_X(dx)$ is the marginal measure of $\mu$ over $X$, and $G_\mu(x) := \int_Y y\mu_Y(dy|x)$ is the regressor function. Thus, the regressor $G_\mu$ is the optimal minimizer, but it generally cannot be computed from this closed form expression since $\mu_Y$ is generally unknown. In practice, the ideal error measure above is replaced with the discrete risk functional

$$E_M(g) := \frac{1}{M} \sum_{i=1}^M \|y_i - g(x_i)\|^2_Y$$

(1.3)
1.1. Background and Review of Literature

that depends on the samples \( \{(x_i, y_i)\}_{i=1}^{M} \). This is the well-known method of empirical risk minimization. Note that the discrete error above can be evaluated for any given \( g \). When minimizers \( G_{M,N} = \arg \inf \{ E_M(g) \mid g \in \mathcal{H}_N \} \) are calculated for some \( N \) dimensional space \( \mathcal{H}_N \) of approximants, it is then possible to address in what sense \( G_{M,N} \to G_\mu \) as the number of samples \( M \) and/or the number of functions used \( N \) approach \( \infty \). The theory for such approximations is mature, and a summary of the state art in these cases can be found [32, 33, 34].

1.1.3 Manifold Learning

In most learning problems, the inputs and outputs come from high-dimensional Euclidean spaces. In regression, the samples needed for an accurate estimate increase exponentially with increases in the dimension of the system [35]. These high-dimensional vector spaces often contain concentrations of data, and, in many cases, we can represent these concentrations by lower dimensional manifolds, \( \mathcal{M} \). Representing the data through these manifolds or utilizing their lower dimension may aid us in addressing the curse of dimensionality.

Manifold learning refers to a class of algorithms with the goal of recovering a low-dimensional manifold embedded in a high-dimensional space. Specifically, the manifold is determined through an embedding function, often denoted \( \psi \). The majority of dimensional reduction algorithms incorporate spectral decomposition in some shape or form. These so-called spectral embedding algorithms of manifold learning can primarily be divided into linear and nonlinear methods. There are a vast number of linear manifold learning that are implemented by projecting the sample data to a reduced set of linearly transformed variables. The most popular projection method is principal component analysis (PCA). PCA focuses on determining linear subspaces, commonly referred to as principle components, that best represent some
measured data. These best fit lines are oriented so that they are parallel to the direction of largest variance of the data.

There are also a vast number of nonlinear manifold learning algorithms such as Isomaps [36], Local Linear Embedding (LLE) [37], Laplacian Eigenmaps [38], Diffusion Maps [39], and Hessian Eigenmaps [40]. Each of these algorithms builds off the assumption that the collected finite random sample of data points, \( \{x_i\}_{i=1}^M \) from the high-dimensional vector space \( X \) actually come from the embedding \( \psi: \mathcal{M} \to X \). The embedding is determined so that each \( x_i \in X \) can be best represented by a point \( p_i \in \mathcal{M} \) where \( x_i = \psi(p_i) \). Nonlinear learning algorithms such as LLE, and Laplacian Eigenmaps and many others utilize assumptions about the behavior of the manifold at a local level. For instance, LLE determines the best lower-dimensional projection which preserves distances within local neighborhoods. Other nonlinear manifold learning methods incorporate PCA on an augmented data matrix that are nonlinear transformations of the original data [41] with kernel PCA [42] as one of the most popular of these variants. A detailed survey of different manifold learning techniques can be found in [43].

As mentioned in [44, 45] one of the primary concerns with most manifold learning techniques comes from the difficulty in learning the manifold from the noisy process of sample collection. These algorithms often demonstrate their effectiveness by comparing the learned estimate to the true function without discussing the collection of potentially bad samples. In this dissertation, distribution-free learning theory over manifolds addresses this issue by determining bounds given a potentially bad collection of samples.
1.1.4 Estimation Methods over Manifolds

This dissertation covers the topic of estimating unknown functions over manifolds. In some sense, this is related to the topic of optimal estimation over manifolds, which is performed using a variety of different techniques. One of the most prevalent estimation methods for nonlinear systems is Extended Kalman Filtering (EKF). EKF is the nonlinear version of the Kalman filter which linearizes about an estimate of the current mean and covariance of the states [46]. There have been a number of studies on EKF over manifolds such as [47, 48] and lie groups [49, 50]. In [48] the authors introduce a variant of the extended Kalman filter that uses stable embedding on the manifold to extend it to an open set in Euclidean space where conventional Kalman filtering on Euclidean space can be performed. In [49] the authors introduce extended Kalman Filters on Lie Groups, which assume that the posterior distribution of the state is a concentrated Gaussian distribution on Lie groups. As mentioned previously, the extended Kalman filter approximates the state distribution through a Gaussian random variable (GRV), which then linearly evolves according to the Jacobian around the current state. This algorithm may quickly diverge if there is significant initial error in the state or if the propagation of the states and covariance are nonlinear.

Another nonlinear estimation method is accomplished using Unscented Kalman Filtering (UKF). In contrast to using a one point mean approximation in EKF, UKF uses several sigma points to build the approximation of the distribution. As mentioned in [51], these sample points better approximate the true mean and covariance of the GRV, and when propagated through the nonlinear system, the posterior mean and covariance are accurately estimated for any nonlinearity via Taylor series expansion up to the third order. Unscented Kalman filtering has been applied to Riemannian manifolds [52] and Lie groups [53]. While UKF has many advantages, this method is still only optimal when making strong assumptions.
about the prior and posterior distributions of the estimate. Additionally, both Kalman filter methods do not guarantee any global convergence of the estimation error to zero.

Similar to the sigma points in UKF, Particle filtering uses a set of samples, called particles, to represent the posterior distribution of a stochastic process given noisy observations. Estimation is performed through a sequential Monte Carlo method. Particle filter techniques provide a well-established methodology for generating samples from the required distribution without requiring assumptions about the state-space model or the state distributions. Particle filters have been applied to Riemannian manifolds [54] and Lie Groups [55, 56, 57]. The main drawback of Particle filtering comes from its expensive computational effort. In contrast to Kalman filtering, which can efficiently estimate a Gaussian distribution with a small number of samples, Particle filters require multiple simulations to generate enough samples to get an accurate representation of the distribution [58]. Particle filters do demonstrate convergence of the error to zero as more samples are collected. However, particle filters, in general, do not perform well when applied to high-dimensional systems. Additionally, the method requires samples be independent and identically distributed (IID). For the cases studied in this defense, systems evolve along some deterministic dynamical system, which we utilize to get some improved convergence characteristics compared to the IID case.

1.1.5 Koopman Theory and Dynamical Systems

The analysis of the uncertain estimation problem in this dissertation is carried out using tools from Koopman theory. Koopman theory is a family of results that studies the behavior of dynamical systems in terms of operator theory, and it constitutes a powerful framework for studying nonlinear dynamics. Very recent accounts of the general theory can be found in book [59] or edited volume [60]. For this introduction and description of contributions, it
1.1. Background and Review of Literature

It suffices to note that, given a discrete and possibly nonlinear evolution law,

\[ x_{i+1} = f(x_i) \]  

(1.4)

the Koopman operator \( U_f \) associated with the discrete evolution is defined by the identity \( U_f g = g \circ f \), with \( f \) the fixed function that appears in the unknown dynamics. Note that even if \( f \) is nonlinear the Koopman operator \( U_f \) acting on observables is linear.

\[ U_f(\alpha g + \beta h) = (\alpha g + \beta h) \circ f = \alpha g \circ f + \beta h \circ f = \alpha U_f g + \beta U_f h \]

Consequently, instead of examining the evolution of the system, Koopman theory focuses instead on the linear evolution of the observables on the system. The main trade-off, however, is that the space of observables and the Koopman operator are infinite-dimensional. This finite-dimensional approximations must be considered for any practical use of the Koopman operator.

Of the hundreds of papers related to Koopman theory, some of the most relevant to this dissertation investigate approximations of the Koopman operator, which includes \([61, 62, 63, 64, 65]\). The specific work in \([61, 62]\) that describes the extended dynamic mode decomposition (EDMD) algorithm for the approximation of the Koopman operator is also quite relevant to this dissertation. Here we summarize the EDMD method using nomenclature that is common in these references. From a set of input samples \( \{x_i\}_{i=1}^M \) generated by Equation 1.4, we define two data matrices \( X \) and \( Y \) as follows,

\[ X = [x_1, \cdots, x_M] \in \mathbb{R}^{d \times M}, \]
\[ Y = [x_2, \cdots, x_{M+1}] := f(X) \in \mathbb{R}^{d \times M}. \]
We want to use these data matrices to construct estimates of the Koopman operator $U_f$. Since the operator $U_f$ is infinite dimensional, we define a finite dimensional subspace for building approximations. Suppose we choose basis functions $\{\psi_i \mid 1 \leq i \leq N\}$ for our approximations. We introduce the vector $\psi = [\psi_1, \ldots, \psi_N]^T$. With the definition of the basis functions in $\psi$, data matrices for the EDMD algorithm are then determined as follows

$$
\Psi(X) = [\psi(x_1), \ldots, \psi(x_M)] \in \mathbb{R}^{N \times M},
\Psi(Y) = [\psi(x_2), \ldots, \psi(x_{M+1})] \in \mathbb{R}^{N \times M}.
$$

Approximations of the Koopman operator $U_f$ are constructed in terms of the matrix solution $A_{N,M} \in \mathbb{R}^{N \times N}$ of the minimization problem

$$
A_{N,M} := \arg \min_{A \in \mathbb{R}^{N \times N}} \| A\Psi(X) - \Psi(Y) \|_F^2
$$

The solution of this minimization problem is given by $A_{N,M} = \Psi(Y) \left( \Psi^T(X) \Psi(X) \right)^{-1} \Psi^T(X)$. Finally, the approximation $U_{N,M}$ of the Koopman operator $U_f$ is given by

$$
(U_{N,M}g)(\cdot) := \alpha^T A_{N,M} \psi(\cdot)
$$

for any function $g(\cdot) := \alpha^T \psi(\cdot)$ with $\alpha \in \mathbb{R}^N$. References [59, 61, 62, 63, 64, 65, 66] give further detailed accounts of how approximations of the Koopman operator $U_f$ or Koopman modes can be used in derivations of forecasting schemes or state estimators, estimation of observable functions, identification of nonlinear state equations.
1.2 Overview of this dissertation

This section gives an overview of each of the chapters in this dissertation. It also highlights the principal contributions to the state-of-art in each respective topic of research. Additionally, it lists the publications associated with each chapter.

1.2.1 Learning Theory to Estimate Animal Motion Submanifolds

The second chapter of this dissertation is concerned with modeling reptilian motion by finding the mapping $\gamma$ that determines the low-dimensional underlying manifold $Q$ that supports a given motion or motion regime. Section 2.2formulates the problem as one of distribution-free learning theory on manifolds. Solving the distribution-free learning problem poses a number of difficulties, some of which are rather standard challenges in the field of distribution-free learning theory. However, this problem is substantially more complicated than the learning problem for functions defined over Euclidean spaces owing primarily to the difficulty in defining appropriate function spaces and spaces of approximants over manifolds. These two topics are covered in Sections 2.3 and 2.3.1, respectively. The primary theoretical result of this chapter gives sufficient conditions to ensure that approximations $\gamma_{n,m} : S \to X$ of the unknown function $\gamma : S \to Q \subset X$ converge to the regressor function $\gamma_\mu : S \to X$, which will be discussed in Section 2.2.2. This chapter also introduces the linear approximation spaces $A^{r,2}(L^2_\mu(S))$ for a smooth manifold $S$, and emphasizes the crucial role these spaces play in deriving rates of convergence; These spaces will be defined in detail in section 2.3. Section 2.3.1 discusses that when the regressor function satisfies the smoothness condition $\gamma_\mu \in (A^{r,2}(L^2_\mu(S)))^d$, the rate of convergence are the same as those for certain approximations of distribution-free learning problems over Euclidean space. Such rates of convergence are
known to be optimal in the Euclidean case, except for the logarithmic factor. [34]. Section 2.4 addresses the problem of empirical risk minimization on manifolds to generate estimates of reptilian motion from inputs that evolve along a well-structured submanifold embedded in a high-dimensional Euclidean space. The analysis in Section 2.4 also shows that the empirical risk minimizer can be calculated through local averages over partitions of the input space. Utilizing the works of [34, 67] the error of the estimate to the true function can be bounded in terms of concentration inequalities and the approximation bias from the finite-dimensional subspace of approximants. Section 2.5 discusses the experimental setup used to collect the motion capture data for this study. Section 2.5.3 gives a comparison between the approximation of the observed motion from the novel approximation space to estimates using a smooth exponential kernel basis is given in Section 2.5.3. Concluding remarks on variations of bone lengths in specimens as well as the computational cost of the algorithm are given in Sections 2.5.5 and 2.5.6 respectively.

There are several distinct contributions to the state-of-the-art in learning and estimation methods for animal motion in this Chapter.

(1) **Manifold Learning:** As opposed to current manifold learning techniques, the works [8][9] cast the manifold learning algorithm in a supervised learning context allowing us to use the tools of distribution-free learning theory. We accomplish this by making the very restrictive assumption that the manifold we seek to estimate comes from inputs of a well-structured manifold. With the approach given by [8] [9], our inputs are manifold valued from our well-structured manifold space denoted $S$. The components of the measured outputs, $y$, which did not necessarily lie on the manifold $Q$, are estimated by regression with functions $\gamma : S \to \mathbb{R}$ where components of the outputs $y^j \approx \gamma(s)$ for pairs $(s, y)$ where $s \in S$ and $y^j \in \mathbb{R}$ and $y^j \approx q^j + \epsilon$, $q \in Q$ and small $\epsilon$.

(2) **Novel Approximation Schemes over Manifolds:** Building approximations over manifolds
is substantially more complicated than the conventional problems for estimating functions
defined over Euclidean spaces. This new learning problem is made complex in no small part
owing to the difficulty of defining appropriate function spaces and spaces of approximants
over manifolds. In particular, we introduce the linear approximation spaces \( A^{r,2}(L^2_\mu(S)) \)
for a smooth manifold \( S \), and these play a crucial role in deriving rates of convergence;
These spaces were defined in detail in \([8],[9]\). We build off of similar work to the papers
\([33, 34]\). The primary theoretical result of this chapter gives sufficient conditions to ensure
that approximations \( \gamma_{N,M} : S \to X \) of the unknown function \( \gamma : S \to Q \subset X \) converge to
the regressor function \( \gamma_\mu : S \to X \). The regressor function between \( S \) and \( X \) is given by
the random variable whose expected value is the corresponding average output \( x \) given the
the corresponding input \( s \),
\[
\gamma_\mu = \int_S x \mu_X(dx|s)
\]  
In this dissertation, we show that that the regressor function \( \gamma_\mu \) is indeed the optimal solution
over the expected error. We show that when the regressor function satisfies the smoothness
condition \( \gamma_\mu \in (A^{r,2}(L^2_\mu(S)))^d \), the rate of convergence is given by
\[
\mathbb{E}(\|\gamma_\mu^j - \gamma_{N,M}^j\|_{L^2_\mu}) \leq C_1 n^{-r}
\]
\[
+ C_2 \frac{n \log(n)}{M}
\]
for constants \( C_1, C_2 > 0 \), where \( \mathbb{E}(\cdot) \) is the expectation over samples, \( N \) is the dimension
of the space of approximants, and \( M \) is the number of samples. This is precisely the rate
achieved for certain approximations of distribution-free learning problems over Euclidean
space.

Much of Chapter 2 has been published in the following peer-reviewed journal article.

Powell, N., & Kurdila, A. J. (2021). Distribution-free learning theory for approximating
submanifolds from reptile motion capture data. Computational Mechanics, 68(2), 337-356.

Other contributions on learning theory in Chapter 2 have appeared in the following conference proceedings:


1.2.2 Koopman Theory to Estimate Functions over Manifolds of Animal Motion

The third chapter analyzes the utilization of Koopman operator theory to estimate functions over a manifold generated by some dynamic evolution. A detailed overview of Koopman theory on deterministic systems is given in Section 3.1.4 followed by a brief discussion on recent literature of direct approximation theorems in Section 3.1.5. The overall strategy used in this chapter is common to many of the approaches outlined in references [59, 61, 62, 63, 64, 65, 66], but in this dissertation, we focus on determining how to ensure error bounds not found in these references. As mentioned in Section 3.1.6, any practical use of the Koopman operator must address finite dimensional approximations. Approximations of the observable forward-kinematics [1] $G$ over the unknown manifold $Q$ are built using the finite dimensional reproducing kernel Hilbert Spaces $H_M$ and the $H$-orthogonal projection $\Pi_M: H \rightarrow H_M$ as discussed in Section 3.2.2. Under certain circumstances, data-dependent approximations $U_j^M$ of the Koopman operator $U_f$ defined by the identity $U_j^M g := \Pi_M((\Pi_M g) \circ f)$ have coordinate representations that are equivalent to those calculated through the so-called method of extended dynamic mode decomposition (EDMD) as detailed in Section 3.2.3. The approximation $U_j^M$ of the Koopman operator $U_f$ can be constructed from the samples $\{(x_{k_i}, y_{k_i})\}_{i=1}^M$.
and the choice of finite dimensional space $H_M$. This dissertation introduces an approximation method in which the distribution of the samples $\{x_k\}_{i=1}^M$, the definition of the space $H_M$, and the choice of the kernel $\mathcal{K}$ determines the rate of convergence of $U^M f \rightarrow U f$. Section 3.3 contains the problem formulation and how the estimation problem can be framed as one of distribution-free learning. Section 3.4 contains results and discussions on several numerical examples with an emphasis on the role of numerical stability on the estimate. This Section also discusses an implementation of the approximation algorithm on experimentally collected motion capture data. Concluding remarks regarding this chapter are given in Section 3.5.

The contributions to the state-of-the-art in Koopman theory and estimation methods for animal motion in this chapter are as follows:

(1) **Reduced Order Models of Animal Motion using Koopman Theory:** Although Koopman methods and manifold learning have been studied for a wide variety of application areas, this research is the first systematic study of the method for characterizing animal motions over submanifolds. The approach should be of general interest to those studying reduced order models of animal motions. The paper describes a rigorous formulation of a general method for approximating such motions, and the theory includes estimates of the rates of convergence in various function spaces as the number of samples $M$ increases. Such estimates are not available in the literature on motion estimation such as in [68, 69, 70, 71, 72, 73].

(2) **Convergence Rates for Koopman Approximations:** In addition, the research describes general, strong rates of convergence of data-dependent approximations of the Koopman operator. This result is of interest in its own right. Full details on the derivation and interpretation of these rates is given [10]. These rates for the Koopman approximation problem are qualitatively similar to rates of convergence that are familiar in regression in Euclidean spaces [32], interpolation over manifolds [74, 75], Chapter 10 of [76], and ambient approximation over manifolds [77]. None of the recent studies of Koopman approximations [61, 62, 63, 64, 65] derive such rates of convergence in terms of fill distances as we have done in [10].
(3) Rates of Convergence of the EDMD Algorithm: The research in this dissertation defines a general data-dependent approximation of the Koopman operator as it acts on a variety of function spaces. In some instances, this data dependent operator can be interpreted as a specific implementation of the extended dynamic mode decomposition (EDMD) algorithm. With this identification it is possible to derive rates of convergence for the EDMD algorithm in certain types of function spaces. This result is described in [10] and also has no precedent in the recent studies of the EDMD algorithm, as in [61, 64].

The work in Chapter 3 has been submitted for publication in a peer-reviewed journal. The reference of the preprint on arxiv is given below.


Other contributions on Koopman theory in Chapter 3 have appeared in the following conference proceedings:


1.2.3 Kernel Methods for Regression in Continuous Time over Subsets and Manifolds

Chapter 4 studies the continuous regression problem over manifolds and subsets. This chapter begins with the motivation behinds the study of the continuous regression problem in Section 4.1.1. Section 4.1.2 formulates the problem associated with regression in continuous time over subsets and manifolds. A detailed discussion of the regression problem in continu-
ous time using an RKHS is given in Section 4.2 with an emphasis on optimal offline estimates in Section 4.2.1 and a discussion on the Galerkin approximation of the estimate in Section 4.2.1. Section 4.3 introduces the persistency of excitation (PE) condition and its effects on the error of the estimates from certain spaces of functions. As detailed in Section 4.3.2 improved error estimates can be determined when the PE condition holds for a particular subspace 4.3.2 of the RKHS. Two different approximation methods of the optimal regressor are detailed in Section 4.3.2 and 4.3.2. We then relate the continuous time regression problem and its approximations to the classical learning theory scenario in Section 4.3.3. In contrast to finite-dimensional subspaces, Section 4.3.4 focuses on identifying an important error bound on estimates from an infinite-dimensional closed subspace that still satisfies the PE condition. Several numerical results on the characteristics of the two approximations of the optimal continuous regressor are given in Section 4.4. Additionally, this section includes a implementation of the two approximation schemes on motion capture data and a discussion on the respective characteristics of each. Finally, Concluding remarks are given regarding this chapter on continuous time regression in Section 4.5. There are three specific new results derived in this chapter that are not addressed in any of the previous papers by the authors in [78, 79, 80, 81, 82], or in the literature at large.

(1) **Sufficient Conditions for Operator Used in Continuous Regression:** Suppose that \( \phi : t \mapsto \phi(t) \in X \) is a trajectory of either an autonomous or nonautonomous flow on the manifold. The regression problem is solved using the operator \( T_\phi(s, t) : \mathcal{H} \to \mathcal{H} \)

\[
T_\phi(s, t) := \int_s^t K_{\phi(\tau)} \otimes K_{\phi(\tau)} \nu(d\tau).
\]

where \( K_{\phi(\tau)} := K(\phi(\tau), \cdot) \) is the kernel basis function centered at \( \phi(\tau) \) and \( \nu \) is a finite measure on \([s, t]\). The tensor product operator \( K_{\phi(\tau)} \otimes K_{\phi(\tau)} \) satisfies \( K_{\phi(\tau)} \otimes K_{\phi(\tau)} g = K_{\phi(\tau)} \langle K_{\phi(\tau)} g \rangle \mathcal{H} \) for all \( g \in \mathcal{H} \). The first new result gives sufficient conditions to ensure that this operator is
compact, positive, and self-adjoint.

(2) **A New Persistency of Excitation (PE) Condition For Flows Over Manifolds:** The second new result is the introduction of a new persistency condition for flows over a manifold that generalizes the one in our earlier papers. It defines persistency for a general closed subspace \( \mathcal{V} \subseteq \mathcal{H} \), where the norm on \( \mathcal{V} \) that can be different than the norm on \( \mathcal{H} \). The publications \([83, 84]\) always make the special choice \( \mathcal{V} := \mathcal{H}_S \) where \( \mathcal{H}_S \) is the native space generated by a subset \( S \subseteq \mathcal{X} \). The generalization in this chapter is essential to prove convergence of estimates in certain spectral approximation spaces \( \mathcal{A}^r \subseteq \mathcal{H} \), which depend on a trajectory \( t \mapsto \phi(t) \in \mathcal{X} \).

(3) **Novel Error Analysis for PE condition over Certain Subspaces:** In this chapter, we show that, if a PE condition holds for the subspace \( \mathcal{V} \subseteq \mathcal{H} \), then

\[
\|\hat{g}_\mathcal{V}(t, \cdot) - \Pi_\mathcal{V} G\|_{\mathcal{H}} \leq \frac{\bar{\mathcal{R}}^2 m \Delta}{\gamma_1 m + \gamma}(I - \Pi_\mathcal{V}) G\|_{\mathcal{H}} + \frac{\gamma}{\gamma_1 m + \gamma} \|\Pi_\mathcal{V} G\|_{\mathcal{H}}
\]

where \( t \mapsto \hat{g}_\mathcal{V}(t, \cdot) \in \mathcal{V} \) is the optimal solution of the (offline) regression problem, and \( \Pi_\mathcal{V} \) is the \( \mathcal{H} \)-orthogonal projection of \( \mathcal{H} \) onto \( \mathcal{V} \). The constant \( \bar{\mathcal{R}} \) is a bound on the reproducing kernel \( \mathcal{R} \) that defines \( \mathcal{H} \), the constants \( \gamma_1 \) and \( \Delta \) arise in the PE condition, \( \gamma \) is the regularization parameter in the continuous regression error functional, and the time \( t := m \Delta \) for the positive integer \( m > 0 \). Note that as time \( t = m \Delta \to \infty \), the error

\[
\|\hat{g}_\mathcal{V}(t, \cdot) - \Pi_\mathcal{V} G\|_{\mathcal{H}} \lesssim \mathcal{O}\left( \|(I - \Pi_\mathcal{H}) G\|_{\mathcal{H}} \right).
\]

(4) **Derived Error Bounds for Irregular Subsets:** In some cases, when \( N \) samples are used to define certain finite dimensional spaces of approximants \( \mathcal{V} := \mathcal{H}_N \) and these spaces are PE, we can bound estimates from \( \mathcal{H}_N \) with a term on the order of the power function \( \mathcal{P}_N(x) \),
over the set $S$, that is defined as

$$\mathcal{P}_N(x) := |\mathcal{R}(x, x) - \mathcal{R}_N(x, x)| \quad \text{for all } x \in S.$$

In this expression $\mathcal{R}_N$ is the kernel that defines the native space $\mathcal{H}_N$ of approximants, and $S$ is the closure of the trajectory $\tau \mapsto \phi(\tau)$ in $X$. The kernel $\mathcal{R}_N(x, y) := (\Pi_N k_x)(y)$ by definition [85]. Similar to the works [86, 87], in which are derived pointwise error bounds for discrete regression or Bayesian estimation for discrete time processes, this chapter examines integrated error bounds for systems in continuous time. The relationship of the solution of the continuous time regression problem to the more familiar discrete IID, stochastic, case is discussed in detail in Section 4.3.3.

The work in Chapter 4 has been submitted for publication in a peer-reviewed journal. The reference of the preprint on arxiv is given below.


**Bibliography**


Chapter 2

Distribution-Free Learning Theory for Approximating Submanifolds from Reptile Motion Capture Data

2.1 Introduction

The related problems of understanding the underlying dynamics of animal locomotion, constructing (multibody) dynamics models of animals, or building robots that emulate them have occupied researchers having diverse backgrounds. Research in [1, 2, 3, 4, 5, 6] is representative and includes investigations related to, or based upon, the motions of humans, bats, birds, lizards, geckos, sheep, frogs, beetles, and cheetahs. Some studies such as in [7, 8, 9, 10] seek to understand or utilize the underlying geometry of the mechanics of human or other animal motion. Other studies are expressly interested in the construction of low dimensional models of complex motion [11, 12, 13, 14]. Still others are dedicated to finding how qualitative understandings of animal motion can be used to build robots that emulate some observed motion [15, 16, 17, 18, 19, 20], with a notable emphasis on locomotion in quadrupeds [21, 22, 23, 24, 25, 26, 27] and bipeds [28, 29, 30, 31, 32, 33].

A fundamental question that is either motivated or directly addressed by these studies is how
CHAPTER 2. DISTRIBUTION-FREE LEARNING THEORY FOR APPROXIMATING SUBMANIFOLDS FROM REPTILE MOTION CAPTURE DATA

to come up with a principled, low-dimensional model of a complex animal motion. In the above studies, many assumptions are made about the underlying constraints and generalized coordinates of the observed motion. We would like our method to work in as general a case as possible and minimize the assumptions about the geometry of the system. Furthermore, we want to give some rigorous degree of confidence in our low-dimensional model. The aim of our research is to provide further insight into the motion of animals by deriving such a general method to discover and approximate an underlying manifold on which the dynamics evolves.

Historically, theories used to derive models of animal motion have been around for some time. Multibody dynamics formulations [34, 35, 36] derived in terms of analytical mechanics are well-known, and constitute the foundation of general simulation packages like MSC Adams®, OpenSim®, Solidworks®, and NX®. These approaches give rise to models that allow for rigid or flexible bodies, linear or nonlinear deformation models of individual bodies, and either ideal or more general types of constraints among bodies. See the reference [34] for a good overview. While the use of analytical mechanics as the basis of such formulations can be traced back several decades, one recent trend in the field of multibody dynamics has been the study of geometric methods based on manifolds and differential geometry [37, 38, 39]. It is fair to say that while such differential-geometric descriptions are couched in a language that is substantially more rigorous than the older approaches derived from analytical mechanics, the latter have given rise to a wider class of models that are now commonly used for more general systems. Interestingly, most of the recent expositions such as [37, 38, 39] that overview geometric methods for multibody systems assume that the constituent bodies are rigid and constraints are ideal. In one sense, this chapter can be viewed as addressing the problem of how to define a submanifold from observations for motion models that could subsequently be used in a geometric formulation of multibody dynamics.
2.1. Introduction

Figure 2.1: Abstract illustration of the relationship between the known manifold, $S$, and the unknown submanifold, $Q$, where the observed motion occurs.

As discussed carefully in [39], one of the primary theoretical advantages of geometric methods is their abstraction: states of the system are known to be elements of manifolds, which are by their nature coordinate-free. This generality is particularly attractive in the problem at hand. Intuitively, we want to approximate or identify some finite “lower-dimensional mathematical object”, which we denote as the submanifold $Q$, that underlies an observed motion of a complex, higher order multibody system in a $d$-dimensional Euclidean space denoted, $X$. We want to ensure that any associated algorithm converges to a coordinate independent entity.

The main focus of this chapter is the analysis and experimental testing of a method to identify and approximate motions over our finite dimensional submanifold $Q$ that underlies a particular motion regime observed during animal motion. We work to formulate the problem in such a way that we can guarantee that approximations converge, in some appropriate sense, over the underlying manifold. We assume this manifold $Q$ is smooth, compact, and homeomorphic to a known manifold $S$, which is selected a priori. Figure 2.1 illustrates the abstract relationship between these two homeomorphic manifolds. The structure of $S$ is specifically chosen to ease the difficulty in determining the submanifold $Q$.

We begin in Section 2.2.2 with a careful description of the problem of estimating a mo-
CHAPTER 2. DISTRIBUTION-FREE LEARNING THEORY FOR APPROXIMATING SUBMANIFOLDS FROM REPTILE MOTION CAPTURE DATA

tion manifold as one of distribution-free learning over manifolds. This problem is substantially more complicated than the conventional problems for functions defined over Euclidean spaces. This new learning problem is made complex in no small part owing to the difficulty of defining appropriate function spaces and spaces of approximants over manifolds. These two topics are covered in Sections 2.3 and 2.3.1, respectively. In particular, we introduce the linear approximation spaces $A^{r,2}(L^2_\mu(S))$ for a smooth manifold $S$, and these play a crucial role in deriving rates of convergence; These spaces will be defined in detail in section 2.3. We build off of similar work to the papers [40, 41]. The primary theoretical result of this chapter gives sufficient conditions to ensure that approximations $\gamma_{n,m} : S \to X$ of the unknown function $\gamma : S \to Q \subset X$ converge to the regressor function $\gamma_\mu : S \to X$. The regressor function between $S$ and $X$ is given by the random variable whose expected value is the corresponding average output $x$ given the corresponding input $s$,

$$\gamma_\mu = \int_S x \mu_X (dx|s)$$

(2.1)

We will show that the regressor function $\gamma_\mu$ is indeed the optimal solution over the expected error also commonly referred to as the expected risk, $E_\mu$. We show that when the regressor function satisfies the smoothness condition $\gamma_\mu \in (A^{r,2}(L^2_\mu(S)))^d$, the rate of convergence is given by

$$\mathbb{E}(\|\gamma_\mu^{(j)} - \gamma_{N,M}^{(j)}\|_{L^2_\mu}) \leq C_1 N^{-r} + C_2 \frac{N \log(N)}{M}$$

(2.2)

for constants $C_1, C_2 > 0$, where $\mathbb{E}(\cdot)$ is the expectation over samples, $N$ is the dimension of the space of approximants, and $m$ is the number of samples. This is precisely the rate achieved for certain approximations of distribution-free learning problems over Euclidean space. Such rates of convergence are known to be optimal in the Euclidean case, except
for the logarithmic factor. The first term on the right of 2.2 is commonly referred to as the bias or approximation error, while the second term is typically labeled the sample error or sample variance. We note that as the dimension $n$ of the system increases the approximation error decreases but the sample error increases. We make one comment that while our approximation using the newly introduced space $A_r^{\nu}(L^2(\Sigma))$ is discontinuous, the error between this estimate and the underlying regressor is bounded by 2.2. Thus, we have some degree of certainty between our estimate and the underlying regressor provided it meets certain regularity criteria. With bounds on this smooth process we can, therefore, have some rigorous description of certainty for our estimate from the underlying regressor. Indeed, when we compare our approximation to estimates using a smooth exponential kernel basis in section 2.5.3, it seems to indicate there exist a bound between these two approximations. Finally, this chapter concludes in Section 2.5 with a study of the performance of the proposed method using samples from recent reptile motion studies.

2.2 Problem Formulation

2.2.1 Kinematics and Inherent Geometry

For purposes of ensuring convergence of approximations, we assume in this chapter that motions are supported on a configuration manifold, $Q$, that is contained in the ambient space $X \approx \mathbb{R}^d$ for some $d > 0$. Of course this space suffices to describe the dynamics of multibody systems comprised of assemblies of lumped masses where motion is specified in terms of the mass centers of the (inertia-free) bodies. Since any Riemannian manifold can be (isometrically) embedded in $\mathbb{R}^d$ for a large enough $d$ via the Nash embedding theorem, in principle this assumption also allows for some of the other standard models where the
configuration manifold includes SO(3) or SE(3) or a product of them as described in [38, 39]. It is further assumed that the manifold $Q$ is a smooth, compact, connected, Riemannian manifold that is regularly embedded in the ambient space $X \approx \mathbb{R}^d$. This manifold is taken to be homeomorphic to some known smooth manifold $S$. We denote the homeomorphism by $\gamma : S \rightarrow Q$ where $\gamma := [\gamma^1, \ldots, \gamma^{(d)}]^T$ and $\gamma^i : S \rightarrow \mathbb{R}$, and we discuss its smoothness in our discussion of approximation spaces in Section 2.3.

The manifold $S$ is equipped with the topology induced by its intrinsic Riemannian metric and $Q$ is equipped with the topology inherited as a regularly embedded submanifold $Q \subseteq X \approx \mathbb{R}^d$.

### 2.2.2 The Distribution-Free Learning Problem on Manifolds

This chapter is concerned with finding the mapping $\gamma$ that determines the low-dimensional underlying manifold $Q$ that supports a given motion or motion regime. We choose to pose this problem as one of distribution-free learning theory on manifolds. In distribution-free learning theory, it is assumed that we are given a collection of independent and identically distributed samples $\{z_i\}_{i=1}^M := \{(s_i, q_i)\}_{i=1}^M \subset \mathcal{Z} := S \times X$ that are generated by some unknown probability measure $\mu$ on the manifold $\mathcal{Z}$. In order to determine the mapping from the known manifold $S$ to the configuration space $Q$, it would be ideal if we could determine an optimal mapping $\gamma^* : S \rightarrow Q \subset X$ such that

$$
\gamma^* = \arg \min_{\gamma \in \Gamma} E_\mu(\gamma) := \arg \min_{\gamma \in \Gamma} \int_{\mathcal{Z}} \|q - \gamma(s)\|^2_X \mu(dz) \quad (2.3)
$$

where $q$ would be a noise-free representation of the motion and $\Gamma$ is a set of admissible operators that map from $S \rightarrow Q$. We refer to $E_\mu$ as the expected risk between the measured samples and our mapping. It is also commonly referred to as the expected error. Solving the optimization problem in Equation 2.3 poses a number of difficulties, some of which are
2.2. Problem Formulation

rather standard challenges in the field of distribution-free learning theory. See [42] for a detailed discussion. For example, it is well-known that the ideal solution $\gamma^*$ above cannot be computed from a number of difficulties. One main issue comes from the minimization problem in Equation 2.3 since the measure $\mu$ on $Z$ is unknown. For this reason, methods of empirical risk minimization are used that substitute a discrete proxy for the expression in Equation 2.3. We discuss this step in detail in Section 2.4.

Beyond these conventional challenges to building approximations of $\gamma^*$, there are additional substantial difficulties that are unique to the problem at hand. In classical treatments of distribution-free learning theory, the rates of convergence of approximations to the solution are often cast in terms of approximation spaces or smoothness spaces (that often end up being Sobolev spaces) of real-valued functions over subsets of Euclidean space $\mathbb{R}^p$. That is, the space $\Gamma$ is usually selected to be some subset of the real-valued functions over subsets of Euclidean space. Here, however, the functions $\gamma \in \Gamma$ are mappings from the manifold $S$ to the manifold $Q \subset X$.

The definition of Sobolev mappings between manifolds is a complicated and delicate subject. For general choices of $S$ or $Q$ it is unclear exactly what definition of smoothness should be selected. We briefly summarize some of these issues. When $\gamma : S \to \mathbb{R}^d$ and $S$ is a manifold, Sobolev functions $W^{s,r}(S, \mathbb{R}^d)$ can be defined intrinsically or in terms of coordinate charts as is described in a number of places [43, 44, 45]. When $\gamma : S \to Q$ and $S$ and $Q$ are both manifolds, one method for defining Sobolev mappings assumes that $Q$ is isometrically embedded in $\mathbb{R}^d$ and then sets

$$W^{s,r}(S, Q) := \{ u \in W^{s,r}(S, \mathbb{R}) \mid u(x) \in Q \text{ a.e.} \} \quad (2.4)$$

This definition enforces a constraint that the mapping $\gamma \in W^{s,r}(S, Q)$ is manifold-valued
almost everywhere, which constitutes an unwieldy side condition that would have to be enforced in optimization problems like that described in this chapter.

For these reasons, we do not enforce an *a priori* requirement in the optimization problems that approximations $\gamma_{n,m}$ take values in $Q$. We emphasize these approximations, $\gamma_{n,m}$, are given subscripts to show they are parameterized by $m$ samples and take values in an $n$-dimensional subspace of approximants. These $\gamma_{n,m}$ are extrinsic in the sense that globally they behave like a smooth $\gamma : S \rightarrow \mathbb{R}^d$, rather than mapping solely to our configuration manifold $Q$. These approximations are constructed from coordinate functions that lie in certain approximation spaces, $A^{r,2}(L^2_\mu(S))$, over the manifold $S$. These spaces are characterized by functions whose error are bounded when projected into a finite dimensional function space. A significant part of this chapter is dedicated to defining approximation spaces, $A^{r,2}(L^2_\mu(S))$, over manifolds to characterize smoothness or approximation properties and subsequently structuring the learning problem so convergence results can be derived in a way that is analogous to the classical case over subsets of Euclidean space.

### 2.3 Function Spaces on Manifolds

In the setup of the problem above we have assumed that $S$ is a known, compact, connected, smooth Riemannian manifold, and that $Q$ is a regularly embedded submanifold of $X := \mathbb{R}^d$. This means that the unknown function $\gamma : S \rightarrow Q$ is constructed of component functions $\gamma^i : S \rightarrow \mathbb{R}$ for $i = 1, \ldots, d$ over the smooth manifold $S$. In this section we define the various function spaces over manifolds that will be used to approximate the functions $\gamma^i$. The function spaces over the manifold $S$ that is used in this chapter consist of certain spaces of square-integrable functions or native spaces of a reproducing kernel. When $\mu_S$ is a measure over the manifold $S$, the usual space of real-valued, $\mu_S$-square integrable functions
is denoted by $L^2_{\mu} := L^2_{\mu} S(S)$, which is a Hilbert space with respect to the usual inner product 
$$(f, g)_{L^2_{\mu}} := \int_S f(\xi)g(\xi)\mu_S(d\xi).$$

Following the methodology of [46, 47], we define a real reproducing kernel Hilbert (RKH) space $H^A$ over an arbitrary set $A$ is defined to be

$$H^A := \text{span} \{ R^A_a \mid a \in A \}$$

where $R^A_a := R^A(a, \cdot)$ and $R^A : A \times A \rightarrow \mathbb{R}$ is a real, positive semi-definite, symmetric, continuous admissible kernel function. In this work, we are concerned with multiple domains, so we use the superscripts to emphasize the domain of our approximation. The kernel is positive definite when $\sum_{i,j} R^A(a_i, a_j)\alpha_i\alpha_j \geq 0$ for any collection of points $\{a_i\}_{i=1}^M \subseteq A$ and coefficients $\{\alpha_i\}_{i=1}^M$ for a finite integer $m$. For any element $a \in A$, we define the kernel function $R^A_a \in H^A$ with $R^A_a(\cdot) = R^A(a, \cdot)$ This definition of an RKH space makes sense over a general set $A$, and in particular makes sense for the specification of functions over manifolds. We denote by $R^X : X \times X \rightarrow \mathbb{R}$ the kernel of the RKH Space $H^X$ over the ambient space $X$. The restriction of functions in $H^X$ to the configuration space $Q$ always defines an RKH space over $Q$, and we set $H^Q := \{ g|_Q \mid g \in H^X \}$. The kernel $R^Q : Q \times Q \rightarrow \mathbb{R}$ that defines the space of restrictions $H^Q$ is given by $R^Q(x, y) := R^X(x, y)|_Q$, for all $x, y \in Q$, which is a classical result from the theory of RKH spaces [46, 47]. Finally, we define the pullback space $H^S := \gamma(H^Q)$ on $S$ as

$$\gamma^S(H^Q) := \{ g \circ \gamma \mid g \in H^Q \},$$

and the kernel $R^S$ defined on the manifold $S$ is written as $R^S(\alpha, \beta) = R^Q(\gamma(\alpha), \gamma(\beta))$.

The approximations in this chapter are constructed in terms of two different types of finite dimensional approximation spaces, those that are contained in $L^2_{\mu}(S)$ and those contained
Figure 2.2: Illustrating our uniform refinement of the manifold $S$. As the refinement level $\ell$ increases, which is illustrated as moving down the figure, the dimension of the space, $N$, which is represented by characteristic functions defined over the partitions represented by the connected black lines, increases by a factor of 2.

in the RKH space $\mathcal{H}^S$

### 2.3.1 Approximants in $L^2_\mu(S)$

When we build approximations in $L^2_\mu(S)$, we make use of a nested sequence $\{S_\ell\}_{\ell=0}^\infty$ of measurable partitions $S_\ell$ of $S$, where each partition $S_\ell := \{S_{\ell,k}\}_{k=1}^{N(\ell)}$. That is, these sets satisfy $S = \bigcup_{k=1}^{N(\ell)} S_{\ell,k}$, $S_{\ell,k} \cap S_{\ell,p} = \emptyset$ for all $\ell \neq k$, and $S_{\ell,k} = \bigcup_{v \in \Lambda_{\ell,k}} S_{\ell+1,v}$ for some finite set of indices $\Lambda_{\ell,k}$. Here $N(\ell) = \#(S_\ell)$ is the number of sets in the $\ell^{th}$ partition $S_\ell$. In other words, for each level of refinement $\ell$ we divide the manifold $S$ into $N$ sets. For this chapter, the uniform refinement $\ell$ scales up the dimension of the space $N(\ell)$ by a factor of 2 for each refinement so that $N(\ell) = 2^\ell$, but other possible choices of refinement could be used. Figure 2.2 illustrates the relation of the refinement $\ell$ to the dimension of the space $N$.

We define the space of approximants $H^S_N$ as the span of the characteristic functions $1_{S_{\ell,k}}$ of the sets in $S_\ell$,

$$H^S_N := \text{span}\{1_{S_{\ell,k}} \mid 1 \leq k \leq N(\ell)\} \subset L^2_\mu(S)$$ (2.5)
2.3. Function Spaces on Manifolds

where \( \{1_{S_{\ell,k}}\}_{k=1}^{N(\ell)} \) is the associated partition of unity over \( S \). We associate to the partition \( S_\ell \) a set of representatives,

\[
\Xi_N := \{\xi_{\ell,k} \mid \xi_k \in S_{\ell,k}, k = 1, \ldots, N(\ell)\}.
\]

These representative points are assumed to fill the manifold in the sense that the fill distance

\[
h_{\Xi,\ell} := \max_{s \in S} \min_{\xi_k \in \Xi_\ell} d(s, \xi_{\ell,k}) \to 0
\]
as \( n \to \infty \). For any \( g \in L^2_\mu(s) \) we define the \( L^2_\mu \)-orthogonal projection \( \Pi^S_N \) in the usual way,

\[
\Pi^S_N g(\cdot) = \sum_{k=1}^{N(\ell)} \int_{S_{\ell,k}} g(x) \frac{1_{S_{\ell,k}}(s) \mu_S(ds)}{\sqrt{\mu_S(S_{\ell,k})}} \frac{1_{S_{\ell,k}}(\cdot)}{\sqrt{\mu_S(S_{\ell,k})}},
\]

We define the linear approximation space \( A^{r,2}(L^2_\mu) \subseteq L^2_\mu(S) \) of functions defined over the manifold \( S \) as

\[
A^{r,2}(L^2_\mu) = \left\{ g \in L^2_\mu(S) \mid \|I - \Pi^S_N\|_{L^2_\mu} \leq Cn^{-r} \right\}
\]

for some constant \( C \in \mathbb{R} \). Note that \( \Pi^S_N \) represents the orthogonal projection onto a finite dimensional space over the domain \( S \). The infimum of the constants \( C \), for which in the inequality above holds, defines a semi-norm on \( A^{r,2}(L^2_\mu) \). We note that if \( S \) is a compact subset of Euclidean space and \( \mu \) is a Lebesgue measure, these spaces can be understood in some cases to be Sobolev spaces. Here the expression above defines approximation spaces over the manifold \( S \). For the interested reader, we have included more information regarding Lebesgue and Sobolev Spaces in the appendix Section 2.8. Additionally, [40] includes a
detailed discussion of approximation spaces.

2.3.2 Approximants in $\mathcal{H}^S$

Approximations are also constructed in terms of the finite dimensional spaces of approximants

$$\mathcal{H}^S_N := \text{span}\{ \mathbb{R}^S_{s_i} | 1 \leq i \leq N \} \subset \mathcal{H}^S.$$  

Approximations based on kernels of an RKH Space have become quite popular, particularly in application in problems of Bayesian estimation in machine learning [48].

Specifically, approximations in the space $\mathcal{H}^S_N$ can be generated by standard linear regression. We give a brief background discussion on this approximation method. Classical regression is a data-driven approximation method which involves collecting sample pairs $z_i$ of the system inputs denoted $s_i \in S$ and outputs $q_i$ and time $t_i$. These samples of data are collected along the evolution of a system.

In an ideal sense, we want our output to come from the submanifold $Q$. From a practical standpoint, our measurement is perturbed by some noise that could potentially lead to output samples outside of $Q$. However, the submanifold $Q$ lies in the Euclidean ambient space $X$. Therefore, we are guaranteed our outputs $q$ are in the space $X \approx \mathbb{R}^d$, and for the remainder of this section we only need to assume our outputs $q$ are d-dimensional.

Given our output $q_i$, we assume that the jth dimension of our output, denoted $q_i^{(j)} \in \mathbb{R}$ is actually related by a nonlinear $\gamma^{(j)} : S \rightarrow \mathbb{R}$ with respect to our input measurements $s_i$.

We can extend the classical linear regression algorithm to the nonlinear case via Koopman approximation. Koopman theory takes advantage of the linearity of the composition opera-
tor. Specifically, the Koopman operator $\mathcal{K}$ can be defined on a so-called observable function from our outputs $g : S \to \mathbb{R}$ where $\mathcal{K}g = g \circ \gamma$. The Koopman operator is linear

$$\mathcal{K}(\alpha g + \beta h) = (\alpha g + \beta h) \circ f = \alpha g \circ f + \beta h \circ f$$

$$= \alpha \mathcal{K}g + \beta \mathcal{K}h$$

but the space, denoted $\mathcal{H}^S$, is infinite dimensional. However, we can restrict the space to an n-dimensional kernel basis $\{\mathcal{K}^S(\zeta_k, s)\}_{k=1}^N$, where the space is denoted $\mathcal{H}_N^S$. These kernels are parameterized by our kernel centers $\{\xi_k\}_{k=1}^N$, so we often denote our kernel functions, $\mathcal{K}(\xi, s) = \mathcal{K}^S(s)$. The kernel functions also rely on some additional hyperparameters, which we discuss in more detail in 2.5.3. We end this section by stating that, in this chapter, we do not study the rates of convergence for these bases. We use them to carry out qualitative comparisons of their performance to the piece-wise constant basis. For a fuller discussion of the kernel basis see [44, 49].

### 2.4 Empirical Risk Minimization over Manifolds

As noted above, in some ideal sense the optimal choice of the mapping $\gamma : S \to X$ is the one that extremizes the cost functional $E_\mu$ in Equation 2.3. Since the measure $\mu$ that generates the samples in $Z$ is unknown, distribution-free learning theory uses a proxy that can be computed to re-cast the optimization problem. The form of the empirical risk can best be motivated by first rewriting the ideal risk $E_\mu$ in terms of the regressor function. Any measure $\mu$ on the space of samples $Z$ can be rewritten (disintegrated) into the expression $\mu(dz) := \mu(dx, ds) := \mu_X(dx|s)\mu_S(ds)$ with $z := (s, x)$ with $\mu_S$ the associated marginal
measure over $S$, and $\mu_X(\cdot|s)$ the conditional probability measure on $X$ given $s \in S$. One way to interpret this conditional dependence is that selecting samples $(s_i, q_i)$ on $Z$ is equivalent to drawing $s_i$ randomly according to $\mu_S$ on $S$ and then drawing samples $q_i$ from $X$ randomly according to $\mu_X(\cdot, x)$ on $X$. From first principles it is straightforward to show that we can rewrite the ideal cost $E_\mu$ in the alternative form

$$E_\mu(\gamma) = \int_S \|\gamma(s) - \gamma_\mu(s)\|^2 \mu_S(ds) + E_\mu(\gamma_\mu)$$  \hspace{1cm} (2.6)$$

where

$$\gamma_\mu(s) := \int_X x \mu(dx|s)$$

is the regressor function associated with the measure $\mu(dz) := \mu_X(dx|s)\mu_S(ds)$. This decomposition shows, in fact, that the optimal mapping $\gamma^*$ that minimizes $E_\mu(\gamma)$ is the regressor function $\gamma^* := \gamma_\mu$ since the last term on the right in Equation 2.6 is a constant that does not depend on $\gamma$. Next, we introduce the empirical risk

$$E_M(\gamma) = \frac{1}{M} \sum_{i=1}^M \|q_i - \gamma(s_i)\|_X^2$$  \hspace{1cm} (2.7)$$

in terms of the samples $\{(s_i, q_i)\}_{i=1}^M \subset Z$. Note that the cost functional $E_M(\gamma)$ can be computed for any admissible function $\gamma$ since the samples are known. We will be interested in two types of approximations of the empirical risk in the discussions that follow. We define

$$\gamma_{N,M} = \arg \min_{\gamma \in \mathcal{F}_N} \frac{1}{M} \sum_{k=1}^M \|q_i - \gamma(s_i)\|_X^2$$ \hspace{1cm} (2.8)$$

when we seek approximations that converge in $L^2_\mu$. This definition has the advantage that it is possible to derive strong convergence rates for the approximations $\gamma_{N,M} \rightarrow \gamma_\mu$. However, the resulting approximations are not very smooth. Note that this minimization can take place
over any finite dimensional family of functions $\mathcal{F}_N$. In this chapter we compare the space $H^S_N$ with $\mathcal{H}^S_N$. The space $\mathcal{H}^S_N$ produces a smoother representations of the mapping $\gamma_{N,M}$. We know that any $\gamma \in \mathcal{H}^S_N$ can be represented by a linear combination, $\gamma = \sum_{k=1}^{N} \alpha_k R_{\xi_k}^S$ of our basis $\{R_{\xi_k}^S\}_{k=1}^{N}$. So we can rewrite 2.8 to be the following.

$$\gamma_{N,M} = \arg \min_{\gamma \in \mathcal{H}^S_N} \frac{1}{M} \sum_{k=1}^{M} \left\| q_i - \sum_{k=1}^{N} \alpha_k R_{\xi_k}^S(s) \right\|_N^2$$

(2.9)

Typical linear regression allows us to determine the optimal coefficients with respect to equation 2.9. Details on the regression process for $\mathcal{H}^S_N$ are included in the appendix in Section 2.9. Despite the smoother approximation, we have not derived such strong rates of convergence $\gamma_{N,M} \rightarrow \gamma_\mu$ in this case. Additionally, calculating the closed form solution to 2.9 requires computing the inverse of a large kernel matrix. We will see shortly that our minimization in $H^S_N$ does not require any matrix inversion. Several numerical examples are included in Section 2.5.3 that compare the performance of the estimates generalized in Equations 2.8.

Before we study the solutions of these distribution-free learning problems over the manifold $S$, we make one last observation regarding the form of minimization problem. By introducing the samples $q_i := \{q_i, \ldots, q_i^{(d)}\}^T$ and mapping $\gamma := \{\gamma^1, \ldots, \gamma^{(d)}\}^T$, we can write

$$\gamma_{N,M} = \arg \min_{\gamma \in H^S_N} \frac{1}{M} \sum_{k=1}^{M} \sum_{j=1}^{N} (q_k^{(j)} - \gamma^{(j)}(s_i))^2$$

$$= \arg \min_{\gamma \in H^S_N} \sum_{j=1}^{N} E_M^{(j)}(\gamma^{(j)})$$

with $E_M^{(j)}(\gamma^{(j)}) := \frac{1}{M} \sum_{k=1}^{M} (q_k^{(j)} - \gamma^{(j)}(s_i))^2$. Note that since each term $E_M^{(j)}(\gamma^{(j)})$ is positive, the optimal $\gamma_{N,M}$ can be obtained by extremizing each of the terms $E_M^{(j)}(\gamma^{(j)})$ for the component functions $\gamma^{(j)}$ for $j = 1, \ldots, d$. The primary theoretical result of this chapter
is summarized in the following theorem, which approximates the solution of the distribution-
free learning problem over manifold $S$

**Theorem 2.1.** The $j$th coordinate, $\gamma_{N,M}^{(j)}$, of the optimal solution, $\gamma_{N,M} := [\gamma_{n,m}^{(1)}, \ldots, \gamma_{n,m}^{(d)}]^T$, of the empirical risk minimization problem in Equation 2.7 is given by

$$\gamma_{N,M}^{(j)}(s) = \sum_{k=1}^{N(\ell)} \frac{\sum_{i=1}^{M} 1_{S_{\ell,k}(s)} q_i^{(j)}(s)}{\sum_{i=1}^{M} 1_{S_{\ell,k}(s)}} 1_{S_i}(s).$$

**Theorem 2.2.** If the regressor $\gamma_\mu \in A^{r,2}(L_\mu^2(S))$, we have the error bound

$$\mathbb{E}(\|\gamma_{\mu}^{(j)} - \gamma_{N,M}^{(j)}\|_{L_\mu^2}) \leq C_1(N(\ell))^{-r} + C_2 \frac{N(\ell) \log(N(\ell))}{M}$$

(2.10)

where the constant $C_1$ is independent of $\gamma$, $C_2 = C_2(\gamma)$, and $\mathbb{E}(\cdot)$ is the expectation on $S^M$ with the product measure $\mu_S^M$.

Detailed proofs of these theorems can be found in the appendix Sections 2.10 and 2.11.

### 2.5 Experiments and Numerical Examples

In this section, we give a rather detailed description of the qualitative performance of the
above methods for estimation of motion submanifolds from experiments. In the subsections
that follow, we summarize the experimental setup, data collection procedure and results of
numerical studies.
2.5.1 Experimental Setup

The experimental setup consisted of three Photron FASTCAM ® cameras surrounding a central tank where the motions of the *anolis sagrei* were recorded, see Figure 2.3. Cameras were spaced around the center of the tank and oriented so that almost every marker placed on the back and side of a specimen could be viewed from at least two cameras, thereby providing enough information for stereo-triangulation of the location of each marker. The master camera faced the dorsal (back) side of the specimen and two supporting cameras were placed to the left and right of the specimen. Photoreflective markers were placed at various physical landmarks, such as the base of the neck, shoulder joint, etc. of the specimens as depicted in Figure 2.5. Motions were recorded while the specimens traversed an inclined, narrow board within the line of sight of the three cameras. The motion of a total of 21 *anolis sagrei* were recorded, of which 15 were male and the remaining 6 were female. Roughly twice as many males as females were recorded because the male specimens were larger on average; thus, the size of the markers with a diameter of around 1 mm were difficult to place on the limbs of the female specimens.

2.5.2 Data collection procedure

Before each recording session, a standard calibration procedure was performed. The calibration process consisted of recording images of a checkerboard pattern of known dimension in different positions and orientations. These images were then imported into MATLAB’s Stereo Camera Calibration application to determine each camera’s orientation and position relative to one another. Calibration also was able to determine intrinsic parameters of the camera.
Figure 2.3: The experimental setup of the motion capture recordings. Three Photron FASTCAM cameras face a central container housing an inclined wooden plank that the specimen crawls up. Because high shutter speeds diminish the brightness of the recording, floodlights were used to illuminate the scene.
Once the calibration was performed specimens of *anolis sagrei* were placed on a wooden inclined plank. Recording began as soon as the *anolis sagrei* started to crawl, and it ended once they reached the top of the plank. For each recording, the three camera views allowed unobstructed tracking of most points along the specimen for the entire recording session. Video was recorded at 500 fps, and Photron’s FASTCAM Viewer application allowed synchronization of the recordings between multiple cameras.

Using the MATLAB toolbox for motion tracking from video images [50], two-dimensional trajectories of markers placed on a lizard’s body were tracked and digitized into camera image coordinates from various camera views. Using the stereo-calibrated direct linear transform (DLT) coefficients, which define extrinsic camera parameters such as position and orientation of cameras as well as intrinsic camera parameters such as lens distortion, these trajectories were transformed to a three-dimensional coordinate frame defined by the master camera’s position and orientation denoted $c_1,c_2,c_3$. Three-dimensional marker trajectories, $\zeta_c$ corresponding to the camera frame, are seen in Figure 2.4, which depicts the paths generated for the body, shoulder, and elbow in the three-dimensional Euclidean space defined by the inertial camera frame. For a detailed explanation of the digitizing toolbox see [50].

The estimation strategy summarized in Section 2.4 assumes that that there is a submanifold that underlies the motion regime. To estimate the submanifold, we process the estimated inertial marker trajectories to define body frames that are local to an animal. With the three-dimensional inertial trajectories of each of the markers and the newly defined body-fixed frame, we can then determine coordinates of the marker trajectories relative to the body-fixed frame. Specifically, the body-fixed frame was constructed using a set of markers placed on the back, as illustrated in Figure 2.5. The origin of the body-fixed frame is located at the intersection of the transverse, sagittal, coronal plane. The $b_1$ axis points anterior (toward the front of the specimen) in the direction of travel. The $b_2$ axis points
Figure 2.4: Coordinates of the trajectories in the inertial frame of the markers placed on a specimen. Note that while the body coordinates move smoothly, the markers on the shoulder, elbow, and wrist were momentarily occluded from either one or both of the cameras leading to small sharp edges in the shoulder, elbow, and wrist trajectories.
medial (away from the body to the side) to the left of the specimen. The $b_3$ axis points ventral (toward the stomach) of the specimen. The definition of the body frame follows the convention commonly used to define vehicle locomotion; the 1-axis is forward, the 2-axis points to the right, and the 3-axis points downward. In order to express the coordinates of

![Body Frame Coordinate System](image)

Figure 2.5: The body frame coordinate system defined on the specimen. Note that $b_1$ points in the anterior direction towards the head, $b_2$ points medially left, and $b_3$ points ventrally towards the stomach of the specimen. Note that all units are in mm.

the measurements in the body-fixed frame, the coordinates designated $\zeta_c$ from the inertial vectors were transformed into body frame components denoted $\zeta_b$ using the following rigid body transformation

$$
\begin{bmatrix}
\zeta_b \\
1
\end{bmatrix} =
\begin{bmatrix}
R^b_c & -R^b_cO_b \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
\zeta_c \\
1
\end{bmatrix}
$$

(2.11)

where $R^b_c$ is an orthonormal matrix defined by the contravariant transformation of the basis vectors in the inertial camera frame $c$ to basis vectors $b$ in the body frame, and $O_b$ is the origin of the body-frame relative to the inertial frame. Using these body-fixed coordinates, articulated motions of the skeletons of the specimens were recreated (see Figure 2.6). Figure
2.7 depicts several typical data trajectories of the elbow relative to a coordinate frame fixed to the body for a single specimen.

In this study, all the observed motions take place during ascent up an incline. We hypothesize that the gaits of the animals during ascent can be understood as taking place on an unknown submanifold \( Q \) that is homeomorphic to a known manifold \( S \). Most of the examples analyze the motions of the elbow and wrist of the specimen relative to its body-frame. We collect our sample outputs with the following correspondence,

\[
q := (\zeta_e^{(1)}, \zeta_e^{(2)}, \zeta_e^{(3)}, \zeta_w^{(1)}, \zeta_w^{(2)}, \zeta_w^{(3)})^T \in \mathbb{R}^6
\]  

(2.12)

with \( \zeta_e = (\zeta_e^{(1)}, \zeta_e^{(2)}, \zeta_e^{(3)})^T \) and \( \zeta_w = (\zeta_w^{(1)}, \zeta_w^{(2)}, \zeta_w^{(3)})^T \) as the locations of the elbow and wrist relative to the body frame, then we are interested in a mapping \( \gamma \) between manifold \( S \) and \( Q \), which from Section 2.2 satisfies

\[
\gamma : S^1 \rightarrow Q \subset \mathbb{R}^6
\]  

(2.13)

Note that each coordinate function \( \gamma^{(j)} : S^1 \rightarrow \mathbb{R} \) for \( j = 1, \ldots, 6 := d \).

Looking at the joint motion relative to the body, we see approximately closed orbital paths traced by the elbow and wrist joint for each gait. Therefore, we chose the underlying manifold to be the one-dimensional unit circle \( S = S^1 \).

We collect our sample inputs \( \{s_i\}_{i=1}^M \) on the manifold by noting the times at which each gait starts, \( t_0 \), and ends, \( t_0 + T \). These measurements are sampled uniformly in time over \( p \) frames. Given a sample measurement at time \( t_i \), our next sample is collected at time \( t_{i+1} = t_i + \delta t \)
where $\delta t = T/p$. The gait evolves along the manifold by the mapping $s : \mathbb{R}^+ \to S$ where

$$s(t) = \left( \cos\left(\frac{2\pi(t - t_0)}{T}\right), \sin\left(\frac{2\pi(t - t_0)}{T}\right) \right)$$

Figure 2.6: The articulated skeleton built from reconstructing the marker coordinates into body-fixed coordinates. The frame is parallel to anatomically defined planes that bisect the body.

We close this section by making some comments regarding assumptions about the underlying dynamics of the observed motion. This estimation only works for this set very specific
Figure 2.7: The elbow motion trajectory over several gaits. Note that the motion of the elbow and wrist follow a periodic trajectory in the body fixed frame. Note that all units are in mm.
dynamics, which, to some extent, is even dependent on the specific specimen itself. One of the difficulties in collecting this data came from getting repeatable motion from the reptile. Fairly different kinematics were observed from recordings where the specimen moved at significantly slower or faster speeds. This is mostly likely because the dynamics are dependent on the specimens movement speed. To address this issue, footage was only used if the specimen’s speed and the periodic motion of the joints were close to the mean values observed over all the motion.

### 2.5.3 Numerical Results

The approximate estimate $\gamma_{N,M}$ is obtained as the solution of the empirical risk minimization problem in Equation 2.7 for two choices of approximant subspaces in the plots that follow. The first method over $H^S_N$ carries out the minimization in terms of the characteristic functions $1_{S_{\ell,k}}$ defined over the uniform partition $S_N$ of the manifold $S^1$. The second method uses the approximate space $H^S_N$ defined over the kernel basis of the manifold $S^1$. Thus, in all the plots, the solutions obtained by minimization over $H^S_N$ are examples of the method studied in this chapter.

The solution over $H^S_N$ has been computed via the linear regression algorithm using the subspaces $H^S_N$ defined in Section 2.3.1 [51]. The centers of the basis that defines $H^S_N$ are given by the centers of the partition $S_{\ell,k}$ used in $H^S_N$. Figure 2.8 illustrates the approximants $\gamma^{(j)}_{N,M}$ for a fixed $j \in \{1, \ldots, d\}$ generated from the two methods over the low dimensional manifold $S$ with $N(\ell) = 32$ for both methods. In the figure, the $q_i$ of the sample $(s_i, q_i)$ are depicted as scattered data over one gait period. It is not clear from the simulations if the approximation obtained by the linear regression algorithm using a smooth kernel basis for $H^S_N$ converges to the solution obtained in this paper via empirical risk minimization over
the manifold $S$. It has been shown in [52] that if the basis for the closely related EDMD algorithm is selected to be the same as the characteristic functions that are in $H_N^S$, then the EDMD method reduces to the solution obtained via empirical risk. But the proof in [52] relies upon the particular $L^2_{\mu}$-orthogonality properties of the basis $H_N^S$ when they are used in the EDMD algorithm. We leave the question of the relationship between EDMD based approximations for more general choices of kernel bases and empirical risk minimization to a more detailed, future study.

Figure 2.8: Comparison of two methods of approximation of the unknown map $\gamma : S \to Q$ for a fixed number of samples, $M = 366$ and a dimension of the approximant space, $N = 32$. This corresponds to a uniform partitioning depth $\ell = 4$ The blue dashed line represents the approximation of the mapping of the wrist coordinates $\gamma^{(4)} = \zeta_{w_1}^1$ corresponding to the $b_1$ vector using the characteristic functions $1_{S_{\ell,k}}$ in the partition $S_\ell$, while the solid red line represents regression approximation using a kernel function basis in $H_N^S$. The known low dimensional manifold $S$ is depicted as a black solid line and the centers of each of the approximations are represented by the crosses on $S$.

In this study, we also have performed families of simulations to illustrate some qualitative
2.5. EXPERIMENTS AND NUMERICAL EXAMPLES

comparisons of convergence with respect to the numbers of centers for the two simulation methods. Figure 2.9 illustrates the effects that the number of centers has on the fidelity of the approximation. For both methods, it is clear that increasing the dimension number $N$ of partitions corresponds to a higher resolution that manages to capture changes in the data over smaller intervals. These figures also illustrate our claim that increasing the dimension $n$ decreases the error of bias term $C_1N^{-r}$ but with a fixed number of samples $m$ the error bound increases through the variance term $C_2N^{log(N)/M}$. We can see the variation appears to increase for the larger approximation with 64 centers (e-f). Since we do not have an analytic expression for the unknown mapping $\gamma$ in this actual experiment, it is impossible to verify rates of convergence directly. We leave this study to future research that uses analytic data instead of actual experimental observations.

The linear regression approximations that are generated from the exponential kernel, given by $k(\xi(s)) = e^{-\beta\|\xi-s\|^2}$, also depend on the hyperparameter $\beta$. Figure 2.10 illustrates the effects of the hyperparameter $\beta$ on the approximation $\gamma^{(j)}_{N,M}$ for the measured scalar value $\zeta^3_w$ which corresponds to the wrist positions from the body-fixed frame axis $b_2$. This coordinate appears to be fairly fixed over the specimen’s gait and may indicate a highly constrained degree of freedom to be considered for a potential control problem. However, careful consideration must be chosen for the hyper-parameters in order to identify such a constraint.

For an exponential basis, larger values of $\beta$ result in basis functions that decay rapidly from their centers, “sharpening” the curve. This results in a more oscillatory approximation. The approximate could therefore be misleading for various hyper-parameters adding more “dynamics” then are actually observed in the system. Depending on the hyperparameter $\beta$, highly oscillatory solutions can result for different choices of $\beta$, even for solutions over approximant spaces having the same dimension $N$. Such oscillations can be addressed in the linear regression methods or Gaussian process models by introducing a regularization
parameter \([51]\). Roughly speaking, the projection onto a lower dimensional subspaces is said to invoke the regularization that is observed in the empirical risk minimization approach. \([40]\). Thus, while the optimization of the empirical risk for functions in \(H^S_N\) yields nonsmooth estimates, there is no analog to the hyperparameter \(\beta\) and there is no concern for this oscillatory behavior.

In the next several figures, we use the approximations to give insight into measures with conditional dependencies. Specifically, we can compare the approximations built from specific specimens with the approximation built using all the specimen data. While the approximation from a specific specimen more accurately reflects the motion of that specific specimen, the approximation for all specimens, which now considers variability between specimens, gives us a sense of an average motion across specimens. Figure 2.11 shows the approximation generated over four male specimens as well as the average generated by concatenating all the specimen data together. Figure 2.11a illustrates the results for the partition basis, while Figure 2.11b shows approximation using the exponential kernel basis with \(\beta = 1\). To further this idea, the same process yielded approximations of motions generated from female specimens. Figure 2.12 displays the approximations from functions in \(H^S_N\) for the male specimens as the blue dashed line and females as the red solid.

As mentioned previously, the size of the female specimens were, on average, smaller than their male counterparts. Consequently, one would expect that the range of motion would be greater for the larger limbs of the male specimens. Looking at Figure 2.12, the range of the approximation is smaller in magnitude for the females than the males, which agrees with our intuition. Future studies may further investigate how the conditional probabilities given by the sex of the specimen affect the generated approximations.

The previous examples depict scalar-valued estimates over the manifold \(S\). Since \(\gamma(s) := (\zeta_e^T(s), \zeta_w^T(s))^T\) it is possible to trace trajectories of the approximations of the elbow and
2.5. Experiments and Numerical Examples

Figure 2.9: The approximations $\gamma_{N,M}^{(j)}$ for various dimension, $N$, of the approximant space. $N = 8$ is used for (a-b) $N = 32$ for (c-d), and $N = 64$ for (e-f). The estimates from $H_N^S$ are on the left (a,c,e) while estimates from $H_N^S$ are on the right (b,d,f). The estimates are mapped over the corresponding manifold given by the black circle and the observations are plotted as the red dots. It is evident the global behavior is consistent given the space dimension, $N$. Note that larger error due to sample variance is evident with the larger dimension, $N = 64$.
Figure 2.10: The approximations $\gamma_{N,M}^{(j)}$ for $j = 5$ corresponding to the wrist coordinate along the $b_2$ vector using $M = 254$ and $N = 16$ for various values of the kernel hyperparameter, $\beta$. The dimension of the approximant space $N$ is demonstrated by the centers marked as crosses on the manifold labeled $S$. For larger values of $\beta$, the evaluation of the kernel basis decays rapidly away from the kernel centers resulting in highly oscillatory approximations. The empirical risk minimization over $H_N^S$ has no corresponding hyperparameter.
wrist over the submanifold $S$. Three of the observations came from the coordinates of the elbow in the three-dimensional body frame while three additional observations came from the coordinates of the wrist in the same frame. To demonstrate these estimates can be used to recreate an average approximation in real-world coordinates for any general motion, the approximations were graphed corresponding to the same coordinate frame as the observations taken. Figure 2.13 illustrates the estimates from the elbow and wrist coordinates from the left arm for one specimen. One can see that the approximation acts as an interpolating closed curve from the sampled data. Note that these curves are homeomorphic to the manifold $S^1$.

Just as for the scalar case, a case study was performed to investigate the effects of the number of centers on the estimated output. Figures 2.14 and 2.15 shows the approximations for $j = 1, 2, 3$ in the basis defined by the body frame using the approximant spaces $H_N^{S}$ and $S_N^{S}$ respectively. From the figures, it is clear that, as seen in the scalar case, the behaviors of multidimensional approximation do not vary greatly based on the number of centers. A larger number of centers can capture changes in the approximation over smaller distances, but it can also lead to unwanted oscillations from a increase in the sample error for higher dimensional function spaces.

### 2.5.4 Statistical Estimates of Error in Stereo-Triangulation

To get a sense of the magnitude of the error induced by the measurement process, a dimensional error analysis was performed. The general strategy follows the approach outlined in [50]. Residuals given in pixel coordinates were measured in the studies to be less than one pixel. To get an understanding of the relationship between error in camera coordinates to errors in the observations, the trajectories measured over a number of runs in camera coordinates were perturbed with Gaussian noise having a standard deviation of one pixel. The
CHAPTER 2. DISTRIBUTION-FREE LEARNING THEORY FOR APPROXIMATING SUBMANIFOLDS FROM REPTILE MOTION CAPTURE DATA

Figure 2.11: The approximates of the observation of the elbow coordinate of the \( b_1 \) vector using 16 centers, \( \gamma_{N,M}^{(j)} \), from four male specimens using a basis of (a) approximation partitions in \( H_N^S \) and (b) exponential kernels in \( H_N^S \). Each approximation generated from an individual specimen is designated by the dashed lines while the average is designated by the solid line in both graphs. Note that the parameter \( M \) differs per specimen, as well as the average approximation, because each approximation uses a different number of samples.

Perturbed camera coordinates were used to reconstruct a perturbed set of three-dimensional coordinates of the elbow and wrist in the inertial camera frame. The standard deviation of the original collected data from the perturbed data was found to be \( \approx 0.086 \) mm over 5535 samples.

As opposed to the inertial frame, results in our studies use sampled data from a body-fixed frame constructed from the inertial markers. Consequently, we also related the error in the camera pixel coordinates to the error of coordinates expressed in the body frame. Based on the transformations between the coordinates in the camera frame to coordinates in the body frame, an exact expression for the error in the body fixed frame was difficult to quantify. Therefore, we resorted to comparing the deviation between measurements of the perturbed coordinates to the original measurements. Over the 5535 samples, the estimated error was \( \approx 0.638 \) mm.
Figure 2.12: The estimates of the observation of the elbow coordinate of the $b_1$ vector using 16 centers $\gamma_{N,M}^{(j)}$ over a number of specimens of different sexes. The solid red line represents the female specimens while the dashed blue line represents the males. Using a basis of approximation partitions in $H_N^S$. It would appear that the average male specimen has a larger, albeit slight, range of motion in the $b_1$ direction.
CHAPTER 2. DISTRIBUTION-FREE LEARNING THEORY FOR APPROXIMATING SUBMANIFOLDS FROM REPTILE MOTION CAPTURE DATA

2.5.5 Variations in Bone Lengths and Skin Artifacts

Estimates of limb lengths from the observations were made to compare average size and variations in skeletal morphology between the sexes. These estimates have been used to understand how well ideal constraints, such as the assumption of constant lengths of bones, are observed during typical motions. Note that these estimates are computed as the average in measured length between markers placed at the joints. These will differ from the physical bone length, but should give a reasonable estimate of the overall dimensions if the joints are ideal, the bones are rigid and skin movement is negligible. Results can be see in Table 2.1. From the table, it is clear that the female specimens are typically smaller than their male counterparts. Note that accurate placement of the markers was more challenging on the joints of the upper limbs and could be an influencing factor in the higher standard deviation of the limb lengths.

It should be noted that the measurements taken from imagery of the specimens are affected by skin movement artifacts which have been well documented in the literature [53]. In

Figure 2.13: Coordinates of the approximations of the motion corresponding to (a) the elbow coordinates, and (b) the wrist coordinates in the body frame using N = 16 basis functions. For each graph, one can see that the mappings act as an interpolating closed curve over the sample data. Note that all units are in mm.
Figure 2.14: The approximations of the elbow trajectory using (a) $\ell = 3$ (b) $\ell = 4$, (c) $\ell = 5$, and (d) $\ell = 6$ refinement depths of the space $S$. Note that the dimension of the space increases with increased refinement (a) $N = 8$ (b) $N = 16$, (c) $N = 32$, and (d) $N = 64$. The samples measured in the body frame are also depicted. Note the oscillatory behavior due to the increase in the variance for a larger number of basis functions. All units are in mm.
Figure 2.15: The approximations of the elbow motion using (a) $N = 8$, (b) $N = 16$, (c) $N = 32$, and (d) $N = 64$ centers with an exponential kernel function basis. The samples measured in the body frame are also depicted. Notice that the behavior does not change drastically while increasing the basis functions. This seems to indicate that the submanifold is well-approximated given a small number of basis functions. Note like the characteristic function basis, the increase in variation due to the increase in the second term of our bound, which we have labeled as the variance. All units are in mm.
2.5. EXPERIMENTS AND NUMERICAL EXAMPLES

Table 2.1: Measured Limb Lengths

<table>
<thead>
<tr>
<th>Limb</th>
<th>Male Population mean ± std. dev. (mm)</th>
<th>Female Population mean ± std. dev. (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Arm</td>
<td>5.18 ± 0.83</td>
<td>4.89 ± 0.84</td>
</tr>
<tr>
<td>Forearm</td>
<td>7.95 ± 0.67</td>
<td>7.05 ± 0.89</td>
</tr>
<tr>
<td>Thigh</td>
<td>9.34 ± 0.54</td>
<td>6.9 ± 0.62</td>
</tr>
<tr>
<td>Shin</td>
<td>11.15 ± 0.31</td>
<td>9.50 ± 0.16</td>
</tr>
</tbody>
</table>

In the ideal case for a rigid body, lengths between markers on the body are assumed to be invariant over the lizard’s gait. With the data measurement error assumed to be small in the inertial frame, we gauged the influence of possible skin artifacts by measuring the variation in the limb lengths. In principle, the variation in these quantities can be attributed to skin movement artifacts or nonideal skeletal joints. The results over the recordings of ten different specimens (6 male, 4 female) show a standard deviation of \( \approx 0.526 \) mm in the forearm length, 0.308 mm in the upper arm length, 0.376 mm in the thigh length, and 0.465 mm in the shin length. Figure 2.16 highlights the deviation between the measured limb lengths from an ideal rigid body assumption where the markers would remain a fixed distance relative to one another.

2.5.6 Computational Cost

Finally, we make a few observations regarding computational cost of the methods introduced in this chapter. It is well known that the analytic expression for the approximations obtained via the linear regression algorithm is given in terms of the pseudoinverse of data matrices. Assuming that the pseudoinversion is realized in terms of standard SVD or QR implementations of least-squares solution solvers, the numerical cost of these solutions will scale like \( O(N^3) \) in terms of the dimension of the space of approximants for a general choice of kernel basis. On the other hand, the empirical risk minimization solution is given in closed form...
Figure 2.16: Measured limb lengths compared to an ideal rigid limb length over a recording session for one specimen in the inertial frame for (a) the upper arm, (b) the forearm, (c) the thigh and (d) the shin. In an ideal case, the distances between markers on these limbs should remain fixed.
2.6. Future Work

This study has derived a formulation of the problem of estimating a submanifold representing animal motion from observations. The general form of the problem has been stated as a type of distribution-free learning over a smooth manifold. The paper defines a closed form expression for approximations $\gamma_{N,M}$ of the map $\gamma : S \to Q$ that defines the unknown manifold $Q$, and derives bounds for the rates of convergence over the manifold that are analogous to those obtained in corresponding distribution-free learning problems over Euclidean spaces. A number of future research questions have been suggested including 1) an analogous study of the rates of convergence of estimates obtained via the linear regression using more general kernel bases, 2) the derivation of adaptive approximation methods that are based on the construction of adaptive partitions of the manifold, and 3) the completion of additional experimental studies that investigate how the identified submanifolds vary with differences between species, specimen morphology within a species, and type of motion regime.

With respect to (3) above, the specimens in this experiment followed a particular motion regime, in this case ascending a plank. From examinations of the recorded video, it seems clear that no two gaits will match exactly. Attempts were made to constrain the specimen’s motion to follow a straight path up the incline, but aspects of the motion regime such as changes in speed, coordination of the limbs, slight veering from the path are difficult to control. Given our experimental process, it is difficult to say with certainty whether variations between gaits is caused by varied kinematics or by the measurement noise itself.
Consequently, these data-driven methods depend heavily on understanding what conditions or constraints are placed on the unknown probability measure generating our samples. For instance, rough symmetry (and anti-symmetry) are expected among various arms and legs, but these constraints are not enforced in the current formulation. Methods for enforcing such symmetry or anti-symmetry should be investigated when the current formulation is extended to full-body motion studies.

In addition to noise in the measurements and motion artifacts from movement of the skin, which has been previously discussed, it seems clear that there are other variables in this experiment which are difficult to understand, control, and quantify. Future work will aim to study these variations by looking at different motion regimes and seek to better control these variables between specimens.

\section{Appendix}

\section{Lebesgue and Sobolev Function Spaces}

The goal of this section is to provide a brief review of the Lebesgue spaces, which we denote \( L^p_{\mu}(\Omega) \), and the Sobolev Spaces that are denoted \( W^{r,p}(\Omega) \). Both are defined over some particular domain \( \Omega \). In order to provide some motivation toward understanding these function spaces, we first introduce the space of continuous functions over the domain \( \Omega \). We denote this space \( C(\Omega) \). The conventional norm \( \| \cdot \|_{C(\Omega)} \) on \( C(\Omega) \) is given by

\[ \| f \|_{C(\Omega)} = \sup_{x \in \Omega} | f(x) | \]
With this norm, the space is complete. In a complete metric space, Cauchy sequences, which are sequences in the space whose elements get closer to one another as the sequence approaches \( \infty \), converge to a limit in the space. We note that the distances between elements, and the convergence, are based on a metric induced by the norm. We can make this space a vector space with the following two operations.

\[
(f + g)(x) = f(x) + g(x) \quad f, g \in C(\Omega)
\]

\[
a(f(x)) = a f(x) \quad f \in C(\Omega), a \in \mathbb{R}
\]

Given these two operations, \( C(\Omega) \) becomes a normed linear space. As mentioned in the paper, our expected risk function is based on convergence using the \( 2 \)-norm, denoted \( \| \cdot \|_2 \) defined by

\[
\|f\|_2 = \left( \int_{\Omega} |f(x)|^2 d\mu \right)^{\frac{1}{2}}
\]

If we instead use the \( 2 \)-norm on the space, \( C(\Omega) \), it is well known that the space is no longer complete. However, the space \( L^2_\mu(\Omega) \), which consists of functions \( f \) where

\[
\|f\|_2 = \left( \int_{\Omega} |f(x)|^2 d\mu \right)^{\frac{1}{2}} < \infty
\]

is complete with the metric induced by the \( 2 \) norm. We can generalize this notion to the so-called \( p \)-norms. If we consider any positive real number \( p \), the \( \mu \)-square integrable Lebesgue functions, denoted \( L^p_\mu(\Omega) \), consist of spaces of measurable functions \( f \) for which

\[
\|f\|_p = \left( \int_{\Omega} |f(x)|^p d\mu \right)^{\frac{1}{p}} < \infty
\]

For functions in \( L^p_\mu(\Omega) \), we can use the norm to get a sense of the "size" of the function. Addi-
tionally, we can utilize topological notions in convergence proofs. For the space \( L^p_\mu(\Omega) \), if we have disjoint subsets \( \{A_k\}_{k=1}^N \) of \( \Omega \) where \( \Omega = \bigcup_{k=1}^\infty A_k \). If we define \( 1_{A_k} \) as the characteristic function of the set \( A_k \), the finite-dimensional space of functions \( f \) of the form

\[
f = \sum_{k=1}^N a_k 1_{A_k}
\]

are dense in \( L^p_\mu(\Omega) \).

While we know that the functions in this space exhibit these nice properties, we are often interested in the regularity of our functions in a particular space. It is well known that taking the derivative of functions in \( L^p_\mu(\Omega) \) can result in functions that are not in \( L^p_\mu(\Omega) \). Thus, we would like a space of functions which when we take the derivative still behave nicely. These are the Sobolev Spaces \( W^{r,p}(\Omega) \), which are functions with derivatives that are in \( L^p_\mu(\Omega) \). There are many different ways to define Sobolev Spaces over various domains.

Conventionally, a norm for a Sobolev Space \( W^{r,p}(\Omega) \) can be defined as

\[
\|f\|_{r,p} = \sum_{0 \leq |a| \leq r} \left( \int_\Omega |D^a f(x)|^p d\mu \right)^{\frac{1}{p}} < \infty
\]

where \( a \) is a multi-index. For functions in \( W^{r,p}(\Omega) \), we have most of the convenient properties of functions in \( L^p_\mu(\Omega) \). In fact, the \( L^p_\mu(\Omega) \) space can be thought of as a special case of a Sobolev space \( W^{r,p}(\Omega) \) when \( r = 0 \). We make a final remark that there are many important details about these spaces omitted because covering all the fundamentals would require a much larger discussion. A much more detailed motivation can begin from a more elementary functional analysis texts such as [54]. Specifically, details and motivation toward Sobolev Spaces can be found in [55].
2.9 Kernel Method Estimation

This section outlines the minimization over $\mathbb{H}_N^S$ and solves for $\gamma_{N,M}$ in terms of the kernel basis. For a particular dimension $j$, any estimate can be represented by a linear combination of $N$ radial basis functions with centers $\xi$. In other words, $\gamma_{N,M}^{(j)}(s) = \sum_{k=1}^{N} \alpha_k \mathcal{R}_k^S(s)$. Given our fixed basis, we see that equation 2.9 can be solved now by determining the optimal coefficients $\{\alpha_k^*\}_{k=1}^{N}$. Specifically, choosing a minimizing $\gamma_{N,M}^{(j)}$ is found by finding optimal coefficients $\alpha^*$ where

$$\alpha_1^*, \ldots, \alpha_N^* = \arg\min_{\alpha_1, \ldots, \alpha_N} \frac{1}{M} \sum_{i=1}^{M} \| q_i^{(j)} - \sum_{k=1}^{N} \alpha_k \mathcal{R}_k^S(s) \|^2_X$$

If we arrange the coefficients, $\alpha_k$, into a $N$-dimensional vector $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$, create an $m$-dimensional vector of measurement outputs of the $j^{th}$ dimension $q^{(j)} = [q_1^{(j)}, q_2^{(j)}, \ldots, q_M^{(j)}]^T$, and denote our $n \times m$ Kernel matrix $K$ where

$$K_{ij} = \mathcal{R}_i^S(s_j)$$

This matrix is a tall matrix with rows equal to the large number of samples $m$ and columns $N$ equal to the dimension of the function space. With these matrices we can reform our minimization with matrix representations to finding the optimal vector $\alpha$ in $\mathbb{R}^N$

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^N} \| q^{(j)} - K\alpha \|^2_X \quad (2.14)$$

for $X \approx \mathbb{R}^d$ is a Hilbert Space, which means we have an inner product related to the norm by

$$\|v\|^2_X = (v, v)_X$$
for any \( v \in X \). Our minimization can now be interpreted as minimizing the least squares error on an \( N \)-dimensional function space

\[
\alpha^* = \arg\min_{\alpha \in \mathbb{R}^N} (q^{(j)} - K\alpha, q^{(j)} - K\alpha)_X
\]

We can determine the optimal coefficients \( \alpha^* \) by taking the partial derivative and setting it equal to zero. The optimal vector \( \alpha^* \) is, therefore,

\[
\alpha^* = (K^T K)^{-1} K^T q^{(j)}
\]

As mentioned before, matrix inversion can prove to be computationally costly. Furthermore, ill-conditioned matrices can lead to poor estimates. We make one remark that the optimal approximation from \( H^S_N \), where we build estimates from piece-wise constants, does not face these same issues.

### 2.10 Proof of Theorem 2.1

**Proof.** When we define \( \gamma_{N,M}^{(j)} := \sum_{k=1}^{N(\ell)} \alpha_k 1_{S_{\ell,k}}(\cdot) \), where \( \alpha_k \) represent the weighted coefficients of our piece-wise constants \( 1_{S_{\ell,k}}(\cdot) \), we have the explicit representation of \( E_M^{(j)}(\gamma_{N,M}^{(j)}) \) given by

\[
E_M^{(j)}(\gamma_{N,M}^{(j)}) = \frac{1}{M} \sum_{i=1}^M \left( x_i^{(j)} - \sum_{k=1}^{N(\ell)} \alpha_k 1_{S_{\ell,k}}(s_i) \right)^2.
\]

We can also write this sum as

\[
E_M^{(j)}(\gamma_{N,M}^{(j)}) = \frac{1}{M} \sum_{k=1}^{N(\ell)} \sum_{i=1}^M \left( 1_{S_{\ell,k}}(s_i) x_i^{(j)} - \alpha_k 1_{S_{\ell,k}}(s_i) \right)^2.
\]
and this summation can be reordered as

\[ E^{(j)}_{N,M}(\gamma^{(j)}_{N,M}) = \frac{1}{M} \sum_{k=1}^{N(\ell)} E^{(j)}_{N,M}(\alpha_k) \]

\[ E^{(j)}_{N,M}(\alpha_k) = \sum_{i=1}^{M}(1S_k(s)x^{(j)}_i - \alpha_k1S_k(s))^2 \]

with each \( E^{(j)}_{N,M}(\alpha_k) \) depending on a single variable \( \alpha_k \). By taking the partial derivative \( D_{\alpha_k}(E^{(j)}_{N,M}) = 0 \), we see that for the optimal choice of coefficients \( \hat{a}_i \)

\[ \hat{a}_i = \frac{\sum_{i=1}^{M} 1S_i(s)x^{(j)}_i}{\sum_{i=1}^{M} 1S_i(s)} \quad \text{for} \quad i = 1, \ldots, N(\ell), \]

which establishes the form of solution given in the claim.

\[ \square \]

**2.11 Proof of Theorem 2.2**

We now turn to the consideration of the error bound in the theorem.

**Proof.** From the triangle inequality

\[ \|\gamma^{(j)}_{\mu} - \gamma^{(j)}_{N,M}\|_{L^2_\mu} \leq \|\gamma^{(j)}_{\mu} - \Pi^S_{N} \gamma^{(j)}_{\mu}\|_{L^2_\mu} + \|\Pi^S_{N} \gamma^{(j)}_{\mu} - \gamma^{(j)}_{N,M}\|_{L^2_\mu}, \]

we can bound the first term above by \( N^{-r} \) from the definition of the linear approximation space \( A^{r,2}(L^2_\mu(S)) \). The bound in the theorem is proven if we can show that there is a constant \( C_2 \) such that \( \|\Pi^S_{N} \gamma^{(j)}_{\mu} - \gamma^{(j)}_{N,M}\|_{L^2_\mu(S)} \leq C_2N(\ell)\log(N(\ell))/M \). We establish this bound by an extension to functions on the manifold \( S \) of the proof in [56], which is given for functions defined on \( \mathbb{R}^p \) for some \( p \geq 1 \). The expression above for \( \gamma^{(j)}_{N,M} \) can be written in the form
\(\gamma_{N,M}^{(j)} := \sum_{k=1}^{\ell} \alpha_k 1_{S_{\ell,k}}\), and that for \(\Pi_N \gamma^{(j)}_{\mu}\) can be written as \(\gamma^{(j)}_{\mu} := \sum_{k=1}^{\ell} \hat{\alpha}_{\ell,k} 1_{S_{\ell,k}}\). In terms of these expansions, we write the error as

\[
\|\Pi_N \gamma^{(j)}_{\mu} - \gamma^{(j)}_{N,M}\|_{L^2(S)}^2 = \sum_{\ell} (\alpha_{\ell,k} - \hat{\alpha}_{\ell,k})^2 \mu_S(S_{\ell,k}).
\]

Let \(\epsilon > 0\) be an arbitrary, but fixed, positive number. We define the set of indices \(\mathcal{I}(\epsilon)\) that denote subsets \(S_{\ell,k}\) that have, in a sense, small measure,

\[
\mathcal{I}(\epsilon) := \left\{ k \in \{1, \ldots, N(\ell)\} \mid \mu_S(S_{\ell,k}) \leq \frac{1}{8N(\ell) \bar{X}^2} \right\}
\]

where \(\bar{X} = \sup_{s \in S} \|\gamma(s)\|_X\). We define the complement \(\tilde{\mathcal{I}}(\epsilon) := \{ i \in \{1 \ldots N(\ell)\} \mid k \notin \mathcal{I}\}\), and then set the associated sums \(S_{\mathcal{I}} := \sum_{k \in \mathcal{I}} (\alpha_{\ell,k} - \hat{\alpha}_{\ell,k})^2 \mu_S(S_{\ell,k})\) and \(S_{\tilde{\mathcal{I}}} := \sum_{k \in \tilde{\mathcal{I}}} (\alpha_{\ell,k} - \hat{\alpha}_{\ell,k})^2 \mu_S(S_{\ell,k})\). The bound in Equation 2.2 follows if we can demonstrate a concentration of formula that has the form

\[
\text{Prob}\left(\|\Pi_N \gamma^{(j)}_{\mu} - \gamma^{(j)}_{N,M}\|_{L^2(S)}^2 > \epsilon^2\right) \leq b e^{c \epsilon^2/N(\ell)}
\]

for some constants \(b, c\). See [40, 56] for a discussion of such concentration inequalities. The fact that such a concentration inequality implies the bound in expectation in Equation 2.2 is proven in [56] on page 1311 for functions over Euclidean space. The argument proceeds exactly in the same way for the problem at hand by integration of the distribution function defined by Equation 2.15 over the manifold \(S\). To establish the concentration inequality, let
us define two events

\[ E_{\mathcal{I} + \tilde{\mathcal{I}}}(\epsilon) := \{ z \in \mathbb{Z}^M \mid S_{\mathcal{I}} + S_{\tilde{\mathcal{I}}} \geq \epsilon^2 \} , \]
\[ E_{\tilde{\mathcal{I}}}(\epsilon) := \{ z \in \mathbb{Z}^M \mid S_{\tilde{\mathcal{I}}} \geq \frac{1}{2} \epsilon^2 \} . \]

We can compute directly from the definitions of the coefficients \( \alpha_k, \hat{\alpha}_{t,k} \), and using the compactness of \( \gamma(S) \subset X \), that \( S_{\mathcal{I}} \leq \epsilon^2 / 2 \) for any \( \epsilon > 0 \). Since we always have

\[
\sum_{k \in \mathcal{I}} (\alpha_k - \hat{\alpha}_{t,k})^2 \mu_S(S_{t,k}) > \epsilon^2 - \sum_{k \in \mathcal{I}} (\alpha_k - \hat{\alpha}_{t,k})^2 \mu_S(S_{t,k}) > \frac{1}{2} \epsilon^2,
\]

we know that \( E_{\mathcal{I} + \tilde{\mathcal{I}}}(\epsilon) \subseteq E_{\tilde{\mathcal{I}}}(\epsilon) \) for any \( \epsilon > 0 \). If the inequality \( S_{\tilde{\mathcal{I}}} > \epsilon^2 / 2 \), then we know there is at least one \( \tilde{k} \in \tilde{\mathcal{I}} \) such that

\[
S_{\tilde{k}}(\epsilon) := (\alpha_k - \hat{\alpha}_{t,k})^2 \mu_S(S_{t,k}) > \frac{1}{2(\#\mathcal{I})} \epsilon^2 > \frac{1}{2N(\ell)} \epsilon^2.
\]

When we define the event \( E_i(\epsilon) := \{ z \in \mathbb{Z}^M \mid S_i(\epsilon) > \epsilon^2 / 2N(\ell) \} \) for each \( i \in \{1, \ldots, N(\ell)\} \), we conclude

\[
E_{\mathcal{I} + \tilde{\mathcal{I}}} \subseteq E_{\tilde{\mathcal{I}}}(\epsilon) \subseteq \bigcup_{i \in \tilde{\mathcal{I}}} E_i.
\]

By the monotonicity of measures, we conclude that \( \text{Prob}(\mathcal{I} + \tilde{\mathcal{I}}) \leq \sum_{i \in \tilde{\mathcal{I}}} \text{Prob}(E_i) \). But we can show, again by a simple modification of the arguments on pages 1310 of [56], that \( \text{Prob}(E_i) \lesssim e^{-cM\epsilon^2 / N(\ell)} \). The analysis proceeds as in that reference by using Bernstein’s inequality for random variables defined over the probability space \((S, \Sigma_S, \mu_S)\) instead of over Euclidean space.

\[ \square \]
Bibliography


[23] Sangok Seok, Albert Wang, Meng Yee Chuah, David Otten, Jeffrey Lang, and Sangbae Kim. Design principles for highly efficient quadrupeds and implementation on the mit


[30] Francesca Negrello, Manolo Garabini, Manuel G Catalano, Przemyslaw Kryczka,


Chapter 3

Koopman Methods for Estimation of Animal Motions over Unknown, Regularly Embedded Submanifolds

3.1 Introduction

Geometric methods that model the dynamics of complex multibody systems, such as those representing the motion of animals, are now well-established. [1, 2]. These techniques have been used in many cases to study the motion of an animal that is modeled as a collection of rigid bodies defined by bones that are interconnected by ideal joints. In these formulations, motions take values in certain types of smooth manifolds, and forward kinematic maps are defined over these manifolds. One attractive feature of these formulations is that such formulations are coordinate-free and intrinsic. [1] Still, for many animal motion studies, the dimension of the full configuration manifold can be large. The study of human motion, bat motion, or the flight of birds are just a few such cases [3, 4]. For these animals, as well as others, it is often important to be able to identify submanifolds that provide accurate representations of particular motion regimes.

This chapter extends our recent work in [5]. In the paper [5], the authors explore methods...
to generate approximations of mappings over certain smooth manifolds in applications that estimate the motion of reptiles. The analysis in [5] determines error bounds for the estimation of functions over the motion manifolds. However, in [5] it is assumed that a known “template manifold” underlies a given motion regime. This paper seeks to generalize these results to the case when the underlying motion manifold is unknown.

As we explain more fully below, this paper defines and estimation algorithm that does not require explicit knowledge of the unknown manifold. However, despite the fact that the manifold is unknown, strong direct theorems of approximation are derived that tell how fast estimates converge over the unknown manifold. The results are derived using Koopman theory, a popular operator-theoretic approach to the study of nonlinear dynamic systems. The analysis in this chapter has the added benefit that it strengthens or refines convergence results in recent papers on the approximation of Koopman operators.

In this chapter, we assume that the motion of the animal is governed by the uncertain equations

\[ x_{i+1} = f(x_i), \]
\[ y_{i+1} = G(x_{i+1}), \]

where \( X := \mathbb{R}^d \) is the state space, \( Y := \mathbb{R}^n \) is the output space, the function \( f : X \to X \) that determines the dynamics is unknown, and the observable function \( G : X \to Y \) is unknown. It is further assumed that this evolution law exhibits some additional “hidden” structure. Specifically, it is assumed that there is a compact, smooth Riemannian manifold \( Q \) that is regularly embedded in \( X \) that is invariant under the flow generated by Equation 3.1. The task is to collect finite samples \( \{(x_{k_i}, y_{k_i})\}_{i=1}^M \in Z = X \times Y \) of state and output pairs at discrete times \( T_M := \{k_i \in \mathbb{N} \mid 1 \leq i \leq M\} \) and use them to build an estimate of the
observable function $G$. Here $\mathbb{N}$ is the collection of nonnegative integers. Is it possible to choose the $k_i$ sequentially so that $k_1 = 1, k_2 = 2, \ldots$ etc, but the approach in this chapter does not require uniform sampling.

Examples 3.1.1 and 3.1.2 described below show that the model class described above is very rich: there are a wide variety of models of the form above that are obtained when geometric methods [1, 2] are discretized in time.

3.1.1 Example 1

This first example illustrates how quite general motion estimation problems of the type governed by Equation 3.1 may arise. For a wide variety of animal models, it is popular to create models that consist of rigid bodies connected by ideal joints. Such a model for human motion can be constructed by choosing as states the rigid body motion of a core body, spherical joints for the hips and shoulders, and revolute joints for all the remaining joints of the body. For human motion, the full configuration manifold might be taken as

$$
\hat{X} := \underbrace{SE(3) \times SO(3)}_{\text{core body}} \times \underbrace{S^1 \ldots S^1}_{\text{left leg}} \times \underbrace{SO(3) \times S^1 \ldots S^1}_{\text{right leg}} \times \underbrace{SO(3) \times S^1 \ldots S^1}_{\text{right leg}}.
$$

In the above expression $SE(3)$ and $SO(3)$ designates the special Euclidean group and special orthogonal group respectively in three dimensions. It is known by the Nash embedding theorem that $\hat{X} \subset X = \mathbb{R}^d$ for a $d$ large enough. Standard geometric methods as in [1, 2] describe how equations of motion generate a flow $\{S(t)\}_{t \in \mathbb{R}^+}$ in continuous time over $\hat{X}$. By discretizing the flow at times $t_p := p\Delta$ with $\Delta$ a fixed time step, we obtain the evolution law $x_{p+1} = S(\Delta)x_p$. The discretization yields a discrete dynamics of the form in
Equation 3.1 with $f = S(\Delta) : \tilde{X} \subset X \rightarrow \tilde{X} \subset X$. If we suppose that the output space $Y$ consists of the inertial positions of $j$ distinguished points on the body, the kinematics map $G : \tilde{X} \rightarrow \mathbb{R}^{3j} := Y$. In motion capture studies that seek to study certain motion regimes, it is then natural to hypothesize an existence of an unknown submanifold $Q \subseteq \tilde{X} \subseteq X$. The estimation problem is then to use the measurements $\{ (x_{k_i}, y_{k_i}) \}_{i=1}^{M}$ to build approximations of $G : Q \rightarrow Y$, where manifold $Q$ is unknown. The strategy in this chapter can be realized, at least in principle, using RKH spaces with bases that are the tensor product of bases for functions over $SO(3)$ or other homogeneous manifolds. Such bases are studied, for example, in [6, 7].

### 3.1.2 Example 2

For this chapter, we perform numerical studies of some simpler problems that are common in motion modeling. Instead of modeling or studying motions in the full configuration space above, it is often of interest to understand the motion for just an appendage, or across a joint. For simplicity, we carry out numerical studies in this chapter when $Q$ is a unknown Riemannian manifold $Q$ that is regularly embedded in $X := \mathbb{R}^d$. Again, if we assume that $Y$ is defined in terms of the inertial positions of $j$ points on the body, we have $G : X \rightarrow Y$ with $Y := \mathbb{R}^{3j}$. The estimation task is to use samples $\{ (x_{k_i}, y_{k_i}) \}_{i=1}^{M} \subset X \times Y$ to build approximations of $G$, even though the manifold $Q$ is unknown.

### 3.1.3 Related Research

There are a wide variety of kinematic studies that evaluate animal motion [8, 9, 10]. Recent studies have focused on discovering fundamental principles of locomotion [11] and lower-dimensional representations of motion models across a vast number of different animals.
3.1. Introduction

[12]. However, applying learning algorithms to obtain kinematic models has been primarily applied to human motion [13, 14]. Recent learning methods have utilized neural networks and deep learning to synthesize and predict human motion [15, 16]. To address the time-dependent nature in synthesizing human motion, recurrent neural networks (RNNs) have been implemented [17, 18].

Koopman operator theory has been one of the primary tools that is used to pose learning problems over time-dependent data, especially with strong nonlinear dynamics are at play [19]. Excellent texts such as [20, 21, 22] as well as a number of survey articles including [23] give ample motivation for the study and use of Koopman theory. References [24, 25, 26] hold that the prevailing viewpoint toward dynamical systems throughout the twentieth century came from that of Poincare's geometric interpretation of the evolution of the states. However, another approach to studying nonlinear dynamics was introduced as early as 1931. Inspired by the operator theoretic approach taken in quantum mechanics, Bernard Koopman began investigating the spectral properties of the composition operator toward the evolution of Hamiltonian systems in classical mechanics [27]. Rather than examining the possibly nonlinear evolution of the states, Koopman theory focuses on the dynamics of the composition of observables along the evolution of states. As mentioned in [26] with increases in computational power, a movement toward high-dimensional data, and breakthroughs in the study of linear dynamics, there has been a surge of renewed interest in the Koopman operator and its spectral decomposition over the last decade. An extensive overview of the spectral properties of both the discrete and continuous Koopman operator have been discussed in [22], [23]. Additionally, a more detailed background on Koopman operator theory will be given in 3.2.3.

As explained in many other studies on Koopman theory such as [26] [22] [28], the linear system that governs the dynamics of the observables is infinite-dimensional. Thus, there
is a fundamental trade-off in utilizing Koopman theory on a typically finite-dimensional nonlinear system. Consequently, Koopman theory often requires a careful consideration of more practical finite-dimensional approximations of the non-tractable infinite-dimensional operator. Specifically, many studies focus on the convergence of approximating quantities associated with the Koopman operator such as the Koopman modes and eigenfunctions [26] [29]. Convergence studies of Koopman theory have also been heavily linked to dynamic mode decomposition (DMD)[30]. A number of studies have used the Koopman operator framework to examine the rates of convergence of the EDMD solution [31] [32] of an Koopman invariant finite dimensional subspace to the Galerkin approximation of the Koopman operator. Additional convergence studies include those of the infinitesimal generator [33], as well as convergence of eigenfunctions of the Koopman operator [28]. The behavior of these eigenfunctions under conjugate transformations were examined in [34]. As powerful and promising as these methods are, none of these references study rates of convergence of their estimates.

To clearly explain the contributions of this paper, we review a few of the most basic definitions that arise in Koopman theory in subsection 3.1.4, and subsequently review the structure of direct approximation theorems in subsection 3.1.5. Finally, these two subsubsections are employed in subsection 3.1.6 to be precise about the nature of approximation theorems for Koopman operators in the recent literature.

### 3.1.4 Deterministic Koopman Operators

A more thorough discussion of the Koopman operator, tailored to the cases considered in this chapter, is given later in subsection 3.2.3 when we introduce and discuss the EDMD method. In general, Koopman theory has been developed for deterministic and stochastic
systems, allowing for either evolutions in discrete or continuous time in both cases. This paper only derives error estimates for deterministic systems, so we restrict attention to this case. Consider first the case when we have a discrete time evolution is governed by the recursion

$$x_{i+1} = f(x_i),$$

with $f : X \to X$ and $X$ a metric space. In this case the Koopman operator $U_f$ is ordinarily defined via the identity $U_f g := g \circ f$. Thus, for the discrete time evolution, the Koopman operator is a constant operator on the state space. A study of the fundamental properties of $U_f$ is given in well-known texts such as [20]. See [20] for a complete description of how this operator yields important insights into the structure of nonlinear systems. We refer to the numerous examples cited above to see how the Koopman operator can be exploited to generate reduced models of complex nonlinear phenomena or to study the qualitative nature of nonlinear systems.

The definition of the Koopman operator for a deterministic system that evolves in continuous time is a bit different. We suppose that the dynamics over the metric space $X$ is defined in terms of a semigroup $\{S(t)\}_{t \geq 0}$ on $X$ defined for continuous time $t \in \mathbb{R}^+$. In this case the time dependent Koopman operator $U^\tau$ is defined for each $\tau \geq 0$ from the identity $(U^\tau g)(\xi) := g(S(\tau)\xi)$ for each initial condition $\xi \in X$. Again, the fundamental properties of the operator $U^\tau$ can be found in [20]. Note that if the semigroup is discretized with a constant step size of $h > 0$, then the definition of the Koopman operator for the continuous time evolution is equivalent to the discrete time definition when we identify $f(\xi) := S(h)\xi$ and set $U^\tau := U^{kh} = (U_f)^{(k)}$. 

3.1. Introduction
3.1.5 Direct Approximation Theorems

The convergence of approximations of the Koopman operator in the recent literature has been studied in both uniform operator topology, the strong operator topology, and the weak operator topology. We only discuss the uniform and strong topology in the review. Let \((W, \| \cdot \|_W)\) be some normed vector space and denote by \(\mathcal{L}(W)\) the collection of bounded linear operators that map \(W\) into itself. By definition, a sequence of bounded linear operators \(\{T_n\}_{n \in \mathbb{N}}\) converges in the uniform operator topology if it holds that \(\|T_n - T\|_\infty \to 0\) as \(n \to \infty\) where the operator norm \(\| \cdot \|_\infty\) is defined as

\[
\|T\|_\infty := \sup_{w \in W, w \neq 0} \frac{\|Tw\|_W}{\|w\|_W}.
\]

The sequence \(\{T_n\}_{n \in \mathbb{N}} \subset \mathcal{L}(W)\) converges to \(T \in \mathcal{L}(W)\) in the strong operator topology if for any \(w \in W\) we have \(\|T_n w - Tw\|_V \to 0\). The uniform operator topology is stronger (finer) than the strong operator topology since convergence of \(\|T - T_n\|_\infty \to 0\) implies that \(\|T w - T_n w\|_W \leq \|T - T_n\|_\infty \|w\|_W \to 0\). It is known that an operator \(T\) is compact if and only if there is a sequence of finite dimensional linear operators \(\{T_n\}_{n \in \mathbb{N}} \subset \mathcal{L}(W)\) that converge to \(T\) in the operator norm. It is a rather special case that Koopman operators are compact, but this case is studied under certain circumstances, for example in references [28] and [35], which are discussed more fully below.

There is a standard framework in approximation theory for characterizing the rates of convergence of approximations that can be used to study the strong convergence of operators. This setup makes use of another space \(S \subseteq W\). We use the symbol \(S\) since it has a stronger topology than that on the space \(W\), which has the weaker topology. It is assumed that \(S\) is dense in \(W\) and that there is a constant \(C > 0\) such that \(\|w\|_W \leq C \|w\|_S\) for all \(w \in S\). Approximations are constructed in terms of a family \(\{W_n\}_{n \in \mathbb{N}}\) that is a sequence of finite
3.1. Introduction

dimensional subspaces whose union is dense in $W$. We suppose that $P_n : W \to W_n$ is a family of uniformly bounded mappings, each $P_n$ onto $W_n$. We say that the pair of sequences $(W_n, P_n)_{n \in \mathbb{N}}$ and the pair of spaces $(W, S)$ satisfies a direct theorem of approximation with approximation rate $n^{-r}$ if there is some other constant $\tilde{C}$ such that we can write

$$
\|w - P_n w\|_W \leq \tilde{C} n^{-r} \|w\|_S
$$

for all $w \in S \subseteq W$. This inequality can be viewed as a statement that functions in the “smaller, more regular” space $S$ converge at a rate $r$ that measures regularity. Often in machine learning theory or approximation theory the membership in the space $S \subseteq W$ is referred to as a prior on a function $f \in W$: it is information that conveys that approximations of such functions converge at a certain rate. The study of how priors and approximation operators $P_n$ can be selected to yield direct theorems of approximation is one of the central problems of approximation theory. This problem has been studied in great detail, see [36, 37], for a wide collection of choices of the spaces $W$ and $S$. In this chapter we use this construction to study the rates of convergence of approximations of the Koopman operator in the strong operator topology. We apply the analysis above when $w$ in equation 3.3 is chosen to be $w := U_f g$, with $U_f$ the Koopman operator of a discrete deterministic system. Thus, we are particularly interested in how certain priors, or choices of $S$, affect the rates of convergence of approximations of the Koopman operator.

3.1.6 Approximations of Koopman Operators

As mentioned above, one of the compelling reasons that motivates the use of Koopman operators is the fact that, while the original dynamics is determined by a nonlinear function $f$ that acts on a finite dimensional space, the Koopman operator is linear and acts on an infinite
dimensional space of functions. Any practical use of a Koopman operator must address questions related to its approximation. Of the recent articles \([28, 28, 32, 32, 34, 35, 38]\) that study approximations of the Koopman operator under various working assumptions, perhaps the most relevant to this paper are \([28, 35, 38, 39]\). We review each of these papers in some detail below to make clear the often subtle difference among the references.

Reference \([28]\) is a detailed analysis of the convergence of approximations \(U_{m,n}\) determined by the EDMD algorithm of the Koopman operator \(U_f\). The integers \(n, m\) denote the number of samples \(m\) and the dimension \(n\) of a subspace of approximants \(\mathcal{H}_n\). The convergence analysis is carried out in the Lebesgue space \(L^2_\mu(M)\) where \(\mu\) is a measure on the manifold \(M\). The paper studies carefully the relationship of approximations of the EDMD algorithm to projections \(P_{\mu,n}\) of the Lebesgue space \(L^2_\mu(M)\) onto subspaces of approximants \(\mathcal{H}_n \subset L^2_\mu(M)\).

A few different types of convergence are studied. Let \(P_{\mu,n} : L^2_\mu \to H_n\) denote the \(L^2_\mu\)-orthogonal projection of \(L^2_\mu\) onto \(\mathcal{H}_n\), and define the projected operator \(U_n := P_{\mu,n} U_f|_{H_n}\) with \(U_f|_{H_n}\) the restriction to the space of approximants \(\mathcal{H}_n\). The paper shows in Theorem 2 that \(\text{for a fixed dimension } n \text{ of the approximant subspace } \mathcal{H}_n\), the EDMD approximations \(U_{m,n}\) converge to the projected operator \(U_n := P_{\mu,n} U|_{H_n}\) as \(m \to \infty\) in any operator norm, which are equivalent due to the assumption that \(n\) is fixed and finite. The paper also shows that \(P_{n,\mu} U_f P_{\mu,n} \to U_f\) as \(n \to \infty\) in the strong operator topology over \(L^2_\mu(M)\). The paper does not address how the choice of different kernels influences or affects the choice of the space \(\mathcal{H} \subset L^2_\mu(M)\), nor how the choice of such a prior determines the rate of convergence of the approximations.

An altogether different strategy is used to derive rates of convergence certain types of Koopman operators in \([35]\). In contrast to \([28]\) above, a regularized kernel Koopman operator \(U^\epsilon\) is introduced in terms of an integral operator that depends on covariance operators associated with the inputs and outputs. In this notation \(\epsilon\) is the regularization parameter. The kernel
Koopman operator $\mathcal{U}^q$ acts on functions. Using $m$ samples, regularized approximations of the Koopman operator $\mathcal{U}^\epsilon$, as well as the associated regularized Frobenius-Perron operators, are defined in terms of integral operators whose integrands are the kernel of an RKH space defined over some compact set. Theorem 3.14 of [35] derives rates of convergence of empirical estimates $\hat{\mathcal{U}}^\epsilon_m$ to the regularized operators $\mathcal{U}^\epsilon$. It is shown that the operator norm error $\|\mathcal{U}^\epsilon - \hat{\mathcal{U}}^\epsilon_m\| = O_p(m^{-1/2}\epsilon^{-1})$ as the number of samples $m \to \infty$. In brief comments following the theorem, it is noted that this bound implies that the the regularization parameter should be chosen to be varying as the number of samples $m$ increases. As above, this convergence guarantee does not explain choice of kernels affects the rate of convergence for the choice of different kernels or priors.

Reference [38] studies kernel analog filtering (KAF) for certain types of forecasting problems. This paper considers deterministic dynamical systems that evolve in continuous time over a set $\Omega$. Moreover, it is assumed that the dynamical systems are measure preserving. The kernel analog filtering problem seeks a function $f_\mathcal{H}$ that minimizes $\| f_\mathcal{H} \circ X - \mathcal{U}^hY \|_{L_\mu^2(\Omega)}$ with $\mu$ the invariant measure of the underlying dynamical system. Approximations in this chapter are studied in terms of eigenfunctions of a compact, self-adjoint operator associated with an integral operator having as integrand a kernel of an RKH space. For theoretical purposes, subspaces of approximants are defined in terms of the span of a finite collection of $n$ eigenfunctions. Finite dimensional approximations of the function $f_\mathcal{H}$ in the paper can be understood as an approximation of a Koopman operator, see Equation (19) of that reference. The convergence of approximations of the Koopman operator are studied in the Lebesgue space $L_\mu^2(\Omega)$ where $\mu$ is the measure that is invariant under the semigroup that determines the dynamical system. Theorem 11 proves convergence of the estimates as the dimension $n$ of the eigenspace used for approximations approaches infinity. In this sense Theorem 11 of [38] qualitatively resembles the convergence result in Theorem 2 in [28]. Theorem 14 of [38]
additionally shows that empirical approximations, based on \( n \) eigenfunctions and \( m \) samples, converge to a regression function as \( m, n \to \infty \). This conclusion is stronger than the analysis in [28] above. However, this paper does not discuss how the choice of the kernel, and a set of priors, affects the rate of convergence.

As opposed to the previous studies, reference [39] does examine rates of convergence for direct approximations of the Koopman operator under a variety of situations. However, the approximations are generated by finite dimensional bases of wavelets, multi-wavelets, and eigenfunctions of the Koopman operator. Such constructions are most useful when the finite-dimensional bases are "well adapted" to the dynamics under study. This study, in contrast, frames the approximation over scattered data and utilizes the many zeros theorem for functions belonging to a particular Sobolev Space. The studies [5][40] examine the rates of convergence of the EDMD estimate by equating it to the empirical risk minimizer as examined in [41] for a particular basis defined by uniform discretization of the submanifold. While these estimates could be calculated efficiently without matrix inversions, the estimates are discontinuous. In this study, we consider smoother estimates generated from reproducing kernel Hilbert (RKH) spaces. This method is well-suited to taking scattered samples over unknown or irregular domains.

### 3.1.7 Overview of New Results

To be precise about the contributions of this paper to the rapidly growing field of Koopman theory, we briefly summarize the overall strategy taken in this chapter. This overall strategy is common to many of the approaches outlined in References [23, 26, 28, 32, 34, 35, 38], but, as mentioned in the introduction, we focus on determining how to ensure error bounds not found in these references. Since \( Q \subset X \) is unknown, but the ambient space \( X \) is known, a
kernel function $\mathcal{K} : X \times X \rightarrow \mathbb{R}$ is selected that defines a reproducing kernel Hilbert space (RKHS) $H$ of real-valued functions over the known set $X$. A kernel subsection, or kernel basis function, centered at a point $x \in X$ is defined in terms of the kernel as $\mathcal{K}_x(\cdot) := \mathcal{K}(x, \cdot)$. Finite dimensional spaces of approximants $H_M$ are defined in terms of bases $\{\mathcal{K}_{\xi_{M,i}}\}_{i=1}^M$ over the full set $X$, which depend on some set of samples $\Xi_M := \{\xi_{M,i} := x_{k_i} \mid 1 \leq i \leq M\}$ collected along trajectories of the uncertain system at the discrete times $T_M := \{k_i \mid 1 \leq i \leq M\}$. Hence, we have the spaces of approximants $H_M := \text{span}\{\mathcal{K}_{\xi_{M,i}} | \xi_{M,i} \in \Xi_M\}$. Approximations of the observable function $G$ for the uncertain system are built using the finite dimensional spaces $H_M$ and the $H$-orthogonal projection $\Pi_M : H \rightarrow H_M$. Data-dependent approximations $U^M_f$ of the Koopman operator $U_f$ are defined by the identity $U^M_f g := \Pi_M((\Pi_M g) \circ f)$, which has a coordinate representation shown in Equation 3.10 in subsection 3.3.4. By inspection of the coordinate expression, it is clear that the approximation $U^M_f$ of the Koopman operator $U_f$ can be constructed from the samples $\{(x_{k_i}, y_{k_i})\}_{i=1}^M$ and the choice of finite dimensional space $H_M$. This paper introduces an approximation method in which the distribution of the samples $\{x_{k_i}\}_{i=1}^M$, the definition of the space $H_M$, and the choice of the kernel $\mathcal{K}$ determines the rate of convergence of $U^M_f g \rightarrow U_f g$.

There are three distinct contributions in this chapter to the state-of-the-art in estimation methods for the system in Equations 3.1 and 3.2.

(1) **Reduced Order Models of Animal Motion**: Although Koopman methods have been studied for a wide variety of application areas, this paper is the first systematic study of the method for characterizing animal motions over submanifolds. The approach should be of general interest to those studying reduced order models of animal motions. The paper describes a rigorous formulation of a general method for approximating such motions, and the theory includes estimates of the rates of convergence in various function spaces as the number of samples $M$ increases. Such estimates are not available in the literature on motion
estimation such as in [13, 14, 15, 16, 17, 18].

(2) **Convergence Rates for Koopman Approximations:** In addition, this paper describes general, strong rates of convergence of data-dependent approximations of the Koopman operator. This result is of interest in its own right. Full details on the derivation and interpretation of these rates is given in the discussion of Theorem 3.2 below, but we briefly summarize the structure and novelty of these rates here. The primary result of this paper defines a regularity assumption in the framework of Koopman theory that is particularly powerful and enables determination of strong error bounds. The regularity assumption states that if the pullback space \( f^*(\mathcal{H}) := \{ g : Q \to \mathbb{R} \mid g = h \circ f, h \in \mathcal{H} \} \) is continuously embedded in a Sobolev space, then we get strong rates of convergence. These rates of approximation are bounded by \( O(h_{\Xi,M,Q}^r) \) where \( h_{\Xi,M,Q} \) is the fill distance of samples \( \Xi \) in the manifold \( Q \), where \( h_{\Xi,M,Q} := \sup_{q \in Q} \inf_{\xi \in M,i} d_Q(q, \xi_{M,i}) \) where \( d_Q \) is the metric over \( Q \) and \( r \) is an index that measures smoothness in a Sobolev space. Thus, even though the manifold \( Q \) is unknown, we obtain a geometric characterization of the error. The convergence depends on how quickly samples fill up the unknown manifold. These rates for the Koopman approximation problem are qualitatively similar to rates of convergence that are familiar in regression in Euclidean spaces [42], interpolation over manifolds [7, 43], Chapter 10 of [37], and ambient approximation over manifolds [44]. None of the recent studies of Koopman approximations [28, 32, 34, 35, 38] derive such rates of convergence in terms of fill distances.

(3) **Rates of Convergence of the EDMD Algorithm** Finally, this paper defines a general data-dependent approximation of the Koopman operator as it acts on a variety of function spaces. In some instances, this data-dependent operator can be interpreted as a specific implementation of the extended dynamic mode decomposition (EDMD) algorithm. With this identification it is possible to derive rates of convergence for the EDMD algorithm in certain types of function spaces. This result is described in Theorem 3.3 and also has no precedent in the recent studies of the EDMD algorithm, as in [32, 34].
3.2 Background Theory

3.2.1 Symbols and Nomenclature

in this chapter, \( \mathbb{R} \) is the set of the real numbers. We write \( a \lesssim b \) to mean that there is a constant \( c \) that does not depend on \( a, b \) such that \( a \leq cb \), and \( \gtrsim \) is defined similarly. We let \( X = \mathbb{R}^d \) be the input space, and \( Y := \mathbb{R}^n \) be the output space. We assume that \( Q \) is a smooth, compact Riemannian manifold that is regularly embedded in \( X = \mathbb{R}^d \). The symbol \( C(Q, Y) \) represents the space of continuous, \( Y \)-valued functions over the manifold \( Q \), equipped with the usual supremum norm. We set \( L^2_{\mu}(Q, Y) \) to be the usual Lebesgue space of \( Y \)-valued, \( \mu \)-integrable functions over the manifold \( Q \), with \( \mu \) a measure on the manifold. This paper also employs the Sobolev space \( W^{s,p}(Q) \) over the compact Riemannian manifold \( Q \), which is defined intrinsically in [7] or extrinsically in [43]. Roughly speaking, for integer \( s > 0 \) this space consists of all functions in \( L^p_{\mu}(Q) \) that have derivatives through order \( s \) that are elements of \( L^p_{\mu}(Q) \). The Sobolev spaces for real smoothness parameter \( s > 0 \) are obtained as interpolation spaces from those that have integer order. [7] We write \( \mathbb{W}^{s,p}(Q) \) for the \( Y \)-valued Sobolev spaces, \( \mathbb{W}^{s,p}(Q) := (W^{s,p}(Q))^n \).

3.2.2 Reproducing Kernel Hilbert Spaces

A number of constructions in reproducing kernel Hilbert spaces are used in this chapter, and the reader is referred to [45, 46] for a good introduction that provides the needed background for the necessarily brief summary given here. Throughout the paper \( \mathcal{K} : X \times X \to \mathbb{R} \) will be an admissible kernel that defines a native space \( \mathcal{H} \) of real-valued functions over \( X \). We always assume that the kernel \( \mathcal{K} \) is continuous, symmetric, and strictly positive definite in
this chapter, which means that for any finite set $\Xi := \{\xi_{M,i} \mid 1 \leq i \leq M\}$ of distinct points, the kernel matrix $[\mathcal{K}(\xi_{M,i}, \xi_{M,j})] := \mathbb{K}(\Xi_M, \Xi_M) \in \mathbb{R}^{M \times M}$ is a positive definite matrix. Kernels of this type include the exponential kernel, inverse multiquadric kernel, or Matern-Sobolev kernel, among others [37]. The space $\mathcal{H}$ is defined as $\mathcal{H} := \text{span}\{\mathcal{R}_x \mid x \in \Omega\}$ where the closure is taken with respect to the inner product that is defined as $(\mathcal{R}_x, \mathcal{R}_y)_\mathcal{H} = \mathcal{K}(x, y)$ for all $x, y \in X$. The RKHS $\mathcal{H}$ is characterized by the fact that it satisfies the reproducing property: we have $(\mathcal{R}_x, g)_\mathcal{H} = g(x)$ for all $x \in X$ and $g \in \mathcal{H}$.

in this chapter, samples are generated along the trajectories of a discrete dynamical system over a subset $Q \subset X$ defined in terms of an unknown function $f$, and the function $f$ defines an RKHS that is important for our analysis. For a mapping $f : Q \subset X \to X$, we define the pullback space $f^*(\mathcal{H})$ to be the set $\{h : Q \to \mathbb{R} \mid h = g \circ f, \text{ for some } g \in \mathcal{H}\}$. The pullback space $f^*(\mathcal{H})$ is itself an RKHS, with its kernel $\mathcal{K}^* : Q \times Q \to \mathbb{R}$ defined as $\mathcal{K}^*(q_1, q_2) = \mathcal{K}(f(q_1), f(q_2))$ for all $q_1, q_2 \in Q$. [46] The norm on the pullback space that is induced by the kernel $\mathcal{K}^*$ is equivalent to $\|h\|_{f^*(\mathcal{H})} := \inf\{\|g\|_{\mathcal{H}} \mid h = g \circ f\}$. For a fixed $f : Q \to Q \subset X$ the Koopman operator $U_f$ is defined as $U_f g = g \circ f$: this definition makes sense for scalar-valued functions $g \in \mathcal{H}$ as well as vector-valued functions $g \in \mathcal{H}^n$. We define the trace or restriction operator $Tg := g|_Q$. The space of restrictions $\mathcal{R} := T(\mathcal{H})$ is itself an RKHS space. It is induced by the kernel $\mathcal{K} : Q \times Q \to \mathbb{R}$ that is the restriction of the kernel $\mathcal{K}$, so that $\mathcal{K}(x, y) := \mathcal{K}(x, y)$ for all $x, y \in Q$. [45]

In this subsection, we review a few well-known observations regarding interpolation and projection for spaces defined in terms of scattered kernel bases. We summarize these properties for the finite-dimensional space $\mathcal{H}_M := \text{span}\{\mathcal{R}_{\xi_{M,i}} \mid \xi_{M,i} \in \Xi_M, 1 \leq i \leq M\}$ that is defined in terms of scattered bases in $\mathcal{H}$ that are located at the centers $\Xi_M := \{\xi_{M,1}, \ldots, \xi_{M,M}\}$. We let $\Pi_M \mathcal{H} \to \mathcal{H}_M$ be the $\mathcal{H}$-orthogonal projection onto the closed subspace $\mathcal{H}_M \subset \mathcal{H}$. Since we have assumed that the kernel $\mathcal{K}$ on $\mathcal{H}$ is strictly positive definite, we have the representation...
The projection operator $\Pi_M$ induces the $\mathcal{H}$-orthogonal decomposition $\mathcal{H} = \mathcal{H}_M \oplus Z_M$ where $Z_M$ is the closed subspace $Z_M := \{f \in \mathcal{H} \mid f(\xi) = 0 \text{ for all } \xi \in \Xi_M\}$. This decomposition implies that the complementary projection $I - \Pi_M$ satisfies $((I - \Pi_M)g)(\xi) = 0$ for all $\xi \in \Xi_M$. In other words, the orthogonal projection $\Pi_Mg$ interpolates the function $g$ at the points in $\Xi_M$. When studying the convergence properties of estimates over the set $Q \subset X$, it is common that approximation errors are described in terms of the fill distance $h_{\Xi_M,Q}$ of the finite samples $\Xi_M \subset Q$ in the set $Q$. The fill distance is defined as $h_{\Xi_M,Q} = \sup_{q \in Q} \min_{\xi \in \Xi_M} d_Q(q, \xi)$ with $d_Q$ the metric on $Q$.

The relationships among the $\mathcal{H}$-orthogonal projection $\Pi_M$, the scattered bases $\mathcal{R}_{\xi_M,i}$ in $\mathcal{H}_M$, and the coordinate representation of $\Pi_M$, have been stated by considering $\mathcal{H}_M$, a subset of the RKHS $\mathcal{H}$ induced by the kernel $\mathcal{R} : X \times X \to \mathbb{R}$ of functions over the ambient space $X$. However, entirely analogous statements can be made for the projections onto finite-dimensional spaces $\mathcal{R}_M$ and $f^*(\mathcal{H}_M)$, which are defined simply by defining scattered bases in terms of the translates of the kernels $\tau_{\xi_M,i}$ and $\mathcal{R}_{\xi_M,i}^*$, respectively, for $\xi_{M,i} \in \Xi_M$.

To simplify notation in this chapter, we use the generic operator $\Pi_M$ to refer to any of the (generally different) projections from $\mathcal{H}$, $\mathcal{R}$, or $f^*(\mathcal{H})$ onto $\mathcal{H}_M, \mathcal{R}_M$, or $f^*(\mathcal{H}_M)$, respectively. The specific choice of the domain and range of $\Pi_M$ is clear from the context in any particular equation that follows.
3.2.3 Koopman Theory and EDMD Algorithm

There are a number of tools of analysis that constitute Koopman theory, and the general philosophy and motivation for its popularity can be found in general treatments like [22, 47]. If we consider only the discrete state evolution Equation 3.1, which defines the recursion \( x_{i+1} = f(x_i) \), recall that the Koopman operator \( U_f \) is defined as the composition operator

\[
U_f g := g \circ f.
\]

There are many good reasons for studying the Koopman operator \( U_f \) to understand the dynamics in Equation 3.1, as summarized in detail [22, 47]. Here we just note that the \( M^{th} \) step of the discrete flow in Equation 3.1 is given by

\[
x_M := ((U_f)^{M-1} f)(x_0).
\]

Thus, obtaining a good approximation of the action \( U_f \) on \( f \) enables finding a good forecast for a future state of the system.

We have noted earlier that the Koopman operator \( U_f \) is defined from the identity \( U_f g := g \circ f \) for some fixed function \( f \). One of the most common ways of approximating \( U_f \) is based on the extended dynamic mode decomposition (EDMD) algorithm. It has been studied carefully in recent work like [28, 32]. Here we summarize the EDMD method using nomenclature that is common in these references so that we can distinguish it from the data-driven approximation studied in this chapter. From a set of input-output samples of Equation 3.1, we define two
data matrices $X$ and $Y$ as follows,

$$
X = [x_1, \ldots, x_M] \in \mathbb{R}^{d \times M},
$$

$$
Y = [x_2, \ldots, x_{M+1}] := f(X) \in \mathbb{R}^{d \times M}.
$$

We want to use these data matrices to construct estimates of the Koopman operator $U_f$. Since the operator $U_f$ is infinite dimensional, we define a finite dimensional subspace for building approximations. Suppose we choose basis functions $\{\psi_i \mid 1 \leq i \leq M\}$ for our approximations. We introduce the vector $\Psi = [\psi_1, \ldots, \psi_N]^T$. With the definition of the basis functions in $\Psi$, data matrices for the EDMD algorithm are then determined as follows

$$
\Psi(X) = [\psi(x_1), \ldots, \psi(x_M)] \in \mathbb{R}^{N \times M},
$$

$$
\Psi(Y) = [\psi(x_2), \ldots, \psi(x_{M+1})] \in \mathbb{R}^{N \times M}.
$$

Approximations of the Koopman operator $U_f$ are constructed in terms of the matrix solution $A_{N,M} \in \mathbb{R}^{N \times N}$ of the minimization problem

$$
A_{N,M} := \arg \min_{A \in \mathbb{R}^{N \times N}} \| A\Psi(X) - \Psi(Y) \|_F^2
$$

The solution of this minimization problem is given by $A_{N,M} = \Psi(Y) \left(\Psi^T(X)\Psi(X)\right)^{-1} \Psi^T(X)$. Finally, the approximation $U_{N,M}$ of the Koopman operator $U_f$ is given by

$$
(U_{N,M}g)(\cdot) := \alpha^T A_{N,M} \psi(\cdot)
$$

(3.4)

for any function $g(\cdot) := \alpha^T \psi(\cdot)$ with $\alpha \in \mathbb{R}^N$. 
3.3 Problem Formulation

The distribution-free learning problem is a general problem that has been studied in a wide variety of contexts, most commonly for mappings between Euclidean spaces. As in we are interested in this chapter in formulating and solving this problem, not over some known compact subset such as $\Omega := [a, b]^d \subset \mathbb{R}^d$, but rather over the unknown manifold $Q$. There are several features of this problem that make it challenging in comparison to the classical problem over $\mathbb{R}^d$. It is first necessary to state the learning problem precisely in terms of operators and functions defined over manifolds, but this seems to follow the form of the Euclidean case fairly closely. One complicating factor for the case at hand, however, is that samples we study in this chapter are not generated from some independent and identically distributed (IID) measurement process. Rather, samples are collected along the trajectory of a discrete, typically nonlinear dynamic system over the unknown manifold, and the evolution law defining the samples is also unknown. Even if it were known, the fact that it defines a dependent process itself is a significant complication over the conventional case. An additional, troublesome issue that this paper addresses is the question of how to define the function spaces defined over the unknown manifold when building approximations of the solution to the distribution-free learning problem. We want to build approximations in such a way that we obtain realizable algorithms that yield rates of convergence that can be proven in a spirit similar to the strategy in Euclidean spaces. This will mean that the regularity of $Q$, as a type of embedded manifold, will play a role in selecting the spaces for the formulation and its approximation.
### 3.3. Problem Formulation

#### 3.3.1 Distribution-Free Learning in Euclidean Spaces

We begin by summarizing the general structure of the distribution-free learning problem for regression over $X := \mathbb{R}^d$, and subsequently, we discuss how the problem is cast in terms of the unknown manifold $Q$ for certain types of discrete evolution laws. When the distribution-free learning problem is applied to regression, we are faced with building approximations of some unknown function over $X$ that takes values in $Y = \mathbb{R}^n$. We are given samples $\{(x_{k_i}, y_{k_i})\}_{i=1}^M \subset X \times Y$. Here $X$ is the inputs space and $Y$ is the output space, and we can view the samples as perhaps noisy measurements of the functional relationship $y = G(x)$ where $G : X \rightarrow Y$ is unknown. As noted previously, it is usually assumed that the samples are independent and identically distributed, with $\nu$ the probability distribution over $X \times Y$ that generates the samples. In the distribution-free learning problem, the probability distribution $\nu$ is unknown. The aim of learning theory is to build an approximation of the unknown function using the samples that makes some error measure small, and here we use the familiar quadratic error for arbitrary functions $g : X \rightarrow Y$,

$$E_\mu(g) := \int_X \int_Y \|y - g(x)\|_Y^2 \nu(dx, dy),$$

where again we interpret $y = G(x)$ in the ideal case. A great deal is known about the structure of this particular problem in the space $L_\mu^2(\mathbb{R}^d)$, or even in $L_\mu^2(\Omega)$ with $\Omega$ a nice compact subset of $X$ like $\Omega = [a, b]^d$. It is known that the minimization of the quadratic error $E_\mu$ above is equivalent to minimizing the expression

$$E_\mu(g) := \int_X \|g(x) - H_\mu(x)\|_Y^2 \mu(x) + E_\mu(H_\mu)$$

(3.6)
where the measure $\nu$ is written as $\nu(dx,dy) = \mu_x(dy)\mu(dx)$, $\mu_x(dy)$ is the conditional probability over $Y$ given $x \in X$, $\mu(dx)$ is the marginal measure of $\nu$ over $X$, and $H_\mu(x) := \int_Y y\mu_x(dy)$ is the regressor function. Thus, the regressor $H_\mu$ is the optimal minimizer, but it cannot be computed from this closed form expression since $\nu$, and therefore $\mu_x$ is unknown. In practice, then, the ideal error measure above is replaced with the discrete
\[
E_M(g) := \frac{1}{M} \sum_{i=1}^M \|y_i - g(x_i)\|_Y^2
\] (3.7)
that depends on the samples $\{(x_i,y_i)\}_{i=1}^M$, which leads to the method of empirical risk minimization. (Traditionally, note that the minimization is over sequentially collected samples in Equation 3.7.) Note that the discrete error above can be evaluated for any given $g$. When estimates $G_{M,N} = \arg\inf_g \{E_M(g) \mid g \in \mathcal{H}_N\}$ that minimize the empirical risk are calculated for some $N$ dimensional space $\mathcal{H}_N$ of approximants, it is then possible to address in what sense $G_{M,N} \to G_\mu$ as the number of samples $M$ and the dimension $N$ approach $\infty$. The theory for such approximations is mature, and a summary of the state art in these cases can be found [41, 42, 48].

### 3.3.2 Distribution-Free Learning over an Unknown Manifold

The manifold estimation problem studied in this chapter modifies the classical learning problem above in a number of ways: some changes are minor, while others lead to subtle issues. Note carefully that the samples generated by Equations 3.1 are defined from a deterministic, nonlinear, dependent measurement process, not one that is IID. It turns out that the assumption that samples are IID is central to many of the most refined estimates of errors in distribution-free learning theory. We will analyze the system in Equation 3.1 in two ways in this and the next few subsections. In this subsection, we assume that the samples
3.3. PROBLEM FORMULATION

\( \{(x_i, y_i)\}_{i=1}^{M} \) are generated by choosing initial conditions randomly according to the unknown measure \( \mu \) on \( Q \), and then observations \( y_i \) are defined as in Equation 3.2. That is, the samples \( \{(x_i, y_i)\}_{i=1}^{M} \) are understood as the initial condition responses of the system in Equations 3.1 and 3.2 as the initial condition is chosen randomly over the unknown manifold \( Q \). (Later we extend our analysis to the case when samples are dependent and determined along a trajectory.) In this case, the measure \( \nu \) that generates the samples has the particular form

\[
\nu(dx, dy) = \delta_{G(f(x))}(dy) \mu(dx),
\]

and the regressor function is \( H_\mu(x) := \int_{Y} y \delta_{G(f(x))}(dy) = G \circ f \).

With these definitions, the ideal error \( E_\mu \) in Equations 3.5 and 3.6 makes sense for the Lebesgue space of functions \( L^2_{\mu}(Q, Y) \) defined over the manifold \( Q \) without change. Likewise, the empirical error \( E_M \) in Equation 3.7 makes sense, since the evaluation operator is well-defined on \( \mathcal{H} \subset C(X) \). In this situation, in contrast to the form of the regression error functional in Equation 3.6, we consider the ideal error

\[
E_\mu(g) = \int_Q \| (U_f(g - G))(x) \|^2_Y \mu(dx) + E_\mu(U_f G),
\]

where \( U_f h = h \circ f \) is the Koopman operator induced by the unknown function \( f \) that defines the unknown dynamics. The corresponding discrete error function is then

\[
E_M(g) := \frac{1}{M} \sum_{i=1}^{M} \| y_i - (g \circ f)(x_i) \|^2_Y = \frac{1}{M} \sum_{i=1}^{M} E_{i}^{(k)}(g),
\]

\[
= \frac{1}{M} \sum_{k=1}^{n} \sum_{i=1}^{M} \| (G^{(k)} \circ f)(x_i) - (g^{(k)} \circ f)(x_i) \|^2,
\]

where we have introduced the component-wise notation \( g := (g^{(1)}, \ldots, g^{(n)}) \) and \( G := (G^{(1)}, \ldots, G^{(n)}) \).
3.3.3 Projection for Regression in the Pullback Space

We start with a straightforward analysis under the assumptions of the last subsection 3.3.2. We interpret the empirical risk minimization in terms of projection or interpolation in a pullback space. We have chosen to start with this elementary case, since it makes clear some of the issues that distinguish the learning problem in this chapter from the usual one in Euclidean space. We consider the minimization \( g_M^k := \arg\min_{f} E_M^{(k)}(g^{(k)}) \mid g^{(k)} \in \mathcal{H}_M \) with \( \mathcal{H}_M := \text{span}\{\mathcal{R}_{\xi_{M,i}} \mid \xi_{M,i} \in \Xi_M\} \) that defines the empirical risk minimizer.

**Theorem 3.1.** Suppose that \( f(\Xi_M) := \{f(\xi_{M,i}) \mid \xi_{M,i} \in \Xi_M\} \) are distinct points. The empirical risk minimizer \( g_M^{(k)} \) satisfies

\[
(U_f g_M^{(k)}(\cdot)) := \sum_{i,j}^M (\mathcal{R}(f(\Xi_M), f(\Xi_M)))^{-1}_{i,j} y_j^{(k)} \mathcal{R}(f(x_i), f(\cdot))
\]  

(3.8)

**Proof.** The proof of this theorem is a direct consequence of scattered data interpolation over the pullback space \( f^*(\mathcal{H}) \). Define the space of approximants

\[
\mathcal{H}^*_M := \text{span}\{\mathcal{R}^*_{\xi_{M,i}} \mid \xi_{M,i} \in \Xi_M\} \subset f^*(\mathcal{H}).
\]

In this case, if we obtain a solution \( h_M^{(k)} \in f^*(\mathcal{H}) \) to a minimization problem over the pullback space \( f^*(\mathcal{H}) \),

\[
h_M^{(k)} = \arg\min_{h \in f^*(\mathcal{H})} \frac{1}{M} \sum_{i=1}^M \left| (y_i^{(k)} - h(x_i))^2 \right|
\]  

(3.9)

then, for any function \( g_M^{(k)} \in \mathcal{H} \) such that \( h_M^{(k)} = U_f g_M^{(k)} \), \( g_M^{(k)} \) solves the empirical risk minimization problem in Equation 3.8. But Equation 3.9 is solved by the function \( h_M^{(k)} \in f^*(\mathcal{H}) \) that interpolates the samples \( \{(x_i, y_i^{(k)})\}_{i=1}^M \), since in this case the right hand side
3.3. PROBLEM FORMULATION

of Equation 3.9 is precisely zero. The right hand side of Equation (4) is precisely the coordinate expression for the $f^*(\mathcal{H})$-orthogonal projection onto $\mathcal{H}_M^* := f^*(\mathcal{H}_M)$. Since the kernel $\mathcal{K}$ is assumed to be positive definite over $X \times X$, it follows that the pullback kernel $\mathcal{K}^*(\cdot, \cdot) = \mathcal{K}(f(\cdot), f(\cdot))$ is also strictly positive definite over $Q \times Q$. The assumption that $f(\Xi)$ are distinct points ensures that the Grammian in Equation 3.8 is invertible.

Overall, Theorem 3.1 is quite informative. It shows that the distribution-free learning problem at hand, which is expressed in terms of an unknown manifold $Q$, measures $\mu$ on $Q$, and function $f : Q \to Q$, can in principle be solved as a simple regression problem on the pullback space $f^*(\mathcal{H})$. Unfortunately, the solution in Equation 3.8 is not of practical use in the problem at hand, since the solution depends on the unknown function $f$. This makes clear a point that distinguishes the learning problem on manifolds that we study here, where the dynamics propagates according to some unknown function $f$, from the classical regression problem.

3.3.4 Data-Driven Approximations

Finally, we study the uncertain system in Equations 3.1 and 3.2 in its full generality. The samples $\{(x_{k_i}, y_{k_i})\}_{i=1}^M$ are collected along the deterministic, nonlinear system at discrete times $\mathbb{T}_M := \{k_i \mid 1 \leq i \leq M\}$. Now the measure $\mu$ is taken to be some convenient probability measure over $Q$ for defining average error, but it does not generate the samples randomly. In this subsection we define a data-driven approximation $U^M_j$ of the Koopman operator, which is used to bound the excess risk, which is defined by the left hand side of
the inequality

$$E_\mu(g) - E_\mu(U_f G) = \int_Q \| (U_f (G - g))(x) \|_Y^2 \mu(dx),$$

$$\lesssim \left( \max_{1 \leq k \leq n} \| U_f(G^{(k)}) - U_f(g^{(k)}) \|_{C(Q,R)} \right)^2.$$ 

We make the structural assumption that the pullback space $f^*(H)$ of functions over the manifold $Q$ is continuously embedded in the space of restrictions $\mathcal{R} := T(H), f^*(H) \hookrightarrow \mathcal{R}$, and furthermore, that the space of restrictions $\mathbb{R}$ is embedded continuously in the Sobolev space $W^{t,2}(Q)$. Following the strategy in [49], we define the data dependent approximation $U^M_f g$ of $U_f g$ to be given by

$$U^M_f g := \Pi_M((\Pi_M g) \circ f),$$

where this equation is interpreted componentwise for the vector-valued function $g$. That is, $U^M_f g := (U^M_f g^{(1)}, \ldots, U^M_f g^{(n)})$ for $g = (g^{(1)}, \ldots, g^{(n)})$. Based on the discussion in subsection 3.2.2, the coordinate expression for each component of this operator is

$$U^M_f g^{(k)} := \sum_{i,j=1}^M (\mathbb{K}(\Xi_M, \Xi_M))^{-1}_{i,j}$$

$$\times \sum_{p,q=1}^M (\mathbb{K}(\Xi_M, \Xi_M))^{-1}_{p,q} g^{(k)}(\xi_{M,q}) \mathbb{K}(\Xi_M, f(\Xi_M))_{p,j} \mathbb{R}_{M,i}.$$

(3.10)

**Theorem 3.2.** Suppose that $f^*(H) \hookrightarrow R \approx W^{t,2}(Q)$. Then there are constants $C, H > 0$ such that for all $\Xi_M \subset Q$ that satisfy $h_{\Xi_M,Q} \leq H$, we have

$$\| (U_f G)(x) - (U^M_f G)(x) \|_Y \leq C h_{\Xi_M,Q}^t \| g \|_{W^{t,2}(Q)}$$

for all $g \in W^{t,2}(Q)$ and $x \in Q$. 
3.3. Problem Formulation

Proof. We define the approximation \( U^M_f G \) as above, \( U^M_f G = \Pi_M((\Pi_M G) \circ f) \). Note that the coordinate expression for \( U^M_f G \) can be calculated using the observations \( \{(x_i, y_i)\}_{i=1}^M = \{(x_i, G(f(x_i)))\}_{i=1}^M \). (Here we assume that we have collected the one additional measurement \( x_{M+1} = f(x_M) \) also that is needed in the coordinate expression.) The proof of this theorem follows that strategy in [49], generalizing it to the case when \( Y = \mathbb{R}^n \), where

\[
\|(U_f G)(x) - (U^M_f G)(x)\|_Y
\leq \max_{1 \leq k \leq n} |(G^{(k)} \circ f)(x) - \Pi_M((\Pi_M G^{(k)}) \circ f)|,
\]

\[
\leq \max_{1 \leq k \leq n} \left\{ \| (I - \Pi_M) G^{(k)} \|_{C(\Omega, \mathbb{R})} \right\}_1 + \| (I - \Pi_M)((\Pi_M G^{(k)}) \circ f) \|_{C(\Omega, \mathbb{R})} \right\}_2.
\]

Both term 1 and term 2 above are bounded using the many zeros Theorem 3.4 on the manifold \( Q \). Term 1 is bounded using the fact that \( \mathcal{R} \approx W^{t,2}(Q) \hookrightarrow W^{m,2}(Q) \) and subsequently

\[
\|(I - \Pi_M) G^{(k)} \|_{C(\Omega, \mathbb{R})} \lesssim \|(I - \Pi_M) G^{(k)} \|_{W^{m,2}(Q)},
\]

\[
\lesssim h^{t-m}_{\mathcal{R},Q} \|(I - \Pi_M) G^{(k)} \|_{W^{t,2}(Q)}
\]

\[
\lesssim h^{t-m}_{\mathcal{R},Q} \|G^{(k)}\|_{W^{t,2}(Q)}.
\]

In the above we have applied Theorem 3.4 to the term \((I - \Pi_M) G^{(k)}\), which vanishes on \( \Xi_M \). We also use that fact that \( \Pi_M \) is a bounded operator on \( \mathcal{R} = W^{t,2}(Q) \). We can similarly bound the second term: (term 2) \( \lesssim h^{t-m}_{\mathcal{R},Q} \| U_f \Pi_M G^{(k)} \|_{W^{t,2}(Q)} \). By virtue of the assumption that \( f^*(\mathcal{H}) \hookrightarrow \mathcal{R} \approx W^{t,2}(Q) \), we know (term 2) \( \lesssim h^{t-m}_{\mathcal{R},Q} \|G^{(k)}\|_{W^{t,2}(Q)} \) also. Collecting the different inequalities for each of the components \( 1 \leq k \leq n \) finishes the proof.

We close this subsection by showing that the analysis of this subsection enables similar error
bounds to be defined for certain cases of the EDMD algorithm.

**Theorem 3.3.** Suppose that $U_{M,M}$ is the approximation of the Koopman operator derived by the EDMD algorithm in Equation 3.4 using the samples $\{(x_{k_i}, y_{k_i})\}_{i=1}^{M}$, so that the basis vector $\psi$ is selected as $\psi := \{\mathcal{R}_{x_{k_1}}, \ldots, \mathcal{R}_{x_{k_M}}\}^T$. Under the hypotheses of Theorem 3.2, we have

$$
\| (U_{fG})(x) - (U_{M,M}G)(x) \|_Y \leq C h_{\Xi_M,Q}^t \| g \|_{W^{1,2}(Q)}
$$

for the EDMD approximation $U_{M,M}$ of the Koopman operator $U_f$.

**Proof.** In general, the EDMD approximation of the Koopman operator depends on two parameters, the number of samples $M$ and the number of basis functions $N$. It is a more general method of building an approximation of $U_f$ than $U_f^M$ in this chapter. However, when $M = N$ and the basis vector $\psi$ consists of $\psi := \{\mathcal{R}_{x_1}, \ldots, \mathcal{R}_{x_M}\}^T$, then the coordinate expression in Equation 3.10 for $U_f^M$ is precisely the same as that in Equation 3.4. Since $U_f^M = U_{M,M}$, the error bound follows.

\[ \square \]

3.4 **Examples and Numerical Results**

This subsection summarizes numerical results for some examples of estimates given a family of observations $\{(x_i, y_i)\}_{i=1}^{M}$ collected over an unknown manifold $Q$. We set the centers $\Xi_M := \{x_1, \ldots, x_M\}$ in this subsection.

3.4.1 **Example: The Pendulum**

For our first example, we consider a simple closed form case to illustrate the concepts behind our empirical estimate. Our evolution $f : Q \to Q$ corresponds to the dynamics of a pendulum...
following the Stőrmer-Verlet scheme as given in Example 1.4 in [50]. In this example, $X_1$ and $X_2$ correspond to the pendulum’s momentum and position respectively. We build estimates for an unknown function $G : X := \mathbb{R}^2 \to \mathbb{R}$ from samples collected over $Q = \subseteq \mathbb{R}^2$. The function $G$ is defined for each $x := \{x_1, x_2\} \in \mathbb{R}^2$ by $G(x) = 1.5 - \sin(x_2) + \frac{1}{5}x_1^2$. Figure 3.1 depicts the estimates of the Koopman operator acting on the observable function $G$. In these studies, we use the Matern-Sobolev kernel,

$$
\hat{K}^{MS}(\xi_{M,j}, x) = \hat{K}^{MS}_{\xi_{M,j}}(x) = \left(1 + \frac{\sqrt{3}\|x - \xi_{M,j}\|_2}{\beta}\right)\exp\left(-\frac{\sqrt{3}\|x - \xi_{M,j}\|_2}{\beta}\right)
$$

with $\| \cdot \|$ as the standard Euclidean metric over $\mathbb{R}^2$, $\xi_{M,j}$ corresponds to the $j$th kernel center, and the hyperparameter $\beta$ corresponds to the decay of the kernel function away from the center. From the figure, it is evident that the approximation closely matches the intersubsection between the true function surface $U_f G$ and the surface generated from the RKHS basis, $U^M_f G$. This is a result of the interpolating properties of the RKHS approximation. Approximations are also generated from compactly supported kernel functions as discussed in [37]. Some examples of these types of kernels include

$$
\hat{K}^{CS}_1(\mathfrak{d}) = (1 - \mathfrak{d})^4_+(4\mathfrak{d} + 1)
$$

$$
\hat{K}^{CS}_2(\mathfrak{d}) = (1 - r)^6_+(35\mathfrak{d}^2 + 18\mathfrak{d} + 3)
$$

$$
\hat{K}^{CS}_3(\mathfrak{d}) = (1 - \mathfrak{d})^8_+(32\mathfrak{d}^3 + 25\mathfrak{d}^2 + 8\mathfrak{d} + 1)
$$

where the radial distance $\mathfrak{d} = \|x - \xi_{M,j}\|$ is defined from the kernel center $\xi_{M,j}$, and the operator $(\cdot)_+$ returns the argument in the parenthesis if it is positive and zero otherwise. The positive-semidefinite property of the $(\cdot)_+$ operator ensures each kernel is compactly supported. Figure 3.2 compares the estimate using the Matern-Sobolev kernel with $\beta = 1$
Figure 3.1: An example of the $U_fG$ approximation in red of $G$, represented by the light green surface. Using $M$ samples collected over the composition $U_fG$ the finite dimensional approximation $U_f^MG$ is generated from the RKHS represented by the mesh blue surface. Minimizing the empirical risk over the samples corresponds to interpolating the function over the samples as represented by the red dashed line.
with the estimate using the compactly supported kernel $R_2^{CS}$. Notice how the estimate itself quickly decays to zero away from the manifold as a consequence of its bases being compactly supported.

As noted in [5] and [40], oscillations in the estimate can be evident depending on how the hyperparameter, $\beta$, from the Matern-Sobolev kernel function. Here we present a careful study of the improvement that can be achieved using beta than the (rather standard) regularization in the learning problem over the manifold, we have found that the results are qualitatively similar to the case of kernel estimation in Euclidean spaces. Figure 3.3 illustrates estimates generated by the Matern-Sobolev kernel for various $\beta$. In these cases, larger values of $\beta$ decrease the decay of the kernel centers and lead to a larger spread of each basis. From the figures it is clear that estimates generated from the larger $\beta$ are more "flat" and see less drastic changes in the estimate over small distances. As the $\beta$ parameter is decreased the decay rate increases from the center and lowers the spread of each kernel function. This leads to the oscillations seen in the estimates for smaller $\beta$ values.

Another factor in the solution $U^M f = \sum_{i=1}^M \alpha^* \xi_{M,i}$ of Equation 3.8 comes from numerical stability in projecting the true solution $G$ onto a subspace $H_M = \text{span}\{\xi_{M,i} | \xi_{M,i} \in \Xi_M\}$ defined over our set of samples $\Xi_M$. This solution satisfies the equation

$$K_M \alpha^* = G(\Xi_M)$$

(3.11)

where we define the kernel matrix $K_M := K(\Xi_M, \Xi_M)$, output vector $G(\Xi_M) := [G(\xi_{M,1}), \ldots, G(\xi_{M,M})]^T$ and $\alpha^* \in \mathbb{R}^M$ is a vector of coefficients corresponding to the $H_M$ projection. However, noise in both the sample inputs and the output measurements as well as computer precision results in perturbations, $\Delta K$ and $\Delta G$ of the kernel matrix $K_M$ and the output vector $G(\Xi_M)$.
Figure 3.2: The estimate $U^M_f G$ using the Matern-Sobolev kernel (a) with $\beta = 1$ and the estimate using the compactly supported kernel $\mathcal{K}_{2}^{CS}$ (b). Both estimates decay in magnitude away from the manifold. However, the estimate (a) is still nonzero everywhere while the estimate (b) itself quickly decays to zero away from the manifold as a consequence of its bases being compactly supported.
3.4. Examples and Numerical Results

Figure 3.3: The approximations \( U^M_f \) for various dimension, \( \beta \), of the Matern kernel with \( \beta = 5 \) (a), \( \beta = 1 \) (b), \( \beta = 0.5 \) (c), \( \beta = 0.2 \) (d). Regardless of the hyperparameter selected, it is clear that each estimate is still determined by an algorithm which fits the estimate to the empirical samples collected along the red line over the manifold \( Q \).
respectively. In practice, a numerical solution \( \hat{\alpha} \) is calculated by

\[
(K_M + \Delta K)\hat{\alpha} = G(\Xi_M) + \Delta G
\]

(3.12)

The bounds between the exact solution and the numerical calculation are determined as follows

\[
\|\alpha^* - \hat{\alpha}\| \leq C\text{cond}(K_M)\|\Delta K\|\|\Delta G\|
\]

(3.13)

where the condition number \( \text{cond}(K_M) = \|K_M\|\|K^{-1}_M\| = \frac{\lambda_{\text{Max}}(K_M)}{\lambda_{\text{Min}}(K_M)} \). For a set of samples \( \Xi_M \), the separation

\[
s_{\Xi_M} = \frac{1}{2} \min_{\xi_{M,i}, \xi_{M,j}} d_Q(\xi_{M,i}; \xi_{M,j})
\]

corresponds to the largest possible radius of disjoint balls centered around the samples \( \Xi_M \).

Figure 3.4 demonstrates the condition number of kernel matrices using various compactly supported kernels as defined in [37] as the spacing between the kernel centers is decreased. The slope increases for each curve demonstrating that for increasing order of kernels, which belong to more restricted smooth spaces, the computation becomes more unstable compared to the kernels which belong to lower order smoothness spaces. Similarly, Figure 3.5 examines the condition number given the fill distance for kernel matrices defined by the Matern-Sobolev kernels with various hyperparameters \( \beta \). As mentioned previously, larger \( \beta \) leads to a slower decay of the kernel function and, therefore, a larger "spread" from the kernel center. This larger spread increases the condition number for Kernel matrices generated from samples and increases the numerical instability evidenced by the higher condition number for curves of increasing \( \beta \).

For Figure 3.4, the separation is relatively uniform amongst the samples and, therefore, decreases along with the fill distance. As mentioned in [37], lower bounds for \( \lambda_{\text{Min}}(K_M) \) can be
determined as functions of the separation distance $s_{Xi}$. For instance, the minimum eigenvalue associated with a kernel matrix defined by compactly supported kernels as defined in [37] decreases exponentially. In order to demonstrate that the minimum eigenvalue decreases with decreases in the separation alone, the minimum eigenvalue of the kernel matrix is examined over a uniformly spaced distribution of samples with the exception of two samples with decreasing distance between them. Consequently, the fill distance remains constant for each curve while decreasing the separation.

Figure 3.6 illustrates the effect of decreasing the separation distance between just two samples points while keeping the fill distance constant. Each curve differs by the fill distance of the set of samples. While decreases in the fill distance do increase the condition number as demonstrated by changes in the condition number between curves, it can be clearly observed that the separation, which increases for each curve, will rapidly increase the condition number of the matrix and lead to poor numerical results.

As mentioned in our introduction and our theoretical results, the solution to the distribution free learning problem satisfies particular rates of convergence. Figure 3.7 demonstrates the calculated error convergence of the estimate of the Koopman approximation to the true function as the number of samples increased. An increase in the number of samples collected on the manifold reduces the fill distance that bounds the error of the approximation. The decay rate is determined by selecting $t$ and $m$ as described in [49]. From the figure, it is clear that the error is bounded by $O(h_{Xi}^{t-m})$, which is represented by the dashed line above the curve.
Figure 3.4: Three cases of the condition number of the Kernel Matrix $K$ given the minimum separation distance of the collected centers from our simulated data using kernels belonging to specific smoothness spaces. In this example, the samples are roughly uniformly distributed and the separation distance $q_X$ is approximately decrease along with the fill distance $h_{\Xi_M,Q}$. Note that as the separation distance decreases the condition number increase exponentially. Additionally, the condition number of the Kernel matrix associated with higher order kernels is larger for equivalent separation distances.
3.4. Examples and Numerical Results

Figure 3.5: This figure demonstrates the relationship between the fill distance and the condition number for the Matern-Sobolev kernel $K_{MS}$ for various values of the hyperparameter $\beta$. The smaller spread of the kernel function corresponding to smaller $\beta$ lowers the condition number for estimates of similar fill distances. However, the rate of increase given by the slope is determined by the kernel type as is, therefore, relatively the same for each curve in this example.
Figure 3.6: A figure illustrating the behavior of the minimum eigenvalue of the Kernel Matrix $K$ given the minimum separation distance of the collected centers from the simulate data. For each curve, the separation, $q_X$ is decreased between only two centers while keeping a constant fill distance $h_{\Xi M, Q}$. Note that, just like the experimental data the condition number increases exponentially as the separation distance decreases.
3.4.2 Example: Human Kinematics Study

Our second example uses three-dimensional motion capture data collected in [51] during a human walking experiment on a treadmill. The experiment tracks 5000 marker coordinates relative to a fixed inertial frame defined by the camera’s fixed position. For this simulation, a small candidate kinematic model is defined from the full collection of experimental trajectories. The marker coordinates of the hip, knee, and ankle in the full data set are projected to the sagittal plane that divides the left and right half of the body see Figure 3.8. With this projection, the joint angle $\theta_1$ roughly corresponds to hip flexion. It is calculated from the projected vector $v_1$ that connects the hip to knee from the inertial $b^{(1)}$ vector in the plane. The knee flexion angle $\theta_2$ is calculated by taking the dot product between the vector $v_1$ and the vector connecting the knee to the ankle $v_2$ in the sagittal plane. We then take these joint angles of the left leg as defining our configuration manifold $Q$. The measured outputs corresponded to the marker positions themselves and could be of arbitrary dimension $n$. For illustrative purposes, we chose to examine the kinematic maps $G^{(1)}$ and $G^{(2)}$ from the
joint angles $\theta^{(1)}$ and $\theta^{(2)}$ to the ankle $y^{(1)}$ and $y^{(2)}$ associated with the $b^{(1)}$ and $b^{(2)}$ vectors in the body-fixed frame. During the experiment described in this chapter, subjects walk on a treadmill with various markers placed along the body. For our study, the functions we seek to approximate are the $b^{(1)}$ and $b^{(2)}$-coordinates of the left ankle. In this case, we do not have an analytical expression for our functions, but rather just our collection of sample points $\{q_i\}_{i=1}^M$ of the ankle coordinates over $M$ measurements.

Figure 3.9 shows the estimate of part of the forward kinematics from the empirical risk minimizer using 64 samples and the Matern-Sobolev kernel basis with $\beta = 2$. Both estimates occur over the same submanifold $Q$ of the input space from measurements of $\theta^{(1)}$ and $\theta^{(2)}$. However, from the figure we can see that the $b^{(1)}$ coordinate, which corresponds to the height of the ankle from the hip remains relatively constant throughout the motion while the $b^{(2)}$ coordinate associated with forward movement relative to the body sees significant changes over the submanifold $Q$.

## 3.5 Conclusion

This paper has derived methods for estimation of animal motions. We have derived a closed form expression for the empirical risk minimizer in a pullback RKHS space that depends on the unknown dynamics. We also have shown that approximations of the unknown function $G : Q \rightarrow Y$ can be constructed in terms of a data-dependent approximation of the the Koopman operator. For the kernel based method, we have derived error bounds based on scattered samples and kernel dependent bases defined over $Q$. Finally, we have presented numerical results to better illustrate the qualitative behavior of the algorithms.
Figure 3.8: An illustration of the inputs $\theta^{(1)}$ and $\theta^{(2)}$, which roughly correspond to hip and knee flexion. These input variables can be mapped to measured marker coordinates placed on joints such as the knee or ankle.
Figure 3.9: The estimates of the Koopman operator acting on observable functions $G^{(1)}$ and $G^{(2)}$ corresponding to the $b^{(1)}$ coordinate (a) and $b^{(2)}$ coordinate (b) of the ankle over the submanifold $Q$ of the input space given by the leg coordinates $\theta^{(1)}$ and $\theta^{(2)}$. 
Appendix

The proof of convergence for the approximation of the Koopman operator in subsection 3.3.4 uses a version of the many zeros theorem that holds over compact, smooth, regularly embedded submanifolds $S \subset \mathbb{R}^d$.

**Theorem 3.4 ([43]Lemma 10).** Let $Q$ be a $k$ dimensional smooth manifold, $t \in \mathbb{R}$ with $t > k/2$, and $m \in \mathbb{N}$ satisfy $0 \leq m \leq \lfloor t \rfloor - 1$. There are constants $C, H$ such that if $h_{\Xi, T} \leq H$ and $f \in W^{t, 2}(Q)$ satisfies $f|_{\Xi} = 0$, then

$$|f|_{W^{m, 2}(Q)} \leq Ch_{\Xi}^{t-m}|f|_{W^{t, 2}(Q)}.$$

To apply this theorem in applications, we will need to choose the kernel $K$ that is defined on all of $X$ in such a way that the space of restrictions $\mathcal{R}$ to $Q$ is equivalent to a Sobolev space over $Q$. This can be achieved using the strategy outline on page 1759 of [43]. The kernel $K$ on $X \times X$ is said to have algebraic decay if its Fourier transform $\hat{K}$ has algebraic decay in the sense that $\hat{K}(\xi) \sim (1 + \|\xi\|_2^2)^{-\tau}$ for $\tau > d/2$. If the kernel satisfies this algebraic decay condition, then it follows that $\mathcal{H} = W^{t, 2}(\mathbb{R}^d)$. We then can use the following theorem, again from [43].

**Theorem 3.5 ([43] Theorem 5).** If the kernel $K$ that induces $\mathcal{H}$ satisfies the algebraic growth condition with exponent $\tau$, then we have $\mathcal{R} = T(H) = W^{-(d-k)/2, 2}(Q)$.

Bibliography


Stefan Klus, Feliks Nuske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schutte. Data-Driven Approximation of the Koopman Operator: Model


Chapter 4

Kernel Methods for Regression in Continuous Time over Subsets and Manifolds

This paper derives error bounds for regression in continuous time over subsets of certain types of Riemannian manifolds. The regression problem is typically driven by a nonlinear evolution law taking values on the manifold, and it is cast as one of optimal estimation in a reproducing kernel Hilbert space (RKHS). A new notion of persistency of excitation (PE) is defined for the estimation problem over the manifold, and rates of convergence of the continuous time estimates are derived using the PE condition. We discuss and analyze two approximation methods of the exact regression solution. We then conclude the paper with some numerical simulations that illustrate the qualitative character of the computed function estimates. Numerical results from function estimates generated over a trajectory of the Lorenz system are presented. Additionally, we analyze an implementation of the two approximation methods using motion capture data.
4.1 Introduction

4.1.1 Motivation

The study of machine or statistical learning theory, and its application to regression problems, has been a topic of interest for years. [1]. These techniques have had a lasting impact in Bayesian estimation and estimation using Gaussian processes. Many of these efforts in machine or statistical learning theory, Bayesian estimation, and Gaussian processes have theoretical foundations that exploit formulations cast in terms of reproducing kernel Hilbert spaces (RKHS), which are also known as native spaces. [2] While some recent efforts including [3, 4, 5] have sought to further understand learning theory in the context of dynamical systems theory, and vice-versa, it is accurate to say that most of the above work to date has focused on cases where the samples used for learning or regression are generated from some independent and identically distributed (IID), stochastic, discrete measurement process. A good account on the state-of-the-art in distribution-free learning theory and its focus on discrete processes can be found in [1, 2, 6].

This paper seeks to use RKHS formulations in continuous time estimation problems, in the spirit of [7, 8, 9, 10, 11, 12], to achieve some of the advantages that are so clear in the above applications of learning theory to IID discrete systems. The theory and algorithms in the references [7, 8, 9, 10, 11, 12] describe many of the working tools used by specialists in the field of adaptive estimation and control theory as it is applied to ordinary differential equations (ODEs). As described in [7, 8, 9, 10, 11, 12], it is standard that the regression problem in finite dimensional spaces is often used to motivate, explain, and study adaptive estimation and control theory for ODEs. The regression problem arises then when the ODE is characterized by a finite linear combination of known regressor functions. In recent
papers, the authors have introduced adaptive estimation problems in RKHS formulations in [13, 14, 15, 16, 17, 18], where the evolution is described by a distributed parameter system (DPS) over a native space. Here we study the related regression problem in continuous time in a native space, which plays an analogous role in the RKHS/DPS framework to that when a finite dimensional collection of regressors appear in an ODE.

One way to view this paper is as an exploration of what features or properties of the well-studied regression problem that underlies adaptive estimation and control of ODEs also hold, or can be extended to, the regression problem over a native space. This paper is also an attempt to address some of the open questions summarized in [3], for instance, that relate learning theory and dynamical systems theory.

4.1.2 Problem Description

in this chapter we study a regression problem in continuous time where an approximating agent traverses the configuration space $X$ along a trajectory $t \mapsto \phi(t) \in X$ making observations $y(t) = G(x(t)) \in \mathbb{R}$ of some unknown function $G : X \to \mathbb{R}$. The configuration space $X$ is always a complete metric space, but it need not be compact. The most important cases discussed in the paper choose $X := \mathbb{R}^d$ or some other smooth Riemannian manifold. The system that generates the trajectory can be quite general. Any generally nonlinear autonomous or nonautonomous system may generate the trajectory $t \mapsto \phi(t) \in X$. We only require that the trajectory is continuous.

in this chapter, we assume that the input/output history $\{x(\tau), y(\tau)\}_{\tau \in [0,t]}$ is observed without noise. We concentrate on this paper on how the choice of a function space and geometric properties of the flow influence the rate of convergence of approximations of the solution of the regression problem in continuous time, which is challenging enough for a single paper.
We address the effects of uncertainty in the continuous time regression problem using the theory of inverse problems in a forthcoming paper.

While the flow is defined on $X$, which may not be compact, approximations of the regression problem will be carried out over some typically compact subset $S \subseteq X$. Two interpretations of the set $S \subseteq X$ are possible. It may be that the compact set $S$ is some known, prescribed subdomain over which approximations are sought. In the most difficult problem setting, however, the set $S$ is not known \textit{a priori} but represents an emergent structure. Over time, samples along the trajectory accumulate in $S \subseteq X$. As time progresses, we obtain more and more information about the structure of $S$, but initially we may not have any idea about its structure. In the problem at hand, it can be the case that $S$ is highly irregular.

Two examples are typical of the abstract situation above. In the first, $X := \mathbb{R}^d$ and $S \subseteq X$ is some compact subset. In the numerical examples in Section 4.4, the Lorenz system is of this type. We have $X = \mathbb{R}^3$ and $S \subset \mathbb{R}^3$ is an irregular, unknown, compact, positively invariant set.

Another important example arises when $S \subseteq \mathcal{M}$ and $\mathcal{M}$ is a compact, smooth, Riemannian manifold that is regularly embedded in $X$. Of course, if $X$ is compact, it is always possible to choose $S \subseteq \mathcal{M} \equiv X$. We emphasize that we reserve the notation $\mathcal{M}$ for a compact manifold in this chapter. Again, in the most difficult form of this problem, the manifold $\mathcal{M}$ may not be known \textit{a priori}. That is, we may not have explicit knowledge of the specific coordinate charts that define $\mathcal{M}$, but rather only that it is embedded in the larger manifold $X, \mathcal{M} \subseteq X$.

One important underlying goal should be clear in view of the comments above describing $S, \mathcal{M}, X$. In a sense, we seek estimation methods that are robust with respect to uncertainty in the knowledge about the underlying unknown subset or manifold supporting the dynamics. The assumption where the form of $\mathcal{M}$ is unknown has recently been studied by the authors.
in [19, 20, 21] when seeking data-dependent approximations of the Koopman operator.

Other progress on a related problem using online, gradient learning laws has been reported by the authors in [14, 15, 22, 23]. In these papers the method of native space or RKHS embedding is used to generate online estimates in applications to adaptive estimation and control theory. Here in contrast we study an offline, optimal estimation approach. In comparison to the now familiar approaches for parametric estimation in finite dimensional Euclidean spaces, this paper derives convergence results for regression estimates in a reproducing kernel Hilbert space $\mathcal{H}$ that is generated by a known, admissible kernel $\mathcal{K}: X \times X \to \mathbb{R}$. The native space $\mathcal{H}$ can be interpreted as the Hilbert space that contains all functions that can be represented as the limit of the translates of a certain template function. Given the kernel $\mathcal{K}$ that defines $\mathcal{H}$, we define the kernel section or basis function centered at $x$ as $\mathcal{K}_x(\cdot) := \mathcal{K}(x, \cdot)$. Then, $\mathcal{H}$ is defined to be $\mathcal{H} = \text{span}\{\mathcal{K}_x \mid x \in X\}$ the closed linear space as the kernel basis moves around in $X$. This paper can be viewed as the extension of standard results in Euclidean spaces as in [7, 8, 9], see page 48 of [7] for Chapter 4 of [9] for instance, to the case when an agent generates estimates in continuous time of a function in the native space $\mathcal{H}$ defined over a subset $S$ manifold of a manifold $\mathcal{M}$ or $X$.

A primary contribution of this paper is the characterization of the error in continuous time using methods from scattered data approximation in kernel spaces. Another contribution is the introduction of a new PE condition that is well-defined over manifolds and that enables the analysis of convergence of the time-varying regression estimate. We review these contributions in some detail next.
4.1.3 Summary of New Results

There are three specific new results derived in this chapter that are not addressed in any of the previous papers by the authors in [14, 15, 19, 22, 23], or in the literature at large. Suppose that \( \phi : t \mapsto \phi(t) \in X \) is a trajectory of either an autonomous or nonautonomous flow on the manifold. The regression problem described above is solved using the operator \( T_\phi(s, t) : \mathcal{H} \to \mathcal{H} \)

\[
T_\phi(s, t) := \int_s^t \mathbf{K}_{\phi(\tau)} \otimes \mathbf{K}_{\phi(\tau)} \nu(d\tau).
\]

where \( \mathbf{K}_{\phi(\tau)} := \mathbf{K}(\phi(\tau), \cdot) \) is the kernel basis function centered at \( \phi(\tau) \) and \( \nu \) is a finite measure on \([s, t]\). The tensor product operator \( \mathbf{K}_{\phi(\tau)} \otimes \mathbf{K}_{\phi(\tau)} \) satisfies \( \mathbf{K}_{\phi(\tau)} \otimes \mathbf{K}_{\phi(\tau)} g = \mathbf{K}_{\phi(\tau)} (\mathbf{K}_{\phi(\tau)} g) \mathcal{H} \) for all \( g \in \mathcal{H} \). The first new result is summarized in Theorem 4.1 where sufficient conditions are given that ensure that this operator is compact, positive, and self-adjoint. This generalizes a result in the Appendix in [24] to the time-dependent case, which is essential to the study of the regression problem in continuous time. The second new result is the introduction of a new persistency condition in Equation 4.5 for flows over a manifold that generalizes the one in our earlier papers. It defines persistency for a general closed subspace \( \mathcal{V} \subseteq \mathcal{H} \), where the norm on \( \mathcal{V} \) that can be different than the norm on \( \mathcal{H} \). The publications [25, 26] always make the special choice \( \mathcal{V} := \mathcal{H}_S \) where \( \mathcal{H}_S \) is the native space generated by a subset \( S \subset X \). The generalization in this chapter is essential to prove convergence of estimates in certain spectral approximation spaces \( \mathcal{A}^r \subseteq \mathcal{H} \), which depend on a trajectory \( t \mapsto \phi(t) \in X \).

Finally, when the new PE condition holds for the subspace \( \mathcal{V} \subseteq \mathcal{H} \), we show that

\[
\| \hat{g}_\mathcal{V}(t, \cdot) - \Pi_\mathcal{V}G \|_{\mathcal{H}} \leq \frac{\bar{R}^2 m \Delta}{\gamma_1 m + \gamma} \| (I - \Pi_\mathcal{V}) G \|_{\mathcal{H}} + \frac{\gamma}{\gamma_1 m + \gamma} \| \Pi_\mathcal{V}G \|_{\mathcal{H}}
\]

where \( t \mapsto \hat{g}_\mathcal{V}(t, \cdot) \in \mathcal{V} \) is the optimal solution of the (offline) regression problem, and \( \Pi_\mathcal{V} \) is
the $\mathcal{H}$-orthogonal projection of $\mathcal{H}$ onto $\mathcal{V}$. The constant $\tilde{K}$ is a bound on the reproducing kernel $\mathcal{R}$ that defines $\mathcal{H}$, the constants $\gamma_1$ and $\Delta$ arise in the PE condition in Equation 4.5, $\gamma$ is the regularization parameter in the continuous regression error functional, and the time $t := m\Delta$ for the positive integer $m > 0$. Note that as time $t = m\Delta \to \infty$, the estimate above implies that
\[
\|\hat{g}_V(t, \cdot) - \Pi_V G\|_{\mathcal{H}} \lesssim O\left(\|(I - \Pi_H)G\|_{\mathcal{H}}\right).
\]
Intuitively, the solution of the regression problem in continuous time under the new PE condition implies that it asymptotically approaches the projection over the PE subspace.

We further refine this estimate in some cases to show that, when $N$ samples are used to define certain finite dimensional spaces of approximants $\mathcal{V} := \mathcal{H}_N$ and these spaces are PE, we have
\[
\|\hat{g}_N(t, \cdot) - \Pi_N G\|_{\mathcal{H}_S} \leq \left(\left(\frac{\tilde{R}^2 m \Delta}{\gamma_1 m + \gamma}\right) \|P_N\|_{L^2(S)}\right) \|G\|_{\mathcal{H}_S} + \frac{\gamma}{\gamma_1 m + \gamma} \|\Pi_N G\|_{\mathcal{H}_S},
\]
for all unknown functions $G$ that are smooth enough. This bound makes use of the power function $P_N(x)$, over the set $S$, that is defined as
\[
P_N(x) := |\mathcal{R}(x, x) - \mathcal{R}_N(x, x)| \quad \text{for all } x \in S.
\]
In this expression $\mathcal{R}_N$ is the kernel that defines the native space $\mathcal{H}_N$ of approximants, and $S$ is the closure of the trajectory $\tau \mapsto \phi(\tau)$ in $X$. The kernel $\mathcal{R}_N(x, y) := (\Pi_N k_x)(y)$ by definition [27]. It is worth noting that this error bound for the regression problem in continuous time has some similarity to that in [28, 29]. These papers derive pointwise error bounds for discrete regression or Bayesian estimation for discrete time processes, in contrast to the integrated error bound above for systems in continuous time. The relationship of the solution
of the continuous time regression problem to the more familiar discrete IID, stochastic, case is discussed in detail in Section 4.3.3.

### 4.1.4 Notation, Symbols, and Background

In this chapter the state space $X$ is a complete metric space. The most important examples choose $X$ to be the Euclidean space $X := \mathbb{R}^d$, a smooth and compact Riemannian manifold $X := M$, or certain measurable subsets of these. We denote by $K : X \times X \rightarrow \mathbb{R}$ a symmetric, nonnegative, continuous kernel that induces the scalar-valued native space $H$ of functions defined over $X$. Throughout the paper $H$ is a reproducing kernel Hilbert space (RKHS) of real-valued functions over the set $X$ that is given by $H := \text{span}\{K(x, \cdot) \mid x \in X\}$.

In this chapter, we often must refer to time-varying functions that take values in $H$. We write $f(t, \cdot)$ to represent the spatial function $x \mapsto f(t, x)$ for fixed time $t$. That is $f(t, \cdot)$ for each fixed time $t$.

We write $\mathbb{E}_X : H \rightarrow \mathbb{R}$ for the evaluation functional at $x \in X$, which satisfies $\mathbb{E}_xf := f(x)$ for each $f \in H$. The adjoint $\mathbb{E}_x^* : \mathbb{R} \rightarrow H$ can be understood as the multiplication operator given by $\mathbb{E}_x^*\alpha := \mathbb{K}(x, \cdot)\alpha$ for all $\alpha \in \mathbb{R}$. We denote by $L(H)$ the linear and bounded operators that map from $H$ to $H$. The notation $\mathcal{B}(L(H))$ denotes the Borel $\sigma$-algebra on $L(H)$.

For any subset $S \subseteq X$ we define the native space $H_S \subseteq H$ generated by $S$ as $H_S := \text{span}\{\mathbb{K}(x, \cdot) \mid x \in S\}$. We emphasize that $H_S$ is not the Hilbert space that consists of restrictions of functions to $S$: since $H_S \subseteq H$ functions in $H_S$ are supported on $X$. The space $H_S$ is a native space having kernel $\mathbb{K}_S(x, y) := \langle \Pi_S\mathbb{K}(x, \cdot), \Pi_S(\mathbb{K}(y, \cdot)) \rangle_H$ for all $x, y \in X$, with $\Pi_S$ the $H$-orthogonal projection of $H$ onto $H_S$.

We denote by $T_S$ the trace or restriction operator $T_Sg := g|_S$. The space of restrictions $\mathcal{R}_S = T_S(H)$ is an RKHS with the restricted kernel $\mathcal{K}(x, y) := \mathbb{K}(x, y)$ for all $x, y \in S \subseteq X$. There
is a canonical minimum norm extension operator $E_S : \mathcal{R}_S \rightarrow \mathcal{H}$ that satisfies $E_ST_S = \Pi_S$. This operator is an isometry $E_S : \mathcal{R}_S \rightarrow \mathcal{H}_S$ and satisfies

$$\|E_Sg\|_\mathcal{H} = \|g\|_{\mathcal{R}_S} = \inf \{\|f\|_\mathcal{H} \mid g = T_Sf, f \in \mathcal{H}\}$$

for all $g \in \mathcal{R}_S$.

### 4.2 Regression in Continuous Time

In this section we study in detail the problem of regression in continuous time in a RKHS $\mathcal{H}$ of real-valued functions over the state space $X$, where $X$ is a complete metric space. The trajectory $t \mapsto \phi(t)$ of the system is assumed to be continuous. The overall situation is depicted graphically in Figure 4.1. In this problem we are given a trajectory $t \mapsto \phi(t) \in X$, and we make measurements $y(\tau) = G(\phi(\tau))$ of an unknown function $G \in \mathcal{H}$ at $\phi(\tau) \in X$ for each $\tau \in [0, t]$. The goal is to use the continuous collection of samples $\{(\phi(\tau), y(\tau))\}_{\tau \in [0, t]}$ to build a time-varying estimate $\hat{g}(t, \cdot) \in \mathcal{H}$ of the unknown function $G$.

#### 4.2.1 The Offline Optimal Regression Estimate in an RKHS

We define the integral error functional $E : \mathbb{R}^+ \times \mathcal{H} \times C([0, t], X) \rightarrow \mathbb{R}$ to be

$$E(t, g; \phi) := \frac{1}{2} \int_0^t |y(\tau) - \mathbb{E}_{\phi(\tau)}g|^2 d\tau + \frac{1}{2} \gamma \|g\|^2_\mathcal{H}$$

for the state space $X$, some measure $\nu$ on $\mathbb{R}^+$, and a regularization parameter $\gamma > 0$. In this equation $\mathbb{E}_{\phi(\tau)} : \mathcal{H} \rightarrow \mathbb{R}$ is the evaluation functional at $\phi(\tau)$, which satisfies $\mathbb{E}_{\phi(\tau)}g = g(\phi(\tau))$. 
Figure 4.1: (Top) An illustration of regression in continuous time. The orbit $\Gamma(\phi_0) = \bigcup_{\tau \geq 0} \phi(\tau)$ is a subset of the state space $X$. The output is determined by a function $G$ represented by the green surface over $X$. (Bottom) An illustration of the estimate $\Pi_\nu G$ represented by the blue mesh is generated by a subspace $\mathcal{V}$ of $\mathcal{H}$. Here the subspace consists of kernel functions illustrated by the “bumps” with centers at various points along the orbit. In this figure, one center is represented by $\phi(\tau)$. 
for any $g \in \mathcal{H}$. The error functional $E(t, g; \phi)$ can be rewritten as

\[
E(t, g; \phi) = \frac{1}{2} \int_0^t |\mathbb{E}_{\phi(\tau)}(G - g)|^2 \nu(d\tau) + \frac{1}{2} \gamma \|g\|^2_{\mathcal{H}},
\]

\[
= \frac{1}{2} \int_0^t \langle \mathbb{E}^*_{\phi(\tau)} \mathbb{E}_{\phi(\tau)}(G - g), G - g \rangle_{\mathcal{H}} \nu(d\tau) + \frac{1}{2} \gamma \|g\|^2_{\mathcal{H}}.
\]

The adjoint $\mathbb{E}^*_{\phi(\tau)} : \mathbb{R} \to \mathcal{H}$ is given by $\mathbb{E}^*_{\phi(\tau)} \alpha := \mathfrak{K}_{\phi(\tau)} \alpha$ for any $\alpha \in \mathbb{R}$. The study of the error functional $E(t, g; \phi)$ takes a familiar structure when we introduce an operator $T_\phi(s, t) : \mathcal{H} \to \mathcal{H}$ via the identity

\[
T_\phi(s, t)g := \int_s^t \mathbb{E}^*_{\phi(\tau)} \mathbb{E}_{\phi(\tau)} g \nu(d\tau).
\]

Note that $T_\phi(s, t)$ is a time-varying operator that depends on the trajectory $t \mapsto \phi(t)$. In the following arguments, and later at several places in the text, the properties of the operator $T_\phi(s, t)$ are important. We summarize some of its properties in the following theorem.

**Theorem 4.1.** Suppose that $\mathfrak{K} : X \times X \to \mathbb{R}$ is a continuous admissible kernel that induces the native space $\mathcal{H}$ of continuous real-valued functions over $X$. The operator $T_\phi(s, t)$ above is an integral operator

\[
(T_\phi(s, t)g)(\xi) = \int_s^t \mathfrak{K}(\xi, \phi(\tau)) \langle \mathfrak{K}_{\phi(\tau)}, g \rangle_{\mathcal{H}} \nu(d\tau)
\]

\[
= \int_s^t \mathfrak{K}(\xi, \phi(\tau)) g(\phi(\tau)) \nu(d\tau).
\]

If there is a constant $\mathfrak{K} > 0$ such that $\sqrt{\mathfrak{K}(x, x)} \leq \mathfrak{K}$ for all $x \in X$ and the trajectory $\phi \in C([0, t], X)$ is continuous in time, then the operator-valued map

\[
\tau \mapsto \mathfrak{K}_{\phi(\tau)} \otimes \mathfrak{K}_{\phi(\tau)} \in \mathcal{L}(\mathcal{H})
\]
is continuous from $\mathbb{T} := [s,t]$ to $\mathcal{L}(\mathcal{H})$ and measurable as a map from $(\mathbb{T}, \mathcal{B}_\mathbb{T})$ into $(\mathcal{H}, \mathcal{B}_{\mathcal{L}(\mathcal{H})})$.

The operator $T_\phi(s,t)$ can be understood as the Bochner integral

$$T_\phi(s,t) = \int_T \mathfrak{K}_{\phi(t)} \otimes \mathfrak{K}_{\phi(t)} \nu(d\tau) = \int_T \mathcal{E}_{\phi(t)}^* \mathcal{E}_{\phi(t)} \nu(d\tau). \quad (4.1)$$

The operator $T_\phi(s,t)$ is compact, self-adjoint, positive, and trace class.

**Proof.** Since $\mathfrak{K}$ is continuous and the trajectory $\tau \mapsto \phi(\tau)$ is continuous, the map $\tau \mapsto \mathfrak{K}_{\phi(\tau)}$ is continuous. This fact follows from the identity

$$\| \mathfrak{K}_{\phi(\tau)} - \mathfrak{K}_{\phi(t)} \|_{\mathcal{H}}^2 := \mathfrak{K}(\phi(\tau), \phi(t)) - 2\mathfrak{K}(\phi(\tau), \phi(t)) + \mathfrak{K}(\phi(t), \phi(t)),$$

and the righthand side goes to zero as $\tau \to t$ by the continuity of $\tau \mapsto \mathfrak{K}(\phi(\tau), \phi(\tau))$. This identity can then be used to show that the curve $\tau \mapsto \mathfrak{K}_{\phi(t)} \otimes \mathfrak{K}_{\phi(t)}$ is continuous as a map from $\mathbb{T} \to \mathcal{L}(\mathcal{H})$ from the expression

$$\| (\mathfrak{K}_{\phi(\tau)} \otimes \mathfrak{K}_{\phi(\tau)} - \mathfrak{K}_{\phi(t)} \otimes \mathfrak{K}_{\phi(t)}) h \|$$

$$\leq \| \mathfrak{K}_{\phi(\tau)} \otimes \mathfrak{K}_{\phi(\tau)} h - \mathfrak{K}_{\phi(t)} \otimes \mathfrak{K}_{\phi(t)} h \|$$

$$+ \| \mathfrak{K}_{\phi(\tau)} \otimes \mathfrak{K}_{\phi(t)} h - \mathfrak{K}_{\phi(t)} \otimes \mathfrak{K}_{\phi(t)} h \|,$$

$$\leq 2\mathfrak{K}\|h\| \| \mathfrak{K}_{\phi(\tau)} - \mathfrak{K}_{\phi(t)} \|.$$

This collection of inequalities above makes repeated use of the bound $\| \mathfrak{K}_x \| \leq \mathfrak{K}$. The measurability of the map from $(\mathbb{T}, \mathcal{B}_\mathbb{T})$ to $(\mathcal{H}, \mathcal{B}_{\mathcal{L}(\mathcal{H})})$ then follows from the continuity of this map.

Finally, the fact that the Bochner integral in Equation 4.1 exists in $\mathcal{L}(\mathcal{H})$ is a consequence
of the bound

$$\int_T \| \mathcal{R} \phi(\tau) \otimes \mathcal{R} \phi(\tau) \|_{L(H)} \nu(d\tau) \leq \bar{R}^2 \nu(T).$$

Next, we consider the compactness of $T \phi(s, t)$. This proof essentially follows the same line of reasoning as that in Proposition 14 of [24], which is carried out for the operator $\mathcal{R}_x \otimes \mathcal{R}_x$ and a spatial measure $\mu$ on $X$. Since the operator $\mathcal{R} \phi(\tau) \otimes \mathcal{R} \phi(\tau)$ is finite rank for each $\tau$, it is trace class for each $\tau$, with

$$\text{Tr} (\mathcal{R} \phi(\tau) \otimes \mathcal{R} \phi(\tau)) = \mathcal{R}(\phi(\tau), \phi(\tau)) \leq \bar{R}^2.$$

The trace operator is a continuous linear operator on the trace norm class, and we know that

$$\text{Tr} \left( \int_T \mathcal{R} \phi(\tau) \otimes \mathcal{R} \phi(\tau) \nu(d\tau) \right) = \int_T \text{Tr} (\mathcal{R} \phi(\tau) \otimes \mathcal{R} \phi(\tau)) \nu(d\tau) \leq \bar{R}^2 \nu(T)$$

by the mapping property of a Bochner integral under continuous linear operators. The Bochner integral exists and is therefore trace class.

We conclude this proof by showing that $T \phi(s, t)$ is a positive operator. This is a modification of the analysis in [30], which treats a different problem where again the integral is over space, not time. For completeness, we give its outline. Let $\mathbb{T}_{N,k}$ be a family of measurable subsets of $\mathbb{T}$ with $\mathbb{T} = \bigcup_{k=1}^N \mathbb{T}_{N,k}$ and $\nu(\mathbb{T}_{N,k}) = \nu(\mathbb{T})/N$. Fix a quadrature point $t_{N,k} \in \mathbb{T}_{N,k}$ from
4.2. Regression in Continuous Time

Each set $T_{N,k}$ and define $\xi_{N,k} = \phi(t_{N,k})$. We then have

\[
(T_\phi(s,t)h, h) = \\
\lim_{N \to \infty} \sum_{i,j=1}^{N} \int_T \int_T \mathcal{R}(\xi_{N,j}, \xi_{N,i})h(\xi_{N,i})h(\xi_{N,j})\chi_{N,i}(\tau)\chi_{N,j}(\tau)\nu(d\tau)
\]

\[
= \lim_{N \to \infty} \frac{1}{N^2} \sum_{i,j=1}^{N} \mathcal{R}(\xi_{N,j}, \xi_{N,i})h(\xi_{N,i})h(\xi_{N,j}) \geq 0,
\]

where $\chi_{N,i}$ is characteristic function of the subset $T_{N,i}$. The positivity of $T_\phi(s,t)$ follows from the semidefiniteness of the kernel $\mathcal{R}$. \hfill \square

With these properties in hand, we can derive the optimal offline regression estimate in continuous time. In the usual way we can “complete the square” and write

\[
E(t, g; \phi) = \frac{1}{2} \langle (T_\phi(0, t) + \gamma I)g, g \rangle_{\mathcal{H}} - \langle T_\phi(0, t)G, g \rangle_{\mathcal{H}} + \frac{1}{2} \langle T_\phi(0, t)G, G \rangle_{\mathcal{H}}.
\]

Now we define the best approximation, i.e. the regressor solution in continuous time,

\[
\hat{g}(t, \phi) := \arg\min_{g \in \mathcal{H}} E(t, g; \phi).
\]

But it is relatively easy to calculate the Gateaux derivative of this functional in the direction $h \in \mathcal{H}$. By definition it is given by

\[
\langle DE(t, g; \phi), h \rangle_{\mathcal{H}} := \lim_{\epsilon \to 0} \frac{E(t, g + \epsilon h; \phi) - E(t, g; \phi)}{\epsilon} = \langle (T_\phi(0, t) + \gamma I)g - T_\phi(0, t)G, h \rangle_{\mathcal{H}}.
\]
Local minima to the above minimization problem must satisfy

\[ \hat{g}(t, \cdot) = (T_\phi(0, t) + \gamma I)^{-1} T_\phi(0, t) G, \]

because the operator \( T_\phi(0, t) + \gamma I \) is invertible as a map from \( \mathcal{H} \to \mathcal{H} \).

### 4.2.2 Galerkin Approximations of the Regression Estimate

The regression estimate \( \hat{g}(t, \cdot) \) is the solution of an operator equation in the generally infinite dimensional space \( \mathcal{H} \). Practical algorithms must consider approximations of this solution. In this section we discuss one method for obtaining approximations based on Galerkin's method. A review of some of the common properties of Galerkin approximations in Hilbert spaces is given in the Appendix in Section 4.6.1. The regression solution \( \hat{g}(t, \cdot) \) satisfies the equation

\[ \langle (T_\phi(0, t) + \gamma I) \hat{g}(t, \cdot), h \rangle_{\mathcal{H}} = \langle T_\phi(0, t) G, h \rangle_{\mathcal{H}} \quad \text{for all } h \in \mathcal{H}. \quad (4.3) \]

Let \( \mathcal{H}_N \subseteq \mathcal{H} \) be some \( N \)-dimensional subspace of \( \mathcal{H} \) that is used to build approximations. The Galerkin approximation \( \hat{g}_N(t, \cdot) \in \mathcal{H}_N \) is the solution of the analogous equation

\[ \langle (T_\phi(0, t) + \gamma I) \hat{g}_N(t, \cdot), h_N \rangle_{\mathcal{H}} = \langle T_\phi(0, t) G, h_N \rangle_{\mathcal{H}} \quad \text{for all } h_N \in \mathcal{H}_N. \quad (4.4) \]

For each \( t \in \mathbb{R}^+ \) we define the bilinear form \( a(t)(\cdot, \cdot) : \mathcal{H} \times \mathcal{H} \to \mathbb{R} \) to be \( a(t)(g, h) := \langle (T_\phi(0, t) + \gamma I) g, h \rangle_{\mathcal{H}} \) for each \( g, h \in \mathcal{H} \). The bilinear form \( a(t)(\cdot, \cdot) \) is bounded and coercive.
as defined in Section 4.6.1 since

\[ a(t)(g, h) \leq (\tilde{k}^2 \nu([0, t]) + \gamma)\|g\|\|h\| \quad \text{for all } g, h \in \mathcal{H}, \quad \text{and} \]

\[ a(t)(g, g) \geq \gamma\|g\|^2 \quad \text{for all } g \in \mathcal{H}. \]

By the Lax-Milgram Theorem 4.5 there is a unique solution \( \hat{g}(t, \cdot) \) of Equation 4.3 and \( \hat{g}_N(t, \cdot) \) of Equation 4.4. From Theorem 4.6 we know that

\[ \|\hat{g}(t, \cdot) - \hat{g}_N(t, \cdot)\| \leq \frac{(\tilde{k}^2 \nu([0, t]) + \gamma)}{\gamma}\|(I - \Pi_N)G\|. \]

In the analysis so far, we have \( \hat{g}(t, \cdot) \in \mathcal{H}, \hat{g}_N(t, \cdot) \in \mathcal{H}_N \subseteq \mathcal{H} \), under the assumption that \( G \in \mathcal{H} \), with \( \mathcal{H} \) a native space of functions supported on the configuration space \( X \). The error in approximating the optimal regression estimate \( \hat{g}(t, \cdot) \) by the Galerkin estimate \( \hat{g}_N(t, \cdot) \) is bounded by the norm on the best approximation of \( G \) from the subspace \( \mathcal{H}_N \), but there are a number of standard techniques to build sharp bounds on the error \( \|(I - \Pi_N)G\| \), which depend on how regular the function \( G \) is. These methods can be based on spectral analysis of integral operators and Mercer kernels, properties of the power function, or versions of the many zeros theorems [31, 32, 33]. We discuss such specific cases in the examples in Section 4.4.

## 4.3 Persistency of Excitation (PE) in Native Spaces

The basic error estimate for the Galerkin approximation described in Section 4.2.2 can be refined in several ways. In this section, we show how introducing priors on \( G \), which are certain assumptions that enforce restrictions or constraints on the unknown function, can yield improved error estimates. In analogy to the case of parametric estimation in
Euclidean space, we introduce a persistency of excitation condition for flows over a manifold. The PE condition can be used to derive alternative terms of an error bound on Galerkin approximations. Define the operator

$$T_\phi(t, t + \Delta) := \int_t^{t+\Delta} \mathbb{E}^{*}_{\phi(\tau)} \mathbb{E}_{\phi(\tau)} \nu(d\tau).$$

### 4.3.1 A New Persistence of Excitation Condition

In [15], we say that a persistence of excitation condition holds over the closed subspace $V \subseteq \mathcal{H}$ if there exist constants $\gamma_1, \gamma_2, \Delta > 0$ such that

$$\gamma_1 \|g\|^2_V \leq \int_t^{t+\Delta} \langle \mathbb{E}^{*}_{\phi(\tau)} \mathbb{E}_{\phi(\tau)} g, g \rangle_{\mathcal{H}} \nu(d\tau) \leq \gamma_2 \|g\|^2_V,$$

or in other words

$$\gamma_1 \|g\|^2_V \leq \langle T_\phi(t, t + \Delta) g, g \rangle_{\mathcal{H}} \leq \gamma_2 \|g\|^2_V, \quad (4.5)$$

for all $t \in \mathbb{R}^+$ and $g \in V$. Note that the above PE condition uses the operator as given in Equation 4.1. We analyze two different cases below:

1. The space $V \subseteq \mathcal{H}$ is the native space $\mathcal{H}_S$ generated by a subset $S \subseteq X$,

$$V := \mathcal{H}_S := \text{span}\{ \mathbb{R}_x \mid x \in S \}. $$

Note that in this case $V := \mathcal{H}_S$ equipped with the norm it inherits as a closed subspace of $\mathcal{H}$.

2. The space $V$ is selected to be the closed subspace $A^s(\mathcal{H})$ that is defined in terms of a
fixed, compact, self-adjoint, positive operator $T^s$ and its spectral decomposition.

4.3.2 Persistency in $\mathcal{V} := \mathcal{H}_S$ with $S \subseteq X$

Note that, if the above PE condition in Equation 4.5 holds, we obtain upper and lower bounds on $T_\phi(0, t) + \gamma I$. For simplicity, suppose that $t = m\Delta$ for some integer $m \in \mathbb{N}$. Then, if the PE condition holds for $\mathcal{V} := \mathcal{H}_S$, we know that

$$(\gamma_1 m + \gamma)\|g\|_{\mathcal{H}}^2 \leq \langle (T_\phi(0, t) + \gamma I)g, g \rangle_{\mathcal{H}} \leq (\gamma_2 m + \gamma)\|g\|_{\mathcal{H}}^2$$

for all $g \in \mathcal{H}_S$. This means that for all $g \in \mathcal{H}_S$ we have the upper bound

$$\|(T_\phi(0, t) + \gamma I)^{-1}g\|_{\mathcal{H}} \leq \frac{1}{\gamma_1 m + \gamma} \|g\|_{\mathcal{H}}.$$

The Optimal Regressor $\hat{g}(t, \cdot)$ in $\mathcal{H}$

We can use the above bound to find an error bound for the best offline estimate $\hat{g}(t, \cdot) \in \mathcal{H}$. Suppose that $\Pi_S : \mathcal{H} \rightarrow \mathcal{H}_S$ is the $\mathcal{H}$-orthogonal projection onto the closed subspace $\mathcal{H}_S$. We set

$$\Delta G := G - \Pi_S G = (I - \Pi_S)G,$$

$$y(t) = \mathbb{E}_{\phi(t)} G = (\Pi_S G)(\phi(t)) + \Delta G(\phi(t)).$$
We define the error between the offline estimate $\hat{g}(t, \cdot)$ and $\Pi_S G$ to be $\tilde{g}(t, \cdot) := \hat{g}(t, \cdot) - \Pi_S G$.

It follows that

$$
\tilde{g}(t, \cdot) = (T_\phi(0, t) + \gamma I)^{-1} T_\phi(0, t) G - \Pi_S G \\
= (T_\phi(0, t) + \gamma I)^{-1} T_\phi(0, t) (\Pi_S G + \Delta G) - \Pi_S G \\
= (T_\phi(0, t) + \gamma I)^{-1} (T_\phi(0, t) + \gamma I) \Pi_S G - \gamma \Pi_S G \\
\quad + (T_\phi(0, t) + \gamma I)^{-1} T_\phi(0, t) \Delta G - \Pi_S G \\
= (T_\phi(0, t) + \gamma I)^{-1} \left( \Pi_S (T_\phi(0, t) \Delta G - \gamma G) \\
\quad + (I - \Pi_S) T_\phi(0, t) \Delta G \right).
$$

Now we apply the bound on $(T_\phi(0, t) + \gamma I)^{-1}$, and we get

$$
\|\tilde{g}(t, \cdot)\|_H \leq \|(T_\phi(0, t) + \gamma I)^{-1} \Pi_S T_\phi(0, t) \Delta G\|_H \\
+ \gamma \|(T_\phi(0, t) + \gamma I)^{-1} \Pi_S G\|_H \\
+ \|(T_\phi(0, t) + \gamma I)^{-1} \|(I - \Pi_S) T_\phi(0, t) \Delta G\|_H \\
\leq \frac{1}{\gamma_1 m + \gamma} \|T_\phi(0, t)\| \|(I - \Pi_S) G\|_H \\
+ \frac{\gamma}{\gamma_1 m + \gamma} \|\Pi_S G\|_H + \frac{1}{\gamma} \|T_\phi(0, t)\| \|(I - \Pi_S) G\|_H.
$$

But, by virtue of $\sqrt{\mathcal{R}(x, x)} \leq \bar{R}$, we know that $\|T_\phi(0, t)\| \leq \bar{R}^2 \nu([0, t])$. Assuming that $\nu$ is Lebesgue measure on $\mathbb{R}^+$ and $t = m \Delta$, we obtain

$$
\|\tilde{g}(t, \cdot) - \Pi_S G\|_H \leq \frac{\bar{R}^2 m \Delta}{\gamma_1 m + \gamma} \|(I - \Pi_S) G\|_H \\
+ \frac{\gamma}{\gamma_1 m + \gamma} \|\Pi_S G\|_H \\
+ \frac{1}{\gamma} \bar{R}^2 m \Delta \|(I - \Pi_S) G\|_H. \tag{4.6}
$$
In particular, if we have \((I - \Pi_S)G = 0\), we conclude that
\[
\lim_{t \to \infty} \| \hat{g}(t, \cdot) - \Pi_S G \|_{\mathcal{H}} = 0.
\]

### The Optimal Regressor in \(\mathcal{H}_S\)

Above we characterized the optimal regressor \(\hat{g}(t, \cdot) \in \mathcal{H}\) when the subspace \(\mathcal{H}_S\) is PE. It is also possible to pose the original regression problem in \(\mathcal{H}_S\) and seek the optimal regressor \(\hat{g}_S(t, \cdot) \in \mathcal{H}_S\) when \(\mathcal{H}_S\) is PE. In this case we seek the approximation \(\hat{g}_S(t, \cdot) \in \mathcal{H}_S\) that satisfies the equation
\[
\langle (T_\phi(0, t) + \gamma I)\hat{g}_S(t, \cdot), h_S \rangle_{\mathcal{H}} = \langle T_\phi(0, t)G, h_S \rangle_{\mathcal{H}} \quad \text{for all } h_S \in \mathcal{H}_S.
\]

Then the solution \(\hat{g}_S(t, \cdot)\) can also be written as
\[
\hat{g}_S(t, \cdot) = (\Pi_S(T_\phi(0, t) + \gamma I)\Pi_S)^{-1} \Pi_ST_\phi(0, t)G.
\]

We can apply the PE condition
\[
\langle \Pi_S(T_\phi(0, t) + \gamma I)\Pi_S g, g \rangle_{\mathcal{H}} \geq (\gamma_1 m + \gamma)\| g \|_{\mathcal{H}}^2
\]
for all \(g \in \mathcal{H}_S\). This gives an upper bound
\[
\| (\Pi_S(T_\phi(0, t) + \gamma I)\Pi_S)^{-1} g \|_{\mathcal{H}} \leq \frac{1}{\gamma_1 m + \gamma} \| g \|_{\mathcal{H}}.
\]
for all \( g \in \mathcal{H}_S \). In this case, \( \hat{g}_S(t, \cdot) \in \mathcal{H}_S \), so that the error \( \tilde{g}_S(t, \cdot) = \hat{g}_S(t, \cdot) - \Pi_S G = \Pi_S \hat{g}_S(t, \cdot) - \Pi_S G \) can then be expressed as follows

\[
\tilde{g}_S(t, \cdot) = \hat{g}_S(t, \cdot) - \Pi_S G = (\Pi_S T_\phi(0, t) + \gamma I) \Pi_S)^{-1} [\Pi_S T_\phi(0, t)(I - \Pi_S) G - \gamma \Pi_S G].
\]

We can bound each of the two terms on the right hand side of the equality by writing

\[
\|((\Pi_S T_\phi(0, t) + \gamma I) \Pi_S)^{-1} [\Pi_S T_\phi(0, t)(I - \Pi_S) G]\|_{\mathcal{H}} \leq \frac{R^2 m \Delta}{\gamma_1 m + \gamma} \|(I - \Pi_S) G\|_{\mathcal{H}},
\]

\[
\|((\Pi_S T_\phi(0, t) + \gamma I) \Pi_S)^{-1} [\gamma \Pi_S G]\|_{\mathcal{H}} \leq \frac{\gamma}{\gamma_1 m + \gamma} \|G\|_{\mathcal{H}}.
\]

We now have the error bound

\[
\|\tilde{g}_S(t, \cdot) - \Pi_S G\|_{\mathcal{H}} \leq \frac{R^3 m \Delta}{\gamma_1 m + \gamma} \|(I - \Pi_S) G\|_{\mathcal{H}}
\]

\[
+ \frac{\gamma}{\gamma_1 m + \gamma} \|\Pi_S G\|_{\mathcal{H}}.
\]

(4.7)

**Observations:**

1. The analysis in Sections 4.3.2 and 4.3.2 shows that the error in the optimal offline estimate can be controlled by a PE condition satisfied by the operator \( T_\phi(t, t + \Delta) \). The PE condition establishes that the best offline regressor estimate \( \hat{g}(t, \cdot) \) of \( G \in \mathcal{H}_S \) converges to \( G \).

2. Note that the bound on \( \tilde{g} \) given by Equation 4.6 consists of three terms that each behave differently as \( m \to \infty \). The first term converges to a constant proportional to the projection error \( (I - \Pi_S) G \) as \( m \to \infty \). The second term decays to zero as \( m \to \infty \). However, the last term, referred to as the drift term, grows indefinitely as \( m \to \infty \). If we seek to compute the optimal estimate \( \hat{g}(t, \cdot) \in \mathcal{H} \) via continuous regression when
only $\mathcal{H}_S$ is PE, this drift term results. The primary issue is that we cannot apply the PE condition
\[
\|(T_\phi(0, t) + \gamma I)^{-1}g\|_{\mathcal{H}} \leq \frac{1}{\gamma m + \gamma} \|g\|_{\mathcal{H}}.
\]
on $\Delta G$ because $\Delta G \notin \mathcal{H}_S$. This problem is addressed by seeking the optimal regressor $\hat{g}_S(t, \cdot) \in \mathcal{H}_S$ that filters out this drift term. This error $\tilde{g}_S(t, \cdot)$ is bounded by only two terms as given in Equation 4.7 where the first converges to a constant proportional to $(I - \Pi_S)G$ and the second decays to zero as $m \to \infty$.

3. In a typical situation, in applications to finite dimensional approximations, it is frequently the case that $S$ consists of a finite number of samples, $S := \{\xi_1, \ldots, \xi_N\}$, and then
\[
\mathcal{H}_S := \mathcal{H}_N := \text{span}\{\mathfrak{R}_{\xi_i} \mid 1 \leq i \leq N\}.
\]
In this case, the error between the best offline estimate $\hat{g}(t, \cdot)$ and the projection $\Pi_N G$ can be bounded above by the rate of convergence of $(I - \Pi_N)G$ to zero. This is carried out in detail in the example in Section 4.4.

4. Since the operator $T_\phi(t, \Delta)$ is compact, if the PE condition in Equation 4.5 holds for $\mathcal{V} := \mathcal{H}_S$, it must be the case that $\mathcal{H}_S$ is finite dimensional. Otherwise, the PE condition would imply that the compact operator $T_\phi(t, \Delta) : \mathcal{H} \to \mathcal{H}$ has a bounded inverse, which is impossible on an infinite dimensional space $\mathcal{H}_S$. It follows that the primary application of case (1) will be to understand convergence of the optimal regressor when finite dimensional subspaces of approximants are PE.
Approximation: Method (1)

The exact optimal solution of the regression problem in continuous time, given by $\hat{g}(t, \cdot) = (T_{\phi}(0, t) + \gamma I)^{-1}T_{\phi}(0, t)G$, or correspondingly $\hat{g}_S(t, \cdot)$ defines a function of time and space, $\hat{g}(t, x)$ for $t \geq 0$ and $x \in X$. Practical implementations and algorithms must employ approximations of the exact regression solution. As in methods for parametric estimation in Euclidean spaces, recursive methods to approximate the solution of this problem are often used. We have studied one recursive method for the problem in this chapter in \[18\]. Here, and in the numerical examples in Section 4.4, we comment on implementations of offline approximations. Let $\hat{g}_N(t, \cdot)$ be the approximation of either $\hat{g}(t, \cdot) \in \mathcal{H}$ or $\hat{g}_S(t, \cdot) \in \mathcal{H}_S$.

When we define the approximation $\hat{g}_N(t, x) := \sum_{j=1}^N \alpha_{N,j}(t)$ using the continuous regression error functional, we obtain an equation that has the form

$$\sum_{j=1}^N \left( \int_0^t \mathcal{R}(\xi_{N,j}, \phi(\tau))\mathcal{R}(\phi(\tau), \xi_{N,i})\nu(d\tau) + \gamma \mathcal{R}(\xi_{N,j}, \xi_{N,i}) \right) \alpha_{N,j}(t) = \int_0^t \mathcal{R}(\xi_{N,i}, \phi(\tau))y(\tau)\nu(d\tau).$$

(4.8)

Note also that the optimal estimate in the continuous case requires integrating the kernel functions along the orbit, which ordinarily cannot be directly calculated in closed form. Consequently, the approximation requires approximating the integral term. For a function $\rho : [0, t] \to \mathbb{R}$, a general form of a quadrature rule builds the approximation

$$\int \rho(\tau)d(\tau) \approx \sum_{k=1}^{N_w} w_k \rho(\tau_k),$$

where $N_w$ is the number of quadrature points, $\{w_k\}_{k=1}^{N_w}$ are the quadrature weights, and $\{\tau_k\}_{k=1}^{N_w}$ are the quadrature points. The approximation of the integral can be determined from a multitude of different quadrature techniques. Standard examples include the trapezoidal...
4.3. Persistency of Excitation (PE) in Native Spaces

rule, Simpsons rule, or Gaussian quadratures. Here we use a particularly simple quadrature rule. Recall the above definitions of $T = [0, t]$, the subintervals $T_{N,k}$, and the quadrature points $\xi_{N,k} := \phi(t_{N,k})$ for $t_{N,k} \in T_{N,k}$ from the proof of Theorem 4.1 where the weights \{w_k\}_{k=1}^{N_w}$ are given by the time intervals $\{t_{N,k+1} - t_{N,k}\}_{k=1}^{N_w}$. One form of approximating the integral via a one point quadrature rule generates the equation

$$
\sum_{j=1}^{N} \left( \sum_{k=1}^{N_w} w_k \mathcal{R}(\xi_{N,j}, \xi_{N,k}) \mathcal{R}(\xi_{N,k}, \xi_{N,i}) + \gamma \mathcal{R}(\xi_{N,j}, \xi_{N,i}) \right) \alpha_{N,j}(t) = \sum_{k=1}^{N_w} w_k \mathcal{R}(\xi_{N,i}, \xi_{N,k}) y(t_{N,k}).
$$

(4.9)

We end this section with the pseudo-code implementation in Algorithm 4.2 to explicitly determine the estimate from this approximation method. Note that the initialization steps require that we select samples that have sufficient distance between one another ensuring the calculation of stable numerical estimates. [19, 31]
Algorithm 4.2.  \textbf{input:} kernel function \( \mathcal{R} \), kernel hyperparameters \( \beta \), regularization parameter \( \gamma \), and desired kernel separation \( \eta \)

\textbf{select samples for kernel centers:} Collect data \( \{t_i, x_i, y_i\}_{i=1}^M \)

\( S_N = \xi_{N,1} = x_1, \ N = 1 \)

Select placement of kernel centers

\textbf{for} \( x_i \) in \( \{x_i\}_{i=2}^M \) \textbf{do}

\hspace{1em} if \( \|\xi_{N,j} - x_i\| > \eta, \ \forall \xi_{N,j} \in S_N \)

\hspace{1em} Add sample \( x_i \) to the set of centers

\hspace{1em} \( S_N = S_N \cup \{x_i\} \)

\hspace{1em} Add sample time \( t_i \) to the set of indexed times

\hspace{1em} \( t_{N,i} = t_i, \ T_{N+1} = T_N \cup \{t_{N,i}\}, \ N = N + 1 \)

\textbf{end for}

Construct Kernel Matrix

\( \mathbb{K}(S_N, S_N)_{i,j} = [\mathcal{R}(\xi_{N,i}, \xi_{N,j})] \)

Determine Weighting Matrix

\( \mathbf{W} = \text{diag}(w_k), \ w_k = t_{k+1} - t_k \)

Construct Output Vector

\( \mathbf{y} = \{y_1, \ldots, y_N\}, \ y_i = y(\xi_{N,i}) = y(\phi(t_{N,i})) \)

Calculate coefficients \( \mathbf{\alpha}_N = \{\alpha_{N,1}, \ldots, \alpha_{N,N}\}^T \) of estimate

\( \mathbf{\alpha}_N = (\mathbb{K}(S_N, S_N)\mathbf{W}\mathbb{K}(S_N, S_N) + \gamma \mathbb{K}(S_N, S_N))^{-1} (\mathbb{K}(S_N, S_N)\mathbf{W}\mathbf{y}) \)

\textbf{return}  Estimate \( \hat{g}_N(t, \cdot) = \sum_{j=1}^N \alpha_{N,j} \mathcal{R}(\xi_{N,j}, \cdot) \).

\textbf{Approximation: Method (2)}

The system of Equations 4.9 above requires the introduction of quadratures over \([0, t]\). In this section, we introduce a second method of approximation that eliminates the calculation
of quadratures. The coefficients \( \{\alpha_{N,i}\}_{i=1}^{N} \) of the second method of approximation satisfy the following equation

\[
\sum_{j=1}^{N} \int_{0}^{t} \mathcal{R}(\xi_{N,j}, \phi(\tau)) \mathcal{R}(\phi(\tau), \xi_{N,k}) \alpha_{N,j}(\tau) \nu(d\tau) + \gamma \mathcal{R}(\xi_{N,j}, \xi_{N,i}) \alpha_{N,j}(t) = \int_{0}^{t} \mathcal{R}(\xi_{N,i}, \phi(\tau)) y(\tau) \nu(d\tau).
\] (4.10)

In the above equation the unknown coefficients \( \{\alpha_{N,i}(t)\}_{i=1}^{N} \) are now inside the integrand and are integrated along the orbit. Taking the time derivative of Equation 4.10 we can get an evolution equation for the coefficients \( \{\alpha_{N,i}(t)\}_{i=1}^{N} \). It is given by

\[
\dot{\alpha}_{N,j}(t) = (\gamma \mathcal{R}(\xi_{N,j}, \xi_{N,i}))^{-1} \times \left( \mathcal{R}(\xi_{N,j}, \phi(t)) y(t) - \sum_{j=1}^{N} \mathcal{R}(\xi_{N,j}, \phi(t)) \mathcal{R}(\phi(t), \xi_{N,i}) \alpha_{N,j}(t) \right).
\] (4.11)

As opposed to the previous approximation method, estimates constructed according to Equation 4.11 evolve continuously according to the ODE given by Equation 4.11, rather than approximations generated by Equation 4.8. As opposed to the previous offline optimization problem, this approximation method could, in principle, generate and update estimates in real-time. We emphasize that the theory presented in this chapter only applies to Method (1), and we leave the theoretical study of Method (2) for a future paper. However, for completeness, we examine the performance and convergence behavior of both methods in the numerical results of this study.

Like the Method (1), we present Algorithm 4.3 to outline the steps needed to implement this approximation method.
Algorithm 4.3.  **input:** kernel function \( K \), kernel hyperparameters \( \beta \), regularization parameter \( \gamma \), and desired kernel separation \( \eta \)

**kernel placement steps:** Collect data \( \{t_i, x_i, y_i\}_{i=1}^M \)

\( \Xi_N = \xi_{N,1} = x_1, N = 1 \)

Select placement of kernel centers

\( \text{for } x_i \text{ in } \{x_i\}_{i=2}^M \text{ do} \)

\( \text{if } \|\xi_{N,j} - x_i\| > \eta, \forall x_i \in S_N \)

Add sample \( x_i \) to the set of centers

\( S_{N+1} = S_N \cup \{x_i\} \)

Add sample time \( t_i \) to the set of indexed times

\( t_{N,i} = t_i, T_{N+1} = T_N \cup \{t_{N,i}\}, N = N + 1 \)

\( \text{end for} \)

Construct Kernel Matrix

\( \mathbb{K}(S_N, S_N)_{i,j} = [K(\xi_{N,i}, \xi_{N,j})] \)

Choose initial conditions

\( \phi(0) = x_0 \)

Denote \( \alpha_N = \{\alpha_{N,1}, ..., \alpha_{N,N}\}^T \) coefficient vector of estimate

Denote \( K(S_N, \cdot) = \{k(\xi_{N,1}, \cdot), ..., k(\xi_{N,N}, \cdot)\}^T \) vector of kernel functions

Numerically integrate ODEs

\( \dot{\phi}(t) = f(\phi(t)) \)

\( \dot{\alpha}_N(t) = (\gamma \mathbb{K}(S_N, S_N))^{-1} \left( K(S_N, \phi(t))y(t) - K(S_N, \phi(t))K(S_N, \phi(t))^T \alpha_N(t) \right) \)

**return** Estimate \( \hat{g}_N(t, \cdot) = \sum_{j=1}^N \alpha_{N,j}(t)K_{N,j}(\cdot) \).
4.3.3 Learning Theory and the Regression Estimate

In this section, we give an expanded discussion of the similarities and differences between the approximations described in Sections 4.3.2 and 4.3.2 in this chapter and related techniques in distribution-free learning theory, statistical learning theory, and machine learning theory. For the most part, these learning theory approaches focus on estimates generated from samples of discrete, independent and identically distributed (IID) stochastic systems. See [1, 2] for popular summaries of the state-of-the-art in these fields. Learning theory in general [34] is concerned with a number of distinct problems including pattern recognition, classification, and function estimation. The learning problem for function estimation involves approximating a mapping $G$ from a set of inputs $x \in X$ to elements $y$ in an output space $Y$. It is commonly assumed that the data is a collection of $M$ noisy sample pairs 
\[ \{z_i\}_i^M = \{x_i, y_i\}_i^M \subset Z = X \times Y \] 
that are generated from a discrete IID stochastic process defined by the probability measure $\mu$ on $Z$. Ideally, optimal estimates of $G$ are defined to be minimizers of the functional $E_\mu$, commonly referred to as the expected risk,

\[ E_\mu(g) := \int |y - g(x)|^2 \mu(dz), \]

where $\mu$ is a joint measure on the sample space $Z$. Note that the measure $\mu$ can be rewritten, 
\[ \mu(dz) = \mu(dy, dx) = \mu_Y(dy|x)\mu_X(dx), \]
where $\mu_Y(dy|x)$ is the conditional measure on $Y$ given $x \in X$ and $\mu_X$ is the marginal measure on the input space $X$. In principle, the ideal minimizer of $E_\mu$ is given by 
\[ G_\mu(x) = \int y\mu_Y(dy|x) \]
with $G_\mu$ referred to as the regressor function. However, the measures $\mu$, $\mu_Y$, and $\mu_X$ are generally unknown, and the ideal solution, $G_\mu$ above cannot be computed in practice. It
is this reason that the above problem is said to define a type of distribution-free learning problem [1].

Since the regressor \( G_{\mu} \) cannot be computed in general, standard approaches in machine learning theory replace the error functional above with its regularized, discrete counterpart

\[
E_M(g) := \frac{1}{M} \sum_{i=1}^{M} |y_i - g(x_i)|^2 + \gamma \|g\|_U^2
\]

that is defined in terms of the samples \( \{(x_i, y_i)\}_{i=1}^{M} \) of a discrete IID stochastic process. Here \( \gamma \) is the regularization parameter and \( U \) is a space of functions that has a norm that measures smoothness. When some finite dimensional space \( \mathcal{H}_N = \text{span}\{\psi_j \mid 1 \leq j \leq N\} \) is used to construct approximations, the method of empirical risk minimization (ERM) seeks the function \( \hat{g}_{N,M}(\cdot) = \sum_{j=1}^{N} \alpha_{N,j} \psi_j(\cdot) \) that is the minimizer

\[
\hat{g}_{N,M} = \min_{g \in \mathcal{H}_N} E_M(g_N)
\]

Note that the minimizer \( \hat{g}_{N,M} \) depends on the number of samples \( M \) and the number of basis functions \( N \). The convergence of \( \hat{g}_{N,M} \to G \) as \( N \) and \( M \) increase is a well-studied topic, certainly one of the most well-known in learning theory. Again, see [1, 2] for a complete description of the myriad of approaches to this problem. The relationship of the approach in this chapter to the standard learning problem can be made more precise by assuming that the basis is taken to be \( \mathcal{H}_N = \text{span}\{\xi_{N,i} \mid 1 \leq i \leq N\} \) that is the scattered basis as we use in Equations 4.10 and 4.11. In this case, it is well-known that \( \hat{g}_{N,M}(\cdot) = \sum_{j=1}^{N} \alpha_{N,j} \xi_{N,j} \) where the coefficients satisfy

\[
\sum_{j=1}^{N} \left( \sum_{i=1}^{M} \langle \xi_{N,j}(x_i), \xi_{N,k}(x_i) \rangle + \gamma \langle \xi_{N,j}, \xi_{N,k} \rangle_U \right) \alpha_{N,j} = \sum_{i=1}^{M} y_i \xi_{N,k}(x_i) \quad (4.12)
\]
These equations should be carefully compared to Equations 4.8 and 4.10. In Equation 4.12 the inner summation is over the samples from an IID process see [35, 36]. For the regression problem in continuous time in Equation 4.8, the inner summation above is replaced with an integration in time along a trajectory. We see that the use of one point quadrature rule in time, which yields Equation 4.2, generates a set of algebraic equations that have a similar structure to that which arises in learning theory for discrete stochastic processes in Equation 4.12. In fact, if the sample times are uniformly distributed, the weights of integration are constant and can be cancelled in Equation 4.8. In such a case, the two sets of equations have identical form. Although the form of the equations is the same, the error analysis for the two cases differs substantially. One significant difference is that, in the typical learning theory scenarios, it is assumed that samples are dense in \( X \), while this is a rather special case for deterministic dynamical systems. For dynamical systems, the set over which the error analysis is performed is typically unknown. The set of samples along a trajectory can be dense in a very irregular set. This fact is emphasized in the numerical example in Section 4.4. Additionally, the error analysis for Equation 4.8 relies on a PE condition that is not part of the stochastic framework. The error analysis of Equation 4.12 usually results from taking the expectation. For deterministic dynamical systems, however, there is no definition of expectation. This work instead considers when the inputs are generated along a trajectory \( t \mapsto \phi(t) \) governed by some underlying, generally unknown evolution. For a more in-depth discussion of learning theory for unknown discrete samples, see the work of Cucker and Zhou in [30] or Devito et. al. in [37].

4.3.4 Subspaces \( \mathcal{V} \) Chosen as a Spectral Space \( A^S \)

The case studied above suffices to derive rates of convergence of approximations in finite dimensional spaces \( \mathcal{H}_S \) where \( S \) is a finite set of points. From a practical point of view, the
results apply to many important cases that can be implemented. However, from a theoretical point of view, we would like to be able to identify a closed subspace $\mathcal{V}$ that is “as large as possible” in the definition in Equation 4.5. It is of interest therefore to find a space $\mathcal{V} \subseteq \mathcal{H}$ that is infinite dimensional and satisfies the PE condition. We do this by introducing spectral approximation spaces $A^s$ associated with a fixed, compact, self-adjoint operator $T : \mathcal{H} \to \mathcal{H}$.

By the spectral theorem for compact, self-adjoint operators, this means that the operator can be expressed as

$$Th := \sum_{k=1}^{\infty} \lambda_k \langle h, h_k \rangle \mathcal{H} h_k,$$

where $\{\lambda_k\}_{k \in \mathbb{N}}$ is the sequence of eigenvalues arranged in nonincreasing order and repeated as needed for multiplicity, and $\{h_k\}_{k \in \mathbb{N}} \subseteq \mathcal{H}$ is a corresponding $\mathcal{H}$-orthonormal collection of eigenfunctions. The only possible accumulation point of the eigenvalues is zero. In the following we always assume that $\lambda_k \to 0$ as $k \to \infty$, since if the sum above terminates after a finite number of terms, all the approximation spaces introduced below degenerate and are equivalent. We also assume that the kernel of $T$ is equal to $\{0\}$. By [37] Proposition 8, the eigenvectors of $T$ span $\text{nullspace}(T)^\perp$, so in the case at hand $\{h_k\}_{k \in \mathbb{N}}$ are an orthonormal basis for $\mathcal{H}$.

By virtue of the functional calculus for compact, self-adjoint operators, the operator $T^s$ is well-defined for all $s \geq 0$ by the expansion

$$T^s h := \sum_{k=1}^{\infty} \lambda_k^s \langle h, h_k \rangle \mathcal{H} h_k, \quad \text{strongly in } \mathcal{H}.$$

We define the spectral approximation space $A^s := A^s(\mathcal{H})$ to be

$$A^s := \left\{ h \in \mathcal{H} \mid \|h\|_{A^s} < \infty \right\}.$$
where the norm is given by

\[ \|h\|_{A^s}^2 := \sum_{k=1}^{\infty} \left( \lambda_k^{-s} |\langle h, h_k \rangle_{\mathcal{H}}| \right)^2 := \sum_{k=1}^{\infty} \lambda_k^{-2s} |\langle h, h_k \rangle_{\mathcal{H}}|^2 \]

The spaces \( A^s \) have a long history and are closely related to approximation spaces. [38, 39]. Since \( \lambda_k \to 0 \), the weight \( \lambda_k^{-2s} \) grows as \( k \to \infty \). The space \( A^s \) consists of functions in \( \mathcal{H} \) whose generalized Fourier coefficients \( \{ \langle h, h_k \rangle_{\mathcal{H}} \}_{k \in \mathbb{N}} \) converge faster than \( \{ \lambda_k^{-2s} \}_{k \in \mathbb{N}} \) increases. It can be shown that these spaces are nested with \( A^s \subseteq A^r \) whenever \( r \leq s \). It should also be noted that \( A^0 = \mathcal{H} \). So, in particular we have \( A^{s+1} \subseteq A^s \subseteq \ldots \subseteq \mathcal{H} \). In the language of approximation theory, the \( A^s \) define a scale of spaces for \( s \geq 0 \) containing functions of increased (generalized) smoothness as \( s \) increases.

The following theorem provides the technical connection between the spaces \( \mathcal{H} \) and spectral space \( A^1 \) in terms of the operator \( T \).

**Theorem 4.4.** We have the equivalence

\[ \|h\|_{A^s}^2 \approx \langle T^{-s}h, T^{-s}h \rangle_{\mathcal{H}}. \]

for all \( h \in \text{domain}(T^{-s}) \equiv A^s \) and \( T^s : \mathcal{H} \to A^s \) is an isometry.

**Proof.** Suppose \( h \in \mathcal{H} \). Then

\[ \langle T^{-s}h, T^{-s}h \rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \lambda_k^{-2s} |\langle h, h_k \rangle_{\mathcal{H}}|^2 = \|h\|_{A^s}^2. \]

It is also immediate that \( ||T^s h||_{A^s} = ||h||_{\mathcal{H}} \), so \( T^s \) is an isometry from \( \mathcal{H} \) onto \( A^s \). In particular \( T : \mathcal{H} \to A^1 \) is an isometry.

In view of Theorem 4.4, when the PE condition holds, we have constants \( \gamma_1, \gamma_2 > 0 \) such
that
\[ \gamma_1 h \| h \|_{A^1}^2 \leq \langle T_\phi(t, t + \Delta) h, h \rangle_\mathcal{H} \leq \gamma_2 h \| h \|_{A^1}^2 \]
for all \( h \in A^1 \subseteq \mathcal{H} \) and \( t \in \mathbb{R}^+ \). Note that this equivalence holds uniformly for the family \( \{ T_\phi(t, t + \Delta) \}_{t \in \mathbb{R}^+} \) for all \( t \in \mathbb{R}^+ \). This pair of inequalities can also be interpreted as the statement that \( T_\phi(t, t + \Delta) \approx T \) on \( A^1 \).

Now we return to the study of the error when we choose \( V := A^1 \), and we seek the optimal \( \hat{g}_V(t, \cdot) \in A^1 \). In analogy to Case 1, we set
\[
\begin{align*}
\Delta G &:= (I - \Pi_V)G, \\
y(t) &= (\Pi_V G)(\phi(t)) + \Delta G(\phi(t)), \\
\tilde{g}_V(t, \cdot) &= \hat{g}_V(t, \cdot) - \Pi_V G.
\end{align*}
\]
(4.13)

Following the same plan of attack as in Case 1, we obtain
\[
\tilde{g}_V(t, \cdot) = (\Pi_V (T_\phi(0, t) + \gamma I) \Pi_V)^{-1} (\Pi_V T_\phi(0, t) \Delta G - \gamma \Pi_V G)
\]
(4.14)

In this case, in contrast, we can write
\[
\langle (T_\phi(0, t) + \gamma I) h, h \rangle_\mathcal{H} \geq \sum_{k=1}^{m} \langle T_\phi((k - 1)\Delta, k\Delta) h, h \rangle_\mathcal{H} \\
+ \gamma \langle h, h \rangle_\mathcal{H} \\
\geq (\gamma_1 m + \gamma) h \| h \|_{A^1}^2
\]
for all \( h \in A^1 \subseteq \mathcal{H} \). In other words \( (T_\phi(0, t) + \gamma I)^{-1} \) restricted to \( A^1 \subseteq \mathcal{H} \) is a bounded
linear operator that satisfies

\[ \| (T_\phi(0,t) + \gamma I)^{-1} h \|_\nu \leq \frac{1}{\gamma m + \gamma} \| h \|_\mathcal{H}. \] (4.15)

By definition, \( \Pi_\nu G \in A^1 \) and \( \Pi_\nu T_\phi(0,t) \Delta G \in A^1 \). Since \( (\Pi_\nu T_\phi(t) \Delta G - \gamma \Pi_\nu G) \in A^1 \), we can apply the bound in Equation 4.15 to Equation 4.14. The remainder of the proof is unchanged and we conclude that, if \( G \in A^1 \),

\[ \limsup_{t \to \infty} \| \hat{g}_\nu(t, \cdot) - \Pi_\nu G \|_\mathcal{H} = 0. \]

**Observations:**

1. It should be emphasized that the operator \( T_\phi(t, t + \Delta) : \mathcal{H} \to \mathcal{H} \) is compact, as described in Theorem 4.1. But when the PE condition holds with \( \nu := A^1 \), it is not compact as an operator \( \mathcal{H} \to A^1 \). It is boundedly invertible as a map from \( \mathcal{H} \to A^1 \).

2. Intuitively, the PE condition can be understood as a statement that the local approximation space defined over the small time-span \([\tau, \tau + \Delta]\) in terms of the operator \( T_\phi(\tau, \tau + \Delta) \) is spectrally equivalent to the global approximation space defined over \([0, t]\) in terms of \( T_\phi(0, t) \).

### 4.4 Numerical Examples

The error estimates above apply to quite general situations. Since some of our earlier works in [14, 19, 22] have included numerical examples with evolutions on compact manifolds, here we model a trajectory that is dense in a complicated, unknown subset in \( \mathbb{R}^n \). Consider the
Lorenz system

\[
\begin{align*}
\dot{x}(t) &= \sigma(y(t) - z(t)), \\
\dot{y}(t) &= r(x(t) - y(t) - x(t)z(t)), \\
\dot{z}(t) &= x(t)y(t) - bz(t)
\end{align*}
\]

for \( t \in \mathbb{R}^+ \). Figures 4.2a and 4.2b illustrate orbits of the system for various initial conditions. Set \( X = \mathbb{R}^3 \) and denote by \( \phi(t) := \{x(t), y(t), z(t)\}^T \). The complex nature of the trajectories of this system has been studied and commented on so extensively that it is now understood as an exemplar of what chaos and complexity is, even in the popular press.

Figure 4.2: The dynamics of the Lorenz system for several initial conditions (a) in three-dimensions (b) projected onto the \( x - y \) plane.

There are a number of Lyapunov functions that have been introduced to study the long-term behavior of this system. One common choice is

\[
V = rx^2 + \sigma y^2 + \sigma(z - 2r)^2 \geq 0
\]
for all \((x, y, z) \in X\). Its derivative along trajectories is given by

\[
\dot{V} = -2\sigma(rx^2 + y^2 + bz^2 - 2brz),
\]

which is negative outside of the compact set

\[
\Omega := \{(x, y, z) \in X \mid \dot{V} \geq 0\}.
\]

The fact that this set is compact follows by demonstrating that the boundary of the set \(\{(x, y, z) \in X \mid \dot{V} = 0\}\) is an ellipse given by

\[
1 = \frac{x^2}{(\sqrt{br})^2} + \frac{y^2}{(\sqrt{br})^2} + \frac{(z - r)^2}{r^2}.
\]

Given that \(\dot{V}\) is continuous and that points inside the ellipse satisfy the following inequality

\[
0 \geq rx^2 + y^2 + bz^2 - 2brz,
\]

it is clear that the set \(\Omega\) is the closure of the interior of the ellipse. Since \(\Omega\) is compact and \(V\) is continuous, the maximum \(\bar{V}\) of \(V\) over \(\Omega\) is achieved, \(\bar{V} = \max_{\phi \in \Omega} V(\phi)\). Define the dilation of the set \(\Omega\) by some parameter \(\epsilon > 0\) to be

\[
\Omega_\epsilon := \{\phi \in X \mid V(\phi) \leq \bar{V} + \epsilon\}.
\]

The set \(\Omega_\epsilon\) is positive invariant. Any trajectory starting at \(\phi_0 \not\in \Omega\) is guaranteed to enter \(\Omega_\epsilon\) in finite time and never leave this set. This means that for any initial condition \(\phi_0 \not\in \Omega\), the orbit \(\Gamma(\phi_0) := \cup_{t \geq 0} \phi(t)\) is precompact, that is, \(\overline{\Gamma(\phi_0)}\) is compact. From standard results on dynamical systems [40], it is known that the positive limit set \(\omega^+(\phi_0)\) of a precompact
trajectory $t \mapsto \phi(t)$, which is defined by

$$\omega^+(\phi_0) := \{ y \in X \mid \exists t_k \to \infty \text{ such that } \phi(t_k) \to y \},$$

is compact. In fact, we have

$$\phi(t) \to \omega^+(\phi_0) \subseteq \overline{\Gamma^+(\phi_0)}.$$

Both $\omega^+(\phi_0)$ and $\overline{\Gamma^+(\phi_0)}$ are guaranteed to be compact sets, but they can be highly irregular.

For the case at hand, where we study the Lorenz system, this fact is well-known.

We want to use the results of this chapter to understand what can be said about the regression problem in continuous time for this system, and we are particularly interested in what the error bounds imply for estimates of an observable function for this system. We would like to understand how the trajectory $t \mapsto \phi(t)$ affects convergence of approximations, and to determine in what spaces the continuous time regression problem converges. We can use either of the sets $\omega^+(\phi_0)$ or $\overline{\Gamma^+(\phi_0)}$ to study the convergence properties.

In this chapter, we study the case when the set $S = \overline{\Gamma^+(\phi_0)}$. We define different sets $S_N \subset S$ that have $N$ samples in $\Gamma^+(\phi_0)$, $S_N := \{ \xi_1, \ldots, \xi_N \} := \{ \phi(t_1), \phi(t_2), \ldots, \phi(t_N) \}$. We assume that these are nested, $S_N \subset S_{N+1}$. Associated with $S_N$ we define the space of approximants in terms of a scattered basis with $\mathcal{H}_N := \text{span}\{ \mathbf{h}_{\xi_1}, \ldots, \mathbf{h}_{\xi_N} \}$. These finite dimensional spaces of approximants are data driven: they are generated along a trajectory of the system. For each $N$, suppose that the PE condition holds for $\mathcal{H}_N$. Whether or not the PE condition holds in the case that $V = \mathcal{H}_N$ can be verified by conditions related to the visitation, or time of occupation, of the trajectory $\tau \mapsto \phi(\tau)$ in neighborhoods of the samples in $S_N$. See [15, 22, 26] for a discussion. If the trajectory persistently excites the
subspace $\mathcal{H}_N$, we have the estimate from Equation 4.7

$$
\|\hat{g}_N(t, \cdot) - \Pi_N G\|_{\mathcal{H}_N} \\
\leq \left( \frac{\tilde{R}^2 m \Delta}{\gamma_1 m + \gamma} \right) \| (I - \Pi_N) G \|_{\mathcal{H}_N} + \frac{\gamma}{\gamma_1 m + \gamma} \| \Pi_N G \|_{\mathcal{H}_N}.
$$

Concrete estimates of the rate of convergence of this expression can be obtained using the power function $P_N(x)$ over the set $S = \Gamma^+(\phi_0)$,

$$
P_N(x) := |\mathcal{K}(x, x) - \mathcal{K}_N(x, x)| \quad \text{for all } x \in S := \Gamma^+(\phi_0)
$$

where $\mathcal{K}_N$ is the kernel that defines the native space $\mathcal{H}_N$. It is well-known [31, 41] that the power function $P_N(x)$ enables the pointwise bound

$$
|(I - \Pi_N)h(x)| \leq P_N(x)\|h\|_{\mathcal{H}} \quad \text{for all } x \in S := \Gamma^+(\phi_0).
$$

This pointwise bound can be used to derive a corresponding bound on $\| (I - \Pi_N)h \|_{\mathcal{H}}$, for smooth enough $h$. The details exceed the length of this brief paper and are given in [25] in a different application to approximation of Koopman operators, or this bound can be inferred from the proof of Theorem 11.23 in [31]. Ultimately, we obtain a bound

$$
\| (I - \Pi_N)h \|_{\mathcal{H}} \lesssim \left( \int_S |P_N(x)|^2 dx \right)^{1/2} \|h\|_{\mathcal{H}}
$$

for all $h \in \mathcal{H}$ that are smooth enough, which means that the optimal regression estimate in
CHAPTER 4. KERNEL METHODS FOR REGRESSION IN CONTINUOUS TIME OVER SUBSETS AND MANIFOLDS

continuous time satisfies

\[ \| \hat{g}_N(t, \cdot) - \Pi_N G \|_{\mathcal{H}_S} \]
\[ \leq \left( \left( \frac{\tilde{R}^2 m \Delta}{\gamma_1 m + \gamma} \| P_N \|_{L^2(S)} \right) \| G \|_{\mathcal{H}_S} + \frac{\gamma}{\gamma_1 m + \gamma} \| \Pi_N G \|_{\mathcal{H}_S} \right) \]  \hspace{1cm} (4.17)

for all \( G \) smooth enough.

We begin with an assessment of the numerical implementation of Method (1). We illustrate the performance of an estimate \( \hat{g}_N(t, \cdot) \) defined in Equation 4.8 of \( \Pi_N G \) generated over an orbit \( \Gamma(\phi_0) \) of the Lorenz system. We only pose the regression problem over a projection of the orbit \( \Gamma(\phi_0) \) onto the \( x - y \) plane so that the results are easy to visualize. The function \( G \) we estimate is given by

\[ G(x, y) = -10 \sin \left( \frac{y}{10} \right) + \frac{(x + y)^3}{2000} + 200. \]

and we choose the initial condition \( \phi_0 = \{1, 1, 1\}^T \). In this example, we use the Matern-Sobolev kernel,

\[ \mathcal{K}(\xi_{N, j}, x) = \left( 1 + \frac{\sqrt{3}\|x - \xi_{M,j}\|_2}{\beta} \right) e\left( -\frac{\sqrt{3}\|x - \xi_{M,j}\|_2^2}{\beta} \right) \]

with \( \| \cdot \|_2 \) the standard Euclidean norm over \( \mathbb{R}^2 \), \( j \in 1, ..., N \), and the hyperparameter \( \beta = 5 \).

In Figure 4.3, the output \( G(\phi(t)) \) for \( t \geq 0 \) is represented by the red curve that hovers over the dynamics of the input orbit labeled by the black curve \( \phi(t) \) in the \( x - y \) plane. In this figure, 126 kernel centers are selected quasi-uniformly along the orbit with a separation distance of around 2 and the regularization parameter \( \gamma = 0.1 \).

From the figure, it is clear that the approximation \( \hat{g}_N(t, \cdot) \) represented by the colored mesh yields, qualitatively speaking, a good estimate of the true function \( G \) (the green surface) over the orbit.
Figure 4.3: An illustration of the kernel estimate $\hat{g}_N(t, \cdot)$ of $G$ for $t = 200$ seconds defined over the orbit $\phi(t)$ starting at some initial conditions $\phi_0 = \{1, 1, 1\}^T$. The orbit is generated by projecting the Lorenz system dynamics onto the x-y plane. The output $G(\phi(t))$ for $t \geq 0$ is represented by the red curve hovering over the dynamics of the input orbit $\phi(t)$. The estimate $\hat{g}_N(t, \cdot)$ represented by the colored mesh minimizes its error from the projection of the true function $G$ (the curved surface) over the orbit $\phi(t)$. 
When interpreting this result, it is important to keep several facts in mind.

**Observations:**

1. The theory in this chapter uses the compact subset $S := \overline{\Gamma(\phi_0)}$, whose regularity is not easy to characterize. The set $\overline{\Gamma(\phi_0)}$ defines the space $\mathcal{H}_S$ in which regression approximations in continuous time converge.

2. The convergence of estimates is in the space $\mathcal{H}_S$, which is an RKHS space of functions over the set $X$. Even though $S$ is quite irregular, the functions in $\mathcal{H}_S$ are supported on the whole set $X$, not just $S$. This means that estimate is, in a sense, “naturally extended” to the whole state space $X$. In the case at hand, the set $S$ has zero Lebesgue measure. Even though the trajectory or orbit may not reach some points or subsets of $\mathbb{R}^2$, the function estimates are well-defined everywhere nonetheless.

### 4.4.1 Example: Characteristics of Approximation Method (1)

The next set of results examines the estimates of approximation Method (1) over orbits spanning different intervals of time. Using the same underlying input dynamics, kernel function, hyperparameter $\beta$, regularization parameter $\gamma$, and unknown function $G$ from the previous results, each of the estimates are generated from an orbit starting at an initial condition $\phi_0 = \{1, 1, 1\}^T$. The centers are placed quasi-uniformly along the orbit with a separation distance of around 2. Figures 4.4a through 4.4d illustrate the estimates calculated over different spans of time. In Figure 4.4a, we can see that, even for small time intervals, the error between the estimate $\hat{g}_N(t, \cdot)$ and the true function $G$ begins to diminish significantly over the orbit. Additionally, there is significant decrease in the error over the trajectory as seen in Figures 4.4a and 4.4b. While the error continues to decrease as $t \to \infty$, it is evident the error reduction between Figures 4.4c and 4.4d occurs at a much smaller rate than the
previous time intervals. This is a consequence of diminishing rate of reduction in the power function $p_N(x)_{x \in \Gamma(\phi_0)}$ as $t \to \infty$ and more samples are collected.

![Numerical Examples](image)

Figure 4.4: An illustration of the various approximations from Equation 4.8 over an orbit after spanning different amounts of time from the same system dynamics and initial conditions. It is evident that the error $\|\Pi_N G - \hat{g}_N(t, \cdot)\|$ converges to zero over the domain of attraction as more time is passed and more samples are collected.

For the next set of figures, we examine the effects of the regularization penalty term of $E(t, g; \phi)$ used in approximation Method (1) by varying the choices of $\gamma$. In order to examine the influence of the regularization parameter on the long-term convergence behavior, each simulation is run over a sufficiently long time interval of 200 seconds. Using the same Lorenz
system, kernel function, hyperparameter $\beta$, unknown function $G$, initial conditions, and center spacing as the previous results, Figures 4.5a through 4.5d illustrate the estimates for different values of $\gamma$. Overall, these four graphs depict qualitative behavior that is well-known in the field of inverse problems. The error functional $E$ introduced in Section 4.2.1 balances two terms

$$E(t, g; \phi) := \frac{1}{2} \int \left| g(\tau) - \mathbb{E}_{\phi(\tau)} g \right|^2 \nu(d\tau) + \frac{1}{2} \gamma \|g\|^2_{\mathcal{H}}.$$ 

Minimizing term 1 decreases the error over the orbit $\Gamma(\phi_0)$, while term 2 penalizes the size of the estimate as measured in the $\| \cdot \|_{\mathcal{H}}$ norm. Increasing $\gamma$ generally leads to smoother estimates. Additionally, it also reduces the chance of over-fitting, which can lead to poor estimates in the presence of a noisy or perturbed data set. This classic phenomenon is studied in great depth in texts like [42]. With these considerations in mind, proper estimates can be generated by selecting a particular choice of $\gamma$ to effectively regularize the estimate.

4.4.2 Example: Characteristics of Approximation Method (2)

The next set of results examine how the estimates generated using approximation Method (2) vary in time and converge as $t \to \infty$. Using the same conditions as the previous results, the estimate was initialized with all coefficients $\alpha_i = 0$ for $1 \leq i \leq N$ and the states initial condition $\phi_0 = \{1, 1, 1\}^T$. Figures 4.6a through 4.6d illustrate the estimate generated after different amounts of time. In contrast to the previous approximation method, these figures depict the evolution of a single estimate as $t \to \infty$. As mentioned previously, this approximation is determined by an evolution of coefficients $\{\hat{\alpha}_{N,i}(t)\}_{i=1}^N$. The evolution law minimize the integrated error for $\hat{g}_N(\tau) = \sum_{i=1}^{N} \hat{\alpha}_{N,i}(\tau) R_{\xi_{N,i}}(\cdot)$ over the orbit for time $\tau \in [0, t]$ rather than computed an offline optimal solution at a specific time. At early
Figure 4.5: Estimates $\hat{g}_N(t, \cdot)$ of the function $G$ for different values of $\gamma$ using the same time-span, initial condition, and underlying dynamics. Notice that for larger $\gamma = 20$, the estimate has a slower convergence over the domain of attraction. While estimates using smaller $\gamma$, may have reduced error over the orbit, the estimate can potentially over-fit noisy training data.
stages of the evolution such as Figure 4.6a, samples predominantly aggregate near one of the unstable equilibrium points. However, as the trajectory approaches the second equilibrium, it begins to influence the coefficients of the nearby centers and decrease the error between the estimate $\hat{g}_N(t, \cdot)$ and $G$. Figures 4.6c and 4.6d suggest that the estimate converges to the projection $\Pi_N G$ over the domain of attraction as more time passes and more samples are collected. Similar to the estimate from approximation Method (1), there is an initial rapid change in the error over the trajectory. While the error appears to decrease for longer periods of time, it is evident that the error reduction between Figures 4.6c and 4.6d occurs at a much lower rate than the reduction seen from Figure 4.6a to 4.6b and Figure 4.6b to 4.6c.

Using the same conditions as those used in the regularization study of approximation Method (1), we also examine the effects of the regularization term on the estimate from approximation Method (2) by building estimates using different choices of $\gamma$. Figures 4.7a through 4.7d illustrate the estimates for different values of $\gamma$. In these estimates, the $\gamma$ term also plays a role in the transient response of the coefficients’ evolution. From the figures, it is evident that a larger $\gamma$ decreases the sensitivity to changes in the estimate as new samples are collected over time. Consequently, the numerical study suggests better convergence of the estimates may require a larger number of samples when $\gamma$ is large. By decreasing $\gamma$, the estimates are more responsive to changes in the data as seen in Figure 4.7c. However, even without noise or disturbances, smaller values of $\gamma$ can yield large oscillatory behavior in the estimate as seen in Figure 4.7d. This example indicates that $\gamma$ must be carefully selected for a desired transient response in the second approximation method.
Figure 4.6: An illustration of the evolution approximation from Equation 4.10 over an orbit as time $t \to \infty$. The error $\| \Pi_N \mathbf{G} - \hat{g}_N(t, \cdot) \|$ converges over the domain of attraction as more time is passed and more samples are collected.
Figure 4.7: Demonstration of estimates $\hat{g}_N(t, \cdot)$ of the function $G$ for different values of $\gamma$ using the same time-span, initial condition, and underlying dynamics. Notice that the larger $\gamma = 20$ has a slower convergence over the domain of attraction. For smaller $\gamma$, such the estimate generated using $\gamma = 0.001$, the estimate suffers from less penalty on the regularization of the estimate. However, the estimate can over-fit noisy training data.
4.4.3 Example: Human Kinematics Study

This example uses three-dimensional motion capture data from a subject running along a treadmill [43]. From the experiment, 17,000 marker coordinates are collected relative to a fixed inertial frame defined by the camera’s position. For this example, a small candidate kinematic model is defined in terms of the full collection of experimental trajectories. The marker coordinates of the hip, knee, and ankle in the full data set are projected onto the sagittal plane that divides the left and right half of the body, see Figure 4.8.

With this projection, the first input is defined to be the joint angle $\theta^{(1)}$. It is measured between the projected vector $v^{(1)}$ that connects the hip to knee and the body-fixed $b^{(1)}$ vector in the plane, and it roughly corresponds to hip flexion. The second input comes from the knee flexion angle $\theta^{(2)}$. It is measured between the vector $v^{(1)}$ and $v^{(2)}$, the vector connecting the knee to the ankle projected to the sagittal plane. These choices of projections and associated degrees of freedom are taken to define a low-dimensional but unknown dynamic system. We seek to estimate outputs of the unknown dynamic system.

For illustrative purposes, we chose to estimate the projection $\Pi_N G$ of the kinematic map $G$ from the joint angles $\theta^{(1)}$ and $\theta^{(2)}$ to the ankle coordinate associated with the body-fixed $b^{(2)}$ vector. Figure 4.9a and 4.9b show the approximations $\hat{g}_N(t, \cdot)$ of $\Pi_N G$ over an orbit in the input space defined by the coordinates $\theta^{(1)}$ and $\theta^{(2)}$. The estimates are generated for both approximation Method (1) and (2), respectively. In both estimates, the Matern-Sobolev kernel with $\beta = 10$ is used. Additionally, the kernel centers for both estimates are chosen so that there is sufficient separation distance of at least 10 radians between each of the kernel centers, $\Xi_N$. As mentioned previously, the two estimates respond differently to the regularization term. A relatively small regularization parameter $\gamma = 0.001$ was selected for the estimate from approximation Method (1). The second approximation exhibited
oscillations over the data for small values of $\gamma$. Consequently, we increased the regularization parameter to $\gamma = 2$ for the estimate from the second approximation method.

When comparing these approximations there are a couple of key things to note. The estimate using the second approximation method in this study is generated using a continuous evolution law as given by Equation 4.11. However, the collected data consists of discrete sets of motion capture data collected at discrete times. In order to utilize the collected data and also have a continuous evolution update, we use MATLAB’s built-in functions to fit splines that are continuous in time with knots at the discrete states. This builds a continuous approximation of the orbit over small time intervals. While the approximation given by Method (1) does not need this step, we must approximate the integral over the orbits given the discrete data and a particular quadrature rule. Consequently, the estimate of Method (1) is susceptible to error associated with quadrature approximations.

![Sagittal Plane Treadmill](image)

Figure 4.8: An illustration of the inputs $\theta^{(1)}$ and $\theta^{(2)}$, which roughly correspond to hip and knee flexion respectively. These input variables can be mapped to measured marker coordinates placed on joints such as the knee or ankle.
4.5 Conclusions

Figure 4.9: The estimates $\hat{g}_N(t, \cdot)$ of the ankle coordinate in the $b^{(2)}$ direction over the input space given by the angles $\theta^{(1)}$ and $\theta^{(2)}$ associated with hip and knee flexion respectively using (a) Approximation Method (1), and (b) Approximation Method (2). It must be noted that the first Approximation method shows relatively small error for the given sample size compared to Method (2).

4.5 Conclusions

In this chapter, an optimal, offline estimation problem is formulated for continuous time regression over state spaces that include certain types of smooth manifolds. A new persistency of excitation condition is introduced that is well-defined for flows on manifolds, and it is used to obtain sufficient conditions for convergence. Error estimates are derived that characterize the rate of convergence of finite dimensional estimates of the solution of the regression problem in continuous time over manifolds. We then discuss two methods to generate finite-dimensional approximations of the optimal regression estimate. Numerical simulations are presented to better illustrate the qualitative behavior of the two approximation methods. Finally, we discuss and analyze results on estimating functions over motion capture data to demonstrate how to implement the algorithm on experimental data.
4.6 Appendix

4.6.1 Background on Galerkin Approximations

Let $U$ be a real Hilbert space, $A \in \mathcal{L}(U)$ be a bounded linear operator on $U$, $b \in U$ be a fixed element of $U$, and suppose we seek to find $u \in U$ that satisfies the operator equation

$$Au = b.$$ 

It is customary that the existence and uniqueness of the solution of this equation is established by studying the associated bilinear form $a(\cdot, \cdot) : U \times U \to \mathbb{R}$ given by $a(u, v) := (Au, v)$ for all $u, v \in U$. Then the operator equation above is equivalent to finding the $u \in U$ for which

$$a(u, v) = \langle b, v \rangle_U \quad \text{for all } v \in U. \quad (4.19)$$

The Lax-Milgram Theorem given below stipulates a concise pair of conditions that ensure the well-posedness of the operator equation.

**Theorem 4.5** (Lax-Milgram Theorem [44]). *The bilinear form $a(\cdot, \cdot) : U \times U \to \mathbb{R}$ is bounded if there is a constant $C_1 > 0$ such that

$$|a(u, v)| \leq C_1 \|u\|_U \|v\|_U \quad \text{for all } u, v \in U,$$

and it is coercive if there is a constant $C_2 > 0$ such that

$$|a(u, u)| \geq C_2 \|u\|_U^2 \quad \text{for all } u \in U.$$*

If the bilinear form $a(\cdot, \cdot)$ is bounded and coercive, then $A^{-1} \in \mathcal{L}(U)$ and there is a unique
solution $u \in U$ to Equation 4.19.

Proof. The first condition above, continuity of the bilinear form, ensures that $A \in \mathcal{L}(U)$ by definition. The coercivity condition implies that the nullspace of $A$ is just $\{0\}$. As a result, we know that $A$ is one-to-one. This means that the operator $A^{-1} : \text{range}(A) \to U$ is well-defined. From the coercivity condition we also conclude that

$$\|A^{-1}b\|^2 \leq \frac{1}{C_2} |(A \cdot A^{-1}b, A^{-1}b)_U| \leq \frac{1}{C_2} \|b\|_U \|A^{-1}b\|_U$$

for every $b \in \text{range}(A)$. This means that $A^{-1} \in \mathcal{L}(\text{range}(A), U)$ and $\|A^{-1}\| \leq 1/C_2$.

One implication of the fact that $A^{-1} \in \mathcal{L}(\text{range}(A), U)$ is that range$(A)$ is closed. Suppose that $\{b_k\}_{k \in \mathbb{N}} \subset \text{range}(A)$ and $b_k \to \bar{b}$. By construction there is a sequence $\{u_k\}_{k \in \mathbb{N}} \subset U$ such that $Au_k = b_k$. But we have

$$\|u_m - u_n\| = \|A^{-1}(y_m - y_n)\|_U \leq \|A^{-1}\| \cdot \|y_m - y_n\| \to 0,$$

and $\{u_k\}_{k \in \mathbb{N}}$ is a Cauchy sequence in the complete space $U$. There is a limit $u_k \to \bar{u} \in U$.

By the continuity of the operator $A$, we know that $A\bar{u} = \bar{b}$, hence $\bar{b} \in \text{range}(A)$. The range of $A$ is consequently closed.

It only remains to show that range$(A) = U$. Suppose to the contrary there is a $\bar{b} \neq 0$ with $\bar{b} \in (\text{range}(A))^\perp$. Since $\mathcal{N}(A^*) = (\text{range}(A))^\perp$, we know that

$$\langle A^*b, w \rangle_U = \langle b, Aw \rangle_U = 0$$

By the coercivity condition, we must have $0 = \langle A\bar{b}, \bar{b} \rangle_U \geq C_2 \|ar{b}\|_U^2 \neq 0$. But this is a contradiction and range$(A)$ is all of $U$. □
Next, we discuss how error bounds are derived for Galerkin approximations $u_N$ of the solution $u$ of the operator equations above. Let $U_N \subseteq U$ be a finite dimensional subspace of $U$. By definition, the Galerkin approximation $u_N \in U_N$ is the unique solution of the equation

$$a(u_N, v_N) = \langle b, v_N \rangle_U \quad \text{for all } v_N \in U_N. \tag{4.20}$$

The theorem below summarizes one of the well-known bounds on the error $u - u_N$ between the Galerkin approximation $u_N \in U_N$ and the true solution $u \in U$.

**Theorem 4.6** (Cea’s Lemma, [44]). Suppose that the hypotheses of the Lax-Milgram Theorem 4.5 hold. There is a unique solution $u_N \in U_N$ of the Galerkin Equation 4.20. The error $u - u_N$ is $a$-orthogonal to the subspace $U_N$ in the sense that

$$a(u - u_N, v_N) = 0 \quad \text{for all } v_N \in U_N.$$

We also have the error bound

$$\|u - u_N\|_U \leq \frac{C_1}{C_2} \min_{v_N \in U_N} \|u - v_N\|_U = \frac{C_1}{C_2} \|(I - \Pi_N)u\|_U$$

where $\Pi_N$ is the $U$-orthogonal projection of $U$ onto $U_N$.

**Proof.** First, note that when $a(\cdot, \cdot)$ satisfies the boundedness and coercivity conditions, its restriction $a : U_N \times U_N \to \mathbb{R}$ to $U_N$ satisfies the boundedness and coercivity conditions with the same constants relative to $U_N$. This means that the Galerkin equations have a unique solution $u_N \in U_N$. Since Equation 4.19 holds for all $v \in U$, it holds for all $v_N \in U_N$. We can subtract Equations 4.19 and 4.20 for each $v_N \in U_N$ and obtain

$$a(u - u_N, v_N) = 0$$
for each $v_N \in U_N$. Using the boundedness and coercivity of the bilinear form, as well as the $a-$orthogonality of the error, we have
\[
C_2\|u - u_N\|_U^2 \leq a(u - u_n, u - u_N) = a(u - u_n, u - v_N) \\
\leq C_1\|u - u_N\|_U\|u - v_N\|_U
\]
for any $v_N \in U_N$. The theorem now follows after canceling the common term on the right and left. □

Bibliography


[23] Jia Guo, Sai Tej Paruchuri, and Andrew J Kurdila. Approximations of the reproducing


Chapter 5

Conclusion

This dissertation focuses on the recent work in [1], [2] and [3]. In Chapter 2, methods were explored to generate estimates of the motion of reptiles through approximations of mappings over certain smooth manifolds. The analysis in [1] determines error bounds for the estimation of functions over the motion manifolds. However, in [1] it is assumed that a known “template manifold” underlies a given motion regime.

In Chapter 3, investigations were done into estimating a function over the unknown manifold that does not require explicit knowledge of the manifold itself. Despite the fact that the manifold is unknown, strong direct theorems of approximation are derived that tell how fast estimates converge over the unknown manifold. The results are derived using Koopman theory, a popular operator-theoretic approach to the study of nonlinear dynamic systems. The analysis in this dissertation has the added benefit that it strengthens or refines convergence results in recent papers on the approximation of Koopman operators.

In Chapter 4, error bounds for regression in continuous time over subsets of certain types of Riemannian manifolds were derived. A new notion of persistency of excitation (PE) was defined for the estimation problem over the manifold, and rates of convergence of the continuous time estimates are derived using the PE condition. We discussed and analyzed two approximation methods of the exact regression solution. Additionally, the chapter contains several numerical simulations that illustrate the qualitative character of the computed function estimates. The analysis in this dissertation develops the framework for regression over
continuous time.

Bibliography


Chapter 6

Future Work

The estimation methods in [1] [2] [3] use kernels $K$ defined over the entire ambient space $X$, which when restricted to the embedded manifold $Q$ yields convergent approximations. The future work proposes a method that attacks this problem in another way altogether. In the approach described in this section, the embedded manifold $Q$ is approximated by a nonuniform grid, and estimates are defined in terms of an adaptively refined tree approximation over the nonuniform grid. This method has a long history for approximation of functions via regression in Euclidean spaces, see [4, 5] and the references therein. Future work focuses on a related method for manifolds. This is interesting in its own right, but it also clarifies some of the positive and negative features of the earlier kernel-based approaches. Additionally, this method needs a rigorous error analysis of the tree-approximation method over manifolds similar to past works. Here we give a brief review of multilevel meshes over Euclidean space followed by a discussion of multilevel meshes over a submanifold.

6.1 Multilevel Meshes and Trees over $\Omega \subset X = \mathbb{R}^d$

In this section, let $\Omega$ denote a compact subset of $\mathbb{R}^d$ that contains the embedded manifold $Q \subset \Omega \subset X$. In the spirit of [4, 5], define a sequence $\{G_j\}_{j \in \mathbb{N}}$ of nested partitions of $\Omega := G_0$, each partition $G_{j+1}$ representing a uniform refinement of each set in $G_j$ into a subsets. The index $j$ is referred to the refinement level of the grid, and the cardinality $\#(G_j) := a^j$. We
denote by $I \in \mathcal{G}_j(X)$ a particular but otherwise generic element in the grid having refinement level $j$. We define $\mathcal{G} = \bigcup_{j=0}^{\infty} \mathcal{G}_j$, which can be thought of a multilevel set of grids over $\Omega$.

We define a tree $\mathcal{T}$ on the multilevel structure $\mathcal{G}$ in an intuitive way. Any tree $\mathcal{T}$ includes the root node $I = \Omega \in \mathcal{G}_0$, which has children $J \in C(I) \subset \mathcal{G}_1$. Proceeding recursively, we obtain successive generations of children and parents. An arbitrary element $I \in \mathcal{T}$ belongs to some $\mathcal{G}_j$, and it always has one parent $J = P(I) \in \mathcal{G}_{j-1}$. If $I \in \mathcal{G}_j$ has children, they are elements $J \in C(I) \subset \mathcal{G}_{j+1}$.

Approximations will be constructed in terms of some basic operations that define subtrees of some given tree. A proper subtree $\tilde{\mathcal{T}}$ of a tree $\mathcal{T}$ is a collection of nodes such that the root $\Omega \in \tilde{\mathcal{T}}$ is in $\mathcal{T}$, and for any other node $I \neq \Omega$ that is in $\tilde{\mathcal{T}}$ its parent $P(I) \in \tilde{\mathcal{T}}$. From any proper subtree $\tilde{\mathcal{T}}$ we can build an associated (potentially nonuniform) partition $\Lambda = \Lambda(\tilde{\mathcal{T}})$ of the set $\Omega$. This partition is defined in terms of the outer leaves of a proper subtree. The outer leaves of a subtree $\tilde{\mathcal{T}}$ consist of those nodes $J \in \mathcal{T}$ for which $J \notin \tilde{\mathcal{T}}$, but the parent $P(J)$ is contained in the subtree $\tilde{\mathcal{T}}$.

Given a partition $\Lambda$, denote $H_\Lambda$ the space of piecewise constant functions over partitions $I \in \Lambda$. Each $h \in H_\Lambda$ can be expressed as $h = \sum_{I \in \Lambda} a_I \chi_I$ where $\chi_I$ is the indicator function, $\chi_I(x) = 1$ for $x \in I$ and $\chi_I(x) = 0$ for $x \notin I$. Tree approximation over Euclidean spaces as studied in [4, 5] studies approximation in $L^2_{\mu}(\Omega)$, where $\mu$ is a measure supported on the set $\Omega$ and samples are IID. For a general $f \in L^2_{\mu}(\Omega)$, the best approximation $f_\Lambda \in H_\Lambda$ to $f$ is $f_\Lambda = \sum_{I \in \Lambda} c_I \chi_I$ where the constant $c_I$ is given by $\int f \mu(dx)/\mu(I)$, the $\mu$-local average of $f$ over $I$. Tree approximations in $L^2_{\mu}(\Omega)$ are constructed in these references using an error indicator function $\varepsilon_I(f)$. This function is used to decide whether a particular node in the tree should be subdivided into children. For any node in the tree $\mathcal{T}$ and any $f \in L^2_{\mu}(\Omega)$ the
potential decrease in the projection error $\varepsilon_I(f)^2$ is given by

$$\varepsilon_I(f)^2 = \|f - c_I\|_{L^2_\mu}^2 - \sum_{J \in C(I)} \|f - c_J\|_{L^2_\mu}^2$$

A recursive method of defining a subtree, and consequently defining a nonuniform partition over $\Omega$, is based on empirical estimates of $\varepsilon_I(f)$. Given a tolerance $\eta > 0$ we can define the smallest proper subtree $\mathcal{T}_\eta$ such that for all nodes $J \in \mathcal{T}_\eta$ there exist an $I \subset J$ such that $\varepsilon_I(f) \geq \eta$. The subtree $\mathcal{T}_\eta$ generates the partition $\Lambda_\eta$ consisting of its leaves. Even though the measure $\mu$ on $\Omega$ is unknown, estimates of $\varepsilon_I(f)$ are generated by averaging the samples that hit the node $I$ and its children $J \in C(I)$. See [4, 5] for the details of building such estimates of functions when IID samples are generated by a measure $\mu$ is supported on $\Omega \subset \mathbb{R}^d$.

## 6.2 Multilevel Meshes and Trees over a Submanifold $Q$

In the problem at hand, we assume that the (dependent) samples are generated by a discrete nonlinear dynamics over the unknown manifold $Q \subset \Omega \subset \mathbb{R}^d$. It is unclear that the theory that guarantees convergence of approximations when IID samples generated by a measure $\mu$ supported on $\Omega := [0, 1]^d$ can be applied in the case at hand with dependent samples concentrated on the manifold $Q$. Among other things, the proofs rely on versions of Bernstein’s inequality that hold for IID samples. In addition to the difference in the processes that generate samples, there are other technical details too related to the choice of spaces. For example, if we try to relate the approximation spaces introduced in [4, 5] to Sobolev spaces, there are nontrivial issues in passing from $W^{s,p}(\Omega)$ to $W^{s,p}(Q)$. Among other things, $Q$ is submanifold that generally has Lebesgue measure zero in $X$, and definitions of the Sobolev
spaces on $X$ lose some of their smoothness when restricted to $Q$. This is discussed in [6], for instance.

Nevertheless, formally we can proceed as in [4, 5] to construct estimates over the tree when the samples are generated by the discrete dynamical system over the manifold $Q$. This amounts to replacing the empirical estimates of the error indicator with ones that depend on the samples generated by the discrete dynamics. We leave open questions about convergence, and associated rates of convergence, of this method for future consideration of this proposed dissertation.

6.3 Koopman Theory for Model Predictive Control Systems

While most research has focused on Koopman theory on uncontrolled systems, there have been a number of studies in applying Koopman operator-theoretic approach to controlled dynamical systems [7]. These studies have shown improved performance from Koopman based MPC [8] and that model predictive control (MPC) can be used to design controllers for the nonlinear dynamical with the computational complexity of the underlying optimization problem comparable to that of MPC for a linear dynamical system with the same number of control inputs and the same dimension of the state-space [7]. However, there is a pressing need to perform the theoretical error analysis for rates of convergence, similar to our previous results work, on these types of system with control inputs.
6.4 Error Analysis for Adaptive Estimate of the Continuous Regressor

As mentioned in my study of the continuous regression problem, there were two methods of approximation investigated. The first approximation method builds estimates of the continuous integration of the optimal regression function through quadrature approximations at a particular time. The second approximation method generates an evolution of coefficients that minimizes the integration of the error over the time span of the generated estimate. While the convergence behavior was analyzed for the first approximation method, the study did not investigate the error analysis of the second method.

Bibliography


