Reachability Analysis-based Safety-Critical Control using Online Fixed-Time Reinforcement Learning

Nick-Marios T. Kokolakis NMKOKOLAKIS@GATECH.EDU
Kyriakos G. Vamvoudakis KYRIAKOS@GATECH.EDU
Wassim M. Haddad WM.HADDAD@AEROSPACE.GATECH.EDU
The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

In this paper, we address a safety-critical control problem using reachability analysis and design a reinforcement learning-based mechanism for learning online and in fixed-time the solution to the safety-critical control problem. Safety is assured by determining a set of states for which there does not exist an admissible control law generating a system trajectory reaching a set of forbidden states at a user-prescribed time instant. Specifically, we cast our safety-critical problem as a Mayer optimal feedback control problem whose solution satisfies the Hamilton-Jacobi-Bellman (HJB) equation and characterizes the set of safe states. Since the HJB equation is generally difficult to solve, we develop an online critic-only reinforcement learning-based algorithm for simultaneously learning the solution to the HJB equation and the safe set in fixed time. In particular, we introduce a non-Lipschitz experience replay-based learning law utilizing recorded and current data for updating the critic weights to learn the value function and the safe set. The non-Lipschitz property of the dynamics gives rise to fixed-time convergence, whereas the experience replay-based approach eliminates the need of satisfying the persistence of excitation condition provided that the recorded data is sufficiently rich. Simulation results illustrate the efficacy of the proposed approach.

Keywords: Adaptive learning, fixed-time stability, safety-critical control, reachability analysis, reinforcement learning.

1. Introduction

In control engineering, the term *autonomy* refers to controlled systems that can operate without involving a supervisor (Vamvoudakis and Kokolakis, 2020). Systems possessing this property are known as intelligent autonomous systems (IASs). IASs include unmanned aerial vehicles (UAVs), unmanned underwater vehicles, robotic manipulators, humanoid robots, and self-driving cars, to name but a few examples. However, in many cases, IASs have accidentally crashed, thereby testifying that IASs are *safety-critical systems*. Therefore, providing safety guarantees becomes imperative, giving rise to *safe autonomy* (Herbert, 2020; Royo, 2020). To enable safe autonomy, the control systems community can exploit the benefits of *reinforcement learning* (RL) (Sutton and Barto, 2018) to develop IASs with learning-enabled control mechanisms that run in real-time and adapt to changes in the environment while providing guarantees of performance and safety. To assure the efficient and safe operation of IASs without human intervention, it is necessary for the decision-making mechanism to generate optimal safe policies along with estimates of the safe sets in fixed time rather than in an infinite or finite time.

Related Work. The control systems community has invested considerable effort toward developing safety verification tools for safety-critical control systems. *Hamilton-Jacobi* (HJ) *reachability analysis* (Mitchell et al., 2005; Lygeros, 2004; Margellos and Lygeros, 2011; Bansal et al., 2017; Chen and Tomlin, 2018) is a *formal verification method* for assuring the optimal performance and safety of dynamical systems by computing a backward reachable set (BRS), which encompasses the set of states from which the system can potentially violate given safety specifications. The work of (Chen et al., 2018) introduces a general system decomposition method for efficiently computing BRSs. The method of "warm-start" reachability is proposed in (Herbert et al., 2019), which exhibits computational benefits. The work of (Herbert et al., 2021) merges the methods of decomposition, warm-starting, and a simple adaptive grid to accelerate the computation of safe sets. Real-world applications of HJ reachability analysis for safety verification include collision avoidance (Chen et al., 2019), vehicle platooning (Chen et al., 2015), administering anesthesia (Kaynama et al., 2012), motion planning (Chen et al., 2021; Bansal et al., 2020; Bajcsy et al., 2019), as well as other applications (Bayen et al., 2007; Huang et al., 2011; Ding et al., 2008). Nevertheless, the aforementioned approaches and the references therein typically involve offline procedures that are computationally intensive as they suffer from the curse of dimensionality.

Adaptive dynamic programming (ADP) (Lewis and Vrabie, 2009; Lewis et al., 2012b; Zhang et al., 2012; Lewis and Liu, 2013; Liu et al., 2017; Kiumarsi et al., 2017; Jiang and Jiang, 2017; Kamalapurkar et al., 2018; Jiang et al., 2020) unifies *optimal* (Lewis et al., 2012a) and *adaptive* (Ioannou and Fidan, 2006) control by constructing adaptive learning algorithms to learn the solutions to optimal control problems online by using measured data along the system trajectories while efficiently dealing with system dimensionality (Powell, 2007). ADP algorithms leverage an *actor-critic structure* composed of two function approximators. In particular, a *critic network* that evaluates the performance of a control policy and an *actor network* that computes the control policy. The vast majority of the existing adaptive learning algorithms for solving optimal control problems (Vrabie et al., 2013; Vamvoudakis and Kokolakis, 2020) converge to a near-optimal control law as long as a *persistence of excitation* (PE) (Ioannou and Fidan, 2006) condition is satisfied. Alternatively, *concurrent learning/experience replay*-based ADP algorithms (Chowdhary and Johnson, 2010; Lin, 1992) allow the learning of a solution to the optimal control problem by requiring a weaker form of a PE condition to be satisfied (Modares et al., 2014; Vamvoudakis et al., 2016; Kokolakis and Vamvoudakis, 2022a,b,c; Kokolakis et al., 2023). These algorithms are data-driven and leverage recorded and instantaneous data concurrently for the adaptation of the critic weights.

In real-world applications, IASs may experience abrupt changes in the system dynamics and operating environment, thereby necessitating the design of decision-making mechanisms featuring fast adaptability to new safety tasks while establishing optimal performance. To this end, it is necessary to bring together HJ reachability analysis, ADP, and stability analysis of dynamical systems (Haddad and Chellaboina, 2011) to develop an *online fixed-time convergent reinforcement learning*-based algorithm for learning the solution to a safety-critical control problem in fixed time. To the best of our knowledge, an ADP approach enabling the learning of a BRS online and in a fixed time is absent from the literature.

Contributions. The contributions of the present paper are fourfold. First, we address a *safety-critical control* problem using reachability analysis formulated as a Mayer optimal control problem. In particular, safety is ensured by determining a set of states for which there does not exist an admissible controller steering a system trajectory to a set of forbidden states at a given time instant.

Then, a RL-based framework is developed for learning online and in fixed-time the value function, the safe set, and the safe control policy. Subsequently, a non-Lipschitz experience replay-based adaptive learning law for updating the critic weights is introduced while ensuring fixed-time stability properties provided that the recorded data is sufficiently rich. Finally, the proposed scheme relies on the use of only a critic network, allowing the simultaneous learning of the value function, the safe set, and the safe strategy, thus yielding a less computationally expensive learning mechanism.

Structure. The rest of the paper is organized as follows. Section 2 presents the safety-critical control problem utilizing HJ reachability analysis. In Section 3, a critic-only learning mechanism is constructed for learning online and in fixed time the solution to the safety-critical control problem. Section 4 provides an illustrative numerical example. Finally, Section 5 provides conclusions and outlines future research directions.

Notation. The notation used in this paper is standard. Specifically, $\|\cdot\|_p \triangleq \left[\sum_{i=1}^{n} |x_i|^p\right]^{1/p}$, $1 \leqslant p < \infty$, denotes the Hölder p-norm of a vector. The induced 2-norm for the matrix $Q \in \mathbb{R}^{m\times n}$ is defined as $\|Q\| \triangleq \sqrt{\lambda_{\max}\left(Q^T Q\right)} = \sigma_{\max}(Q)$, with $\lambda_{\max}$ (resp., $\lambda_{\min}$) denoting the maximum (resp., minimum) eigenvalue and $\sigma_{\max}$ (resp., $\sigma_{\min}$) denoting the maximum (resp., minimum) singular value. The gradient of a scalar-valued function V with respect to a vector-valued variable x at x is defined as a row vector and is denoted by $V_x(x)$. We define the open ball $\mathcal{B}_\varepsilon\left(x_{\mathrm{e}}\right) \triangleq \{x \in \mathbb{R}^n : \|x - x_{\mathrm{e}}\| < \varepsilon\}$ centered at x_{e} with radius ε in the Euclidean norm, while the corresponding closed ball is denoted as $\mathcal{B}_\varepsilon[x_{\mathrm{e}}]$. Let $\lceil\cdot\rfloor^\eta \triangleq |\cdot|^\eta \operatorname{sign}(\cdot)$, where $|\cdot|$ and $\operatorname{sign}(\cdot)$ operate componentwise and $\eta > 0$. The distance of a point $x_0 \in \mathbb{R}^n$ to a closed set $C \subseteq \mathbb{R}^n$ in the norm $\|\cdot\|$ is defined as $\operatorname{dist}\left(x_0, C\right) \triangleq \inf_{x\in C}\left\{\|x_0 - x\|\right\}$. The notation $\mathcal{X} \times \mathcal{Y}$ denotes the Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$. Finally, $\partial\mathcal{S}$ and $\mathcal{S}^{\mathrm{c}}$ denote the boundary and the complement of the set $\mathcal{S}$, respectively.

2. Safety Verification using Hamilton-Jacobi Reachability Analysis

In this section, we address a safety-critical control problem for determining a set of states for which there does not exist an admissible controller that steers the system trajectory to the set of forbidden states at a user-prescribed time instant. In particular, we cast our safety-critical problem as an optimal control problem, thereby allowing the characterization of the set of safe states as a zero-strict superlevel set of a function satisfying the HJB equation.

2.1. Safety-Critical Control Problem Formulation

Consider the continuous-time controlled nonlinear time-invariant dynamical system given by

$$\dot{x}(t) = F(x(t), u(t)), \quad x(t_0) = x_0, \quad t \geqslant t_0, \tag{1}$$

where, for every $t \geqslant t_0$, $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in U \subset \mathbb{R}^m$ is the control input with U being a compact set, and $F : \mathbb{R}^n \times U \to \mathbb{R}^n$ is jointly Lipschitz continuous in x and u, and is such that, for every $x \in \mathbb{R}^n$, the set of velocities $\mathcal{V}(x) = \{F(x, u) : u \in U\}$ is convex.

Let $t_{\mathrm{f}} \geqslant t_0$ and let

$$\mathcal{U} \triangleq \{u : [t_0, t_{\mathrm{f}}] \to U : u(\cdot) \text{ is Lebesgue measurable}\}$$

be the set of *admissible* controls. We assume that the required properties for the existence and uniqueness of solutions to (1) are satisfied, and we write $s\left(t, t_0, x_0, u(\cdot)\right)$, $t_0 \leqslant t \leqslant t_{\mathrm{f}}$, to denote the solution to (1) with initial time t_0, initial condition x_0, and $u(\cdot) \in \mathcal{U}$.

Let $l : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function bounded from below and let $\mathcal{L} \subset \mathbb{R}^n$ be the set of *forbidden states* defined as

$$\mathcal{L} \triangleq \{x \in \mathbb{R}^n : l(x) \leqslant 0\}. \tag{2}$$

Note that $\mathcal{L}$ is a closed set defined as a zero-sublevel set of $l(\cdot)$ characterizing the states that must be avoided; $\mathcal{L}$ can, for example, encode the region occupied by an obstacle.

Before stating our control problem formulation, the following definitions introducing the concepts of a *safe state*, a *safe set*, and a *safe controller* are needed.

Definition 1 *A state $x \in \mathbb{R}^n \backslash \mathcal{L}$ is a* safe state *of the dynamical system* (1) *with respect to the set of forbidden states $\mathcal{L}$ at time t_f if there does not exist an admissible control law $u(\cdot) \in \mathcal{U}$ such that the solution $s\left(t, t_0, x_0, u(\cdot)\right),\ t_0 \leqslant t \leqslant t_\mathrm{f}$, to* (1) *satisfies* $\operatorname{dist}\left(s\left(t_\mathrm{f}, t_0, x, u(\cdot)\right), \mathcal{L}\right) = 0$. *A state $x \in \mathbb{R}^n \backslash \mathcal{L}$ is an* unsafe state *if it is not a safe state.* □

Definition 2 *A set $\mathcal{S}(\mathcal{L}, t_\mathrm{f}) \subset \mathbb{R}^n \backslash \mathcal{L}$ is a* safe set *for the dynamical system* (1) *with respect to the set of forbidden states $\mathcal{L}$ at time t_f if every state $x \in \mathcal{S}(\mathcal{L}, t_\mathrm{f})$ is a safe state. A set $\mathcal{S}(\mathcal{L}, t_\mathrm{f}) \subset \mathbb{R}^n \backslash \mathcal{L}$ is an* unsafe set *if it is not a safe set.* □

Definition 3 *An admissible controller $u(\cdot) \in \mathcal{U}$ is a* safe control law *at $x \in \mathbb{R}^n \backslash \mathcal{L}$ for the dynamical system* (1) *with respect to the set of forbidden states $\mathcal{L}$ at time t_f if the solution $s\left(t, t_0, x_0, u(\cdot)\right),\ t_0 \leqslant t \leqslant t_\mathrm{f}$, to* (1) *satisfies* $\operatorname{dist}\left(s\left(t_\mathrm{f}, t_0, x, u(\cdot)\right), \mathcal{L}\right) \neq 0$. *Furthermore, $u(\cdot) \in \mathcal{U}$ is a* safe control law *on $\mathcal{G} \subseteq \mathbb{R}^n \backslash \mathcal{L}$ if $u(\cdot) \in \mathcal{U}$ is a safe control law at every $x \in \mathcal{G}$. An admissible controller $u(\cdot) \in \mathcal{U}$ is an* unsafe control law *at $x \in \mathbb{R}^n \backslash \mathcal{L}$ if $u(\cdot)$ is not a safe control law at $x \in \mathbb{R}^n \backslash \mathcal{L}$. Furthermore, $u(\cdot) \in \mathcal{U}$ is an* unsafe control law *on $\mathcal{G} \subseteq \mathbb{R}^n \backslash \mathcal{L}$ if $u(\cdot) \in \mathcal{U}$ is not a safe control law on $\mathcal{G} \subseteq \mathbb{R}^n \backslash \mathcal{L}$.* □

It follows from Definitions 1 and 2 that in order to ensure safety with respect to the set of forbidden states $\mathcal{L}$ at a time horizon t_f, we need to determine the set of unsafe states $\mathcal{B}(\mathcal{L}, t_\mathrm{f}) \subset \mathbb{R}^n \backslash \mathcal{L}$ such that, for every $x_0 \in \mathcal{B}(\mathcal{L}, t_\mathrm{f})$, there exists an admissible control law $u(\cdot) \in \mathcal{U}$ such that the solution $s\left(t, t_0, x_0, u(\cdot)\right),\ t_0 \leqslant t \leqslant t_\mathrm{f}$, to (1) reaches the set of forbidden states $\mathcal{L}$ at the time instant $t = t_\mathrm{f}$, i.e., $s\left(\cdot, t_0, x_0, u(\cdot)\right)$ satisfies $\operatorname{dist}\left(s\left(t_\mathrm{f}, t_0, x_0, u(\cdot)\right), \mathcal{L}\right) = 0$. In other words, the safety objective amounts to determining the *backward reachable set* $\mathcal{B}(\mathcal{L}, t_\mathrm{f})$ of (1) with respect to $\mathcal{L}$ at time t_f, which is defined as follows.

Definition 4 *The* backward reachable set $\mathcal{B}(\mathcal{L}, t_\mathrm{f})$ *of the dynamical system* (1) *with respect to $\mathcal{L}$ at time t_f is defined as*

$$\mathcal{B}(\mathcal{L}, t_\mathrm{f}) \triangleq \{x \in \mathbb{R}^n :\ \textit{there exists}\ u(\cdot) \in \mathcal{U}\ \textit{such that}\ \operatorname{dist}\left(s\left(t_\mathrm{f}, t_0, x, u(\cdot)\right), \mathcal{L}\right) = 0\}.$$

□

Note that the safe set $\mathcal{S}(\mathcal{L}, t_\mathrm{f})$ is the complement of the backward reachable set $\mathcal{B}(\mathcal{L}, t_\mathrm{f})$ and is given by $\mathcal{S}(\mathcal{L}, t_\mathrm{f}) \triangleq \mathcal{B}^\mathrm{c}(\mathcal{L}, t_\mathrm{f})$. Furthermore, note that the backward reachable set $\mathcal{B}(\mathcal{L}, t_\mathrm{f})$ is an unsafe set.

In light of the above, we now state our safety-critical control problem.

Problem 1 *Consider the controlled nonlinear dynamical system* (1), *let $t_\mathrm{f} \geqslant t_0$, and let the set of forbidden states be characterized by $\mathcal{L}$. Then, determine the* safe set $\mathcal{S}(\mathcal{L}, t_\mathrm{f})$ *of* (1) *with respect to $\mathcal{L}$ at time t_f.* □

2.2. Hamilton-Jacobi Reachability Analysis

To evaluate the safety of the controlled nonlinear dynamical system (1) with respect to the set of forbidden states $\mathcal{L}$ at time t_{f}, define the cost functional

$$J\left(t_0, x_0, u(\cdot)\right) \triangleq l\left(s\left(t_{\mathrm{f}}, t_0, x_0, u(\cdot)\right)\right), \quad (t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n. \tag{3}$$

Let $u^\star : [t_0, t_{\mathrm{f}}] \times \mathbb{R}^n \to U$ be a *feedback control law* and let $x(t),\ t_0 \leqslant t \leqslant t_{\mathrm{f}}$, be a solution to (1). If $u^\star(\cdot) = u^\star(\cdot, x(\cdot))$ is Lebesgue measurable, then we call $u^\star(\cdot)$ an *admissible feedback control law*. Next, for every $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$, define $\mathcal{F}\left(t_0, x_0\right) \triangleq \{u : [t_0, t_{\mathrm{f}}] \times \mathbb{R}^n \to U : u(\cdot) = u(\cdot, x(\cdot))$ is admissible and $x(\cdot)$ is a solution to $(1)\} \subseteq \mathcal{U}$ to be the set of admissible feedback controllers.

We now cast the reachability problem as Mayer optimal feedback control problem using a dynamic programming approach, which involves the minimization

$$V(t_0, x_0) \triangleq \min_{u(\cdot) \in \mathcal{F}(t_0, x_0)} J\left(t_0, x_0, u(\cdot)\right), \quad (t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n, \tag{4}$$

subject to the dynamical system (1) and the terminal constraints captured by the closed target set $S = \{t_{\mathrm{f}}\} \times \mathbb{R}^n$. Note that S encodes a *fixed-time, free-endpoint problem* whose value function is given by $V(t_0, x_0)$ and can be viewed as the optimal cost (cost-to-go) from (t_0, x_0). Furthermore, note that we adopt the convention that $V(t_0, x_0) = \infty$, if $\mathcal{F}\left(t_0, x_0\right)$ is empty.

Next, define the Hamiltonian function

$$H\left(t, x, u, V_x^{\mathrm{T}}(t, x)\right) \triangleq V_x(t, x) F(x, u), \quad (t, x, u) \in [t_0, t_{\mathrm{f}}] \times \mathbb{R}^n \times U. \tag{5}$$

Applying the stationarity conditions to the Hamiltonian function (5), we obtain the feedback control law $u^\star(t, x)$ as a global minimizer of the Hamiltonian function since $H(t, x, u, V_x^{\mathrm{T}}(t, x))$ is convex in u for all $(t, x) \in [t_0, t_{\mathrm{f}}] \times \mathbb{R}^n$ owing to the regularity condition on the system dynamics $F(\cdot, \cdot)$. Namely,

$$u^\star(t, x) \triangleq \underset{u \in U}{\arg\min} \left[H\left(t, x, u, V_x^{\mathrm{T}}(t, x)\right)\right], \quad (t, x) \in [t_0, t_{\mathrm{f}}) \times \mathbb{R}^n. \tag{6}$$

Substituting (6) into (5) yields the HJB equation

$$V_t(t, x) + V_x(t, x) F(x, u^\star(t, x)) = 0, \quad (t, x) \in [t_0, t_{\mathrm{f}}) \times \mathbb{R}^n, \tag{7}$$

subject to the boundary condition

$$V(t_{\mathrm{f}}, x) = l(x), \quad x \in \mathbb{R}^n. \tag{8}$$

Alternatively, the HJB equation (7) can be written in compact form as

$$V_t(t, x) + \min_{u \in U} \left[H\left(t, x, u, V_x^{\mathrm{T}}(t, x)\right)\right] = 0, \quad (t, x) \in [t_0, t_{\mathrm{f}}) \times \mathbb{R}^n. \tag{9}$$

The next theorem characterizes the safe set $\mathcal{S}(\mathcal{L}, t_{\mathrm{f}})$ in terms of the value function $V(\cdot, \cdot)$.

Theorem 1 *Consider the controlled nonlinear dynamical system* (1) *with the set of forbidden states* (2) *and performance measure* (3). *For every* $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$, *let* $\mathcal{F}\left(t_0, x_0\right) \subseteq \mathcal{U}$ *be the class of admissible feedback controllers and assume that* $\mathcal{F}\left(t_0, x_0\right)$ *is nonempty. Assume that there exists a* *continuously differentiable function $V : [t_0, t_\mathrm{f}] \times \mathbb{R}^n \to \mathbb{R}$ given by (4) and an optimal feedback controller $u^\star(\cdot) \in \mathcal{F}(t_0, x_0)$ satisfying (6) such that*

$$V(t_0, x_0) = J(t_0, x_0, u^\star(\cdot)), \quad (t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n. \tag{10}$$

Then, the safe set $\mathcal{S}(\mathcal{L}, t_\mathrm{f}) \subseteq \mathbb{R}^n$ takes the form

$$\mathcal{S}(\mathcal{L}, t_\mathrm{f}) = \{x \in \mathbb{R}^n \backslash \mathcal{L} : \; V(t_0, x) > 0\} \tag{11}$$

and $u^\star(\cdot)$ is a safe controller on $\mathcal{S}(\mathcal{L}, t_\mathrm{f})$.

Proof The proof will appear in a future paper due to space limitations. ■

Remark 1 *Note that the backward reachable set $\mathcal{B}(\mathcal{L}, t_\mathrm{f})$ is given by $\mathcal{B}(\mathcal{L}, t_\mathrm{f}) = \mathcal{S}^\mathrm{c}(\mathcal{L}, t_\mathrm{f}) = \{x \in \mathbb{R}^n \backslash \mathcal{L} : \; V(t_0, x) \leqslant 0\}$ and the optimal feedback controller $u^\star(\cdot)$ is an unsafe controller on $\mathcal{B}(\mathcal{L}, t_\mathrm{f})$.* □

In light of the above, the problem of characterizing the safe set $\mathcal{S}(\mathcal{L}, t_\mathrm{f})$ amounts to solving the HJB equation (9), which is in general intractable aside from special cases. In the next section, we present *learning-based techniques* for approximating the solution of the HJB equation.

3. Fixed-Time Stable Online Learning

In this section, we build on the results of (Kokolakis and Vamvoudakis, 2022b; Kokolakis et al., 2023) to develop a *learning-based* algorithm for learning *online* and in *fixed-time* the solution of the HJB equation (9) by utilizing data gathered along the system trajectories. Specifically, we employ a critic structure, i.e., an approximator, allowing us to *simultaneously* approximate the value function (4) and the optimal controller (6), which in turn enables the approximation of the safe set (11) together with the safe control law (6).

By the Weierstrass higher-order approximation theorem (Hornik et al., 1990; Fotiadis et al., 2021), we can locally approximate the value function $V(t, x)$ and the partial derivative of $V(t, x)$ with respect to either t or x over a compact set $\mathcal{X}_\mathrm{c} \triangleq [t_0, t_\mathrm{f}] \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}^n$ with a neural network approximator as

$$V(t, x) = l(x) + W^{\star\mathrm{T}}\phi(t, x) + \varepsilon(t, x), \quad (t, x) \in \mathcal{X}_\mathrm{c}, \tag{12}$$

$$V_x^\mathrm{T}(t, x) = l_x^\mathrm{T}(x) + \phi_x^\mathrm{T}(t, x)W^\star + \varepsilon_x^\mathrm{T}(t, x), \quad (t, x) \in \mathcal{X}_\mathrm{c}, \tag{13}$$

and

$$V_t(t, x) = W^{\star\mathrm{T}}\phi_t(t, x) + \varepsilon_t(t, x), \quad (t, x) \in \mathcal{X}_\mathrm{c}, \tag{14}$$

where $W^\star \in \mathbb{R}^N$ is an ideal constant weight vector satisfying $\|W^\star\| \leqslant W_\mathrm{m}$ for some $W_\mathrm{m} > 0$, $\phi : [t_0, t_\mathrm{f}] \times \mathcal{X} \to \mathbb{R}^N$ is a time-varying activation function vector with components $\varphi_i(t, x)$, $i = 1, \ldots, N$, such that $\phi(t_\mathrm{f}, \cdot) = 0$ so that the boundary condition (8) is satisfied, N is the number of neurons in the hidden layer of the neural network, and $\varepsilon : [t_0, t_\mathrm{f}] \times \mathcal{X} \to \mathbb{R}$ is an approximation error. Note that one has to pick the basis functions $\varphi_i(t, x)$, $i = 1, \ldots, N$, properly so that they form a complete independent basis set for every $(t, x) \in [t_0, t_\mathrm{f}] \times \mathcal{X}$ (Lewis et al., 2020).

Remark 2 *Note that the critic neural network approximates the function $V(t, x) - l(x)$ over $\mathcal{X}_\mathrm{c}$, which is unknown since $V(t, x)$ is unknown.* □

Since the ideal weights $W^\star$ are unknown, we consider a critic with estimates $\hat{W} \in \mathbb{R}^N$ of the form

$$\hat{V}(t,x) \triangleq l(x) + \hat{W}^{\mathrm{T}}\phi(t,x), \quad (t,x) \in \mathcal{X}_{\mathrm{c}}, \tag{15}$$

an estimate of the safe set given by

$$\hat{\mathcal{S}}(\mathcal{L},t_{\mathrm{f}}) \triangleq \{x \in \mathbb{R}^n \backslash \mathcal{L} : \; \hat{V}(t_0,x) > 0\}, \tag{16}$$

and an approximate optimal (safe) controller of the form

$$\hat{u} \triangleq \arg\min_{u \in U}\left[H\left(t,x,u,l_x^{\mathrm{T}}(x) + \phi_x^{\mathrm{T}}(t,x)\hat{W}\right)\right], \quad (t,x) \in [t_0,t_{\mathrm{f}}) \times \mathcal{X}. \tag{17}$$

Substituting the approximate value function (15) and the approximate optimal control law (17) into (9), we obtain an estimate of the HJB equation as

$$\hat{h}\left(t,\phi_t^{\mathrm{T}}(t,x)\hat{W},x,\hat{u},l_x^{\mathrm{T}}(x) + \phi_x^{\mathrm{T}}(t,x)\hat{W}\right) \triangleq \hat{W}^{\mathrm{T}}\left(\phi_t(t,x) + \phi_x(t,x)F(x,\hat{u})\right) + l_x(x)F(x,\hat{u}), \\ (t,x) \in [t_0,t_{\mathrm{f}}) \times \mathcal{X}, \tag{18}$$

which is available for measurement, unlike the parameter error $\tilde{W} \triangleq \hat{W} - W^\star$, which is not since $W^\star$ is unknown.

Define the HJB estimation error corresponding to the data collected at the current time $t \in [t_0,t_{\mathrm{f}})$ as

$$e(t) \triangleq \; \hat{h}\Big(t,\phi_t^{\mathrm{T}}(t,x(t))\hat{W}(t),x(t),\hat{u}(t),l_x^{\mathrm{T}}(x(t)) + \phi_x^{\mathrm{T}}(t,x(t))\hat{W}(t)\Big), \quad t \in [t_0,t_{\mathrm{f}})\,, \tag{19}$$

and the HJB estimation error associated with the recorded data at the time instants $t_0 \leqslant t_1,\ldots,t_k < t < t_{\mathrm{f}}$ as

$$\begin{aligned} e\left(t_i,t\right) &\triangleq \hat{h}\Big(t_i,\phi_t^{\mathrm{T}}(t_i,x(t_i))\hat{W}(t),x\left(t_i\right),\hat{u}\left(t_i\right),l_x^{\mathrm{T}}(x(t_i)) + \phi_x^{\mathrm{T}}(t_i,x(t_i))\hat{W}(t)\Big) \\ &= \hat{W}^{\mathrm{T}}(t)\omega\left(t_i\right) + l_x(x\left(t_i\right))F(x\left(t_i\right),\hat{u}\left(t_i\right)), \end{aligned}$$

where $\omega(t) \triangleq \phi_t(t,x(t)) + \phi_x(t,x(t))F(x(t),\hat{u}(t)),\; t \in [t_0,t_{\mathrm{f}})$.

The *fixed-time convergent data-driven* learning law for updating the critic weights is given by

$$\begin{aligned} \dot{\hat{W}}(t) = &-\alpha\frac{\omega(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\left\lceil\frac{e(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\right\rfloor^{\gamma_1} - \alpha\frac{\omega(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\left\lceil\frac{e(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\right\rfloor^{\gamma_2} \\ &-\alpha\sum_{i=1}^{k}\frac{\omega(t_i)}{\omega^{\mathrm{T}}(t_i)\omega(t_i)+1}\left\lceil\frac{e\left(t_i,t\right)}{\omega^{\mathrm{T}}\left(t_i\right)\omega\left(t_i\right)+1}\right\rfloor^{\gamma_1} \\ &-\alpha\sum_{i=1}^{k}\frac{\omega(t_i)}{\omega^{\mathrm{T}}(t_i)\omega(t_i)+1}\left\lceil\frac{e\left(t_i,t\right)}{\omega^{\mathrm{T}}\left(t_i\right)\omega\left(t_i\right)+1}\right\rfloor^{\gamma_2}, \quad \hat{W}\left(0\right) = \hat{W}_0, \quad t \geqslant 0, \end{aligned} \tag{20}$$

where $\alpha > 0$ is a constant gain that dictates the learning rate.

Using (18) and (20), the parameter error dynamics are given by

$$\begin{aligned}\dot{\tilde{W}}(t) = &-\alpha\frac{\omega(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\left[\frac{\omega^{\mathrm{T}}(t)\tilde{W}(t)+\varepsilon_{\mathrm{H}}(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\right]^{\gamma_1}-\alpha\frac{\omega(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\left[\frac{\omega^{\mathrm{T}}(t)\tilde{W}(t)+\varepsilon_{\mathrm{H}}(t)}{\omega^{\mathrm{T}}(t)\omega(t)+1}\right]^{\gamma_2}\\ &-\alpha\sum_{i=1}^{k}\frac{\omega(t_i)}{\omega^{\mathrm{T}}(t_i)\omega(t_i)+1}\left[\frac{\omega^{\mathrm{T}}(t_i)\tilde{W}(t)+\varepsilon_{\mathrm{H}}(t_i)}{\omega^{\mathrm{T}}\left(t_i\right)\omega\left(t_i\right)+1}\right]^{\gamma_1}\\ &-\alpha\sum_{i=1}^{k}\frac{\omega(t_i)}{\omega^{\mathrm{T}}(t_i)\omega(t_i)+1}\left[\frac{\omega^{\mathrm{T}}(t_i)\tilde{W}(t)+\varepsilon_{\mathrm{H}}(t_i)}{\omega^{\mathrm{T}}\left(t_i\right)\omega\left(t_i\right)+1}\right]^{\gamma_2}, \quad \tilde{W}\left(0\right)=\tilde{W}_0, \quad t\geqslant 0, \end{aligned}\tag{21}$$

where $\varepsilon_{\mathrm{H}} \triangleq \hat{h}\left(t, \phi_t^{\mathrm{T}}(t,x)W^{\star}, x, \hat{u}, \phi_x^{\mathrm{T}}(t,x)W^{\star}\right)$ behaves as a disturbance stemming from the value function approximation error.

Before proceeding to our main theorem establishing the *fixed-time stability properties* of our concurrent learning law, the following definition introducing the concept of a *sufficiently rich* data set is needed.

Definition 5 *(Chowdhary and Johnson, 2010). The recorded data set* $\{\omega\left(t_i\right)\}_{i=1}^{k}$ *is* k-sufficiently rich *if the matrix* $\Omega \triangleq \left[\omega\left(t_1\right)\ldots\omega\left(t_k\right)\right]$ *has* $\operatorname{rank}(\Omega) = N$. □

It follows from Definition 5 that a recorded data set is k*-sufficiently rich* if and only if the set $\{\omega\left(t_i\right)\}_{i=1}^{k}$ contains N linearly independent vectors.

Theorem 2 *Consider the weight parameter error dynamics given by* (21)*. Define* $\bar{\omega}(\cdot) \triangleq \frac{\omega(\cdot)}{\omega^T(\cdot)\omega(\cdot)+1}$, $\bar{\Omega} \triangleq \left[\bar{\omega}\left(t_1\right)\ldots\bar{\omega}\left(t_k\right)\right]$, *and* $\bar{\varepsilon}_H(\cdot) \triangleq \frac{\varepsilon_H(\cdot)}{\omega^T(\cdot)\omega(\cdot)+1}$, *let* $\bar{\varepsilon}_{\mathrm{Hm}}, \ \bar{\omega}_m, \ c > 0$, *and* $\theta \in (0,1)$, *and assume that the recorded data set* $\{\bar{\omega}\left(t_i\right)\}_{i=1}^{k}$ *is k-sufficiently rich. Then the following statements hold.*

i) If $\varepsilon_H \equiv 0$, *then the zero solution* $\tilde{W}(t) \equiv 0$ *to* (21) *is globally fixed-time stable with a settling-time function*

$$T\left(\tilde{W}\left(0\right)\right) \leqslant \frac{1}{\sigma_{\min}^{\gamma_1+1}(\bar{\Omega})\left(2\alpha\right)^{\frac{\gamma_1+1}{2}}\left(\frac{1-\gamma_1}{2}\right)} + \frac{1}{c^{\gamma_2+1}\sigma_{\min}^{\gamma_2+1}(\bar{\Omega})\left(2\alpha\right)^{\frac{\gamma_2+1}{2}}\left(\frac{\gamma_2-1}{2}\right)}, \qquad \tilde{W}\left(0\right) \in \mathbb{R}^N.$$

ii) If $\varepsilon_H \not\equiv 0$ *and there exists* $N_0 > 0$ *such that*

$$\sup_{(t,x)\in\mathcal{X}_{\mathrm{c}}} |\varepsilon_H| < \bar{\varepsilon}_{\mathrm{Hm}}, \quad N \geqslant N_0,$$

and

$$\left|\bar{\omega}^T\tilde{W}\right| \geqslant \left|\bar{\varepsilon}_H\right|, \quad N \geqslant N_0,$$

then the compact set $\mathcal{B}_\mu[0]$, *where* $\mu \triangleq \left(\frac{(k+1)\left(\bar{\varepsilon}_{\mathrm{Hm}}^{\gamma_1}+\bar{\varepsilon}_{\mathrm{Hm}}^{\gamma_2}\right)\bar{\omega}_m}{\theta\sigma_{\min}^{\gamma_1+1}(\bar{\Omega})}\right)^{\frac{1}{\gamma_1}}$, *is globally fixed-time attractive; that is, for every initial condition* $\tilde{W}(0) \in \mathbb{R}^N$, *the solution* $\tilde{W}(t), \ t \geqslant 0$, *to* (21) *satisfies* $\operatorname{dist}\left(\tilde{W}(t), \mathcal{B}_\mu[0]\right) = 0, \ t \geqslant T_{\max}$, *where* $T_{\max} \triangleq \frac{2}{\delta(\gamma_2-1)} + \frac{\left(\sqrt{2\alpha}\right)^{1-\gamma_1}-\mu^{1-\gamma_1}}{\alpha\sigma_{\min}^{\gamma_1+1}(\bar{\Omega})(1-\theta)(1-\gamma_1)}$ *with* $\delta \triangleq 2^{1-\gamma_2}c^{\gamma_2+1}\sigma_{\min}^{\gamma_2+1}(\bar{\Omega})(2\alpha)^{\frac{\gamma_2+1}{2}} - (k+1)\bar{\omega}_m\left(\bar{\varepsilon}_{\mathrm{Hm}}^{\gamma_1}+\bar{\varepsilon}_{\mathrm{Hm}}^{\gamma_2}\right)\sqrt{2\alpha} > 0.$

Proof The proof will appear in a future paper due to space limitaitons. ■

Remark 3 *In the absence of the value function approximation error, the critic weights will converge to the optimal weights $W^\star$ in fixed-time. However, even though the settling-time function is unknown, it is uniformly upper bounded by a constant number that depends on the parameter α. Thus, the larger the learning rate α is, the faster the convergence of the parameter error to the origin will be.* □

Remark 4 *In the presence of the value function approximation error, the solution $\tilde{W}(t)$, $t \geqslant 0$, to (21), for every initial condition $\tilde{W}(0) \in \mathbb{R}^N$, reaches the compact set $\mathcal{B}_\mu\left[0\right]$ in fixed-time, that is, at most in time $T_{\max}$, and remains therein for all future time. Note that $T\left(\tilde{W}(0)\right) = 0$ if and only if $\tilde{W}(0) \in \mathcal{B}_\mu\left[0\right]$. Finally, one can reduce the parameter error along with the settling time by choosing the parameters γ_1, γ_2, and k properly since they determine the size of μ, that is, the size of the ball $\mathcal{B}_\mu\left[0\right]$, as well as the upper bound of the settling-time $T_{\max}$.* □

Remark 5 *The fixed-time convergence of the critic weights gives rise to a fixed-time estimate of the safe set $\mathcal{S}(\mathcal{L}, t_{\mathrm{f}})$ and the safe control $u^\star(t,x)$ given by (16) and (17), respectively.* □

4. Illustrative Numerical Example

In this section, we provide an illustrative numerical example to demonstrate the proposed safety-critical control and learning framework. Specifically, consider a UAV flying at a constant altitude and airspeed $v > 0$ whose kinematics is captured by a Dubins vehicle given by (Valavanis and Vachtsevanos, 2015)

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{\psi}(t) \end{bmatrix} = \begin{bmatrix} v\cos\psi(t) \\ v\sin\psi(t) \\ u(t) \end{bmatrix}, \quad \begin{bmatrix} x_1(0) \\ x_2(0) \\ \psi(0) \end{bmatrix} = \begin{bmatrix} x_{10} \\ x_{20} \\ \psi_0 \end{bmatrix}, \quad t \geqslant 0, \tag{22}$$

where, for every $t \geqslant 0$, $\left[x_1(t),\ x_2(t)\right]^{\mathrm{T}} \in \mathbb{R}^2$ and $\psi(t) \in \mathbb{R}$ denote the planar position and the heading angle of the UAV, respectively, and the guidance law is the commanded heading rate $u(t) \in [-u_{\max},\ u_{\max}]$ with $u_{\max}$ being a positive constant. The dynamical system (22) can be cast in the form of (1) with $n = 3$, $m = 1$, $x = \left[x_1,\ x_2,\ \psi\right]^{\mathrm{T}}$, $F(x,u) = \left[v\cos\psi,\ v\sin\psi,\ u\right]^{\mathrm{T}}$, and $U = [-u_{\max},\ u_{\max}]$.

Let $r > 0$ and let $l(x) = \sqrt{x_1^2 + x_2^2 + \psi^2} - r$. Then, the set of forbidden states is given by

$$\mathcal{L} = \left\{x \in \mathbb{R}^3 : \sqrt{x_1^2 + x_2^2 + \psi^2} - r \leqslant 0\right\},$$

which represents a sphere of radius r in the state space $\mathbb{R}^3$.

Let $v = 5\ \mathrm{m/s}$, $u_{\max} = 1\ \mathrm{rad/s}$, $r = 1$, $t_{\mathrm{f}} = 1\ \mathrm{s}$, $\alpha = 400$, $\gamma_1 = 0.9$, and $\gamma_2 = 1.1$. For our critic, the initial weights are randomly initialized within the interval $[0, 2]$ and the basis functions are selected as $\phi(t,x) = \tanh\left(t_{\mathrm{f}} - t\right)\left[x_1^2,\ x_2^2,\ \psi^2 - 2\right]^{\mathrm{T}}$. To enable the collection of sufficiently rich data along the closed-loop system trajectories, we inject a dithering excitation to the control input (17) for every $t \in [0,\ 0.1]$.

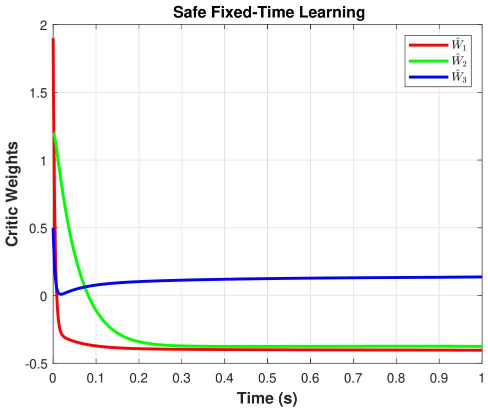

Figure 1: The time evolution of the critic weights $\hat{W}(t),\ t \geqslant 0$. Note that learning is attained in a fixed-time of $t = 0.3$ sec.

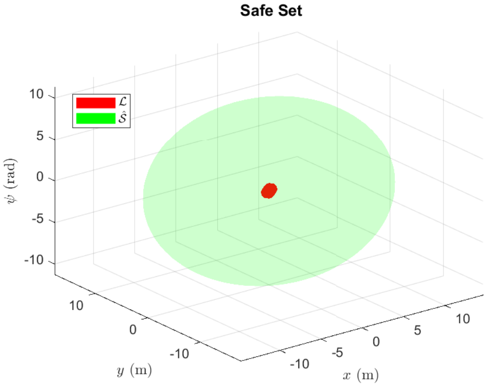

Figure 2: Visualization of the safe set estimate $\hat{\mathcal{S}}(\mathcal{L}, t_\mathrm{f})$ along with the set of forbidden states $\mathcal{L}$. Note that $\hat{\mathcal{S}}(\mathcal{L}, t_\mathrm{f})$ encloses $\mathcal{L}$.

Fig. 1 shows the fixed-time convergence of the critic weights to $[-0.4037,\ -0.3758,\ 0.1373]^\mathrm{T}$, which implies the fixed-time estimate of the safe set $\mathcal{S}(\mathcal{L}, t_\mathrm{f})$ given by

$$\hat{\mathcal{S}}(\mathcal{L}, t_\mathrm{f}) \triangleq \{x \in \mathbb{R}^3 \backslash \mathcal{L} :\ 0.693x_1^2 + 0.714x_2^2 + 1.105\psi^2 - 1.21 > 0\}.$$

Fig. 2 shows the approximate safe set $\hat{\mathcal{S}}(\mathcal{L}, t_\mathrm{f})$ and the set of forbidden states $\mathcal{L}$. Note that $\mathcal{L}$ is enclosed in $\hat{\mathcal{S}}(\mathcal{L}, t_\mathrm{f})$. Furthermore, a video demonstration of the approximate safe set boundary $\partial\hat{\mathcal{S}}(\mathcal{L}, t_\mathrm{f})$ for every t_f in $[0.5,\ 1.5]$ can be found at https://tinyurl.com/5n7ed3sm, whereby we observe that $\hat{\mathcal{S}}(\mathcal{L}, t_{\mathrm{f}_2}) \subseteq \hat{\mathcal{S}}(\mathcal{L}, t_{\mathrm{f}_1})$ with $0.5 \leqslant t_{\mathrm{f}_1} \leqslant t_{\mathrm{f}_2} \leqslant 1.5$. Finally, as expected, the simulations verify that the learning of the safe set is attained in a fixed time of $t = 0.3$ sec $< t_\mathrm{f}$, dictating its potential for a real-time safety-critical application.

5. Conclusion

In this paper, we developed a fixed-time convergent reinforcement learning-based architecture for addressing a safety-critical control problem for nonlinear systems utilizing reachability analysis. Safety is ensured by synthesizing a set of states for which there does not exist an admissible controller generating a system trajectory reaching a set of forbidden states at a user-prescribed time instant. In particular, we showed that the safe set is a zero-strict superlevel set of a function satisfying the HJB equation. In view of the intractability of the latter, we developed a critic-only RL-based algorithm for learning in fixed-time the safe set alongside the optimal safe control policy. Under the assumption of sufficiently rich data, which is an easier condition to satisfy as compared to the traditional PE condition, we developed a non-Lipschitz data-driven learning law for updating the critic weights while establishing fixed-time stability via Lyapunov analysis. We also showed that the proposed learning mechanism is composed only of a critic, and hence, exhibiting a lower complexity than other architectures proposed in the literature whose structure additionally requires an actor. In future research, we will explore discrete-time extensions of this framework.

Acknowledgments

This work was supported in part by ARO grant No. W911NF-$19-1-0270$, ONR Minerva grant No. $N00014-18-1-2160$, NSF grant Nos. CAREER CPS-1851588, CPS-2227185, and CPS-2038589, AFOSR grant No. $FA9550-20-1-0038$, and Onassis Foundation - Scholarship ID: F ZR $025-1/2021-2022$.

References

Andrea Bajcsy, Somil Bansal, Eli Bronstein, Varun Tolani, and Claire J. Tomlin. An efficient reachability-based framework for provably safe autonomous navigation in unknown environments. In *IEEE Conference on Decision and Control*, pages 1758–1765, 2019. doi: 10.1109/CDC40024.2019.9030133.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton-Jacobi reachability: A brief overview and recent advances. In *IEEE Conference on Decision and Control*, pages 2242–2253, 2017.

Somil Bansal, Andrea Bajcsy, Ellis Ratner, Anca D. Dragan, and Claire J. Tomlin. A hamilton-jacobi reachability-based framework for predicting and analyzing human motion for safe planning. In *IEEE International Conference on Robotics and Automation*, pages 7149–7155, 2020.

Alexandre M Bayen, Ian M Mitchell, Meeko MK Oishi, and Claire J Tomlin. Aircraft autolander safety analysis through optimal control-based reach set computation. *Journal of Guidance, Control, and Dynamics*, 30(1):68–77, 2007.

Mo Chen and Claire J Tomlin. Hamilton–jacobi reachability: Some recent theoretical advances and applications in unmanned airspace management. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):333–358, 2018.

Mo Chen, Qie Hu, Casey Mackin, Jaime F. Fisac, and Claire J. Tomlin. Safe platooning of unmanned aerial vehicles via reachability. In *IEEE Conference on Decision and Control*, pages 4695–4701, 2015.

Mo Chen, Sylvia L. Herbert, Mahesh S. Vashishtha, Somil Bansal, and Claire J. Tomlin. Decomposition of reachable sets and tubes for a class of nonlinear systems. *IEEE Transactions on Automatic Control*, 63(11):3675–3688, 2018.

Mo Chen, Somil Bansal, Jaime F. Fisac, and Claire J. Tomlin. Robust sequential trajectory planning under disturbances and adversarial intruder. *IEEE Transactions on Control Systems Technology*, 27(4):1566–1582, 2019.

Mo Chen, Sylvia L. Herbert, Haimin Hu, Ye Pu, Jaime Fernández Fisac, Somil Bansal, SooJean Han, and Claire J. Tomlin. Fastrack: A modular framework for real-time motion planning and guaranteed safe tracking. *IEEE Transactions on Automatic Control*, 66(12):5861–5876, 2021.

Girish Chowdhary and Eric Johnson. Concurrent learning for convergence in adaptive control without persistency of excitation. In *Proc. IEEE Conference on Decision and Control*, pages 3674–3679, 2010.

Jerry Ding, Jonathan Sprinkle, S Shankar Sastry, and Claire J Tomlin. Reachability calculations for automated aerial refueling. In *IEEE Conference on Decision and Control*, pages 3706–3712, 2008.

Filippos Fotiadis, Christos K. Verginis, Kyriakos G. Vamvoudakis, and Ufuk Topcu. Assured learning-based optimal control subject to timed temporal logic constraints. In *Proc. IEEE Conference on Decision and Control*, pages 750–756, 2021.

Wassim M Haddad and VijaySekhar Chellaboina. *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach*. Princeton University Press, Princeton, NJ, 2011.

Sylvia Herbert, Jason J Choi, Suvansh Sanjeev, Marsalis Gibson, Koushil Sreenath, and Claire J Tomlin. Scalable learning of safety guarantees for autonomous systems using hamilton-jacobi reachability. In *IEEE International Conference on Robotics and Automation*, pages 5914–5920, 2021.

Sylvia L. Herbert, Somil Bansal, Shromona Ghosh, and Claire J. Tomlin. Reachability-based safety guarantees using efficient initializations. In *IEEE Conference on Decision and Control*, pages 4810–4816, 2019.

Sylvia Lee Herbert. *Safe Real-World Autonomy in Uncertain and Unstructured Environments*. University of California, Berkeley, 2020.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990.

Haomiao Huang, Jerry Ding, Wei Zhang, and Claire J Tomlin. A differential game approach to planning in adversarial scenarios: A case study on capture-the-flag. In *IEEE International Conference on Robotics and Automation*, pages 1451–1456, 2011.

Petros Ioannou and Baris Fidan. *Adaptive Control Tutorial*, volume 11. SIAM, Philadelphia, PA, 2006.

Yu Jiang and Zhong-Ping Jiang. *Robust Adaptive Dynamic Programming*. John Wiley & Sons, Hoboken, NJ, 2017.

Zhong-Ping Jiang, Tao Bian, and Weinan Gao. *Learning-Based Control: A Tutorial and Some Recent Results*, volume 8. Now Publishers, Boston - Delft, 2020.

Rushikesh Kamalapurkar, Patrick Walters, Joel Rosenfeld, and Warren Dixon. *Reinforcement Learning for Optimal Feedback Control*. Springer, Cham, Switzerland, 2018.

Shahab Kaynama, John Maidens, Meeko Oishi, Ian M Mitchell, and Guy A Dumont. Computing the viability kernel using maximal reachable sets. In *Proceedings of ACM International Conference on Hybrid Systems: Computation and Control*, pages 55–64, 2012.

Bahare Kiumarsi, Kyriakos G Vamvoudakis, Hamidreza Modares, and Frank L Lewis. Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6):2042–2062, 2017.

Nick-Marios T. Kokolakis and Kyriakos G. Vamvoudakis. Safe finite-time reinforcement learning for pursuit-evasion games. In *Proc. IEEE Conference on Decision and Control*, pages 4022–4027, 2022a.

Nick-Marios T. Kokolakis and Kyriakos G. Vamvoudakis. Online learning-based optimal control of nonlinear systems with finite-time convergence guarantees. In *Proc. American Control Conference*, pages 812–817, 2022b.

Nick-Marios T. Kokolakis, Kyriakos G. Vamvoudakis, and Wassim M. Haddad. Fixed-time reinforcement learning for optimal feedback control. In *Proc. ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, 2023.

Nikolaos-Marios T. Kokolakis and Kyriakos G. Vamvoudakis. Safety-aware pursuit-evasion games in unknown environments using gaussian processes and finite-time convergent reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022c.

Frank L Lewis and Derong Liu. *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, volume 17. John Wiley & Sons, Hoboken, NJ, 2013.

Frank L Lewis and Draguna Vrabie. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 9(3):32–50, 2009.

Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. *Optimal Control*. John Wiley & Sons, Hoboken, NJ, 2012a.

Frank L Lewis, Draguna Vrabie, and Kyriakos G Vamvoudakis. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems*, 32(6):76–105, 2012b.

Frank L Lewis, Jagannathan Sarangapani, and Aydin Yesildirak. *Neural Network Control of Robot Manipulators and Non-Linear Systems*. CRC press, Philadelphia , PA, 2020.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3-4):293–321, 1992.

Derong Liu, Qinglai Wei, Ding Wang, Xiong Yang, and Hongliang Li. *Adaptive Dynamic Programming with Applications in Optimal Control*. Springer, Cham, Switzerland, 2017.

John Lygeros. On reachability and minimum cost optimal control. *Automatica*, 40(6):917–927, 2004.

Kostas Margellos and John Lygeros. Hamilton–Jacobi formulation for reach–avoid differential games. *IEEE Transactions on Automatic Control*, 56(8):1849–1861, 2011.

I.M. Mitchell, A.M. Bayen, and C.J. Tomlin. A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 50(7): 947–957, 2005.

Hamidreza Modares, Frank L Lewis, and Mohammad-Bagher Naghibi-Sistani. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica*, 50(1):193–202, 2014.

Warren B Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, volume 703. John Wiley & Sons, Hoboken, NJ, 2007.

Vicenc Rubies Royo. *Assured Autonomy for Safety-Critical and Learning-Enabled Systems*. University of California, Berkeley, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, Cambridge, MA, 2018.

Kimon P Valavanis and George J Vachtsevanos. *Handbook of unmanned aerial vehicles*, volume 1. Springer, 2015.

Kyriakos G. Vamvoudakis and Nick-Marios T. Kokolakis. *Synchronous Reinforcement Learning-Based Control for Cognitive Autonomy*, volume 8. Now Publishers, Boston - Delft, 2020.

Kyriakos G Vamvoudakis, Marcio Fantini Miranda, and João P Hespanha. Asymptotically stable adaptive–optimal control algorithm with saturating actuators and relaxed persistence of excitation. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2386–2398, 2016.

Draguna Vrabie, Kyriakos G Vamvoudakis, and Frank L Lewis. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*, volume 2. IET, London, UK, 2013.

Huaguang Zhang, Derong Liu, Yanhong Luo, and Ding Wang. *Adaptive Dynamic Programming for Control: Algorithms and Stability*. Springer Science & Business Media, London, UK, 2012.