Convergence Rates of Online Critic Value Function Approximation in Native Spaces

Shengyuan Niu[1], Ali Bouland[1], Haoran Wang[1], Filippos Fotiadis[2], Andrew Kurdila[1], Andrea L'Afflitto[3], Sai Tej Paruchuri[4], Kyriakos G. Vamvoudakis[2]

***Abstract*— This paper derives rates of convergence of online critic methods for the estimation of the value function for a class of nonlinear optimal control problems. Assuming that the underlying value function lies in reproducing kernel Hilbert space (RKHS), we derive explicit bounds on the performance of the critic in terms of the kernel functions, the number of basis functions, and the scattered location of centers used to define the RKHS. The performance of the critic is precisely measured in terms of the power function of the scattered bases, and it can be used either in an *a priori* evaluation of potential bases or in an *a posteriori* assessments of the value function error for basis enrichment or pruning. The most concise bounds in the paper describe explicitly how the critic performance depends on the placement of centers, as measured by their fill distance in a subset that contains the trajectory of the critic. To the authors' knowledge, precise error bounds of this form are the first of their kind for online critic formulations used in optimal control problems. In addition to their general and immediate applicability to a wide range of applications, they have the potential to constitute the groundwork for more advanced "basis-adaptive" methods for nonlinear optimal control strategies, ones that address limitations due to the dimensionality of approximations.**

***Index Terms*— Reinforcement learning, Optimal Control, Reproducing Kernel, Native Space**

I. Introduction

Optimal control has become one of the core methodologies in modern control theory for nonlinear systems. One of its main advantages lies in its ability to yield control laws that achieve a compromise between the control effort expended and the time needed to attain regulation. At the heart of nonlinear optimal control design is the Hamilton-Jacobi-Bellman (HJB) equation [1], a partial differential equation (PDE) that is notoriously difficult to solve analytically. Numerous studies describe methods to approximate the solution of the HJB equation, see the reviews in [2], [3] for example. One of the most popular such tools is policy iteration (PI), which is a process that cyclically evaluates the cost function for a given controller, and, subsequently, improves that controller from measurements. Nevertheless, an issue with PI is its need to employ a neural network (NN) for the policy evaluation step, called the "critic," which inherently leads to approximation errors that degrade performance or lead to failure of convergence of the PI process.

Notable early efforts that study Galerkin approximations for PI in a recursive implementation include [4]–[6]. Subsequent papers [7], [8] use some of the theory in [4]–[6] to study various online implementations based on learning theory. The works referenced in [9] and [10], for example, provide comprehensive reviews of contemporary theory underlying many recent online and offline methods. Yet, these results do not derive explicit descriptions of how performance is related *quantitatively* to rates of convergence of value function approximations generated by a critic. On the other hand, some very recent efforts in [11]–[13] emphasize the importance of examining the impact of the approximation error on the performance of reinforcement learning methods.

This work continues the strategy started for offline reproducing kernel Hilbert spaces (RKHS) methods in [14], but now considers online approaches for the critic step in PIs. We describe how the fill distance of the centers used to define the bases for approximation dictates the performance of the critic. In several case, we relate the rate of convergence in the RKHS directly and explicitly to the performance of the critic. To the authors' knowledge, this is the first time that such rates of convergence have been derived for the online critic step. These general results and error rates have a host of potential applications to reduce guesswork by the control designer when using PI techniques. The derived rates also have the potential to serve as the foundation of methods that dynamically add or delete scattered or sparse basis functions to address cases when dimensionality of approximations becomes an issue.

II. Problem Statement

Consider the continuous-time nonlinear system

$$\dot{x}(t) = f(x(t)) + g(x(t))u(x(t)),\ x(0) = x_0,\ t \geq 0, \quad (1)$$

[1]S. Niu, A. Bouland, H. Wang, and A. Kurdila are with the Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA, USA. Email: {syniu97, bouland, haoran9, kurdila}@vt.edu.

[2]F. Fotiadis and K. G. Vamvoudakis are with The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Email: {ffotiadis, kyriakos}@gatech.edu.

[3]A. L'Afflitto is with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA. Email: a.lafflitto@vt.edu.

[4]S. T. Paruchuri is with the Department of Mechanical Engineering and Mechanics, Lehigh University, Bethlehem, PA, USA. Email: saitejp@lehigh.edu.

This work was supported in part, by NSF under grants No. CAREER CPS-1851588, CPS-2227185, S&AS-1849198, and 2137159, and the US Army Research Lab under Grant No. $W911QX2320001$.

where $x : [0,\infty) \to \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$, and $u : \mathbb{R}^n \to \mathbb{R}^m$ represent the state of the system, the drift dynamics, the input dynamics, and the control input, respectively. The problem is to find a continuous control input $t \mapsto u(t)$ that minimizes the cost functional

$$J(x_0,u) = \int_0^\infty \underbrace{\big(Q(x(t)) + u^{\mathrm{T}}(t)Ru(t)\big)}_{r(x(t),u(t))}\,\mathrm{d}t \tag{2}$$

where $Q : x \mapsto Q(x) \geq 0$, and $R \succ 0$. One of the main issues with this problem is that one needs to solve a challenging nonlinear HJB equation. A minimizer $u^\star$ of (2) is called an optimal control input, and $V^\star(\cdot) = J(\cdot,u^\star)$ defines the optimal value function. Then, to find $u^\star$ and $V^\star$, in principle, one needs to find the positive-definite solution $V^\star$ of the HJB equation

$$\begin{aligned}\mathcal{H}_{u^*}(V^*(x)) = -\frac{1}{4}\nabla V^{\star\mathrm{T}}(x)g(x)R^{-1}g^{\mathrm{T}}(x)\nabla V^\star(x)\\ + \nabla V^{\star\mathrm{T}}(x)f(x) + Q(x) = 0,\ V^\star(0) = 0,\ \forall x\in\Omega,\end{aligned} \tag{3}$$

and then calculate $u^\star(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V^\star(x)$ [1], where $\mathcal{H}_u(V)$ is the Hamiltonian function associated with u and V. Nevertheless, (3) is generally difficult, if not impossible, to solve analytically for $V^\star$. For this reason, PI is often employed to solve (3) approximately [4], [6].

The most crucial and computationally demanding step of PI is that of policy evaluation. Given a continuous feedback gain $\mu : \mathbb{R}^n \to \mathbb{R}^m$ that stabilizes (1) on a set $\Omega \subseteq \mathbb{R}^n$, policy evaluation seeks to find the value function $V_\mu(\cdot) \triangleq J(\cdot,\mu)$ associated with that controller. Provided that this function is continuously differentiable, it follows from [1] that it satisfies

$$\begin{aligned}\mathcal{H}_\mu(x) \triangleq \mathcal{H}_\mu(V_\mu(x)) = \nabla V_\mu^{\mathrm{T}}(x)(f(x)+g(x)\mu(x))\\ + Q(x) + \mu^{\mathrm{T}}(x)R\mu(x) = 0,\ V_\mu(0) = 0.\end{aligned} \tag{4}$$

While an analytical solution to (4) is also difficult to obtain, its linearity in V_μ — a property not present in (3) — enables the use of the so-called *critic* NN as a means to approximately solve it over a compact set $\Omega \subset \mathbb{R}^n$.

To that end, note that since V_μ is continuous, it can be expressed on Ω as $V_\mu(x) = W^{\mathrm{T}}\phi(x) + \epsilon_N(x),\ \forall x \in \Omega$, where $\phi : \mathbb{R}^n \to \mathbb{R}^N$ is a suitable vector of N basis functions, $W \in \mathbb{R}^N$ denote the "ideal weights" for that basis, and $\epsilon_N : \mathbb{R}^n \to \mathbb{R}$ denotes the approximation error. The critic NN then uses an estimate $\hat{W}(t) \in \mathbb{R}^N$ of W, and provides an estimate of $\hat{v}_N(t,\cdot)$ of V_μ according to the formula $\hat{v}_N(t,x) = \hat{W}^{\mathrm{T}}(t)\phi(x)$. The purpose of policy evaluation is, thus, to properly train the critic weights $\hat{W}(t)$ so that the norm of the parameter error $\tilde{W}(t) \triangleq W - \hat{W}(t)$ becomes as small as possible. In [7], the online policy evaluation law

$$\dot{\hat{W}}(t) = -\frac{a\sigma(t)}{(\sigma^{\mathrm{T}}(t)\sigma(t)+1)^2}\Big(\sigma^{\mathrm{T}}(t)\hat{W}(t) + r(x(t),\mu(x(t)))\Big) \tag{5}$$

was proposed, where $\sigma(t) \triangleq \sigma(x(t)) = \nabla\phi(x(t))(f(x(t)) + g(x(t))\mu(x(t))$, and $a > 0$ denotes the learning rate. Interestingly, it was proved that, under a persistency of excitation condition, the parameter estimation error $\tilde{W}(t)$ under (5) indeed converges exponentially fast to a neighborhood of the origin, the size of which scales with the size of the approximation error ϵ_N over Ω. Nevertheless, the size of ϵ_N is rarely known beforehand and, to our knowledge, no existing general strategy yet has been able to precisely quantify how the basis influences the performance of the critic.

This paper lifts the analysis of the norm of the parameter error $\|\hat{W}(t) - W\|_{\mathbb{R}^N}$ to an analysis of $\|v_N(t,\cdot) - V_\mu\|_{H(\Omega)}$, which captures estimates of the error of the value function. This analysis makes explicit the contribution of approximation errors in a wide variety of choices of the RKHS $H(\Omega)$ assumed to contain the value function. Our goal is to ultimately use this analysis to quantitatively relate the choice of the basis function ϕ of the critic NN to the error $\|\hat{v}_N(t,\cdot) - V_\mu\|_{H(\Omega)}$ in online critic estimates $v_N(t,\cdot)$ of the value function V_μ. A further goal of the paper is to reduce trial-and-error in realistic implications of the critic for adaptive nonlinear optimal control.

III. Notation and Preliminaries

Denote v as a generic value function, $\hat{v}$ as an estimate of v, and $\tilde{v} \triangleq v - \hat{v}$ as the error.

A. Elements of RKHS Theory

We denote by $H(\Omega)$ an RKHS over the set $\Omega \subseteq \mathbb{R}^n$ that is constructed using a Mercer reproducing kernel $\mathfrak{K} : \Omega\times\Omega \to \mathbb{R}$. A Mercer kernel $\mathfrak{K}(\cdot,\cdot)$ is continuous, symmetric, and of positive type. Being of positive type means that, for any N-point subset $\Xi_N \subset \Omega$, the corresponding Grammian matrix $\mathbb{K}_N \triangleq [\mathfrak{K}(\xi_i,\xi_j)] \in \mathbb{R}^{N\times N}$ is positive semidefinite. The native space $H(\Omega)$ itself is then determined as the closure of the linear span of the kernel sections $\mathfrak{K}_x(\cdot) \triangleq \mathfrak{K}(x,\cdot)$, that is, $H(\Omega) \triangleq \overline{\mathrm{span}\{\mathfrak{K}_x(\cdot) \mid x\in\Omega\}}$, where the closure is taken with respect to the candidate inner product $(\mathfrak{K}_x,\mathfrak{K}_y) \triangleq \mathfrak{K}(x,y)$ for all $x,y\in\Omega$.

Approximations in this paper are constructed using the finite-dimensional subspace $H_N \triangleq \mathrm{span}\{\mathfrak{K}_{\xi_i}(\cdot) \mid \xi_i \in \Xi_N, 1\leq i\leq N\}$. We denote by $\Pi_N : H(\Omega) \to H_N$ the $H(\Omega)$-orthogonal projection of $H(\Omega)$ onto H_N. A key property of orthogonal projections onto a closed subspace of a Hilbert space is that they map an arbitrary input into the closest element of the subspace.

The evaluation functional $E_x : H(\Omega) \to \mathbb{R}$ is defined so that, for each $x \in \Omega$ and every $f \in H(\Omega)$, it holds that $E_x f \triangleq f(x)$. Thus, the evaluation functional defines the bounded linear mapping $H(\Omega) \to \mathbb{R}$. The reproducing property, which is satisfied for any RKHS, implies that $E_x f = f(x) = (f,\mathfrak{K}_x)_H$ for any $f \in H(\Omega)$ and $x \in \Omega$. Furthermore, as E_x is a bounded linear operator between Hilbert spaces, its adjoint operator $E_x^* \triangleq (E_x)^* : \mathbb{R} \to H(\Omega)$ is a bounded linear operator. This adjoint operator is expressed as $E_x^*\alpha \triangleq \mathfrak{K}_x\alpha$ for all $\alpha \in \mathbb{R}, x \in \Omega$. That is, E_x^* can be understood as a multiplication operator since it multiplies any real number by the function $\mathfrak{K}_x$.

If the kernel $\mathfrak{K}(\cdot,\cdot)$ is bounded on the diagonal, then per definition, there exists a constant $\bar{\mathfrak{K}}$ such that, $\mathfrak{K}(x,x) \leq \bar{\mathfrak{K}}^2$ for every $x \in \Omega$. This condition guarantees that every function within the space $H(\Omega)$ is continuous and bounded. Furthermore, it ensures boundedness of the operator norm, that is, $\|E_x\| = \|E_x^*\| \leq \bar{\mathfrak{K}}$. It is worth noting that many commonly

used kernels satisfy this criterion, including the inverse multiquadric, Sobolev-Matérn, Wendland, and exponential kernels [15].

B. Differential Operator A on Native Spaces

We begin by introducing the differential operator A that is defined pointwise as $(Av)(x) \triangleq (f(x)+g(x)\mu(x))^{\mathrm{T}} \nabla v(x)$ for all $x \in \Omega$, whenever v is sufficiently smooth. Note that (4) then corresponds to the operator equation $Av = b$ with $b = -r$, r defined in terms of the kernel r of the cost function in (2), and $v = V_\mu$.

Theorem 1. *Let the kernel $\mathfrak{K} : \Omega \times \Omega \to \mathbb{R}$ that defines the native space $H(\Omega)$ be a $C^{2m}(\Omega,\Omega)$ function with $m \geq 1$, and suppose that μ and f_i, g_i for $1 \leq i \leq d$ are multipliers for $C(\Omega)$ and $H(\Omega)$. Then,*

1) *The operator $A : H(\Omega) \to C(\Omega)$, as well as the operator $A : H(\Omega) \to L^2(\Omega)$, is bounded, linear, and compact.*
2) *The adjoint operator $A^* : L^2(\Omega) \to H(\Omega)$ has representation*

$$A^* = \int_\Omega (\nabla_x \mathfrak{K}(x,y))^{\mathrm{T}} (f(x)+g(x)\mu(x))\, h(x)\mathrm{d}x \triangleq \int_\Omega \ell^*(y,x)h(x)\mathrm{d}x$$

for any $y \in \Omega$ and $h(\cdot) \in L^2(\Omega)$.
3) *Considered as a mapping $A^* : L^2(\Omega) \to H(\Omega)$, or as a mapping $A^* : L^2(\Omega) \to L^2(\Omega)$, the operator A^* is compact.*

Proof: The proof of this theorem can be found in [14], which uses Theorem 1 of [16]. □

Note that the assumptions in Theorem 1 imply that the basis functions that define H_N are continuously differentiable.

C. The DPS Learning Law and Its Approximation

For developing the online learning laws, we introduce the time-varying functional

$$\mathcal{J}(t,\tilde{v}) \triangleq \frac{1}{2}|E_{x(t)}A\tilde{v}|^2 = \frac{1}{2}\left(A^*E^*_{x(t)}E_{x(t)}A\tilde{v},\tilde{v}\right)_H,$$

which is defined for all $\tilde{v} \in H(\Omega)$ that satisfy the additional regularity condition $\tilde{v} \in \{f \in H(\Omega) \mid A\tilde{v} \in H(\Omega)\}$. The analysis in the remainder of this paper always assumes that this regularity condition holds. An elementary calculation shows that the Fréchet derivative of $\mathcal{J}(t,\tilde{v})$ is given by $D\mathcal{J} \triangleq A^*E^*_{x(t)}E_{x(t)}A$. For a fixed time t, let $\hat{v}(t,\cdot) \in H(\Omega)$ be a time-varying approximation of the minimizer v of $\mathcal{J}(t,v)$. An ideal gradient learning law designs the error $\tilde{v}(t,\cdot) \triangleq v - \hat{v}(t,\cdot)$ so that it evolves in the local direction of steepest descent, which is defined in terms of the Fréchet differential in

$$\frac{\partial}{\partial t}\tilde{v}(t,\cdot) = -aA^*E^*_{x(t)}(y(t) - E_{x(t)}A\hat{v}(t,\cdot)) \in H(\Omega),$$

where $y(t) \triangleq E_{x(t)}Av$, and $a > 0$. This ideal gradient law evolves in $H(\Omega)$, and defines a distributed parameter system. In the usual way, we define the ideal evolution law for the estimate $\hat{v}(t,\cdot)$ as

$$\frac{\partial}{\partial t}\hat{v}(t,\cdot) = -aA^*E^*_{x(t)}E_{x(t)}A\hat{v}(t,\cdot) + aA^*E^*_{x(t)}y(t) \in H(\Omega).$$

Note that, in contrast to [7], the critic state evolves in a function space and this evolution law can be understood as a PDE. Finite-dimensional approximations of this PDE are obtained by choosing $\hat{v}_N(t,\cdot) \triangleq \sum_{j=1}^N \hat{W}_j(t)\mathfrak{K}_{\xi_j}(\cdot)$ and seeking a solution of

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\hat{v}_N(t,\cdot)\\ &= -a\Pi_N A^*E^*_{x(t)}E_{x(t)}A\Pi_N\hat{v}_N(t,\cdot) + a\Pi_N A^*E^*_{x(t)}y(t).\end{aligned} \tag{6}$$

These finite-dimensional equations evolve in H_N, and they are equivalent to a system of ODEs.

D. Online Coordinate Realizations

The critical step in deriving coordinate realizations of (6) must examine representations of the operator $\Pi_N A^*E^*_xE_xA\Pi_N$. The finite-dimensional approximation $\Pi_N A^*E^*_xE_xA\Pi_N$ can be deduced by considering $g = \mathfrak{K}_{\xi_j}$ and $h = \mathfrak{K}_{\xi_i}$ to obtain

$$\begin{aligned}[\mathbb{A}_N(x)]_{i,j} &\triangleq ((\Pi_N A^*E^*_xE_xA\Pi_N)\mathfrak{K}_{\xi_j},\mathfrak{K}_{\xi_i})_H,\\ &= \left[\Phi^{\mathrm{T}}(x,\Xi_N)\psi(x)\psi(x)^{\mathrm{T}}\Phi(x,\Xi_N)\right]_{i,j}.\end{aligned}$$

After taking the inner product of (6) with an arbitrary $\mathfrak{K}_{\xi_i} \in H_N$, we therefore obtain the system of ODEs

$$\mathbb{K}_N\dot{\hat{W}}(t) = -a\mathbb{A}_N(x(t))\hat{W}(t) + aY(t),$$

where $\hat{W} \triangleq [\hat{W}_1(t),\ldots,\hat{W}_N(t)]^{\mathrm{T}}$, $y(t) = E_{x(t)}Av = b(x(t))$ denotes the output, $Y_i(t) \triangleq (A^*E^*_{x(t)}y(t),\mathfrak{K}_{\xi_i})_{H(\Omega)}$ and $Y(t) = [Y_1,\ldots,Y_N(t)]^{\mathrm{T}}$.

Remark 1: Interestingly, $Y(\cdot)$ is essentially equivalent to the right-hand-side of (5), with a slight difference being that the normalization with $(\sigma^{\mathrm{T}}\sigma+1)^2$ in (5) is not introduced here.

Remark 2: It is well-known that, in practice, the gradient learning law in (6) must use a robust modification whenever external noise, numerical noise, or approximation error appears in $Y(\cdot)$. This is the reason for the normalization ordinarily used in PI and reinforcement learning. In the following, we discuss a dead-zone robust modification for this purpose. The dead-zone modification is advantageous since it enables a simpler proof of rates of convergence in some cases.

E. Rates of Convergence and Online Performance Bounds

In our first error analysis of online algorithms, we employ the gradient learning law (6). This analysis is based on modifying the approach in [7] and carefully tracking the dependence of expressions on the number of bases N and the approximation error. The theorem below develops an ultimate bound on $\bar{v}_N \triangleq \Pi_N\tilde{v}_N = \Pi_N(v-\hat{v}_N)$.

Theorem 2. *Suppose that the kernel $\mathfrak{K}(\cdot,\cdot)$ that defines the RKHS $H(\Omega)$ is bounded on the diagonal by a constant $\bar{\mathfrak{K}}^2$.*

In addition assume that the family of subspaces $\{H_N\}_{N\in\mathbb{N}}$ *and trajectory* $t \mapsto x(t)$ *are PE in the sense that there are constants* $\Delta(N), \gamma_1(N)$ *depending on* N *and* $\gamma_2 > 0$ *such that*

$$\gamma_1(N) I_{H_N} \leq \underbrace{\int_t^{t+\Delta(N)} \Pi_N A^* E^*_{x(\tau)} E_{x(\tau)} A \Pi_N \mathrm{d}\tau}_{S_N(t)} \leq \gamma_2 I_{H_N}$$

for each $N \in \mathbb{N}$*, where* $S_N(t) : H_N \to H_N$*. Then,*

$$\begin{aligned}\|\bar{v}_N(t,\cdot)\|_{H(\Omega)} &\triangleq \|\Pi_N v - \hat{v}_N(t,\cdot)\|_{H(\Omega)} \\ &\leq \frac{\sqrt{\gamma_2 \Delta(N)}}{\gamma_1(N)} (\bar{\mathcal{Y}}_{N,\max} + \delta\gamma_2 a(\bar{\mathcal{Y}}_{N,\max} + \epsilon_{N,\max})). \quad (7)\end{aligned}$$

where

$$\bar{\mathcal{Y}}_{N,max} \triangleq \sup_{\tau\in[t,t+\Delta(N)]} |E_{x(\tau)} A \Pi_N \bar{v}_N(t,\cdot)|, \quad (8)$$

$$\epsilon_{N,max} \triangleq \sup_{\tau\in[t,t+\Delta(N)]} \epsilon_N(\tau). \quad (9)$$

Proof: The consistent approximation of the gradient law can be written as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bar{v}_N(t,\cdot) = &- a\Pi_N A^* E^*_{x(t)} E_{x(t)} A \bar{v}_N(t,\cdot) \\ &- a\Pi_N A^* E^*_{x(t)} E_{x(t)} A (I - \Pi_N) v.\end{aligned}$$

Following the proof of Technical Lemma 2, part b in [7], we rewrite this equation as the system

$$\begin{aligned}\dot{\mathcal{X}}_N(t) &= B_N(t)\mathcal{U}_N(t), \\ \mathcal{Y}_N(t) &= C^*_N(t)\mathcal{X}_N(t),\end{aligned}$$

where $B_N(t) \triangleq -a\Pi_N A^* E^*_{x(t)}$, $C^*_N(t) \triangleq E_{x(t)} A \Pi_N$, $\mathcal{X}_N(t) \triangleq \bar{v}_N(t,\cdot)$, $\epsilon_N(t) \triangleq E_{x(t)} A(I - \Pi_N)v$, and $\mathcal{U}_N(t) \triangleq -\mathcal{Y}_N(t) + \epsilon_N(t)$. Both $B_N(t)$ and $C_N(t)$ are bounded linear operators, and their bounds can be chosen independently of N sicne the kernel $\mathfrak{K}(\cdot,\cdot)$ is bounded on the diagonal, and, hence, $\|E_{x(t)}\| = \|E^*_{x(t)}\| \leq \bar{\mathfrak{K}}$. Also, it holds that $\mathcal{X}_N(t) \in H_N$ and $\mathcal{Y}_N(t) \in \mathbb{R}$. The proof of Technical Lemma 2 part b in [7] holds for states, controls, and observations in Euclidean spaces, like $\mathbb{R}^d$ or $\mathbb{R}$. Since all the operators above are bounded, each step in the proof of Equation (A.9) in Technical Lemma 2 part b in [7] can also be applied without change in the current setting. In particular, it holds that

$$\begin{aligned}\|\mathcal{X}(t)\|_{H_N} \leq &\frac{\sqrt{\gamma_2\Delta(N)}}{\gamma_1(N)} \bar{\mathcal{Y}}_{N,\max} \\ &+ \frac{\delta\gamma_2}{\gamma_1(N)} \int_t^{t+\Delta(N)} \|B_N(\tau)\| \cdot \|\mathcal{U}_N(\tau)\| \mathrm{d}\tau\end{aligned}$$

for a constant δ of order one. Furthermore, $\|B_N(t)\| \leq a\|A^*\|\bar{\mathfrak{K}}$ and $\|\mathcal{U}_N(t)\| \leq \bar{\mathcal{Y}}_{N,\max} + |\epsilon_N(t)|$. We conclude that the rate in (7) holds with

$$\epsilon_{N,\max} \leq \sup_{\tau\geq 0} E_{x(\tau)} A(I - \Pi_N)v. \quad \square \quad (10)$$

The next theorem bounds the ultimate output error $\tilde{y}_N(t) \triangleq y(t) - \hat{y}_N(t)$, where $y(t) = E_{x(t)} Av$ and $\hat{y}_N(t) \triangleq E_{x(t)} A\hat{v}_N(t,\cdot)$, in terms of the approximation error $\epsilon_{N,\max}$ in the case whereby we use a hard dead-zone version of the learning law with a properly sized dead-zone. We emphasize how the next result does not require a PE condition, and the error bound on performance is more readily tied to just the approximation error $\epsilon_{N,\max}$ as described in Section IV. On the other hand, in principle, an oracle must define a dead-zone that is a tight bound for the approximation error. In practice, the size of the dead-zone is defined iteratively.

Theorem 3. *Consider a learning law for* $\hat{v}_N(\cdot,\cdot)$*, such that if* $\tilde{y}_N(t) \triangleq E_{x(t)} A\tilde{v}_N(t,\cdot) \geq \bar{\epsilon} \geq \epsilon_{N,\max}$ *for some* $t \geq 0$*, then* (6) *is verified, and, if* $\tilde{y}_N(t) < \epsilon_{N,\max}$ *for some* $t \geq 0$*, then* $\frac{d}{dt}\hat{v}_N(t,\cdot) = 0$*. Then, for any arbitrarily small constant* $\eta > 0$*, there exists* $T \triangleq T(\eta) > 0$ *such that* $|E_{x(t)}(\mathcal{H}_\mu - \hat{\mathcal{H}}_N(t,\cdot))| \equiv |\tilde{y}_N(t)| \leq \frac{1+\eta}{a}\bar{\epsilon}$ *for all* $t \geq T(\eta)$*, where the Hamiltonian* $\mathcal{H}_\mu$ *is defined in* (4) *and* $\hat{\mathcal{H}}_N(t,x) \triangleq A\hat{v}_N(t,x) + r(x)$ *denotes the approximate Hamiltonian with* $x \in \Omega$*. If we choose* $\bar{\epsilon} \triangleq M(N)\epsilon_{N,\max}$ *for some (small) integer* $M(N)$*, and* $T_O > 0$ *is the time that the measurement error* $\tilde{y}_N(t)$ *spends outside the dead-zone, then we an ultimate bound on the decrease of the value function error is given by* $\|\tilde{v}_N(t,\cdot)\|^2_{H(\Omega)} \leq \|\tilde{v}_N(t_0,\cdot)\|^2_{H(\Omega)} - 2aT_O(1+M(N))M(N)\epsilon^2_{N,\max}$ *for all* $t \geq 0$ *large enough.*

Proof: In this proof we choose the Lyapunov function $\mathcal{V}(\tilde{v}_N) \triangleq \frac{1}{2}(\tilde{v}_N, \tilde{v}_N)_{H(\Omega)}$. When $|\tilde{y}_N(t)| \geq \bar{\epsilon}$, the derivative of the Lyapunov function along trajectories of the learning law satisfy

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{V}(\tilde{v}_N(t,\cdot)) \\ &= -a\left(E_{x(t)}A\tilde{v}_N(t,\cdot), E_{x(t)}A\tilde{v}_N(t,\cdot)\right)_{\mathbb{R}} \\ &\quad + a\left(E_{x(t)}A\tilde{v}_N(t,\cdot), -E_{x(t)}A(I-\Pi_N)\tilde{v}_N(t)\right)_{\mathbb{R}} \\ &\leq -a|\tilde{y}_N(t)|\left(|\tilde{y}_N(t)| - \epsilon_{N,\max}\right).\end{aligned}$$

Because $|\tilde{y}_N(t)| \geq \bar{\epsilon} \geq \epsilon_{N,\max}$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{V}(\tilde{v}_N(t,\cdot)) &\leq -a|\tilde{y}_N(t)|\left(|\tilde{y}_N(t)| - \epsilon_{N,\max}\right), \\ &\leq -a\bar{\epsilon}\left(\bar{\epsilon} - \epsilon_{N,\max}\right) < 0,\end{aligned}$$

while the trajectory is outside the dead-zone. Following standard convergence arguments, we conclude that the time spent outside the dead-zone is finite, and, thus, the norm of the output $\tilde{y}(t)$ is ultimately bounded by the dead-zone. The bound on the value function error $\|\tilde{v}(t,\cdot)\|_{H(\Omega)}$ can be derived by evaluating the Lyapunov function at the time the observations $\tilde{y}(t)$ enters the dead-zone [17, Ch. 10]. $\square$

IV. Explicit Error Bounds and Fill Distances

In this section, we describe how some techniques used to describe rates of convergence of approximations in a native space can be applied to the bounds in Theorems 2 and 3 on the online critic. Note that a bit more can be said about the errors $\epsilon_N(t)$ and $\epsilon_{N,\max}$ that appear in these theorems. It holds that

$$\begin{aligned}\epsilon_N(t) &\triangleq |E_{x(t)}A(I-\Pi_N)v| = |(\ell(\cdot,x(t)), (I-\Pi_N)v)_{H(\Omega)}| \\ &\leq \sup_{\xi\in\Omega} \|\ell(\cdot,\xi)\|_{H(\Omega)} \|(I-\Pi_N)v\|_{H(\Omega)} \leq \ell_{\max}\|(I-\Pi_N)v\|_{H(\Omega)},\end{aligned}$$

where $\ell(x,y) = \ell^*(y,x)$ and $\ell^*(y,x)$ is defined in Theorem 1.

The remainder of this section describes how $\|(I - \Pi_N)v\|_{H(\Omega)}$ can be explicitly bounded in terms of the placement of centers in Ξ_N. The power function of the subspace H_N in the RKHS $H(\Omega)$ is defined as $\mathcal{P}_N(x) \triangleq \sqrt{\mathfrak{K}(x,x) - \mathfrak{K}_N(x,x)}$ with $\mathfrak{K}_N$ the reproducing kernel of the subspace H_N [15], [18]. It can be proven that $\mathfrak{K}_N(x,y) \triangleq \mathfrak{K}_{\Xi_N}(x)^{\mathrm{T}}\mathbb{K}_N^{-1}\mathfrak{K}_{\Xi_N}(y)$ where $\mathfrak{K}_\Xi(x) = [\mathfrak{K}_{\xi_1}(x), \ldots, \mathfrak{K}_{\xi_N}(x)]^{\mathrm{T}} \in \mathbb{R}^N$ denotes the column vector of N basis functions defined in terms of the set of centers $\Xi_N \subset \Omega$. The power function is useful for generating pointwise bounds on the projection error such as $|E_x(I - \Pi_N)v| \leq \mathcal{P}_N(x)\|(I - \Pi_N)v\|_{H(\Omega)}$ for all $x \in \Omega$ and $v \in H(\Omega)$ and any native space whatsoever [15], [18].

We use this well-known identity to bound the error $\|(I - \Pi_N)v\|_{H(\Omega)}$ that appears in the ultimate bound of the critic.

Theorem 4 (Modification of Theorem 11.23 in [15])**.** *Suppose that v satisfies the regularity condition $v = \mathcal{L}u$, where $\mathcal{L} : L^2(\Omega) \to H(\Omega)$ is the bounded, linear, compact operator $(\mathcal{L}u)(x) \triangleq \int_\Omega \mathfrak{K}(x,y)u(y)\mathrm{d}y$. Then, there is a constant $C > 0$ such that $\|(I - \Pi_N)v\|_{H(\Omega)} \leq C \sup_{\xi\in\Omega} |\mathcal{P}_N(\xi)| \|\mathcal{L}^{-1}v\|_{L^2(\Omega)}$ provides an error bound.*

Proof: This proof is based on that of Theorem 11.23 of [15], and for completeness we summarize the simple modifications here. First note that

$$(w, Lu)_{H(\Omega)} = \int_\Omega (w, \mathfrak{K}_y)_{H(\Omega)} u(y)\mathrm{d}y = (w,u)_{L^2(\Omega)}.$$

Now we can write

$$\begin{aligned}\|(I - \Pi_N)v\|^2_{H(\Omega)} &= ((I - \Pi_N)v, v)_{H(\Omega)}, \\ &= (I - \Pi_N)v, u)_{L^2(\Omega)}, \\ &\leq \|(I - \Pi_N)v\|_{L^2(\Omega)}\|u\|_{L^2(\Omega)}.\end{aligned}$$

But we also have

$$\begin{aligned}\|(I - \Pi_N)v\|^2_{L^2(\Omega)} &= \int_\Omega |E_x(I - \Pi_N)v|^2 \mathrm{d}x \\ &\leq |\Omega| |\sup_{\xi\in\Omega} \mathcal{P}_N(\xi)|^2 \|(I - \Pi_N)v\|^2_{H(\Omega)}\end{aligned}$$

Substituting this bound above completes the proof of the theorem. □

Since the centers Ξ_N, kernel $\mathfrak{K}$, and power function $\mathcal{P}_N$ are known, Theorem 4 can be used, in either *a priori* or *a posteriori* estimation of the value function estimate error that results from using a collection of centers Ξ.

The geometric nature of the bound in Theorem 4 is often emphasized by relating the power function to the fill distance of the centers Ξ_N in the set Ω, which is defined as $h_{\Xi_N,\Omega} \triangleq \sup_{y\in\Omega} \min_{\xi_i\in\Xi_N} \|y - \xi_i\|_2$. For a variety of kernel functions, which can be applied to the problem addressed in this paper, [15], [18] provide bounds on the power function in the form $\mathcal{P}_N(x) \lesssim \sqrt{\mathcal{N}(h_{\Xi_N,\Omega})}$, where $\mathcal{N} : \mathbb{R}^+ \to \mathbb{R}^+$ depends on the kernel function. The following lemma summarizes three common examples of such bounds.

Lemma 1. *Suppose that v is contained in the uncertainty class $\mathcal{C}_{L,R} \triangleq \{g = \mathcal{L}u \in H(\Omega) \mid \|u\|_{L^2(\Omega)} \leq R\} \subseteq H(\Omega)$, and that the assumptions of Theorem 3 holds with the minimum size dead-zone $\bar{\epsilon} \approx \epsilon_{N,\max}$. For the Sobolev-Matérn kernel of a high enough smoothness k in Table 11.1 in [15], then there exists $T > 0$ such that, for all $t \geq T$,*

$$|E_{x(t)}(\mathcal{H}_\mu - \hat{\mathcal{H}}_N(t,\cdot))| \equiv |y(t) - \hat{y}_N(t)| \approx O(h^{k-n/2}_{\Xi_N,\Omega}).$$

For the Wendland compactly supported kernel $\eta_{n,k}$, $|E_{x(t)}(\mathcal{H}_\mu - \hat{\mathcal{H}}_N(t,\cdot))| \approx O(h^{k+1/2}_{\Xi_N,\Omega})$. For the exponential kernel, $|E_{x(t)}(\mathcal{H}_\mu - \hat{\mathcal{H}}_N(t,\cdot))| \approx O\left(\sqrt{e^{-\alpha|h_{\Xi,\Omega}|/h_{\Xi_N,\Omega}}}\right)$ for a constant α that depends on the hyperparameters of the exponential kernel.

Proof: This result follows from Theorems 3 and 4. □

V. Numerical Results

In this section, we present numerical validation studies for the system of the form (1) studied in [7], with $f(x) = \left[-x_1 + x_2, -0.5x_1 - 0.5x_2\left(1 - (\cos(2x_1) + 2)^2\right)\right]^{\mathrm{T}}$ and $g(x) = [0, \cos(2x_1) + 2]^{\mathrm{T}}$. The cost function for this problem sets $R = 1$ and $Q = I_2$, with I_2 the identity matrix in $\mathbb{R}^{2\times 2}$. The optimal value function is $V^\star(x) = 0.5x_1^2 + x_2^2$, which generates the optimal feedback controller $u^\star(x) = -(\cos(2x_1) + 2)x_2$. The numerical validation studies in [7] are based on a very low-dimensional system of polynomial bases, whose span contains the exact optimal value function.

Figure 1 depicts the error norm $\|V^\star - \hat{v}_N(t,\cdot)\|_{L^\infty(\Omega)}$ for two Matérn kernels and an exponential kernel. Since $\|E_x\| \leq \bar{\mathfrak{K}}$, it holds that $|E_x\tilde{v}_N(t,\cdot)| \leq \bar{\mathfrak{K}}\|\tilde{v}_N(t,\cdot)\|_{H(\Omega)}$ and $\|\tilde{v}_N(t,\cdot)\|_{L^\infty(\Omega)} \leq \bar{\mathfrak{K}}\|\tilde{v}_N(t,\cdot)\|_{H(\Omega)}$, and Lemma 1 implies the corresponding convergence in the norm of $L^\infty(\Omega)$. The ultimate approximate value function $\hat{v}_N(t,\cdot)$ closely matches the analytical expression for the optimal value function $V^\star$ as the dimension $N \to \infty$. Note that Theorem 1 only guarantees that $\hat{v}_N(t,\cdot)$ converges to V_μ, not $V^\star$, and, indeed, this plot is a more stringent empirical test of the performance of the critic. Figure 1 illustrates that the online critic estimates $\hat{v}_N(t,\cdot)$ for the Sobolev-Matérn kernels converge at a rate that is theoretically determined by the fill distance as described in the paper in Lemma 1.

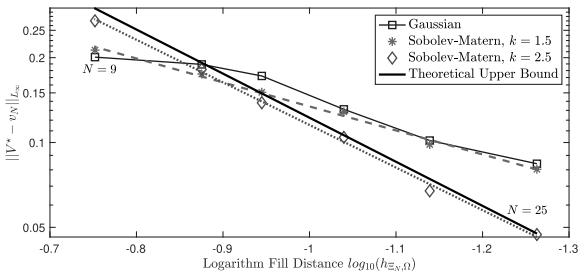

Fig. 1. The $L^\infty(\Omega)$ error norm of the online critic estimates of the value function $V^\star$ using the dead-zone rule described in Lemma 1. The steady-state value function approximations using Sobolev-Matérn kernels of smoothness $k = 1.5, 2.5$ and exponential kernels are plotted above. Note that the rates of convergence for the Sobolev-Matérn kernels closely follow the theoretical bounds derived in Lemma 1.

This article has been accepted for publication in IEEE Control Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2024.3417178

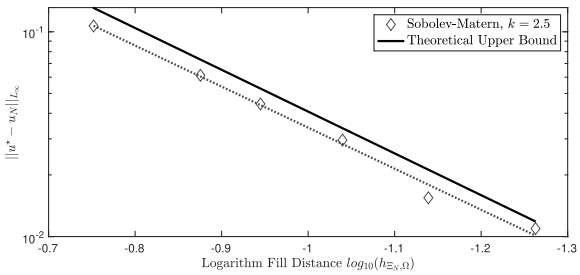

Fig. 2. The $\boldsymbol{L^\infty(\Omega)}$ error norm of the online critic estimates of the control input $\boldsymbol{u^\star}$ with Sobolev-Matérn kernels of smoothness $\boldsymbol{k = 2.5}$. Note that the rates of convergence for the Sobolev-Matérn kernels closely follow the theoretical bounds derived in Theorem 4.

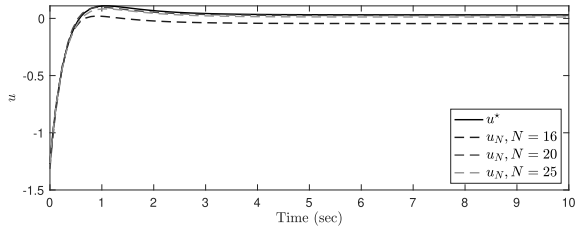

Fig. 3. Feedback control $\boldsymbol{u}$ by learned $\hat{\boldsymbol{W}}$ with Sobolev-Matérn kernels of smoothness $\boldsymbol{k = 2.5}$.

We should emphasize that these studies make use of regular arrays of centers to verify and validate the derived error bounds. The proposed method does not require that centers be selected using regular grids. Suppose for example that the state trajectory of interest is embedded in a high-dimensional state space, but its evolution resides on a low-dimensional submanifold embedded in that high-dimensional space. Since the approach in this paper uses scattered bases that are not confined to any *a priori* grids or triangulations, their locations can be tailored to locations on the submanifold. Such a strategy can be pursued to address issues related to the curse of dimensionality. An in-depth study of how to execute this strategy in practice far exceeds the limits of this introductory paper. However, in principle, the proposed method is not restricted to regular grids of bases that scale like N^d.

A bit more can be deduced about the value function error in $L^\infty(\Omega)$ when the regularity condition in Lemma 1 holds. This is referred to as the "doubling trick" in the literature on approximations in RKHS; see Theorem 11.23 of [15] that enables the conclusion $|E_x(I-\Pi_N)f| \leq O((\sup_{\xi\in\Omega}\mathcal{P}_N(\xi))^2)$. A line having this slope for the Sobolev-Matérn kernel with $k = 2.5$ is labeled in Figure 1 as the "theoretical upper bound."

Often, in implementations, it is of vital concern to establish the rates of convergence of the error $\mu - \hat{\mu}_N$ where $\hat{\mu}_N$ is the control approximation based on $\hat{v}_N(t,\cdot)$ of the ideal control $u^\star$. We can proceed exactly as in the proof of Theorem 3 of [19] in the case at hand to conclude that

$$\begin{aligned}\|u^\star - \hat{u}_N(t,\cdot)\|_{C(\Omega)} &\leq C\|V^\star - \hat{v}_N(t,\cdot)\|_{H(\Omega)} \\ &\leq C\left(\|V^\star - V_\mu\|_{H(\Omega)} + \|\tilde{v}_N(t,\cdot)\|_{H(\Omega}\right)\end{aligned}$$

for some fixed constant $C > 0$. Thus, if $\|V^* - V_\mu\|_{H(\Omega)}$ is sufficiently small, say of $O(\epsilon_{N,\max})$, then we expect the same rate of convergence for the control convergence in $C(\Omega)$ as in Lemma 1 for $\|\tilde{v}_N(t,\cdot)\|_{H(\Omega)}$.

VI. Conclusions

This paper has formulated the online critic for estimating the optimal value function in terms of evolution laws for a wide variety of RKHSs. The paper lifts conventional approaches, which focus on studies of the convergence of parameter errors $\|W - \hat{W}(t)\|_{\mathbb{R}^N}$ in $\mathbb{R}^N$, to instead focus on the norms of the value function error $\|V^\star - \hat{v}(t,\cdot)\|_{H(\Omega)}$. A wide variety of the performance bounds on the error in the value function estimates are derived in terms of the power function of the scattered basis. This basic result is subsequently refined to obtain performance guarantees on the critic in terms of the fill distance of the centers in the subset of interest Ω.

References

[1] F. L. Lewis, D. Vrabie, and V. L. Syrmos, Optimal control. John Wiley & Sons, 2012.
[2] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 6, pp. 2042–2062, 2017.
[3] K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Handbook of reinforcement learning and control. Springer, 2021.
[4] R. W. Bea, "Successive galerkin approximation algorithms for nonlinear optimal and robust control," Int. J. Control, vol. 71, no. 5, pp. 717–743, 1998.
[5] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized hamilton-jacobi-bellman equation," Automatica, vol. 33, no. 12, pp. 2159–2177, 1997.
[6] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach," Automatica, vol. 41, no. 5, pp. 779–791, 2005.
[7] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," Automatica, vol. 46, no. 5, pp. 878–888, 2010.
[8] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," Automatica, vol. 49, no. 1, pp. 82–92, 2013.
[9] F. L. Lewis and D. Liu, Reinforcement learning and approximate dynamic programming for feedback control. John Wiley & Sons, 2013.
[10] R. Kamalapurkar, P. Walters, J. Rosenfeld, and W. Dixon, Reinforcement learning for optimal feedback control. Springer, 2018.
[11] T. Bian and Z.-P. Jiang, "Reinforcement learning and adaptive optimal control for continuous-time nonlinear systems: A value iteration approach," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 7, pp. 2781–2790, 2021.
[12] D. Kalise, S. Kundu, and K. Kunisch, "Robust feedback control of nonlinear pdes by numerical approximation of high-dimensional hamilton–jacobi–isaacs equations," SIAM J. Appl. Dyn. Syst., vol. 19, no. 2, pp. 1496–1524, 2020.
[13] Y. Yang, H. Modares, K. G. Vamvoudakis, W. He, C.-Z. Xu, and D. C. Wunsch, "Hamiltonian-driven adaptive dynamic programming with approximation errors," IEEE Trans. Cybern., vol. 52, no. 12, pp. 13 762–13 773, 2021.
[14] A. Bouland, S. Niu, S. T. Paruchuri, A. Kurdila, J. Burns, and E. Schuster, "Rates of convergence in a class of native spaces for reinforcement learning and control," IEEE Control Syst. Lett., vol. 8, pp. 55–60, 2024.
[15] H. Wendland, Scattered data approximation. Cambridge university press, 2004, vol. 17.
[16] D.-X. Zhou, "Derivative reproducing properties for kernel methods in learning theory," J. Comput. Appl. Math., vol. 220, no. 1-2, pp. 456–463, 2008.
[17] E. Lavretsky and K. Wise, Robust and Adaptive Control: With Aerospace Applications. Springer, 2013.
[18] R. Schaback, "Error estimates and condition numbers for radial basis function interpolation," Adv. Comput. Math., vol. 3, pp. 251–264, 1994.
[19] A. Bouland, S. Niu, S. T. Paruchuri, A. Kurdila, J. Burns, and E. Schuster, "Rates of convergence in certain native spaces of approximations used in reinforcement learning," 2023, arXiv:2309.07383.