State Following (StaF) Kernel Functions for Function Approximation Part I: Theory and Motivation

Joel A. Rosenfeld, Rushikesh Kamalapurkar, and Warren E. Dixon

***Abstract*—Unlike traditional methods that aim to approximate a function over a large compact set, a function approximation method is developed in this paper that aims to approximate a function in a small neighborhood of a state that travels within a compact set. The development is based on universal reproducing kernel Hilbert spaces over the $n-$dimensional Euclidean space. Three theorems are introduced that support the development of this state following (StaF) method. In particular an explicit uniform number of StaF kernel functions can be calculated to ensure good approximation as a state moves through a large compact set. An algorithm for gradient descent is demonstrated where a good approximation of a function can be achieved provided that the algorithm is applied with a high enough frequency.**

I. Introduction

Over the past few decades universal reproducing kernel Hilbert spaces (RKHSs) have been used extensively in applications. The theory of RKHSs allows the use of a wide range of functions for approximation. In particular, the kernel functions provide a basis set for approximating continuous functions. Moreover, since the supremum norm is dominated by the Hilbert space norm, finding the weights for a linear combination of kernel functions that approximate a function becomes easier when using the Hilbert space norm.

Kernel functions for RKHSs can take many forms, such as polynomials, $k(x,y) = (1 + x^T y)^d$, exponential functions, $k(x,y) = \exp(x^T y)$, Gaussian radial basis functions, $k(x,y) = \exp(-\|x-y\|^2/\mu)$, and many others. A RKHS that is dense in the space of continuous functions over any given compact subset, D, of the input space (with respect to the supremum norm) is said to be universal. Given a continuous function V, most approximation schemes aim to achieve a good approximation of V over D. Such methods have been thoroughly investigated, and in particular can be found in the works of Micchelli et al. and others [1]–[4]. While approximations over D are achievable, for large sets the approximation algorithms can become computationally intractable. In this paper, a scheme for local approximations is presented where the approximation of a function is only required to be accurate in a neighborhood of a state in $\mathbb{R}^n$. The state is controlled by a dynamical system, and kernel functions whose centers move with the state are used. These kernels are called *state following* (StaF) kernel functions. In Section V, the developed StaF kernel method is applied to solve a function minimization problem using gradient descent, where a good approximation of a function is achieved provided that the algorithm is applied with a high enough frequency. In Part II of this paper [5], the StaF kernel method is applied to solve an infinite horizon optimal regulation problem online using adaptive dynamic programming (ADP).

Many standard function approximation methods that have been in control systems are designed for approximation on a compact subset of the state space (e.g. neural networks (NN), wavelets, and polynomials). The intuition behind these approaches is that if a state has its initial position inside of D, then the weight estimates should converge fast enough (and to their ideal values with some form of persistence of excitation) to provide good approximation of the function. Such approximation methods ensure that the state does not leave D.

One way to improve the transient response of the function approximation is to use some knowledge about the system to determine the basis that yields accurate function approximation. For instance, for a system that is known to be quadratic in terms of the state components, NN basis functions can be used as quadratic monomials in $\mathbb{R}^n$.

For general nonlinear systems, generic basis functions are used for function approximation. Possible choices are polynomials, Gaussian Radial Basis Functions (RBF), universal kernel functions and others. In adaptive systems that employ parametric function approximation, convergence of the parameter estimates to their true values requires sufficiently rich data. For example, adaptive control techniques (cf. [6]–[9]) and online deterministic optimal control techniques (cf. [**?**], [10]–[12]) require persistence of excitation for parameter convergence, and data-driven adaptive and optimal control techniques (cf. [13], [14]) establish parameter convergence under eigenvalue conditions that quantify richness of the recorded data.

A larger number of unknown parameters require correspondingly larger amounts of excitation and richness of the recorded data. Hence, implementation of adaptive control

Joel A. Rosenfeld, Rushikesh Kamalapurkar, and Warren E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA. Email: {joelar, rkamalapurkar, wdixon}@ufl.edu.

This research is supported in part by NSF award numbers 1161260 and 1217908, ONR grant number N00014-13-1-0151, and a contract with the AFRL Mathematical Modeling and Optimization Institute. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agency.

techniques with smaller number of basis functions is desirable. Since the StaF methodology only aims to approximate a function near the current state, much fewer basis function are required than for approximation over a large compact set. Reducing the number of basis functions reduces computational limitations and improves scalability. The main theoretical result of the paper establishes that there exists a collection of centers c_i that are functions of the state x and weights w_i such that an unknown function of x can be approximated to within ϵ in a ball of radius r at any point in the space. The result (see subsequent theorems) holds for any continuous function over a compact set, but can also be extended to apply to any function that is uniformly continuous on $\mathbb{R}^n$ when an infinitely differentiable dot product kernel is used.

There are several ways that the state-following center methodology may be implemented. The most straightforward approach would place the centers at a permanent offset from the state. That is, $c_i(x) = x + d_i$ for $d_i \in \mathbb{R}^n$. This would be the benchmark test for any other choice of state-following techniques, since it does not add any more dynamics than necessary to the system. However, allowing d_i to be a bounded function of the state may improve the approximation by enabling the centers to move to a more ideal position relative to the state.

In Section II of this paper, the relevant theory of RKHSs is reviewed to provide necessary preliminary details for the development in Section III. Section III introduces and establishes the theoretical foundation of state following (StaF) approximations. In particular, Theorem 1 shows that the ideal weights (with respect to the Hilbert space norm) change smoothly with smooth change of the centers, and Theorem 2 demonstrates that there is a bound on the number of centers required for the approximation as the state moves within a compact set. Section IV further develops Theorem 3 for the exponential dot product kernel, and it demonstrates an explict bound on the number of kernel functions required for function approximation. As an immediate corollary to Theorem 3, the well known result of the universality of exponential kernel is established in a constructive manner. In Section V, a gradient descent algorithm is developed so that approximation of a function is maintained as a state travels through a compact domain.

II. Reproducing Kernel Hilbert Spaces

A RKHS H over a set X is a Hilbert space of real (or complex) valued functions over a set X such that for every $x \in X$ the functional $E_x(f) = f(x)$ (for $f \in H$) is bounded. For each $y \in X$ the Reisz representation theorem [15] guarantees a function $k_y = k(\cdot, y) \in H$ for which $\langle f, k_y \rangle = f(y)$ for all $f \in H$. The function $k(\cdot, y)$ is called the reproducing kernel corresponding to y, and the function $k : X \times X \to \mathbb{C}$ is called the kernel function corresponding to H. Much of the theory of RKHSs can be found in the classic article by Aronszajn [16].

Proposition 1. *[16] Let X be a set. If H and k are as above, then the following properties hold:*

1) $\|k_x\|_H^2 = k(x, x) = \langle k_x, k_x \rangle \geq 0$ *for all* x.
2) *The kernel function satisfies*[1] $k(x, y) = \overline{k(y, x)}$ *and for any finite collection of points* $\{c_1, ..., c_M\} \subset X$*, the matrix* $K = (k(c_i, c_j))_{i,j=1}^M$ *is self-adjoint and positive definite.*
3) *Given any* $V \in H$ *and* $\epsilon > 0$*, there are kernel functions* $k_{c_1}, ..., k_{c_M}$ *(the* $c_i \in X$ *are commonly referred to as centers) and real (or complex) numbers* $a_1, ..., a_M$ *such that*
$$\left\| \sum_{i=1}^{M} a_i k_{c_i} - V \right\|_H < \epsilon.$$

Moreover, given any function $k : X \times X \to \mathbb{C}$ *satisfying (2) there is a unique RKHS* H *for which* k *is its kernel function.*

If a kernel function is continuous on X and D is a compact subset of X, then for all $x \in D$ and $f, g \in H$,

$$|g(x) - f(x)| = |\langle g - f, k_x \rangle| \leq \|g - f\|_H \sqrt{k(x, x)}$$

for all $x \in D$ so that[2] $\|g - f\|_{\sup, D} \leq \|g - f\|_H \sup_{x \in D} \sqrt{k(x, x)}$. Thus the Hilbert space norm dominates the supremum norm over compact subsets of X. Therefore, when a good approximation is achieved with respect to the Hilbert space norm, good approximation is simultaneously achieved in the supremum norm.

Of particular importance are universal RKHSs where the Hilbert space H is dense inside of the space of continuous functions over any compact subset D of X with respect to the supremum norm. Important examples include the exponential kernel functions, $k(\langle x, y \rangle) = \exp(\langle x, y \rangle)$, and the Gaussian RBFs $k(x, y) = \exp(-\|x - y\|/\mu)$ [4]. It has been shown that for any analytic function $h(x) = \sum_{m=0}^{\infty} a_m x^m$ with $a_m > 0$ and radius of convergence $R > 0$, the RKHS over $B_{\sqrt{R}}(0) := \{x \in \mathbb{R}^n : \|x - 0\| < \sqrt{R}\} \subset \mathbb{R}^n$ corresponding to the kernel function $k(x, y) = h(\langle x, y \rangle)$ is universal and strictly positive [1], [17].

III. StaF Kernel Development

The following development demonstrates that the ideal weights corresponding to a collection of centers change smoothly with smooth changes of the centers. In applications, if the ideal weights corresponding of a collection of centers has been found by some technique, say by gradient descent or least squares, then for a small change of the centers the weights need only to be adjusted slightly in order to reach the ideal weights for the new centers. Thus, for sufficiently fast weight updates, good local approximation of a function can be maintained. The motivation for the use of StaF Kernels is that for many applications involving dynamical systems only information about the current state is needed. StaF kernels are designed so that good local approximation around the

[1] The notation $\overline{z}$ denotes the complex conjugate of a complex number z, and $\overline{A}$ denotes the closure of a set A in a metric space.

[2] The notation $\|(\cdot)\|_{\sup, D}$ denotes the supremum norm of the function $(\cdot)$ over the set D and the notation $\|(\cdot)\|_H$ denotes the Hilbert space norm of the function $(\cdot)$ in the Hilbert space H.

current state is maintained, and in this way much fewer kernel (or basis) functions are required, which can reduce the computational load required for a control system. A similar investigation for the continuity of the ideal weights was done with the squared error loss and the radial basis function in support vector machines by [18]. Here more general kernels are investigated and for different error functions.

The continuity of the ideal weights is harder to guarantee for the supremum norm, since there is not necessarily a unique minimum: the function

$$F(w,c) = \left\| \sum w_i k(\cdot, c_i) - V(\cdot) \right\|_{\sup}$$

is only convex, not strictly convex with respect to the weights.

An alternative choice of norm would be that of a Hilbert space, where for each $V \in H$ and closed subspace S, there is a unique function $\hat{V} \in S$ that minimizes the distance from V to the subspace. In this case, $\hat{V}$ is simply the projection of V onto S. In this setting, it is straightforward to prove the continuity of the ideal weights with respect to the centers.

Theorem 1. *Let H be a RKHS over a set $X \subset \mathbb{R}^n$ with a strictly positive kernel $k : X \times X \to \mathbb{C}$ such that $k(\cdot, c) \in C^m(\mathbb{R}^n)$ for all $c \in X$. Suppose $V \in H$. Let C be an ordered collection of M distinct centers, $C = (c_1, c_2, ..., c_M) \in D^M$, with associated ideal weights*

$$W(C) = \arg\min_{a \in R^M} \left\| \sum_{i=1}^{M} a_i k(\cdot, c_i) - V(\cdot) \right\|_H .$$

The function $W(C)$ is m-times continuously differentiable with respect to each component of C.

Proof: Let $C = (c_1, ..., c_M) \in D^M$ and let $\epsilon > 0$. Let $S_C = \text{span}\{k(\cdot, c_i)\}_{i=1}^M$. The ideal weights can be computed by finding the projection of V onto the closed subspace S_C. We will first show that the projection itself is C^m with respect to the change of centers. To compute the projection, an orthonormal basis for S is required.

Strict positivity of k guarantees linear independence of the collection of $k(\cdot, c_i)$'s when $c_i \neq c_j$ for $i \neq j$. From the linearly independent set $\{k_{c_1}, k_{c_2}, ..., k_{c_M}\}$, the Graham-Schmidt process can be used to find an orthonormal basis for S_C.

Let $u_1(\cdot, C) = k_{c_1}/\|k_{c_1}\|_H = k(\cdot, c_1)/\sqrt{k(c_1, c_1)}$. Note that $k(c_1, c_1)$ is nonzero when k is a strictly positive definite kernel. Since $k(\cdot, c_i)$ is C^m in c_i, u_1 is C^m with respect to c_1. Now consider $u_2^*(\cdot, C) = k(\cdot, c_2) - \langle k(\cdot, c_2), u_1\rangle u_1$. The function u_2^* is C^m with respect to C. Moreover, the norm

$$\begin{aligned} \|u_2^*\|_H^2 &= \langle u_2^*, u_2^* \rangle \\ &= \|k_{c_2}\|_H^2 + |\langle k_{c_2}, u_1\rangle|^2 - 2|\langle k_{c_2}, u_1\rangle|^2 \\ &= \|k_{c_2}\|_H^2 - |\langle k_{c_2}, u_1\rangle|^2 \end{aligned}$$

is always positive, since k_{c_2} is not in the span of u_1 and by an application of Bessel's inequality [19]. Therefore, the function $u_2(\cdot, C) = u_2^*(\cdot, C)/\|u_2^*\|_H$ is C^m with respect to the centers as well.

Now suppose that $u_1, u_2, ..., u_l$ is an orthonormal sequence for $2 \leq l < M$, and that each is C^m with respect to the centers. Consider

$$u_{l+1}^*(\cdot, C) = k(\cdot, c_{l+1}) - \sum_{i=1}^{l} \langle k(\cdot, c_{l+1}), u_i\rangle u_i.$$

The function u_{l+1}^* is C^m with respect to the centers by the induction assumption. Moreover,

$$\|u_{l+1}^*\|_H = \|k_{c_{l+1}}\|_H^2 - \sum_{i=1}^{l} |\langle k_{c_{l+1}}, u_i\rangle|^2$$

is positive since $k_{c_{l+1}}$ is not in the span of $u_1, ..., u_l$ and again by Bessel's inequality. Therefore, $u_{l+1} = u_{l+1}^*/\|u_{l+1}^*\|_H$ is C^m with respect to the centers.

Since each u_i and V is C^m with respect to the centers, $\langle V, u_i\rangle = \sum p_i(C)V(c_i)$ is also C^m with respect to the centers (here $p_i(C)$ represent the rational functions resulting from the Gram-Schmidt process). Since $\langle V, u_i\rangle$ is C^m, the projection

$$\text{Proj}_{S_C} V \doteq \sum_{i=1}^{M} \langle V, u_i\rangle u_i(\cdot, C)$$

is also C^m with respect to the centers. $\text{Proj}_{S_c} V$ can also be expressed as $\text{Proj}_{S_c} V = \sum_{i=1}^{M} w_i k(\cdot, c_i)$ where each w_j is a linear combination of $\langle V, u_i\rangle$ for $i = 1, 2, ..., M$. Therefore w_j is C^m as is $W(C)$. ∎

The principle utility of using StaF kernel functions is the reduction of the number of basis functions required to maintain a good approximation of a function. The following proposition justifies this statement.

Theorem 2. *Let D be a compact subset of $\mathbb{R}^n$. Consider a continuous function $V : \mathbb{R}^n \to \mathbb{R}$ and a continuous universal kernel function $k(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Let $\epsilon > 0$ and $r > 0$, then there exists a number M such that for each $x \in D$ there exists a collection of centers $\{c_1, ..., c_M\} \subset B_r(x)$, and weights $\{w_1, ..., w_M\} \subset \mathbb{R}$ such that $|\sum_{i=1}^{M} w_i k(y, c_i) - V(y)| < \epsilon$ for all $y \in \overline{B_r(x)}$.*

Proof: For each closed neighborhood $\overline{B_r(x)}$ with $x \in D$, there exists a finite number of centers, $c_1, ..., c_M$, and weights, $w_1, ..., w_M$, such that

$$\left| \sum_{n=1}^{M} w_i k(y, c_i) - V(y) \right| < \epsilon$$

for all $y \in \overline{B_r(x)}$. Let $M_{x,\epsilon}$ be the minimum such number. The claim of the proposition is that the set $Q_\epsilon = \{M_{x,\epsilon} : x \in D\}$ is bounded. Assume that Q_ϵ is not bounded, and take a sequence $\{x_n\} \subset D$ such that $M_{x_n,\epsilon}$ is a strictly increasing sequence and $x_n \to x$ for some $x \in D$. It is always possible to find such a convergence sequence, since every sequence in a compact set has a convergent subsequence.

Let $c_1, ..., c_{M_{x,\epsilon/2}}$ and $w_1, ..., w_{M_{x,\epsilon/2}}$ be the centers and weights for which,

$$\left| \sum_{i=1}^{M_{x,\epsilon/2}} w_i k(y, c_i) - V(y) \right| < \epsilon/2$$

for all $y \in \overline{B_r(x)}$. For convenience, each $y \in \overline{B_r(x)}$ can be expressed as $x + z$ for $z \in \overline{B_r(x)}$. Let $\eta > 0$ so that

$$\left| \sum_{i=1}^{M_{x,\epsilon/2}} w_i k(x + z, c_i) - \sum_{i=1}^{M_{x,\epsilon/2}} w_i k(\tilde{x} + z, c_i) \right| < \epsilon/4$$

and

$$|V(x + z) - V(\tilde{x} + z)| < \epsilon/4$$

for all $\|x - \tilde{x}\| < \eta$ and all $z \in \overline{B_r(x)}$. This is possible since all continuous functions are uniformly continuous on compact sets. Now suppose $N \in \mathbb{N}$ is such that $\|x - x_n\| < \eta$ for all $n > N$. Then by two applications of the triangle inequality:

$$\left| \sum_{i=1}^{M_{x,\epsilon/2}} w_i k(x_n + z, c_i) - V(x_n + z) \right| < \epsilon$$

for all $n > N$ and $z \in \overline{B_r(x)}$. Hence, $M_{x_n,\epsilon} \leq M_{x,\epsilon/2}$ for all $n > N$, which is a contradiction. ∎

IV. Upper Bound for Number of Kernel Functions Required in StaF Implementation

This section demonstrates that an explicty bound, as in Theorem 2, can be calculated for the exponential kernel function. The universality of the dot-product kernel function is proved via the Weierstass theorem and uses polynomials [4]. The following theorem also uses polynomial functions to determine the number of kernel functions necessary for approximation with the exponential kernel function.

The degree of such a polynomial can be calculated explicitly via Bernstein's proof of the Weirstrass approximation theorem, and therefore the number of centers can also be calculated. This yields a concrete bound on the number of centers required to achieve an accurate approximation of a continuous function.

Theorem 3. *Let $K(x, y) = e^{x^T y}$ be the exponential kernel function, which corresponds to an universal RKHS. Let $D \subset \mathbb{R}^n$ be compact, $V : D \to \mathbb{R}$ be a continuous function, and $\epsilon, r > 0$. For each $y \in D$, there exists a finite number of centers, $c_1, c_2, ..., c_{M_{y,\epsilon}} \in B_r(y)$ and weights $w_1, w_2, ..., w_{M_{y,\epsilon}}$ such that*

$$\left\| V(x) - \sum_{i=1}^{M_{y,\epsilon}} w_i e^{x^T c_i} \right\|_{B_r(y),\infty} < \epsilon.$$

If p is an approximating polynomial that achieves the same approximation over $B_r(y)$ with degree $N_{y,\epsilon}$, then an asymptotically similar bound can be found with $M_{y,\epsilon}$ kernel functions, where $M_{y,\epsilon} < \binom{n+N_{y,\epsilon}+S_{y,\epsilon}}{N_{y,\epsilon}+S_{y,\epsilon}}$ for some constant $S_{y,\epsilon}$. Moreover, $N_{y,\epsilon}$ and $S_{y,\epsilon}$ can be bounded uniformly over D.

Proof: First, consider the ball of radius r centered at the origin. The statement of the theorem can be proven by finding an approximation of monomials by a linear combination of exponential kernel functions. Let $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ be a multi-index, and define $|\alpha| = \sum \alpha_i$. Note that[3]

$$m^{|\alpha|} \prod_{i=1}^{n} \left(e^{x_i/m} - 1 \right)^{\alpha_i} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} + O\left(\frac{1}{m}\right)$$

which leads to the sum

$$m^{|\alpha|} \sum_{l_i \leq \alpha_i, i=1,2,...,n} \prod_{j=1}^{n} \binom{\alpha_j}{l_j} (-1)^{|\alpha| - \sum_i l_i} e^{\sum_{i=1}^{n} x_i \left(\frac{l_i}{m}\right)} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} + O\left(\frac{1}{m}\right). \quad (1)$$

The big-oh constant indicated by $O(1/m)$ can be computed in terms of the derivatives of the exponential function via Taylor's Theorem. The centers corresponding to this approximation are of the form l_i/m where l_i is a nonnegative integer satisfying $l_i < \alpha_i$. Hence, for m sufficiently large, the centers reside in $B_r(0)$.

In order to shift the centers so that they reside in $B_r(y)$, let $y = (y_1, y_2, ..., y_n)^T \in \mathbb{R}^n$, and multiply both sides of (1) by $e^{x^T y}$ to get

$$m^{|\alpha|} \sum_{l_i \leq \alpha_i, i=1,...,n} \prod_{j=1}^{n} \binom{\alpha_j}{l_j} (-1)^{|\alpha| - \sum_i l_i} e^{\sum_{i=1}^{n} x_i \left(\frac{l_i}{m} + y_i\right)} = e^{x^T y} \left(x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}\right) + O\left(\frac{1}{m}\right).$$

For each multi-index, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$, the centers for the approximation of the corresponding monomial are of the form $y_i + l_i/m$ for $0 \leq l_i \leq \alpha_i$. Thus by linear combinations of these kernel functions, a function of the form $e^{x^T y} g(x)$ with g a multivariable polynomial, can be uniformly approximated by exponential functions over $B_r(y)$. Moreover if g is a polynomial of degree β, then this approximation can be a linear combination of $\binom{n+\beta}{\beta}$ kernel functions.

Now suppose that p_y is polynomial with degree N_y such that $p_y(x) = V(x) + \epsilon_1(x)$ where $|\epsilon_1(x)| < \|e^{x^T y}\|^{-1}_{D,\infty} \epsilon/2$ for all $x \in B_r(y)$. Let $q_y(x)$ be a polynomial in $\mathbb{R}^n$ variables of degree $S_{y,\epsilon}$ such that $q_y(x) = e^{-x^T y} + \epsilon_2(x)$ where $\epsilon_2(x) < \|V\|^{-1}_{D,\infty} \|e^{x^T y}\|^{-1}_{D,\infty} \epsilon/2$ for all $x \in B_r(x)$.

By our above construction there is a sequence of linear combinations of kernel functions, $F_m(x)$, (with a fixed number of centers inside $B_r(y)$) for which

$$F_m(x) = e^{x^T y} q_y(x) p_y(x) + O\left(\frac{1}{m}\right).$$

[3]The notation $g_m(x) = O(f(m))$ means that for sufficiently large m, there is a constant C for which $g_m(x) < C f(m)$ for all $x \in \overline{B_r(0)}$.

After multiplication and an application of the triangle inequality, the following is established:

$$|F_m(x) - V(x)| < \epsilon + \left(\frac{\|V\|_{D,\infty}^{-1} \|e^{x^T y}\|_{D,\infty}^{-1}}{4} \right) \epsilon^2 + O\left(\frac{1}{m}\right)$$

for all $x \in B_r(y)$. The degree of the polynomial q_y, $S_{x,\epsilon}$, can be uniformly bounded in terms of the modulus of continuity of $e^{x^T y}$ over D. Similarly, the uniform bound on the polynomial degree of p_y, $N_{y,\epsilon}$, can be described in terms of the modulus of continuity of V over D. Note that the number of centers needed for $F_m(x)$ is determined by the degree of the polynomial $q \cdot p$ (treating the y terms of q as constant), which is sum of the two polynomial degrees. The completes the proof. ∎

Given that V can be well approximated in a neighborhood by a polynomial of degree N, it can also be well approximated by $\binom{n+N+S}{N+S}$ kernel functions where S is the uniform bound of $S_{y,\epsilon}$ described in the Theorem 3. It follows from Bernstein's constructive proof of the Weierstrass theorem [20], that as the approximation neighborhood of V shrinks, so does the degree of the polynomial needed to achieve a good approximation. Furthermore, this degree bound can be calculated.

V. A Numerical Example with Gradient Descent

In this section, an application of StaF kernel functions is presented. Theorem 4 indicates that with sufficiently frequent applications of the gradient descent algorithm a good approximation of a function can be achieved as long as the centers are C^1 with respect to the state variable x. The development of Theorem 4 follows the standard proof of the convergence of the gradient descent algorithm for quadratic functions as can be found in [21], [22].

Theorem 4. *Let $V \in H$, a real valued RKHS over $\mathbb{R}^n$ with kernel $K(x,y) = k_y(x)$, D a compact set subset of $\mathbb{R}^n$, and x a state variable controlled by the dynamical system $\dot{x} = q(x,t)$, where $q : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ is a bounded continuous function. Further suppose that $x \in D$ for all time. Let $\mathbf{c} \in D^M$ where for each $i = 1, ..., M$ we set $c_i(x) = x + d_i(x)$ where $d_i \in C^1(\mathbb{R}^n)$, so c_i depends implicitly on time, and let $\mathbf{a} \in \mathbb{R}^M$. Consider the function*

$$F(\mathbf{a}) = F(\mathbf{a}, \mathbf{c}) = \left\| V - \sum_{i=1}^{M} a_i k(\cdot, c_i(x)) \right\|_H^2.$$

At each time instance t, there is unique $\mathbf{w}(t)$ for which

$$\mathbf{w}(t) = \arg\min_{\mathbf{a} \in \mathbb{R}^M} F(\mathbf{a}, \mathbf{c}(x(t))).$$

Given any $\epsilon > 0$ and initial value $\mathbf{a}^0$ there is a frequency $\tau > 0$, where if the gradient descent algorithm (with respect to $\mathbf{a}$) is iterated at time steps $\Delta t < \tau^{-1}$, then

$$F(\mathbf{a}^k, \mathbf{c}^k) - F(\mathbf{w}^k, \mathbf{c}^k)$$

will approach a neighborhood of radius ϵ as $k \to \infty$.

Proof: Suppose $\dot{x} = q(x,t)$ where $q(x,t)$ is bounded in absolute value by some constant R_0. Let $\bar{\epsilon} > 0$. Note by the Hilbert space structure of H we have

$$\begin{aligned} F(\mathbf{a}, \mathbf{c}) &= \left\| V - \sum_{i=1}^{M} a_i k(\cdot, c_i) \right\|_H^2 \\ &= \|V\|_H^2 - 2\sum_{i=1}^{M} a_i V(c_i) + \sum_{i,j=1}^{M} a_i a_j k(c_i, c_j) \\ &= \|V\|_H^2 - 2V(\mathbf{c})^T \mathbf{a} + \mathbf{a}^T K(\mathbf{c}) \mathbf{a}, \end{aligned}$$

where $V(\mathbf{c}) = (V(c_1), V(c_2), ..., V(c_M))^T$ and $K(\mathbf{c}) = (K(c_i, c_j))_{i,j=1}^M$ is the symmetric strictly positive definite kernel matrix corresponding to the centers.

Let $\mathbf{a}^0$ be the initial condition for the weights. For each time iteration t^k we will write the updated centers and weights as $\mathbf{c}^k$ and $\mathbf{a}^k$ respectively. The ideal weights corresponding to $\mathbf{c}^k$ are denoted by $\mathbf{w}^k$. It can be shown that $\mathbf{w}^k = K(\mathbf{c}^k)^{-1} V(\mathbf{c}^k)$ and $F(\mathbf{w}^k, \mathbf{c}^k) = \|V\|_H^2 - V(\mathbf{c}^k)^T K(\mathbf{c}^k)^{-1} V(\mathbf{c}^k)$.

As was proven in Theorem 1, the ideal weights change continuously with respect to the centers which remain in the compact set D^M, so the collection of all ideal weights is bounded. Let $R > \bar{\epsilon}$ be large enough so that $B_R(0)$ contains both the initial value $\mathbf{a}^0$ and the set of ideal weights. To facilitate the subsequent analysis, consider the constants:

$$\begin{aligned} R_0 &= \max_{x \in D, t \in \mathbb{R}^+} |q(x,t)| \\ R_1 &= \max_{\mathbf{a} \in \overline{B_R(0)}, \tilde{\mathbf{c}} \in \mathbf{c}(D)} |\nabla_{\mathbf{a}} F(\mathbf{a}, \tilde{\mathbf{c}})| \\ R_2 &= \max_{\tilde{\mathbf{c}} \in \mathbf{c}(D)} |\nabla_{\mathbf{c}} F(\mathbf{w}(\tilde{\mathbf{c}}), \tilde{\mathbf{c}})|, \end{aligned}$$

and let $\Delta t < \bar{\epsilon}(2R_0)^{-1} \min\{R_1^{-1}, R_2^{-1}\} \doteq \tau^{-1}$.

The analysis aims to show that by using the gradient descent law for choosing $\mathbf{a}^k$ the following inequality can be achieved:

$$\frac{F(\mathbf{a}^{k+1}, \mathbf{c}^{k+1}) - F(\mathbf{w}^{k+1}, \mathbf{c}^{k+1})}{F(\mathbf{a}^k, \mathbf{c}^k) - F(\mathbf{w}^k, \mathbf{c}^k)} < \delta + \frac{\bar{\epsilon}}{F(\mathbf{a}^k, \mathbf{c}^k) - F(\mathbf{w}^k, \mathbf{c}^k)}$$

for some $0 < \delta < 1$. Here we set

$$\mathbf{a}^{k+1} = \mathbf{a}^k + \lambda \mathbf{g} \tag{2}$$

where $\mathbf{g} = -\nabla_{\mathbf{a}} F = 2V(\mathbf{c}^k) - 2K(\mathbf{c}^k)\mathbf{a}^k$ and λ is selected so that the quantity $F(\mathbf{a}^k + \lambda \mathbf{g}, \mathbf{c}^k)$ is minimized. Consider the λ that minimizes this quantity is

$$\lambda = \left(\frac{\mathbf{g}^T \mathbf{g}}{2\mathbf{g}^T K(\mathbf{c}^k) \mathbf{g}} \right)$$

which yields

$$F(\mathbf{a}^{k+1}, \mathbf{c}^k) = F(\mathbf{a}^k, \mathbf{c}^k) - \frac{(\mathbf{g}^T \mathbf{g})^2}{4\mathbf{g}^T K(\mathbf{c}^k) \mathbf{g}}.$$

Since $F(\mathbf{a}^{k+1}, \mathbf{c}^{k+1})$ is continuously differentiable in the second variable, we have

$$F(\mathbf{a}^{k+1}, \mathbf{c}^{k+1}) = F(\mathbf{a}^{k+1}, \mathbf{c}^{k}) + \nabla_{\mathbf{c}} F(\mathbf{a}^{k+1}, \xi) \cdot (\mathbf{c}^{k+1} - \mathbf{c}^{k}).$$

By another application of the mean value theorem we find

$$F(\mathbf{a}^{k+1}, \mathbf{c}^{k+1}) = F(\mathbf{a}^{k+1}, \mathbf{c}^{k}) + \epsilon_1(t^k),$$

where $|\epsilon_1(t^k)| \leq \bar{\epsilon}/2$, since the time-step is less than τ^{-1}.

The quantity $F(\mathbf{w}^{k+1}, \mathbf{c}^{k+1})$ is also continuously differentiable with respect to $\mathbf{c}$. Thus by applying the mean value theorem again we find: $F(\mathbf{w}^{k+1}, \mathbf{c}^{k+1}) = F(\mathbf{w}^{k}, \mathbf{c}^{k}) + \epsilon_2(t^k)$, for $|\epsilon_2(t^k)| < \bar{\epsilon}/2$. Thus, we have the following:

$$\begin{aligned}
&\frac{F(\mathbf{a}^{k+1}, \mathbf{c}^{k+1}) - F(\mathbf{w}^{k+1}, \mathbf{c}^{k+1})}{F(\mathbf{a}^{k}, \mathbf{c}^{k}) - F(\mathbf{w}^{k}, \mathbf{c}^{k})} \\
&\quad = \frac{F(\mathbf{a}^{k+1}, \mathbf{c}^{k}) - F(\mathbf{w}^{k}, \mathbf{c}^{k}) + (\epsilon_1(t^k) - \epsilon_2(t^k))}{F(\mathbf{a}^{k}, \mathbf{c}^{k}) - F(\mathbf{w}^{k}, \mathbf{c}^{k})} \\
&= 1 - \frac{(\mathbf{g}^T \mathbf{g})^2}{(\mathbf{g}^T K(\mathbf{c}^k)\mathbf{g})(\mathbf{g}^T K(\mathbf{c}^k)^{-1}\mathbf{g})} + \frac{\epsilon_1(t^k) - \epsilon_2(t^k)}{F(\mathbf{a}^{k}, \mathbf{c}^{k}) - F(\mathbf{w}^{k}, \mathbf{c}^{k})}
\end{aligned}$$

For each $\mathbf{c} \in D$ there is an associated bound for the first two terms. In this case, by the Kantorovich inequality [22]

$$1 - \frac{(\mathbf{g}^T \mathbf{g})^2}{(\mathbf{g}^T K(\mathbf{c}^k)\mathbf{g})(\mathbf{g}^T K(\mathbf{c}^k)^{-1}\mathbf{g})} \leq \left(\frac{A_{\mathbf{c}}/a_{\mathbf{c}} - 1}{A_{\mathbf{c}}/a_{\mathbf{c}} + 1} \right)^2$$

where $A_{\mathbf{c}}$ is the largest eigenvalue for $K(\mathbf{c})$ and $a_{\mathbf{c}}$ is the smallest. The quantity on the right is continuous with respect to the largest and smallest eigenvalues, and the largest and smallest eigenvalues are continuous with respect to the matrix $K(\mathbf{c})$ (see Exercise 4.1.6 in [15]) which is continuous with respect to $\mathbf{c}$. Therefore, there is a largest value that this obtains on the compact set $\mathbf{c}(D)$ and this value is less than 1. Moreover, δ is independent of $\bar{\epsilon}$, so $\bar{\epsilon} = \epsilon(1 - \delta)$. Finally we have

$$\begin{aligned}
&\frac{F(\mathbf{a}^{k+1}, \mathbf{c}^{k+1}) - F(\mathbf{w}^{k+1}, \mathbf{c}^{k+1})}{F(\mathbf{a}^{k}, \mathbf{c}^{k}) - F(\mathbf{w}^{k}, \mathbf{c}^{k})} \leq \\
&\qquad \delta + \frac{(\epsilon_1(t^k) - \epsilon_2(t^k))}{F(\mathbf{a}^{k}, \mathbf{c}^{k}) - F(\mathbf{w}^{k}, \mathbf{c}^{k})}.
\end{aligned}$$

Therefore, if $e(k) = F(\mathbf{a}^k, \mathbf{c}^k) - F(\mathbf{w}^k, \mathbf{c}^k)$, then we find that

$$e(k+1) \leq \delta e(k) + \epsilon(1 - \delta).$$

The conclusion follows. ∎

VI. Conclusion

In this paper a new approximation methodology is introduced, the so called StaF kernel method. With this method it is shown that local approximation of a function can be maintained as a state moves through a compact domain, and that much fewer kernel functions are needed than are needed in more traditional function approximation schemes. For exponential dot product kernels, an explicit bound on the number of kernel functions required is calculated. In Section V, a gradient descent algorithm is developed. There it is seen that a function may be well approximated provided that the algorithm is applied with a high enough frequency.

References

[1] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *J. Mach. Learn. Res.*, vol. 7, pp. 2651–2667, 2006.

[2] J. Park and I. Sanberg, "Universal approximation using radial-basis-function networks," *Neural Comput.*, vol. 3, pp. 246–257, 1991.

[3] A. Christmann and I. Steinwart, "Univeral kernels on non-standard input spaces," in *Advances in Nueral Information Processing*, 2010, pp. 406–414.

[4] I. Steinwart and A. Christmann, *Support vector machines*, ser. Information Science and Statistics. New York: Springer, 2008.

[5] R. Kamalapurkar, J. A. Rosenfeld, and W. E. Dixon, "State following (StaF) kernel functions for function approximation part ii: Adaptive dynamic programming," in *Proc. Am. Control Conf.*, 2015, to appear (see also arXiv:1502.02609).

[6] K. Narendra and A. Annaswamy, *Stable Adaptive Systems*. Prentice-Hall, Inc., 1989.

[7] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*. Upper Saddle River, NJ: Prentice-Hall, 1989.

[8] M. Krstic, I. Kanellakopoulos, and P. V. Kokotovic, *Nonlinear and Adaptive Control Design*. New York, NY, USA: John Wiley & Sons, 1995.

[9] P. Ioannou and J. Sun, *Robust Adaptive Control*. Prentice Hall, 1996.

[10] K. Vamvoudakis and F. Lewis, "Online synchronous policy iteration method for optimal control," in *Recent Advances in Intelligent Control Systems*, W. Yu, Ed. Springer, 2009, pp. 357–374.

[11] ——, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[12] S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. L. L. keywords = RISE,Robot,Network,Optimal,NN,theory,learning, and W. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 89–92, 2013.

[13] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Concurrent learning-based approximate optimal regulation," in *Proc. IEEE Conf. Decis. Control*, Florence, IT, Dec. 2013, pp. 6256–6261.

[14] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[15] G. K. Pedersen, *Analysis now*, ser. Graduate Texts in Mathematics. Springer-Verlag, New York, 1989, vol. 118.

[16] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[17] A. Pinkus, "Strictly positive definite functions on a real inner product space," *Adv. in Comput. Math.*, vol. 20, pp. 263–271, 2004.

[18] C. Panchapakesan, D. Ralph, and M. Palaniswami, "Effects of moving the centers in an rbf network," in *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, vol. 2, May 1998, pp. 1256–1260 vol.2.

[19] J. P. Folland and A. G. Williams, "Methodological issues with the interpolated twitch technique," *J. Electromyogr. Kinesiol.*, vol. 17, no. 3, pp. 317–327, Jun 2007.

[20] G. G. Lorentz, *Bernstein polynomials*, 2nd ed. Chelsea Publishing Co., New York, 1986.

[21] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.

[22] M. A. Epelman. (2007) Continuous optimization methods, section 1. [Online]. Available: http://www-personal.umich.edu/~mepelman/teaching/IOE511/511notes.pdf.