2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

An Adaptive Critic Learning Approach for Nonlinear Optimal Control Subject to Excitation and Weight Constraints

Anthony Siming Chen, *Member*, *IEEE* and Guido Herrmann, *Senior Member*, *IEEE*

***Abstract*— We propose a novel adaptive critic learning algorithm for a continuous-time nonlinear system subject to excitation and weight constraints. The algorithm is able to learn the optimal control in real-time under only finite excitation without requiring the *a priori* knowledge of the system model, i.e. the Hamilton-Jacobi-Bellman (HJB) equation is approximately solved online by the adaptive critic learning of a nonlinear Q-function. The main contribution of this paper is twofold: First, we present an optimisation-based approach to the derivation of a weight-error-driven adaptive law that guarantees exponential convergence of the critic weight. Such formulation enables a new $\mathcal{P}$-projection operator to enhance the convergence property, i.e. the weight estimate always stays in a bounded convex set that contains the true weight. Second, we adopt a new measure to build the information matrix that stores its richness over incoming data such that the standard persistent excitation (PE) condition is relaxed to a finite excitation (FE) condition. In this way, the convergence of the critic weight is guaranteed without persistently injecting exploration noise. We show that the method is model-free and can achieve semi-global stability. A numerical example demonstrates the effectiveness of the theoretical result.**

***Index Terms*— adaptive optimal control, adaptive critic, projection operator, finite excitation, Q-learning**

I. INTRODUCTION

Optimal control [1], [2] is commonly used to tackle a minimisation/maximisation problem by offline solving the *Hamilton-Jacobi-Bellman* (HJB) equation or, in a linear quadratic case, the *Riccati* equation. The nonlinear HJB equation is often difficult or impossible to solve due to the requirement of complete knowledge of the system. On the other hand, adaptive control [3] online learns to control unknown systems using data measured in real time along the system trajectories. Recent ideas of incorporating reinforcement learning principles into feedback control have prompted extensive research on adaptive optimal control. This is also referred to as approximate/adaptive dynamic programming (ADP) [4], [5], [6]. Vrabie *et al.* [7] proposed an integral reinforcement learning (IRL) approach in continuous time which generates a large family of algorithms but most of them require at least partial knowledge of the system dynamics. Vamvoudakis [8] developed a model-free IRL algorithm by leveraging the idea of Q-learning. The algorithm employed two neural networks in a critic/actor configuration and was restricted to the LQR case. Chen and Herrmann [9] derived a model-free adaptive optimal controller for general unknown nonlinear systems. Instead of the gradient algorithm with normalisation in [8], the adaptive law in [9] used a sliding mode technique which guarantees the convergence towards the optimal solution in finite time.

In this paper, we improve a model-free adaptive optimal control algorithm via Q-learning for a nonlinear system [9] to consider various practically motivated constraints. Hence, the controller can learn the optimal control solution in real time under only finite excitation without requiring the system model; moreover, adaptation weights strictly remain within some convex set, converging to the optimal weights in that convex set. The main contribution of this paper is twofold: First, we present an optimisation-based approach to the derivation of a recently presented weight-error-driven adaptive law that guarantees exponential convergence of the critic weight. The work [9], [10], [11] used that adaptive algorithm but here we show an alternative derivation of such method based on an integral discounted cost function inspired by a least-squares approach. This formulation then enables the use of a new $\mathcal{P}$-projection operator to enhance the convergence property, i.e., the weight estimate always stays in a bounded convex set in which the true weight lies in. Second, we provide semi-global [12] closed-loop stability analysis under a finite excitation (FE) condition via a Lyapunov theorem. This relaxes the widely-used assumption on the persistent excitation (PE) condition in adaptive control [3] and also in ADP [6], [7], [8]. Unlike [7] which requires *a priori* knowledge of input gain $g(x)$, the proposed algorithm is model-free for the benefits of Q-learning.

Anthony Siming Chen and Guido Herrmann are with the Control and Robotics Group at the Department of Electrical and Electronic Engineering at the University of Manchester, Manchester, M13 9PL, United Kingdom siming.chen@manchester.ac.uk guido.herrmann@manchester.ac.uk

II. PROBLEM FORMULATION

A. Nonlinear System and Cost Function

Given the continuous-time nonlinear affine time-invariant system

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

where $x(t) \in \mathbb{R}^n$ is the measurable state vector, $u(t) \in \mathbb{R}^m$ is the control policy or input vector, and $f(x(t)) \in \mathbb{R}^n$, $g(x(t)) \in \mathbb{R}^{n\times m}$ are the system drift and the input gain functions, respectively.

Assumption 1. *$f(x)+g(x)u$ is Lipschitz continuous in x on a compact set $\Omega \in \mathbb{R}^n$ given a continuous control $u \in U$. The pair (f, g) is stabilisable.*

We define an infinite-horizon integral cost function $V^u(x) \in C^1(\Omega)$ given by

$$V^u(x(t)) := \int_{t_0}^{\infty} r(x(t), u(t))\mathrm{d}t, \quad t \geq t_0 \tag{2}$$

with $r(x, u)$ being the utility function, which is equivalent to the *reward function* in reinforcement learning. We select $r = S(x(t)) + u^\mathsf{T}(t)Ru(t)$ with $0 \preceq S(x(t)) \in \mathbb{R}$ and $0 \prec R = R^\mathsf{T} \in \mathbb{R}^{m\times m}$, where R is set to be a constant diagonal matrix in this paper for the sake of simplicity.

B. Optimal Control and its HJB Solution

The optimal control problem is to minimise the cost function (2) by finding an optimal stabilising control $u^*(t)$. Thus, the optimal cost $V^*(x)$ is defined by

$$V^*(x(t)) := \min_u \int_{t_0}^{\infty} r(x(t), u(t))\mathrm{d}t, \quad t \geq t_0 \tag{3}$$

A general solution to the optimal control problem can be derived using *Pontryagin's minimum principle* or *dynamic programming* as a partial differential equation in terms of

the state gradient of the optimal cost [1][2]. To show this, we first define the Hamiltonian

$$\mathcal{H}(x,u,\nabla V_x^u) := r(x,u) + (\nabla V_x^u)^\mathsf{T}(f(x)+g(x)u) \quad (4)$$

with the (vertical) gradient vector $\nabla V_x^u = \partial V^u/\partial x \in \mathbb{R}^n$. The optimal value function $V^*(x)$ in (3) satisfies the *Hamilton-Jacobi-Bellman* (HJB) equation

$$0 = \nabla V_t^* + \min_u \ \mathcal{H}(x,u,\nabla V_x^*) \quad (5)$$

where $\nabla V_t^* = \partial V^*/\partial t = 0$ as the optimal cost is not an explicit function of time. The optimal control u^* can be found by setting $\partial\mathcal{H}(x,u,\nabla V_x^*)/\partial u = 0$ so that

$$u^* = -\frac{1}{2}R^{-1}g(x)^\mathsf{T}\nabla V_x^* \quad (6)$$

Inserting the optimal control (6) into (5) gives the HJB equation in terms of ∇V_x^* as

$$0 = S(x) + (\nabla V_x^*)^\mathsf{T} f(x) - \frac{1}{4}(\nabla V_x^*)^\mathsf{T} g(x)R^{-1}g(x)^\mathsf{T}\nabla V_x^* \quad (7)$$

In general, solving the HJB equation analytically is difficult due to the nonlinearity and the lack of system knowledge.

III. CONTINUOUS-TIME Q-LEARNING WITHOUT PERSISTENT EXCITATION

This section provides a *model-free* "critic-actor" reinforcement learning method which enables the agent to learn the optimal control *online* without the requirement of the PE condition.

A. Nonlinear Q-Function and Q-Learning Bellman Equation

The idea of the continuous-time Q-learning method for nonlinear systems was initiated in [9], we summarise the basics and the key lemma as follows in preparation for the later design of adaptive critic and control. We define a Q-function by adding the right-hand side of the HJB equation (5) onto the optimal value (3) as

$$\begin{aligned} Q(x,u) :&= V^*(x) + \nabla V_t^* + \mathcal{H}(x,u,\nabla V_x^*) \\ &= \underbrace{V^*(x) + S(x) + (\nabla V_x^*)^\mathsf{T} f(x)}_{F_{xx}(x)} + \\ &\quad \underbrace{(\nabla V_x^*)^\mathsf{T} g(x)u}_{F_{xu}(x,u)} + \underbrace{u^\mathsf{T} R u}_{F_{uu}(u)} \end{aligned} \quad (8)$$

where $F_{xx}(x)$, $F_{xu}(x,u)$, and $F_{uu}(u)$ are the lumped terms. Such parameterisation in terms of x and u allows the proper approximation via neural networks.

Lemma 1. [9] *The Q-function defined in* (8) *is positive definite with the optimisation scheme* $Q^*(x,u^*) = \min_u Q(x,u)$. *The optimal Q-function* $Q^*(x,u^*)$ *has the same optimal cost* $V^*(x)$ (3) *as for the cost function* $V^u(x)$ (2)*, i.e.* $Q^*(x,u^*) = V^*(x)$ *when applying the optimal control* u^*.

Proof. Refer to [9] (*Lemma 3*) for the proof. □

In critic-actor structure, the assumption on smoothness of the cost function $V^u(x)$ is desired for its approximation in Sobolev norm, i.e., approximation of the value function and its gradient. Hence, we now make the following assumption in terms of $Q(x,u)$.

Assumption 2. *Given state* $x \in \Omega$*, an admissible [13] control* $u(x) \in \Psi(\Omega)$ *and* $V^u(x) \in H^{1,2}(\Omega)$ *with* $H^{1,2}(\Omega)$ *denoting a Sobolev space on* Ω*, the Q-function defined as* (8) *is smooth, i.e.,* $Q(x,u) \in C^1(\Omega_Q)$ *with* $\Omega_Q \subset \Omega \times \Psi(\Omega)$ *being a compact set.*

This assumption allows the use of a neural network approximation via the Weierstrass higher-order approximation theorem [13]. To this end, we approximate the Q-function (8) via an adaptive critic with a neural-network-type structure given by

$$Q(x,u) = W^\mathsf{T}\Phi(x,u) + \varepsilon(x,u) \quad (9)$$

where $\Phi(x,u) \in C^1(\Omega_Q)$ is the activation function vector with the number N of neurons in the hidden layer; $W \in \mathbb{R}^N$ is the ideal constant weight vector; $\varepsilon(x,u)$ denotes the neural network approximation error; and $W^\mathsf{T}\Phi(x,u)$ can be explicitly expressed according to the three components $F_{xx}(x)$, $F_{xu}(x,u)$, and $F_{uu}(u)$ in (8) as

$$W^\mathsf{T}\Phi(x,u) = \begin{bmatrix} W_{xx}^\mathsf{T} \mid W_{xu}^\mathsf{T} \mid W_{uu}^\mathsf{T} \end{bmatrix} \begin{bmatrix} \Phi_{xx}(x) \\ vec(\Phi_{xu}(x)\otimes u) \\ \Phi_{uu}(u) \end{bmatrix} \quad (10)$$

where $\otimes$ denotes the Kronecker product and $vec(\cdot)$ is the vectorisation function which stacks the columns of a matrix together; $\Phi(x,u)$ are selected for $\Phi_{xx} \in \mathbb{R}^{N_{xx}}$, $\Phi_{xu} \in \mathbb{R}^{N_{xu}}$ and $\Phi_{uu} \in \mathbb{R}^m$ to provide a complete linearly independent basis such that, under *Assumption 2*, $Q(x,u)$ is uniformly bounded with $N = N_{xx} + m(N_{xu}+1)$ on the compact set Ω_Q. Recall the Weierstrass higher-order approximation theorem [13], the approximation error $\varepsilon(x,u)$ is bounded for a fixed N and as the number of neurons $N \to \infty$, we have $\varepsilon(x,u) \to 0$.

Now we derive the Bellman equation in terms of the Q-function to update the critic. By Bellman's principle of optimality, we have the following optimality equation [7]

$$V^*(x(t-T)) = \int_{t-T}^{t} r(x(\tau),u(\tau))\mathrm{d}\tau + V^*(x(t)) \quad (11)$$

The result from *Lemma 1* showed that $Q^*(x,u^*) = V^*(x)$. Considering the instrumental Lemma [2, p.441], we can rewrite (11) in terms of $Q^*(x,u^*)$ as a Q-learning Bellman equation:

$$\begin{aligned} &\overbrace{-\int_{t-T}^{t} r(x,u)\mathrm{d}\tau}^{-\rho(x,u)} = Q^*(x(t),u^*(t)) \\ &\qquad - Q^*(x(t-T),u^*(t-T)) + \psi \\ &= \underbrace{W^\mathsf{T}\Phi(x(t),u^*(t)) - W^\mathsf{T}\Phi(x(t-T),u^*(t-T))}_{W^\mathsf{T}\Delta\Phi(x,u^*)} \\ &\quad + \underbrace{\varepsilon(x(t),u^*(t)) - \varepsilon(x(t-T),u^*(t-T)) + \psi}_{\varepsilon_B} \end{aligned} \quad (12)$$

with ψ being a residual error as

$$\begin{aligned} \psi = &-\int_{t-T}^{t} (u(\tau)-u^*(\tau))^\mathsf{T} R(u(\tau)-u^*(\tau))d\tau \\ &+ (u(t-T)-u^*(t-T))^\mathsf{T} R(u(t-T)-u^*(t-T)) \\ &- (u(t)-u^*(t))^\mathsf{T} R(u(t)-u^*(t)) \end{aligned} \quad (13)$$

the integral reinforcement $\rho(x,u) \in \mathbb{R}$, the difference $\Delta\Phi(t) = \Phi(x(t),u^*(t)) - \Phi(x(t-T),u^*(t-T))$, $\Delta\Phi(t) \in \mathbb{R}^N$ and the Bellman error $\varepsilon_B = \Delta\varepsilon + \psi$, $\varepsilon_B \in \mathbb{R}$ with $\Delta\varepsilon = \varepsilon(x(t),u^*(t)) - \varepsilon(x(t-T),u^*(t-T))$ being bounded for a bounded ε. The Bellman equation (12) forms the basis for adaptive critic design.

We write the critic neural network as

$$\hat{Q}(x,u) = \hat{W}^{\mathsf{T}}\Phi(x,u) \tag{14}$$

where $\hat{Q}(x,u)$ and $\hat{W}$ denote the Q-function $Q(x,u)$ and the estimate of the weight W, respectively.

B. Adaptive Critic Design

This section shows a novel design of the adaptive critic which is updated from the weight estimation error. We present an alternative way to derive the adaptation scheme that was used similarly in [9], [10] and also later in [11]. Instead of using an auxiliary filter operation, we apply the gradient descent method on a discounted integral cost function. This formulation helps us to further enhance and understand the convergence properties of our algorithm by using projection combined with the gradient.

The gradient descent method that has been widely used in adaptive control often considers an instantaneous cost in terms of the output error, e.g., $e(t) = \hat{W}^{\mathsf{T}}(t)\Delta\Phi(t) - W^{\mathsf{T}}\Delta\Phi(t)$ as in [8]. The weight $\hat{W}(t)$ is to be chosen at each time t to minimise $J_e = \frac{1}{2}e^2$, which is an instantaneous cost function and is convex over the space of e. However, the convexity of J_e guarantees the existence of a single global minimum $e = 0$ but does not address the convergence of the weight $\hat{W}$. In order to ensure the convergence of $\hat{W}$ to W, instead of instantaneous J_e, we consider an integral cost function $J \in \mathbb{R}$ for

$$\bar{e}(t,\tau) = \hat{W}^{\mathsf{T}}(t)\Delta\Phi(\tau) - W^{\mathsf{T}}\Delta\Phi(\tau) \tag{15}$$

on all past data that are exponentially discounted as

$$J = \frac{1}{2}\int_{t_0}^{t} \exp(-\ell(t-\tau))\bar{e}^2(t,\tau)\mathrm{d}\tau \tag{16}$$

where the design constant $\ell > 0$ acts as a forgetting factor, i.e., the effect of history data at time $\tau < t$ is discarded exponentially as time t increases. The cost J (16) penalises all the past errors $\bar{e}$ from time t_0 to t. This is equivalent to the cost for recursive least-squares algorithms with a forgetting factor. The method, however, for developing estimate $\hat{W}$ for W is different. It is clear that the cost J (16) is convex over the space of $\hat{W}$ for each time t. We can apply the gradient descent method for minimising J with respect to $\hat{W}$ as

$$\dot{\hat{W}} = -\Gamma\nabla J_{\hat{W}} \tag{17}$$

where $\Gamma \in \mathbb{R}^{N\times N}$ is the adaptive gain or learning rate with $\Gamma \succ 0$ and $\nabla J_{\hat{W}} = \partial J/\partial\hat{W} \in \mathbb{R}^N$ is the gradient vector of J. By inspection of the Bellman equation (12), we have $-W^{\mathsf{T}}\Delta\Phi = \rho + \varepsilon_B$. Then, expanding the gradient $\nabla J_{\hat{W}}$ at time t yields

$$\begin{aligned}
\nabla J_{\hat{W}} &= \int_{t_0}^{t} \exp(-\ell(t-\tau))(\hat{W}^{\mathsf{T}}\Delta\Phi - W^{\mathsf{T}}\Delta\Phi)\Delta\Phi^{\mathsf{T}}\mathrm{d}\tau \\
&= \int_{t_0}^{t} \exp(-\ell(t-\tau))(\Delta\Phi\Delta\Phi^{\mathsf{T}}\hat{W} + \Delta\Phi(\rho+\varepsilon_B))\mathrm{d}\tau \\
&= \underbrace{\int_{t_0}^{t} \exp(-\ell(t-\tau))\Delta\Phi\Delta\Phi^{\mathsf{T}}\mathrm{d}\tau}_{\mathcal{P}(t)}\hat{W}(t) + \\
&\underbrace{\int_{t_0}^{t} \exp(-\ell(t-\tau))\Delta\Phi\rho\mathrm{d}\tau}_{\mathcal{Q}(t)} + \underbrace{\int_{t_0}^{t} \exp(-\ell(t-\tau))\Delta\Phi\varepsilon_B\mathrm{d}\tau}_{\Lambda(t)}
\end{aligned} \tag{18}$$

Hence, an adaptation law can be implemented as

$$\dot{\hat{W}} = -\Gamma(\mathcal{P}\hat{W} + \mathcal{Q}) \tag{19}$$

with the *information matrix* $\mathcal{P} \in \mathbb{R}^{N\times N}$ and the *reinforcement matrix* $\mathcal{Q} \in \mathbb{R}^N$ written as

$$\dot{\mathcal{P}} = -\ell\mathcal{P} + \Delta\Phi\Delta\Phi^{\mathsf{T}}, \qquad \mathcal{P}(t_0) = 0 \tag{20a}$$

$$\dot{\mathcal{Q}} = -\ell\mathcal{Q} + \Delta\Phi\rho, \qquad \mathcal{Q}(t_0) = 0 \tag{20b}$$

Note that the Bellman error ε_B as well as Λ are to be ignored for implementation. We shall analyse their effect on the adaptation law robustness as follows.

Let $M = \mathcal{P}\hat{W} + \mathcal{Q}$, $M \in \mathbb{R}^N$, then the adaptation law becomes $\dot{\hat{W}} = -\Gamma M$. The vector M contains the explicit information of the weight estimation error $\tilde{W} = \hat{W} - W$ for $\Lambda = 0$. Such adaptation law differs from the common adaptive control where an output error e is often used. The use of the weight estimation error instead of output error will guarantee the convergence of the weights given properly excited regressor signals. From (18) and $-\rho = W^{\mathsf{T}}\Delta\Phi + \varepsilon_B$, we express M as

$$M = \mathcal{P}\hat{W} + \mathcal{Q} = \mathcal{P}\hat{W} - (\mathcal{P}W + \Lambda) = \mathcal{P}\tilde{W} - \Lambda \tag{21}$$

where $\tilde{W} = \hat{W} - W$ is the weight estimation error and $\Lambda \in \mathbb{R}^N$ as defined in (18) is a bounded variable for bounded state x and control u. Note that $M = \mathcal{P}\tilde{W}$ if $\varepsilon_B = 0$.

Remark 1. *This adaptation scheme* (19) *with* (20) *holds better convergence properties than the commonly-used gradient descent method in adaptive optimal control, e.g., [8][14]. Note that the vector M represents the gradient of the integral cost function J* (16) *when $\varepsilon_B = 0$. It is clear from* (21) *that M explicitly contains the information weight estimation error $\tilde{W}$. Hence, the adaptation law* (19)*, i.e., $\dot{\hat{W}} = -\Gamma M$, can update the weight estimate $\hat{W}$ based on weight estimation error $\tilde{W}$ instead of using output error e. We will show that this adaptation scheme can guarantee the critic weight convergence under a more relaxed condition on the regressor signal.*

Remark 2. *The forgetting factor ℓ in the discounted integral cost* (16) *is vital to such adaptation scheme. If $\ell = 0$,* (20) *implies that $\mathcal{P}(t) \geq 0$ at any time t, and therefore, the information matrix $\mathcal{P}(t)$ may grow without bound. We call this information wind-up problem.*

C. A $\mathcal{P}$-Projection Approach for Weight Adaptation

Although the algorithm we propose is model-free (see [9]), we may still have some *a priori* knowledge as to where the optimal weight W is located in $\mathbb{R}^N$. For instance, the knowledge may come in terms of upper or lower bounds for the elements of W or a well-defined subset in $\mathbb{R}^N$. One would like to leverage such *a priori* knowledge to constrain the search of weight estimate $\hat{W}$ or to keep $\hat{W}$ within some possibly "safe" bounds. Adding such constraints may also improve the convergence properties or reduce transients that may occur when $\hat{W}(t_0)$ is initialised to be far away from W. We investigate the use of projection to address this problem. The following facts are provided in relation to the construction of projection.

Definition 1. (*Convexity*) [15] A set $S \subset \mathbb{R}^n$ is a *convex set* if $\lambda x + (1-\lambda)y \in S$ for all $x \in S$, $y \in S$, and $0 \leq \lambda \leq 1$. Likewise, a function $\mathcal{F}(x) : \mathbb{R}^N \to \mathbb{R}$ is a *convex function* if $\mathcal{F}(\lambda x + (1-\lambda)y) \leq \lambda\mathcal{F}(x) + (1-\lambda)\mathcal{F}(y)$ for all $x \in S$, $y \in S$, and $0 \leq \lambda \leq 1$.

Definition 2. (*Coercive Function*) [15] A function $\mathcal{F}(x) : \mathbb{R}^N \to \mathbb{R}$ is *coercive* if $\forall\{x_k\}$, $k \in \mathbb{N}$ with $||x_k|| \to \infty$ such that $\lim_{k\to\infty} \mathcal{F}(x_k) = \infty$.

We first consider S_a, the convex subsets S_0 and S_1 within $\mathbb{R}^N$ for the weight estimate $\hat{W}$ given by

$$S_0 := \{\hat{W} \in \mathbb{R}^N |\ \mathcal{F}(\hat{W}) \leq 0\} \tag{22}$$

$$S_1 := \{\hat{W} \in \mathbb{R}^N |\ \mathcal{F}(\hat{W}) \leq 1\} \tag{23}$$

$$S_a = S_1 \backslash S_0 = \{\hat{W} \in \mathbb{R}^N |\ 0 < \mathcal{F}(\hat{W}) \leq 1\} \tag{24}$$

where $\mathcal{F}(\hat{W}) : \mathbb{R}^N \to \mathbb{R}$ is a smooth function of weight estimate $\hat{W}$ properly chosen to be convex and coercive such that S_0 is nonempty. It is clear from (22) and (23) that $S_0 \subset S_1$.

Assumption 3. *The true critic weight W lies in the convex set S_0 defined as per* (22)*, i.e., $W \in S_0 \subset S_1$.*

Lemma 2. *For a coercive convex function $\mathcal{F}(\hat{W}) : \mathbb{R}^N \to \mathbb{R}$ and a positive constant $\delta > 0$, any nonempty subset $S_\delta = \{\hat{W} \in \mathbb{R}^N |\ \mathcal{F}(\hat{W}) \leq \delta\}$ is convex and bounded.*

Proof. See proof in [15]. □

The projection operator was first introduced in early [16] and a detailed analysis can be found in [17][18]. Taking that work as inspiration, we incorporate the adaptive gain Γ and define a $\mathcal{P}$-projection operator for the adaptation law so that we can also interpret the new adaptation law in terms of a gradient descent approach.

Definition 3. (*$\mathcal{P}$-projection Operator*) The $\mathcal{P}$-projection operator for two vectors $\hat{W}, M \in \mathbb{R}^N$ and a smooth function $\mathcal{F}(\hat{W})$ is defined as

$$Proj_{\mathcal{P}}(\hat{W}, M, \mathcal{F}) = \Gamma \begin{cases} -M + \varrho \dfrac{\nabla\mathcal{F}\nabla\mathcal{F}^\mathsf{T}\mathcal{P}^{-1}}{\nabla\mathcal{F}^\mathsf{T}\Gamma\nabla\mathcal{F}} M\mathcal{F}, \\ \qquad \text{if } \mathcal{F} > 0 \wedge -M^\mathsf{T}\mathcal{P}^{-1}\nabla\mathcal{F} > 0 \\ -M - \varrho \dfrac{\nabla\mathcal{F}\nabla\mathcal{F}^\mathsf{T}\mathcal{P}^{-1}}{\nabla\mathcal{F}^\mathsf{T}\Gamma\nabla\mathcal{F}} M\mathcal{F}, \\ \qquad \text{if } \mathcal{F} > 0 \wedge -M^\mathsf{T}\mathcal{P}^{-1}\nabla\mathcal{F} \leq 0 \\ -M, \qquad \textit{otherwise} \end{cases} \tag{25}$$

for $\mathcal{F}(W) < 0$ and $0 \prec \Gamma = \Gamma^\mathsf{T} \in \mathbb{R}^{N\times N}$ and $\nabla\mathcal{F} = \partial\mathcal{F}/\partial\hat{W} \in \mathbb{R}^N$. Here, $\varrho > 0$ is a scalar that is chosen large enough so that the following inequality holds:

$$M^T\Gamma\nabla\mathcal{F}^\mathsf{T}\nabla\mathcal{F}\Gamma M \leq \varrho M^T\mathcal{P}^{-1}\nabla\mathcal{F}^\mathsf{T}\nabla\mathcal{F}\mathcal{P}^{-1}M$$

It is easily shown that a sufficient choice is $\varrho \geq \lambda_{max}(\Gamma)\lambda_{max}(\mathcal{P})$ for the largest eigenvalues of Γ and $\mathcal{P}$.

Lemma 3. *Given Assumption 3, i.e., $W \in S_0$, and the projection operator as per* (25)*,*

$$\tilde{W}^\mathsf{T}\Gamma^{-1}(Proj_{\mathcal{P}}(\hat{W}, M, \mathcal{F}) + \Gamma M) \leq 0 \tag{26}$$

Proof. *Case 1:* If $\mathcal{F} > 0$ for a smooth convex function $\mathcal{F}$ and $\tilde{W} = \hat{W} - W$, according to [17] (Lemma 4), we can obtain $\tilde{W}^\mathsf{T}\nabla\mathcal{F} \geq 0$. Then we have

$$\begin{aligned} &\tilde{W}^\mathsf{T}\Gamma^{-1}(Proj_{\mathcal{P}}(\hat{W}, M, \mathcal{F}) + \Gamma M) \\ &= \tilde{W}^\mathsf{T}\Gamma^{-1}(\Gamma \frac{\nabla\mathcal{F}\nabla\mathcal{F}^\mathsf{T}}{\nabla\mathcal{F}^\mathsf{T}\Gamma\nabla\mathcal{F}}\mathcal{P}^{-1}M\mathcal{F}) \\ &= \frac{\underbrace{\tilde{W}^\mathsf{T}\nabla\mathcal{F}}_{\geq 0}\underbrace{\nabla\mathcal{F}^\mathsf{T}\mathcal{P}^{-1}M}_{<0}}{\underbrace{\nabla\mathcal{F}^\mathsf{T}\Gamma\nabla\mathcal{F}}_{>0}} \underbrace{\mathcal{F}}_{>0} \leq 0 \end{aligned} \tag{27}$$

Case 2: If $\mathcal{F} > 0$ and $-M^\mathsf{T}\mathcal{P}^{-1}\nabla\mathcal{F} \leq 0$, we also have

$$\tilde{W}^\mathsf{T}\Gamma^{-1}(Proj_{\mathcal{P}}(\hat{W}, M, \mathcal{F}) + \Gamma M) \leq 0 \tag{28}$$

Case 3: Otherwise, i.e., if $\mathcal{F} \leq 0$ or $-M^\mathsf{T}\Gamma\nabla\mathcal{F} \leq 0$, we have $Proj_{\mathcal{P}}(\hat{W}, M, \mathcal{F}) = -\Gamma M$ so that

$$\tilde{W}^\mathsf{T}\Gamma^{-1}(Proj_{\mathcal{P}}(\hat{W}, M, \mathcal{F}) + \Gamma M) = 0 \tag{29}$$

Hence, $\tilde{W}^\mathsf{T}\Gamma^{-1}(Proj_{\Gamma}(\hat{W}, M, \mathcal{F}) + \Gamma M) \leq 0$ holds. □

Then, we can write the adaptation law for updating $\hat{W}$ using the projection operator (25) as

$$\dot{\hat{W}} = Proj_{\Gamma}(\hat{W}, M, \mathcal{F}) \tag{30}$$

Note that the proof of Lemma 3 and the discussion of (21) show that for $\varepsilon_B = 0$, the first case $\mathcal{F} > 0 \wedge -M^\mathsf{T}\mathcal{P}^{-1}\nabla\mathcal{F} > 0$ does not exist since then $-M^\mathsf{T}\mathcal{P}^{-1}\nabla\mathcal{F} = -\tilde{W}^T\nabla\mathcal{F} \leq 0$ for $\mathcal{F} > 0$.

To be specific, the projection operator (25) does not change $-\Gamma M$ if $\hat{W} \in S_0$. If $\hat{W} \in S_a$, it substracts a perpendicular vector that is normal to the boundary $\{\hat{W} \in \mathbb{R}^N |\ \mathcal{F}(\hat{W}) = \delta\}$ with $0 \leq \delta \leq 1$ so that we get a smooth transformation from the original vector $-\Gamma M$ for $\delta = 0$ to a vector for $\delta = 1$ [17], which is either, at worst, a tangent to S_1 or pointing inside S_1.

A popular choice of coercive convex $\mathcal{F}(\hat{W})$ used for projection operator in adaptive control is given by [17]

$$\mathcal{F}(\hat{W}) = \frac{||\hat{W}||^2 - w_{max}^2}{2\epsilon w_{max} + \epsilon^2} \tag{31}$$

where ϵ and w_{max} are positive constants with the true weight $||W|| \leq w_{max}$. It is obvious that $\mathcal{F}(\hat{W}) = 0$ when $||\hat{W}|| = w_{max}$ and $\mathcal{F}(\hat{W}) = 1$ when $||\hat{W}|| = w_{max} + \epsilon$. Hence, we have that S_0, S_1, and also S_a are convex and bounded due to *Lemma 2*. For the choice of $\mathcal{F}$ as (31), the size of margin S_a is designed by a constant ϵ.

We now define the following matrix

$$\mathcal{N} = (\mathcal{P}^{-1} + \varrho H(\mathcal{F})\mathcal{F}\frac{\mathcal{P}^{-1}\nabla\mathcal{F}\nabla\mathcal{F}^\mathsf{T}\mathcal{P}^{-1}}{\nabla\mathcal{F}^\mathsf{T}\Gamma\nabla\mathcal{F}})$$

for the Heaviside step function $H(*)$. The following function follows

$$J_P(t) = \frac{1}{2}\tilde{W}^\mathsf{T}\mathcal{P}\mathcal{N}\mathcal{P}\tilde{W} \tag{32}$$

It is easy to see that for $\varepsilon_B = 0$, (the fist case in (25) does not exist), so

$$\nabla J_P = \Gamma^{-1} Proj_{\mathcal{P}}(\hat{W}, M, \mathcal{F}).$$

Moreover, for $\mathcal{F} < 0$, it follows

$$J_P(t) = J(t). \tag{33}$$

Remark 3. The $\mathcal{P}$-projection operator (25) used in this paper is different from the Γ-projection in [3] (see equation B.27 in [3]) or [17], [18]. The distinctions are i) in (25), an additional $\mathcal{F}$ is multiplied onto the perpendicular vector. This leads to a more relaxed constraint on the weight estimate $\hat{W}$, i.e. $\hat{W}$ can move into a margin subset S_a, whereas in [3] $\hat{W}$ will be restricted to S_0. ii) the first case $\mathcal{F} > 0 \wedge -M^\mathsf{T}\mathcal{P}^{-1}\nabla\mathcal{F} > 0$ is identified considering Λ or the Bellman error $\varepsilon_B \neq 0$, whereas for the Γ-projection in [17], [18], this case becomes void since $\tilde{W}^T\nabla\mathcal{F} \geq 0$ for $\mathcal{F} > 0$. Moreover, we provide an interpretation of the new adaptive law with projection in terms of the function $J_P(t)$.

D. Relaxing Persistent Excitation: Finite Excitation

Another important component of this paper is with regards to the excitation condition for regressor $\Delta\Phi(t)$ which are provided as follows.

Definition 4. (*PE condition*) The signal $\Delta\Phi(t)$ is said to be *persistently excited (PE)* if there exist $\mathcal{T} > 0$ and $\sigma_1 > 0$ such that

$$\int_t^{t+\mathcal{T}} \Delta\Phi(\tau)\Delta\Phi(\tau)^\mathsf{T} \mathrm{d}\tau \geq \sigma_1 I, \quad \forall t \geq t_0 \tag{34}$$

Definition 5. (*FE condition*) The signal $\Delta\Phi(t)$ is said to be *finitely excited (FE)* over a finite time interval $[t_s, t_e]$ if there exist $t_e > t_s \geq t_0$ and $\sigma_2 > 0$ such that

$$\int_{t_s}^{t_e} \Delta\Phi(\tau)\Delta\Phi(\tau)^\mathsf{T} \mathrm{d}\tau \geq \sigma_2 I \tag{35}$$

The levels of excitation of both conditions are indicated by some constants σ_1 and σ_2, respectively. One can differentiate the two conditions by the time interval of excitation. Note that the PE condition in *Definition 4* regards a property over a moving window for all $t+\mathcal{T} > t$, whereas the FE condition in *Definition 5* pertains to a single interval $[t_s, t_e]$. As shown in [10] (*Lemma 2.2*) and [11], if the regressor $\Delta\Phi(t)$ is persistently excited, the information matrix $\mathcal{P}(t)$ should be full rank, i.e., $rank(\mathcal{P}(t)) = N$. Otherwise, $\mathcal{P}(t)$ is only positive semi-definite, i.e., $\mathcal{P}(t) \succeq 0$. The rank deficiency of the information matrix is the fundamental cause of being in breach of the PE condition for weight convergence. By inspection of (18) in terms of the solution $\mathcal{P}(t)$, it is clear that, if the regressor signal $\Delta\Phi(t)$ is excited sufficiently (not persistently), the information matrix $\mathcal{P}(t)$ will be full rank over time. In other words, $\mathcal{P}(t)$ will have full rank after a certain time interval $[t_0, t_a]$ for $t_a > t_0$ unless the regressor $\Delta\Phi(t)$ remains on an affine hyperplane for the entire time $[t_0, t]$. Thus, we can relax the PE condition for the weight convergence by the following assumption.

Assumption 4. *$\exists t_e > t_s \geq t_0$ such that the regressor $\Delta\Phi(t)$ is finitely excited (FE) over a time interval $[t_s, t_e]$.*

Since an FE condition rather than PE condition is imposed in *Assumption 4*, we use only the information in a specific time interval $[t_0, t_r]$, $t_r > t_0$ instead of the entire time $[t_0, t]$ for the weight update in order to avoid the degeneration of the information matrix due the forgetting design $\exp(-\ell(t-\tau))$. In principle, the choice of t_r should retain the *richness* of the information matrix $\mathcal{P}(t)$ for consistent learning performance. One way to determine t_r is

$$t_r(t) = \max\{\arg \sup_{\tau \in [t_0, t]} \lambda_{min}(\mathcal{P}(\tau))\} \tag{36}$$

where $\lambda_{min}(\mathcal{P})$ denotes the minimum eigenvalue of the matrix $\mathcal{P}$. The inner $\sup(\cdot)$ finds a sequence of moments in $[t_0, t]$ that have the largest minimum eigenvalue. The outer $\max(\cdot)$ selects the largest time t_r in that sequence to leverage the finite excitation of the regressor $\Delta\Phi(t)$.

We design a new information matrix $\mathcal{P}_r(t)$ and reinforcement matrix $\mathcal{Q}_r(t)$ such that

$$\mathcal{P}_r(t) = \mathcal{P}(t_r(t)) \tag{37a}$$
$$\mathcal{Q}_r(t) = \mathcal{Q}(t_r(t)) \tag{37b}$$

Lemma 4. *Under the FE condition in Assumption 4, the selection of t_r as* (36) *guarantees for $\mathcal{P}_r(t)$* (37) *that*

$$\lambda_{min}(\mathcal{P}_r(t)) \geq 0, \quad \forall t \geq t_0 \tag{38}$$

$$\lambda_{min}(\mathcal{P}_r(t)) \geq \lambda_{min}(\mathcal{P}(t_e)) \geq \sigma_2 > 0, \quad \forall t \geq t_e \tag{39}$$

Proof. For entire time $t \geq t_0$, it is obvious that $\mathcal{P}_r(t)$ is always positive semi-definite. Hence, $\lambda_{min}(\mathcal{P}_r(t)) \geq 0$ for $\forall t \geq t_e$. This proves (38). The regressor $\Delta\Phi(t)$ is finitely excited over a time interval $[t_s, t_e]$ as stated in *Assumption 4*. For $t \geq t_e$, it is obvious that $\mathcal{P}(t_e) \succeq \sigma_2 I \succ 0$ as *Definition 5* and then $\mathcal{P}_r(t) \succeq \sigma_2 I \succ 0$, i.e., $\mathcal{P}_r(t)$ is positive definite after time $t = t_e$. When $t \in [t_s, t_e]$, $\mathcal{P}_r(t)$ will be updated at least once so that it has a strictly positive minimum eigenvalue. Then $\mathcal{P}_r(t)$ will be updated only when there is an increase in $\lambda_{min}(\mathcal{P}_r(t))$. Hence, $\lambda_{min}(\mathcal{P}_r(t)) \geq \lambda_{min}(\mathcal{P}(t_e)) \geq \sigma_2 > 0$ for $\forall t \geq t_e$. This proves (39). □

Therefore, we propose a new adaptation law based on (30) using the vector $M_r(t) = \mathcal{P}_r(t)\hat{W}(t) + \mathcal{Q}_r(t)$ for the weight $\hat{W}(t)$ update

$$\dot{\hat{W}} = Proj_{\mathcal{P}_r}(\hat{W}, M_r, \mathcal{F}) \tag{40}$$

where Γ is the constant learning rate with $0 \prec \Gamma \in \mathbb{R}^{N\times N}$.

Remark 4. *Having a full rank $\mathcal{P}_r$ relaxes the assumption on the regressor $\Delta\Phi$ from PE to FE. Lemma 4 implies that the information matrix remains positive definite after finite excitation. Such benefit obviates the need to inject perturbation noise continuously to maintain persistent excitation.*

As before in (32), it is again possible to interpret the projection algorithm in terms of an optimization function for $\varepsilon_B = 0$, where now:

$$J_{P_r}(t) = \frac{1}{2}\tilde{W}^\mathsf{T}\mathcal{P}_r\mathcal{N}_r\mathcal{P}_r\tilde{W} \tag{41}$$

and

$$\mathcal{N}_r = (\mathcal{P}_r^{-1} + \varrho H(\mathcal{F})\mathcal{F}\frac{\mathcal{P}_r^{-1}\nabla\mathcal{F}\nabla\mathcal{F}^\mathsf{T}\mathcal{P}_r^{-1}}{\nabla\mathcal{F}^\mathsf{T}\Gamma\nabla\mathcal{F}}).$$

Moreover, $J_{P_r}(t) = J(t)$ for $t_r(t) = t$ and $\mathcal{F} < 1$.

E. On the Optimal Control and Critic Weight Convergence

In this section, we first leverage the parameterisation of the Q-function and its adaptive critic design to determine the optimal control. Then we synthesise the solutions to the nonlinear optimal control problem presented followed by the convergence analysis.

We reconstruct the optimal control u^* from (6) based on the parameterisation of $Q(x,u)$ (8) such that

$$u^* = -\frac{1}{2}diag(W_{uu})^{-1}W_{xu}^\mathsf{T}\Phi_{xu}(x) + \varepsilon_u \tag{42}$$

where ε_u is a bounded approximation error due to ε, $W_{xu}^\mathsf{T}\Phi_{xu}(x)$ accounts for the term $g(x)^\mathsf{T}\nabla V_x^*$, and $diag(W_{uu})$ is essentially the pre-defined R. However, we write the actor in the form of

$$u = -\frac{1}{2}diag(\hat{W}_{uu})^{-1}\hat{W}_{xu}^\mathsf{T}\Phi_{xu}(x) \tag{43}$$

for the sake of theoretical consistency. In practice, the initial value of $\hat{W}_{uu}$ can be simply chosen to be the same as the diagonal elements in R.

The main result of the paper is therefore:

Theorem 1. *The adaptive critic* (14)*, the actor* (43)*, and the adaptation law* (40) *along with Assumptions 1-4 and* (20)(25)(36)(37) *form an adaptive optimal control so that*
a) For $\hat{W}(t_0) = w_0 \in S_1$, $\hat{W}(t) \in S_1$ for all time $t \geq t_0$;
b) the state $x(t)$ and the weight estimation error $\tilde{W}(t)$ are uniformly ultimately bounded in a semi-global sense [12] for all time $t \geq t_e$;

c) the control u will enter and stay in a small compact region around its optimal solution u^* for $\forall t \geq t_e$ if there exists a neural network approximation error, i.e., if $\varepsilon \neq 0$;
d) if there is no neural network approximation error, i.e., if $\varepsilon = 0$, the weight estimation error $\tilde{W}(t)$ will exponentially converge to zero and the control u will exponentially converge towards its optimal solution u^* for $\forall t \geq t_e$.
Proof. The proof is omitted due to space limitations. □

IV. NUMERICAL EXAMPLE

Consider a continuous-time nonlinear system [9][14] with $x = [x_1\ x_2]^\intercal \in \mathbb{R}^2$, $u \in \mathbb{R}$, and

$$f(x) = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1 - (\cos(2x_1) + 2)^2) \end{bmatrix} \quad (44)$$

$$g(x) = \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix} \quad (45)$$

The infinite-horizon cost function is selected as (2) with $S(x) = x_1^2 + x_2^2$ and $R = 1$. We know the optimal control $u^* = -(\cos(2x_1) + 2)x_2$ and the optimal cost $V^* = \frac{1}{2}x_1^2 + x_2^2$. This is verified through a converse HJB approach [19].

We run the numerical simulation of the proposed adaptive optimal control via Bogacki-Shampine (ode23) solver. We also compare the results with the benchmark adaptation $\dot{\hat{W}} = -\Gamma M_r$ (19), i.e. without projection. The convergence of the adaptive critic weight without injecting a persistently excited exploration noise demonstrates the effectiveness of the proposed controller. Fig. 1 shows the comparison of the weight convergence (only 5 weights are displayed for visibility). For the projection approach, it can be found that the weights were always staying in a bounded set. This is also verified in Fig. 2 where the coercive convex function $\mathcal{F}(\hat{W})$ was always capped at 1 when using the projection operator ($w_{max} = 1.15$ and $\epsilon = 0.1$).

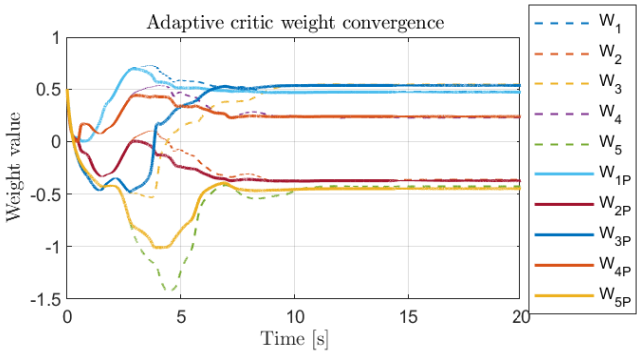

Fig. 1. Weight convergence comparison: without projection (dashed) vs with projection (solid).

V. CONCLUSIONS

The proposed adaptive critic learning approach for nonlinear optimal control is a model-free algorithm that requires only a finite period for excitation assuming the information matrix has become full rank and therefore has finite non-zero eigenvalues. If *a priori* knowledge of weight constraints is given, e.g., $W \in S_0$, convergence properties can be improved by using projection while keeping the weights within a pre-given region of the optimal weights. Simulations have shown that the temporal characteristics can improve through projection, e.g. settling time.

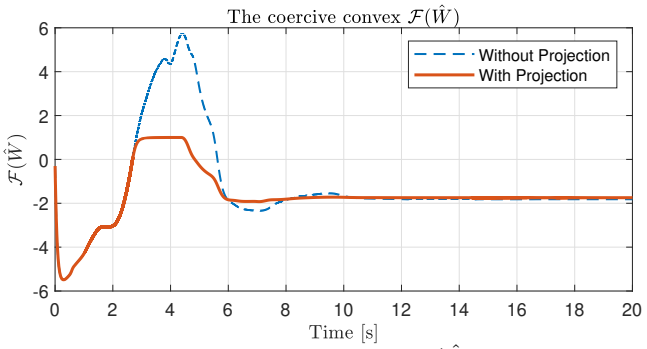

Fig. 2. The coercive convex function of weight $\mathcal{F}(\hat{W})$ chosen as eqn (31).

REFERENCES

[1] D. E. Kirk, *Optimal control theory: an introduction*. Courier Corporation, 2012.
[2] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.
[3] P. A. Ioannou and J. Sun, *Robust adaptive control*. Courier Corporation, 2012.
[4] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 11, pp. 2850–2859, 2012.
[5] Y. Jiang and Z. P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.
[6] D. Liu, X. Yang, D. Wang, and Q. Wei, "Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints," *IEEE transactions on cybernetics*, vol. 45, no. 7, pp. 1372–1385, 2015.
[7] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal adaptive control and differential games by reinforcement learning principles*. IET, 2013, vol. 2.
[8] K. G. Vamvoudakis, "Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach," *Systems & Control Letters*, vol. 100, pp. 14–20, 2017.
[9] A. S. Chen and G. Herrmann, "Adaptive optimal control via continuous-time q-learning for unknown nonlinear affine systems," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 1007–1012.
[10] J. Na, M. N. Mahyuddin, G. Herrmann, X. Ren, and P. Barber, "Robust adaptive finite-time parameter estimation and control for robotic systems," *International Journal of Robust and Nonlinear Control*, vol. 25, no. 16, pp. 3045–3071, 2015.
[11] N. Cho, H. S. Shin, Y. Kim, and A. Tsourdos, "Composite model reference adaptive control with parameter convergence under finite excitation," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 811–818, 2017.
[12] A. Teel and L. Praly, "Global stabilizability and observability imply semi-global stabilizability by output feedback," *Systems & Control Letters*, vol. 22, no. 5, pp. 313–325, 1994.
[13] 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.
[14] 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.
[15] D. P. Bertsekas, *Convex optimization theory*. Athena Scientific Belmont, 2009.
[16] J. B. Pomet, L. Praly *et al.*, "Adaptive nonlinear regulation: Estimation from the lyapunov equation," *IEEE Transactions on automatic control*, vol. 37, no. 6, pp. 729–740, 1992.
[17] E. Lavretsky and T. E. Gibson, "Projection operator in adaptive systems," *arXiv preprint arXiv:1112.4232*, 2011.
[18] J. E. Gaudio, A. M. Annaswamy, E. Lavretsky, and M. A. Bolender, "Parameter estimation in adaptive control of time-varying systems under a range of excitation conditions," *IEEE Transactions on Automatic Control*, vol. 67, no. 10, pp. 5440–5447, 2021.
[19] V. Nevistić and J. A. Primbs, *Constrained nonlinear optimal control: a converse HJB approach*. Tech. rep. CIT-CDS 96-021. California Institute of Technology, 1996.