

Convergence Rates of Online Critic Value Function Approximation in Native Spaces

Shengyuan Niu¹, Ali Bouland¹, Haoran Wang¹, Filippos Fotiadis²,
Andrew Kurdila¹, Andrea L’Afflitto³, Sai Tej Paruchuri⁴, Kyriakos G. Vamvoudakis²

Abstract—In this paper, the evolution equation that defines the online critic for the approximation of the optimal value function is cast in a general class of reproducing kernel Hilbert spaces (RKHSs). Exploiting some core tools of RKHS theory, this formulation allows deriving explicit bounds on the performance of the critic in terms of the kernel and definition of the RKHS, the number of basis functions, and the location of centers used to define scattered bases. The performance of the critic is precisely measured in terms of the power function of the scattered basis used in approximations, and it can be used either in an *a priori* evaluation of potential bases or in an *a posteriori* assessments of 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.

I. INTRODUCTION

Optimal control has become one of the core methodologies in modern control theory for efficient decision-making and policy design. One of its main advantages lies in its ability to yield control laws that achieve a compromise between the control effort expended in the closed loop 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, but whose solution directly yields the optimal control policy for the system. Numerous studies describe methods to approximate the solution of the HJB equation [2], with one of the most popular such tools being policy iteration (PI). This process iteratively evaluates the cost function for a given controller, and, subsequently, improves that controller from measurements. Nevertheless, one issue with PI is its need to employ a neural network (NN) for the policy evaluation step, called “critic,” which inherently leads to approximation errors that

degrade performance or lead to failure of convergence of the PI process.

To study approximation errors and their influence on performance in the context of PI, the use of Galerkin approximations in a recursive implementation has a long history. Notable early efforts include [3], [4]. The authors in [5] build on earlier work on Galerkin approximations to handle saturating actuators. This latter paper derives some basic convergence behavior of recursive methods but does not discuss how approximation errors *quantitatively* affect critic or closed-loop control performance. Subsequent papers [6], [7] use some of the theory in [3]–[5] to study various online implementations based on learning theory. Some basic guarantees of the convergence of the critic are derived in [6] in the style of [5], but the role of basis selection and resulting approximation error on the system’s performance are not studied. To the authors’ knowledge, the overall question of how basis selection and approximation errors affect the controller performance for either offline or online versions of these methods remains largely an open question. The works referenced in [8] and [9], for example, provide comprehensive insights into the modern 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 [10]–[12] emphasize the importance of examining the impact of the approximation error on the performance of reinforcement learning methods.

This work extends the effort started in [13] that rigorously frames the optimal control problem in a reproducing kernel Hilbert space (RKHS), and further fills this gap in the literature on basis selection, approximation error, and performance of PI. Specifically, we address the basic question of how the basis functions of the critic NN ought to be selected in the RKHS setting. The strongest results of this paper are geometric since we describe how the fill distance of the centers used to define the bases for approximation dictates the performance of the critic. In a few different instances, we relate the rate of convergence in the RKHS directly and explicitly to the performance of the critic.

II. PROBLEM STATEMENT

Consider the continuous-time nonlinear system

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

¹S. 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.

²F. 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.

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

⁴S. 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 $x(t) \in \mathbb{R}^n$ represents the state of the system, and $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$, and $u : \mathbb{R}^n \rightarrow \mathbb{R}^m$ represent the drift dynamics, the input dynamics, and the control input, respectively. This work is motivated by NN methods used to approximately solve the infinite-horizon optimal control problem. The problem is to find the continuous control input $t \mapsto u(t)$ that minimizes the cost functional

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

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

$$\begin{aligned} \nabla V^{*T}(x)f(x) - \frac{1}{4}\nabla V^{*T}(x)g(x)R^{-1}g^T(x)\nabla V^*(x) \\ + Q(x) = 0, \quad V^*(0) = 0, \quad \forall x \in \Omega, \end{aligned} \quad (3)$$

and then calculate $u^*(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla V^*(x)$ [1]. Nevertheless, (3) is generally difficult, if not impossible, to solve analytically for V^* . For this reason, PI [3], [5], a process that iteratively evaluates the cost of a stabilizing controller, and then improves that controller, is usually employed as a means to at least solve (3) approximately.

The most crucial and computationally demanding step of PI is that of policy evaluation. Given a continuous feedback gain $\mu : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 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

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

where $\mathcal{H}_\mu(V_\mu)$ is known as the Hamiltonian associated with μ and V_μ . While an analytical solution to (4) is also difficult to obtain, its linearity with respect to 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 uniformly approximated on Ω as $V_\mu(x) = W^T\phi(x) + \epsilon_N(x)$, $\forall x \in \Omega$, where $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^N$ is a suitable vector of N basis functions, $W \in \mathbb{R}^N$ denote the “ideal weights” for that basis, and $\epsilon_N : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $\hat{v}_N(t, x) = \hat{W}^T(t)\phi(x)$. The purpose of policy evaluation is, thus, to properly train the critic weights $\hat{W}(t)$ so that the parameter error $\tilde{W}(t) \triangleq W - \hat{W}(t)$ becomes as small as possible. In [6], the online policy evaluation law

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

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 choice of basis influences the performance of the critic.

This paper lifts the analysis of the parameter error $\tilde{W}(t) \triangleq \|\hat{W}(t) - W\|_{\mathbb{R}^N}$ to instead generate estimates of the error of the value function $\|v_N(t, \cdot) - V_\mu\|_{H(\Omega)}$ in a way that makes explicit the contribution of approximation errors in a wide variety of choices of the space $H(\Omega)$. 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 potential 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

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 $\mathfrak{K}(\cdot, \cdot) : \Omega \times \Omega \rightarrow \mathbb{R}$. 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 Gramian 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, $H(\Omega) \triangleq \overline{\text{span}\{\mathfrak{K}_x(\cdot) \mid x \in \Omega\}}$, 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 $H_N \triangleq \text{span}\{\mathfrak{K}_{\xi_i}(\cdot) \mid \xi_i \in \Xi_N, 1 \leq i \leq N\}$, where $\Xi_N \subset \Omega$ is a collection of N “centers of approximation” contained in Ω . We denote by $\Pi_N : H(\Omega) \rightarrow H_N$ the $H(\Omega)$ -orthogonal projection of $H(\Omega)$ onto H_N .

Many of the new results in this paper follow from some standard properties of the evaluation functional $E_x : H(\Omega) \rightarrow \mathbb{R}$ over an RKHS. By definition, for each $x \in \Omega$ and every $f \in H(\Omega)$, it holds that $E_x f \triangleq f(x)$, and this property defines a bounded linear mapping $H(\Omega) \rightarrow \mathbb{R}$. 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 $E_x^* \triangleq (E_x)^* : \mathbb{R} \rightarrow H(\Omega)$ 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 positive 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 [14].

B. Differential Operator A on Native Spaces

We begin by introducing the differential operator A that is defined pointwise as $(Av)(x) \triangleq (f(x) + g(x)\mu(x))^T \nabla v(x)$ 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 $\mathfrak{K} : \Omega \times \Omega \rightarrow \mathbb{R}$ 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 $A : H(\Omega) \rightarrow C(\Omega)$, as well as the operator $A : H(\Omega) \rightarrow L^2(\Omega)$, is bounded, linear, and compact.
- 2) The adjoint operator $A^* : L^2(\Omega) \rightarrow H(\Omega)$ has the representation

$$\begin{aligned} A^* &= \int_{\Omega} (\nabla_x \mathfrak{K}(x, y))^T (f(x) + g(x)\mu(x)) h(x) dx \\ &\triangleq \int_{\Omega} \ell^*(y, x) h(x) dx \end{aligned}$$

for any $y \in \Omega$ and $h(\cdot) \in L^2(\Omega)$.

- 3) Considered as a mapping $A^* : L^2(\Omega) \rightarrow H(\Omega)$, or as a mapping $A^* : L^2(\Omega) \rightarrow L^2(\Omega)$, the operator A^* is compact.

Proof: The proof of this theorem can be found in [13], which uses Theorem 1 of [15]. ■

C. The DPS Learning Law and Its Approximation

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

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

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 Frechet derivative of $\mathcal{J}(t, \tilde{v})$ is simply $D\mathcal{J} \triangleq A^* E_{x(t)}^* E_{x(t)} A$. 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 Frechet differential in

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

where $y(t) = E_{x(t)} A v$, and $a > 0$. This ideal gradient law evolves in $H(\Omega)$, and it defines a distributed parameter

system. In the usual way, we define the ideal evolution law for the estimate $\hat{v}(t, \cdot)$ as

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

Note that in contrast to [6], the critic state evolves in a function space: it can be understood as a PDE. Finite-dimensional approximations 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

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

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 the finite-dimensional equations above must examine representations of the operator $\Pi_N A^* E_{x(t)}^* E_{x(t)} A \Pi_N$. The finite-dimensional approximation $\Pi_N A^* E_{x(t)}^* E_{x(t)} A \Pi_N$ can be deduced by considering $g = \mathfrak{K}_{\xi_j}$ and $h = \mathfrak{K}_{\xi_i}$ to obtain

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

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

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

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

Observation 1: 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 $Y(t)$. Below we discuss a dead zone robust modification for this purpose.

Observation 2: Interestingly, this expression for $Y(\cdot)$ is essentially the same as the expression for the right-most term in (5), with a slight difference being that the normalization with $(\sigma^T \sigma + 1)^2$ in (5) is not introduced here. We will show that the dead zone robust modification suffices as an alternative to the usual normalization in the reinforcement learning literature.

E. Rates of Convergence and Online Performance Bounds

In our first error analysis of online algorithms, we employ the gradient learning law in (6). The analysis is based on modifying the approach in the paper [6] 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

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

for each $N \in \mathbb{N}$ where $S_N(t) : H_N \rightarrow H_N$. Then we have the error bound

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

where $\bar{\mathcal{Y}}_{N, \max}$ is defined in (8) and $\epsilon_{N, \max}$ is given in (9).

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

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

Following the proof of Technical Lemma 2, part b in [6], 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 $\mathcal{U}_N(t) \triangleq -\mathcal{Y}_N(t) + \epsilon_N(t)$. Carefully note that each of the operators $B_N(t)$ and $C_N(t)$ are bounded linear operators, and the bounds can be chosen independently of N . This holds owing to the assumption that the kernel \mathfrak{K} is bounded on the diagonal, so $\|E_{x(t)}\| = \|E_{x(t)}^*\| \leq \bar{\mathfrak{K}}$. Also we have $\mathcal{X}_N(t) \in H_N$ and $\mathcal{Y}_N(t) \in \mathbb{R}$. The proof of Technical Lemma 2 part b in [6] is carried out 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 [6] can also be applied without change in the current setting. We have

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

for a constant δ of order one where

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

But we also have $\|B_N(t)\| \leq a \|A^*\| \bar{\mathfrak{K}}$ and $\|\mathcal{U}_N(t)\| \leq \bar{\mathcal{Y}}_{N, \max} + |\epsilon_N(t)|$. We conclude that the rate in (7) holds where

$$\epsilon_{N, \max} \triangleq \sup_{\tau} \epsilon_N(\tau) \leq \sup_{\tau \geq 0} E_{x(\tau)} A (I - \Pi_N) v. \quad (9)$$

The next theorem bounds the ultimate output error $\tilde{y}_N(t) \triangleq y(t) - \hat{y}_N(t)$, where $y(t) = E_{x(t)} A v$ 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 when we use a hard dead-zone version of the learning law with a properly designed dead-zone size. We emphasize that the result below 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 the next Section IV. On the other hand, in principle an oracle must define a deadzone that is a tight bound for the approximation error. In practice this requires iteration.

Theorem 3: Employ the hard dead zone learning law that uses the gradient law in (6) whenever $\tilde{y}_N(t) : y(t) - \hat{y}_N(t) \triangleq E_{x(t)} A \tilde{v}_N(t, \cdot) \geq \bar{\epsilon} \geq \epsilon_{N, \max}$, and otherwise use $\hat{v}_N(t, \cdot) = 0$ inside the deadzone. Then for any arbitrarily small constant $\eta > 0$ there is a time $T \triangleq T(\eta)$ such that $|E_{x(t)}(\mathcal{H}_\mu - \tilde{\mathcal{H}}_N(t, \cdot))| \equiv |\tilde{y}_N(t)| \leq \frac{1+\eta}{a} \bar{\epsilon}$ for all $t \geq T(\eta) > 0$, where the Hamiltonian \mathcal{H}_μ is defined in (4) and the approximate Hamiltonian is $\tilde{\mathcal{H}}_N(t, x) \triangleq A \hat{v}_N(t, x) + r(x)$ for all $x \in \Omega$. 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 deadzone, we obtain an ultimate bound on the decrease of the value function error $\|\tilde{v}_N(t, \cdot)\|_{H(\Omega)}^2 \leq \|\tilde{v}_N(t_0, \cdot)\|_{H(\Omega)}^2 - 2aT_O(1 + M(N))M(N)\epsilon_{N, \max}^2$ 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

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

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

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

while the trajectory is outside the deadzone. Following standard telescoping sum arguments, as in Chapters 4 or 6 of [16], we conclude that the time spent outside the deadzone is finite, and thus, the norm of the output $\tilde{y}(t)$ is ultimately bounded by the deadzone. The bound on the value function error $\|\tilde{v}(t, \cdot)\|_{H(\Omega)}$ is likewise derived using a telescoping sum of the Lyapunov function defined in terms of the times that the observations $\tilde{y}(t)$ repeatedly enters and leaves the deadzone, as described in Chapter 6 of [16]. ■

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 above bounds 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 the theorems above. We have

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

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 bounded explicitly in terms of the placement of centers in Ξ_N . Recall that the power function \mathcal{P}_N [14], [17] of the subspace H_N in the RKHS $H(\Omega)$ is given by $\mathcal{P}_N(x) \triangleq \sqrt{\mathfrak{K}(x, x) - \mathfrak{K}_N(x, x)}$, with \mathfrak{K}_N the reproducing kernel of the subspace H_N . It is easy to show that $\mathfrak{K}_N(x, y) \triangleq \mathfrak{K}_{\Xi_N}(x)^T \mathbb{K}_N^{-1} \mathfrak{K}_{\Xi_N}(y)$ where $\mathfrak{K}_{\Xi}(x) = \{\mathfrak{K}_{\xi_1}(x), \dots, \mathfrak{K}_{\xi_N}(x)\}^T \in \mathbb{R}^{N \times 1}$ is the column vector of N basis functions defined in terms of the set of centers $\Xi_N \subset \Omega$. It is well-known that the power function is useful for generating pointwise bounds on the projection error, and we have $|E_x(I - \Pi_N)v| \leq \mathcal{P}_N(x) \|(I - \Pi_N)v\|_{H(\Omega)}$ for all $x \in \Omega$ and $v \in H(\Omega)$. [14], [17] This identity holds for any native space whatsoever.

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 [14]):

Suppose that v satisfies the regularity condition $v = Lu$ where $L : L^2(\Omega) \rightarrow H(\Omega)$ is the bounded, linear, compact operator $(Lu)(x) \triangleq \int_{\Omega} \mathfrak{K}(x, y)u(y)dy$. Then there is a constant $C > 0$ such that we have the error bound $\|(I - \Pi_N)v\|_{H(\Omega)} \leq C \sup_{\xi \in \Omega} |\mathcal{P}_N(\xi)| \|L^{-1}v\|_{L^2(\Omega)}$.

Proof: This proof is based on that of Theorem 11.23 of [14], and for completeness we summarize the simple modifications here. First note that

$$\begin{aligned}(w, Lu)_{H(\Omega)} &= (w, \int_{\Omega} \mathfrak{K}(\cdot, y)u(y)dy)_{H(\Omega)} \\ &= \int_{\Omega} (w, \mathfrak{K}_y)_{H(\Omega)} u(y)dy \\ &= \int_{\Omega} w(y)u(y)dy = (w, u)_{L^2(\Omega)}.\end{aligned}$$

Now we can write

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

But we also have

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

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

Since the centers Ξ_N , kernel \mathfrak{K} , and power function \mathcal{P}_N is known, the above theorem 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 above is often emphasized by relating the power function to the fill distance $h_{\Xi_N, \Omega}$ of the centers Ξ_N in the set Ω . Define $h_{\Xi_N, \Omega} \triangleq \sup_{y \in \Omega} \min_{\xi_i \in \Xi_N} \|y - \xi_i\|_2$. References such as [14] and [17] summarize upper bounds for the power function $\mathcal{P}_N(x)$ for a variety of common kernels. These bounds have the form $\mathcal{P}_N(x) \lesssim \sqrt{\mathcal{N}(h_{\Xi_N, \Omega})}$ for a function $\mathcal{N} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$. In this inequality, the function \mathcal{N} may depend on the dimension n , the set Ω , and the kernel \mathfrak{K} , but not on the samples Ξ_N . The tables in [14], [17] describe many families of kernels that can provide alternatives for the ultimate bounds on the performance of the critic methods in this paper. Here, in the following lemma, we just summarize three common examples.

Lemma 1: Suppose that v is contained in the uncertainty class $\mathcal{C}_{L,R} \triangleq \{g = Lu \in H(\Omega) \mid \|u\|_{L^2(\Omega)} \leq R\} \subseteq H(\Omega)$, and that the hypotheses of Theorem 3 holds with the minimum size deadzone $\bar{\epsilon} \approx \epsilon_{N,\max}$. If we choose the Sobolev-Matérn kernel for a high enough smoothness k in Table I, then there is a time $T > 0$ such that for all $t \geq T$ we have the ultimate performance bound

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

If we select the Wendland compactly supported kernel $\eta_{n,k}$ the right-hand side above is $O(h_{\Xi_N, \Omega}^{k+1/2})$, while if we choose the exponential kernel the right-hand side is $O(\sqrt{e^{-\alpha|h_{\Xi_N, \Omega}|/h_{\Xi_N, \Omega}}})$ 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 carry out numerical validation studies for the system of the form (1) studied in [6], with

$$\begin{aligned}f(x) &= \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2 \left(1 - (\cos(2x_1) + 2)^2\right) \end{bmatrix} \\ g(x) &= \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}\end{aligned}$$

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 $V^*(x) = 0.5x_1^2 + x_2^2$, which generates the optimal feedback controller $u^*(x) = -(\cos(2x_1) + 2)x_2$. The numerical validation studies in [6] 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 $\|V^* - \hat{v}_N(t, \cdot)\|_{L^\infty(\Omega)}$ for two Matérn kernels and an exponential kernel. Since we have $\|E_x\| \leq \bar{\mathfrak{K}}$, it follows that $|E_x \tilde{v}_N(t, \cdot)| \leq \bar{\mathfrak{K}} \|\tilde{v}_N(t, \cdot)\|_{H(\Omega)}$ and $\|\tilde{v}_N(t, \cdot)\|_{L^\infty(\Omega)} \leq \bar{\mathfrak{K}} \|\tilde{v}_N(t, \cdot)\|_{H(\Omega)}$, so Lemma 1 implies the corresponding convergence in the norm of $L^\infty(\Omega)$.

Type	Kernel	$\mathcal{N}(\cdot)$
Gaussian	$e^{-\alpha r^2}, \alpha > 0$	$e^{-a \log h /h}$
Multiquadric	$(-1)^{\lceil \beta \rceil} (c^2 + r^2)^\beta, \beta > 0, \beta \notin \mathbb{N}$	$e^{-a/h}$
Inverse multiquadric	$(c^2 + r^2)^\beta, \beta < 0$	$e^{-a/h}$
Compactly supported functions	$\eta_{d,k}$	h^{2k+1}
Sobolev-Matérn	$\frac{2\pi^{d/2}}{\Gamma(k)} K_{k-d/2}(r/2) r^{k-d/2}, d, k \in \mathbb{N}$	h^{2k-d}

TABLE I

FUNCTIONS $\mathcal{N}(\cdot)$ THAT BOUND THE POWER FUNCTION AS $\mathcal{P}_N(x) \lesssim \sqrt{\mathcal{N}(h_{\Xi_N, \Omega})}$ FOR $x \in \Omega$. IN THIS TABLE $h \triangleq h_{\Xi_N, \Omega}$. THE CONSTANT a IS A GENERIC CONSTANT THAT VARIES WITH THE KERNEL CHOICE. THE SYMBOL K_ν IS THE MODIFIED BESSEL FUNCTION OF THE 3^{rd} KIND OF ORDER ν , WHILE THE SYMBOL $\eta_{d,k}$ DENOTES THE COMPACTLY SUPPORTED WENDLAND KERNEL OF DIMENSION d AND SMOOTHNESS k GIVEN IN [14].

The ultimate approximate value function $\hat{v}_N(t, \cdot)$ closely matches the analytical expression for the optimal value function V^* as the dimension $N \rightarrow \infty$. Note that Theorem 1 only guarantees that $\hat{v}_N(t, \cdot)$ converges to V_μ , not V^* , so in fact 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.

In fact, 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 [14] 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^* . We can proceed exactly as in the proof of Theorem 3 of [18] in the case at hand to conclude that

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

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 $\|\hat{v}_N(t, \cdot)\|_{H(\Omega)}$.

VI. CONCLUSIONS

This paper has formulated the online critic for the estimation of the optimal value function in terms of evolution laws for a wide variety of RKHSs. Essentially, the approach in the paper can be viewed as a strategy to lift 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 $\|V^* - \hat{v}(t, \cdot)\|_{H(\Omega)}$ when the critic evolves in some appropriate RKHS $H(\Omega)$. A wide variety of results are derived in the paper that hold over a large family of RKHSs. Performance bounds are derived that are explicit in the number N of basis functions, the choice of the kernel \mathfrak{K} that defines the RKHS, and the placement of centers Ξ_N in a subset Ω of interest that define

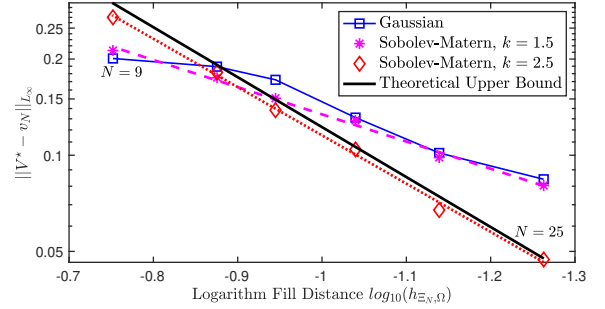


Fig. 1. The $L^\infty(\Omega)$ error norm of the online critic estimates of the value function V^* using the deadzone rule described in Lemma 1. The steady-state value function approximations using Sobolev-Matérn kernels of smoothness $k = 1.5, 2.5$ and exponential kernels are plotted above. Note that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

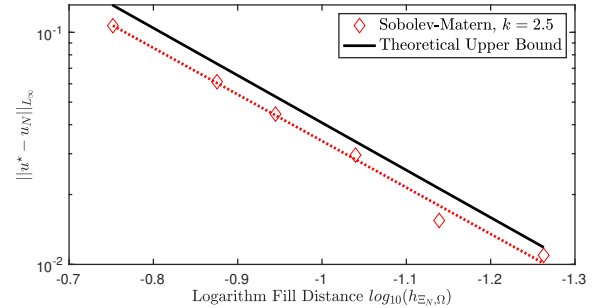


Fig. 2. The $L^\infty(\Omega)$ error norm of the online critic estimates of the control input u^* with Sobolev-Matérn kernels of smoothness $k = 2.5$. Note that the rates of convergence for the Sobolev-Matérn kernels closely follow the theoretical bounds derived in Theorem 3 of [18].

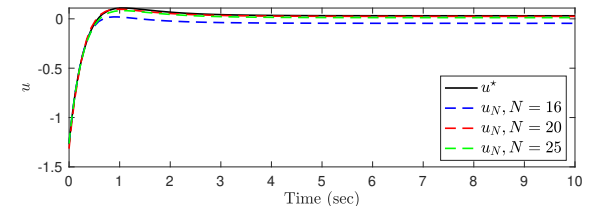


Fig. 3. Feedback control u by learned \hat{W} with Sobolev-Matérn kernels of smoothness $k = 2.5$.

the scattered bases. The final form of the performance bounds is given in terms of the power function of the scattered

basis, which 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] R. W. Bea, "Successive galerkin approximation algorithms for nonlinear optimal and robust control," Int. J. Control, vol. 71, no. 5, pp. 717–743, 1998.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] F. L. Lewis and D. Liu, Reinforcement learning and approximate dynamic programming for feedback control. John Wiley & Sons, 2013.
- [9] R. Kamalapurkar, P. Walters, J. Rosenfeld, and W. Dixon, Reinforcement learning for optimal feedback control. Springer, 2018.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] A. Boulund, 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.
- [14] H. Wendland, Scattered data approximation. Cambridge university press, 2004, vol. 17.
- [15] 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.
- [16] J. A. Farrell and M. M. Polycarpou, Adaptive approximation based control: unifying neural, fuzzy and traditional adaptive approximation approaches. John Wiley & Sons, 2006, vol. 48.
- [17] R. Schaback, "Error estimates and condition numbers for radial basis function interpolation," Adv. Comput. Math., vol. 3, pp. 251–264, 1994.
- [18] A. Boulund, 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.