2019 American Control Conference (ACC)
Philadelphia, PA, USA, July 10-12, 2019

Safety-Aware Reinforcement Learning Framework with an Actor-Critic-Barrier Structure

Yongliang Yang[1], Kyriakos G. Vamvoudakis,[2], Hamidreza Modares[3], Wei He[1], Yixin Yin[1], Donald C.Wunsch[4]

***Abstract*— This paper considers the control problem with constraints on full-state and control input simultaneously. First, a novel barrier function based system transformation approach is developed to guarantee the full-state constraints. To deal with the input saturation, the hyperbolic-type penalty function is imposed on the control input. The actor-critic based reinforcement learning technique is combined with the barrier transformation to learn the optimal control policy that considers both the full-state constraints and input saturations. To illustrate the efficacy, a numeric simulation is implemented in the end.**

***Index Terms*— reinforcement learning, full-state constraints, input saturation, safe control**

I. INTRODUCTION

Due to the constraints in real-world applications, control problem with constraints has attracted numerous attention recently. In addition to stability and control performance, the constraints on state and/or control is critical for safety purpose. The constrained control problem has wide applications in the field of robotics [1]. This paper presents a novel adaptive optimal learning algorithm to consider the constraints on states and control input simultaneously.

Related work

To consider the input saturation constraints, [2] employs the hyperbolic function to represent the penalty on the control input. In addition to closed-loop stability, the control performance is considered using optimal control theory [2]. Another type of input saturation problem is analyzed in [3], where an auxiliary design system is introduced for constrained adaptive control design. However, these results only considers input saturations, whereas the state constraints is not analyzed. In this paper, the control problem with full-state constraints is taken into consideration.

In contrast to the linear quadratic control, the state constraints in control problem is achieved by some penalty on the state variable [4]. The state penalty is extended to the case of Lyapunov stability analysis, where a novel type of Lyapunov function, named barrier Lyapunov function, is proposed to deal with the state constrained control problem [5]. The barrier Lyaounov function based method is initially designed to deal with the output constraints [5], which can be viewed as the case of partial state constraints. Later, it is extended to deal with the full-state constrained control problem [6]. However, these results does not consider the input constraints. In some cases, both the state and input constraints are critical for safety issue. Therefore, we solve the constrained control problem with consideration of the state and input constraints simultaneously.

Adaptive dynamic programming is an efficient algorithm to solve the optimal decision-making problem, such as optimal regulation problem [7], [8], optimal tracking problem [9], multiagent systems [10], [11], robust control problem [12], differential games [13], [14] and intermittent feedback design [15]. ADP has also been successfully applied to the input constrained control problem [16], [17]. In this paper, we extend the ADP method to the full-state constrained optimal control problem.

In this paper, we first propose a novel barrier function based system transformation method. Different from the barrier Lyapunov function method, the full-state constrained system can be transformed to an equivalent system without state constraints. It is guaranteed that if the initial state is within the prescribed bound, the state constraints can be guaranteed. Then, the input saturation is considered for transformed system. The hyperbolic penalty function is designed for the control input. Moreover, we optimize the nonlinear system using an online algorithm.

Notation and Background

Definition 1. *(Persistent Excitation)* A vector signal $y(t) \in \mathbb{R}^p$ is exciting over the interval $[t, t+T]$ with $T \in \mathbb{R}^+$ if there exists $\beta_1 \in \mathbb{R}^+$ and $\beta_2 \in \mathbb{R}^+$ such that for $\forall t$,

$$\beta_1 I_{p\times p} \leqslant \int_t^{t+T} y(\tau)\, y^{\mathrm{T}}(\tau)\, d\tau \leqslant \beta_2 I_{p\times p} \qquad \square$$

This work was supported in part by the Fundamental Research Funds for the China Central Universities of USTB under grant No. FRF-TP-18-031A1 and No. FRF-BD-17-002A, in part by the China Post-Doctoral Science Foundation under Grant 2018M641197, in part by the National Science Foundation under grant NSF CAREER CPS-1851588, in part by NATO under grant No. SPS G5176, in part by ONR Minerva under grant No. N00014-18-1-2160, in part by the Mary K. Finley Endowment, in part by the Missouri S&T Intelligent Systems Center and in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-18-2-0260. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for overnment purposes notwithstanding any copyright notation herein.

[1]Y. Yang, W. He and Y. Yin are with the School of Automation & Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China. yangyongliang@ieee.org; weihe@ieee.org; yyx@ies.ustb.edu.cn

[2]K. G. Vamvoudak is with the Daniel Guggenheim School of Aerospace Engineering, Georgia Tech, GA 30332, USA. kyriakos@gatech.edu

[3]H. Modares is with the Mechanical Engineering Department, Michigan State University, East Lansing, MI 48824, USA. modares@msu.edu

[4]D. C. Wunsch is with the Department of Electrical and Computer Engineering, Missouri University of Science & Technology, Rolla, MO 65401, USA. wunsch@ieee.org

II. Problem Formulation

In this paper, we consider the following continuous-time affine nonlinear dynamical systems

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= x_3\\ &\vdots\\ \dot{x}_{n-1} &= x_n\\ \dot{x}_n &= f(x) + g(x)\,u\end{aligned} \tag{1}$$

where $x = \left[\begin{array}{ccc} x_1 & \cdots & x_n \end{array}\right]^T \in \mathbb{R}^n$ with $x_i \in \mathbb{R}$ is the system state, $u \in \mathbb{R}$ is the control input, $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are lipschitz continuous functions. The constrained control problem of system (1) with full state constraints can be formulated as follows.

Problem 1. Consider the system (1), find a policy $u : \mathbb{R}^n \to \mathbb{R}^m$ such that the closed-loop system has an asymptotic stable equilibrium while the control input satisfies

$$\|u_i\| \leqslant \lambda, \forall i = 1, ..., m \tag{2}$$

and the state $x = \left[\begin{array}{ccc} x_1 & \cdots & x_n \end{array}\right]^T$ satisfies

$$\begin{aligned}x_1 &\in (a_1, A_1)\\ &\vdots\\ x_n &\in (a_n, A_n),\end{aligned} \tag{3}$$

□

A. Barrier Function

Definition 2. The function $b(\cdot) : \mathbb{R} \to \mathbb{R}$ defined on (a, A) is referred to as barrier function if

$$b(z; a, A) = \log\left(\frac{A}{a}\frac{a - z}{A - z}\right), \forall z \in (a, A) \tag{4}$$

where a and A are two constants satisfying $a < A$. Moreover, the barrier function is invertible on interval (a, A), i.e.,

$$b^{-1}(y; a, A) = aA\frac{e^{\frac{y}{2}} - e^{-\frac{y}{2}}}{ae^{\frac{y}{2}} - Ae^{-\frac{y}{2}}}, \forall y \in \mathbb{R} \tag{5}$$

with the derivative

$$\frac{db^{-1}(y; a, A)}{dy} = \frac{Aa^2 - aA^2}{a^2e^y - 2aA + A^2e^{-y}} \tag{6}$$

Consider the barrier function based state transformation as

$$\begin{aligned}s_i &= b(x_i; a_i, A_i),\\ x_i &= b^{-1}(s_i; a_i, A_i),\end{aligned} \quad i = 1, ..., n \tag{7}$$

then,

$$\frac{dx_i}{dt} = \frac{dx_i}{ds_i}\frac{ds_i}{dt} \tag{8}$$

which can yield the following based on (6)

$$\begin{aligned}\dot{s}_i &= \frac{x_{i+1}(s_{i+1})}{\left.\frac{db^{-1}(y; a_i, A_i)}{dy}\right|_{y=s_i}}\\ &= F_i(s_i, s_{i+1}), \qquad i = 1, ..., n-1\end{aligned} \tag{9}$$

$$\begin{aligned}\dot{s}_n &= \frac{f(x) + g(x)\,u}{\left.\frac{db^{-1}(y; a_n, A_n)}{dy}\right|_{y=s_n}}\\ &= F_n(s) + g_n(s)\,u\end{aligned} \tag{10}$$

where

$$\begin{aligned}F_n(s) &= \frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2}\times\\ &\quad f\left(\left[\begin{array}{ccc} b_1^{-1}(s_1) & \cdots & b_n^{-1}(s_n) \end{array}\right]\right)\\ g_n(s) &= \frac{A_n^2e^{-s_n} - 2a_nA_n + a_n^2e^{s_n}}{A_na_n^2 - a_nA_n^2}\times\\ &\quad g\left(\left[\begin{array}{ccc} b_1^{-1}(s_1) & \cdots & b_n^{-1}(s_n) \end{array}\right]\right)\end{aligned} \tag{11}$$

The above system of $s = \left[\begin{array}{ccc} s_1 & \cdots & s_n \end{array}\right]^T$ can be expressed in a compact form as

$$\dot{s} = F(s) + G(s)\,u \tag{12}$$

with

$$F(s) = \left[\begin{array}{c} F_1(s_1, s_2)\\ \vdots\\ F_n(s) \end{array}\right], G(s) = \left[\begin{array}{c} 0\\ \vdots\\ 0\\ g_n(s) \end{array}\right] \tag{13}$$

Assumption 1. The system dynamics satisfies

1) $F(s)$ is Lipschitz with $F(0) = 0$, and there exists a constant b_f such that, for $x \in \Omega$, $\|F(s)\| \leqslant b_f\|s\|$ where Ω is a compact set containing the origin.
2) $G(s)$ is bounded on Ω, i.e., there exists a constant b_g such that $\|G(s)\| \leqslant b_g$
3) The system (5) is controllable over the compact set Ω.
4) The performance functional (7) satisfies zero-state observability. □

In the following, in order to solve Problem 1 with input saturation and full-state constraints, a novel problem with constraints of input saturation is presented.

Problem 2. (*Optimal Control Problem*) find a policy $u(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ such that the performance

$$V(s_0) = \int_{t_0}^{\infty}\left[Q(s(t)) + \Theta(u(t))\right]dt, \tag{14}$$

is minimized with the constraints of input saturation,

$$\|u_i\| \leqslant \lambda, \forall i = 1, ..., m \tag{15}$$

where $\lambda > 0$ is the bound of control input, $Q(s)$ is a positive definite monotonically increasing function and $\Theta(u)$ is a positive definite integrand function. For simplicity, we denote the reward function $U(s, u) = Q(s) + \Theta(u)$. □

Lemma 1. *Suppose that $u^*(\cdot)$ solves Problem 2 for system system* (12)*. Then, $u^*(\cdot)$ can also solve Problem 1 provided that the initial state x_0 of system* (1) *satisfies the constraints in* (3)*.*

Proof. This proof follows from [5, Lemma 1]. □

To deal with input saturation, the nonquadratic penalty function in [2] is adopted,

$$\Theta(u) = 2\int_0^u\left[\theta^{-1}(v)\right]^{\mathrm{T}}Rdv, \tag{16}$$

with $R = \text{diag}\left(\begin{bmatrix} r_1 & \cdots & r_m \end{bmatrix}\right)$ with the positive penalty weight $r_i \in \mathbb{R}^+$ for $\forall i \in \{1, ..., m\}$, and $\theta(\cdot)$ is the hyperbolic tangent, i.e.,

$$\theta(v) = \lambda \tanh\left(\frac{v}{\lambda}\right) \tag{17}$$

Then, the term $\Theta(u)$ in (16) can be rewritten as

$$\Theta(u) = 2\int_0^u \left[\lambda \tanh^{-1}\left(\frac{v}{\lambda}\right)\right]^{\mathrm{T}} R dv \tag{18}$$

Definition 3. *(Admissible Policy)* A control policy $\mu(s)$ is said to be admissible with respect to (14) on $\Omega \in \mathbb{R}^n$, denoted by $\mu(s) \in \pi(\Omega)$, if

- $\mu(s)$ is continuous on Ω,
- $\mu(0) = 0$,
- $u(s) = \mu(s)$ stabilizes (12) on Ω,
- $V(s)$ is finite for $\forall s \in \Omega$. □

Given an admissible policy u, the Hamiltonian is defined as

$$\mathcal{H}\left(s, u, \frac{\partial V}{\partial s}\right) = \left(\frac{\partial V}{\partial s}\right)^{\mathrm{T}} \left[F(s) + G(s)u\right] + U(s, u), \tag{19}$$

Inserting (16) into (19) yields the Bellman equation

$$\begin{aligned} 0 &= \mathcal{H}\left(s, u, \frac{\partial V}{\partial x}\right) \\ &= \left(\frac{\partial V}{\partial s}\right)^{\mathrm{T}} \left[F(s) + G(s)u\right] + U(s, u). \end{aligned} \tag{20}$$

Consider the optimal value function defined as

$$V^*(s(t)) = \min_{u(\cdot)\in\pi(\Omega)} \int_t^\infty \left[Q(s(\tau)) + \Theta(u(\tau))\right] d\tau. \tag{21}$$

Based on the optimal control theory, the stationary condition in the Hamiltonian (19) yields the optimal control policy

$$u^*(s) = -\lambda \tanh\left(\frac{1}{2\lambda} R^{-1} G^{\mathrm{T}}(s) \frac{\partial V^*(s)}{\partial s}\right) \tag{22}$$

Inserting the optimal control policy (22) yields the Hamilton-Jacobi-Bellman equation

$$\begin{aligned} 0 = &\left[\frac{\partial V^*(s)}{\partial s}\right]^{\mathrm{T}} F(s) + \lambda^2 \bar{R}^{\mathrm{T}} \ln\left[\mathbf{1_m} - \tanh^2(D^*)\right] \\ &+ Q(s) \end{aligned} \tag{23}$$

As shown in (22), the optimal constrained control solution $u^*(s)$ depends on solving the HJB equation (23) for the optimal value function $V^*(s)$. However, the HJB equation (23) is a nonlinear partial differential and extremely difficult to solve. In the following, an online algorithm is presented to find an approximate solution to the HJB equation (23).

III. Online Actor-Critic-Barrier Learning

In this section, we present a novel online algorithm with an actor-critic-barrier structure to learn the optimal control policy. First, the barrier function is employed to transform the origin system (1) to system (12). Then, based on system (12), the learning structure uses the value function approximator with two NNs, an actor NN and a critic NN. The critic NN is used to approximate the Bellman equation (20) and the actor NN is used to approximate the optimal policy (22). Both networks are updated simultaneously in an asynchronous fashion. The actor-critic-barrier structure can be shown in Figure 1.

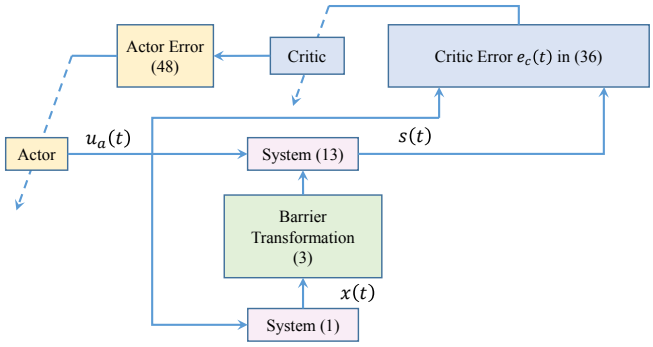

Fig. 1. Actor-critic-barrier transformation.

Assumption 2. There exists positive definite and smooth solution $V(x)$ to the Bellman equation (20). □

Assumption 3. There exists a single-layer NN such that the value function $V(s)$ and its gradient $\nabla V(s)$ can be uniformly approximated with a critic NN, within a set $\Omega \subseteq \mathbb{R}^n$ that contains the origin, as

$$V^*(s) = W^{\mathrm{T}}\phi(s) + \varepsilon(s) \tag{24}$$

$$\nabla V^*(s) = \left[\nabla\phi(s)\right]^{\mathrm{T}} W + \nabla\varepsilon(s) \tag{25}$$

where $W \in \mathbb{R}^N$ is the critic weight, $\phi(\cdot) : \mathbb{R}^n \to \mathbb{R}^N$ is the critic basis, $\varepsilon(s)$ and $\nabla\varepsilon(s)$ are the bounded approximation errors satisfying $\|\varepsilon(s)\| \leqslant b_\varepsilon$ and $\|\nabla\varepsilon(s)\| \leqslant b_{d\varepsilon}$, $\phi(s)$ and $\nabla\phi(s)$ satisfying $\|\phi(s)\| \leqslant b_\phi$ and $\|\nabla\phi(s)\| \leqslant b_{d\phi}$ for $\forall s \in \Omega$. □

For the optimal control policy $u^*(s)$, the Bellman equation (20) approximation error using the value function approximation (24) can be expressed as

$$\varepsilon_B = U(s, u^*) + W^{\mathrm{T}}\sigma, \tag{26}$$

where σ is a N-dimensional vector signal defined as

$$\sigma := \nabla\phi(s)\left[F(s) + G(s)u^*\right] \tag{27}$$

Considering the value gradient approximation (25), one can obtain that the Bellman approximation error ε_B results from the value gradient approximation error $\nabla\varepsilon(s)$, i.e.

$$\varepsilon_B = -\left[\nabla\varepsilon(s)\right]^{\mathrm{T}}\left[F(s) + G(s)u^*\right] \tag{28}$$

From Assumption 2, there exists a constant b_B such that $\|\varepsilon_B\| \leqslant b_B$.

The HJB equation (23) can be approximated using the value function approximation (24) as

$$\begin{aligned} \varepsilon_{hjb}(s) = &W^{\mathrm{T}}\nabla\phi(s)F(s) + \lambda^2 R \ln\left[\mathbf{1_m} - \tanh^2(D)\right], \\ &+ Q(s) \end{aligned} \tag{29}$$

with

$$D = \frac{1}{2\lambda} R^{-1} G^{\mathrm{T}}(s) (\nabla\phi)^{\mathrm{T}} W \tag{30}$$

It is assumed that the ideal value function approximation (24) guarantees $\|\varepsilon_{hjb}\| \leqslant b_{hjb}$.

A. Value Function Approximation

The ideal weight, W in (24), provides the best approximate to the optimal value function $V^*(s)$ on the compact set Ω and is unknown. Therefore, the estimation of W is implemented by the critic network with the approximations of the value function and value gradient

$$\hat{V}(s) = W_c^{\mathrm{T}} \phi_c(s) \tag{31}$$

$$\nabla\hat{V}(s) = [\phi_c(s)]^{\mathrm{T}} W_c \tag{32}$$

Then, for a given policy $u(\cdot)$, the residual of Bellman equation approximation using the identifier NN and the critic NN, can be determined as

$$e_c(t) := U(s(t), u(t)) + \hat{W}_c^{\mathrm{T}}(t)\sigma(t) \tag{33}$$

Define the critic weight approximation error

$$\tilde{W}_c = W - W_c \tag{34}$$

Then, from (26), the relation between Bellman residual e_c and the Bellman equation approximation error ε_B can be written in terms of the critic weight error $\tilde{W}_c$ as

$$e_c = \varepsilon_B - \tilde{W}_c^{\mathrm{T}}\sigma. \tag{35}$$

As discussed in [7], [16], the Bellman residual e_c using critic NN (31) is the counterpart of the temporal difference for continuous-time system. The policy evaluation for an admissible control policy $u(\cdot)$ can be formulated as adapting the critic weight $\hat{W}_c$ to minimize the objective function [8]

$$E_c = \frac{1}{2}\frac{[e_c(t)]^2}{(1+\sigma^{\mathrm{T}}\sigma)^2} \tag{36}$$

Then $e_c \to \varepsilon_B$ as $W_c \to W$. Using the chain rule yields the gradient descent algorithm for minimizing E_c given by [8]

$$\dot{W}_c = -\alpha_c \frac{\partial E_c}{\partial W_c} = -\alpha_c \frac{\sigma}{(1+\sigma^{\mathrm{T}}\sigma)^2}\left[\sigma^{\mathrm{T}} W_c + U(s,u)\right]. \tag{37}$$

Theorem 1. *Let u be any admissible control policy. Let the critic NN* (31) *with the adaptive tuning law* (37) *be used to evaluate the given control policy. Suppose that the signal* $\frac{\sigma(t)}{1+\sigma^{\mathrm{T}}(t)\sigma(t)}$ *satisfies the PE condition. Then,* $\tilde{W}_c$ *is uniformly ultimately bounded.* □

Proof. Then, from (34) and (35), the dynamics of $\tilde{W}_c$ is

$$\dot{\tilde{W}}_c = -\alpha_c \left[\frac{\sigma(t)\sigma^{\mathrm{T}}(t)}{[1+\sigma^{\mathrm{T}}(t)\sigma(t)]^2}\right]\tilde{W}_c + \alpha_c \left[\frac{\sigma(t)}{[1+\sigma^{\mathrm{T}}(t)\sigma(t)]^2}\right]\varepsilon_B(t), \tag{38}$$

which can be viewed as a linear time-varying system with the control input $\varepsilon_B(t)$. Note that the signal $\frac{\sigma(t)}{1+\sigma^{\mathrm{T}}(t)\sigma(t)}$ satisfies the PE condition. Then, $\tilde{W}_c$ can be expressed as

$$\begin{aligned}\tilde{W}_c(t) &= \varphi(t,t_0)\tilde{W}_c(t_0) + \int_{t_0}^{t} \varphi(t,\tau)\frac{\alpha_c \sigma(\tau)\varepsilon_B(\tau)}{[1+\sigma^{\mathrm{T}}(\tau)\sigma(\tau)]^2} d\tau \\ &\leqslant \kappa_1 e^{-\kappa_2 t} + \frac{\alpha_c}{\kappa_2[1+\sigma^{\mathrm{T}}(\tau)\sigma(\tau)]}\|\varepsilon_B(t)\|\end{aligned}$$

where

$$\frac{\partial\varphi(t,t_0)}{\partial t} = -\alpha_c \frac{\sigma(\tau)\sigma^{\mathrm{T}}(\tau)}{[1+\sigma^{\mathrm{T}}(\tau)\sigma(\tau)]^2}\varphi(t,t_0), \varphi(t,t) = I$$

Note that $\|\varepsilon_B\| \leqslant b_B$. Then, according to [8], the critic weight error $\tilde{W}_c$ is uniformly ultimately bounded. This completes the proof. ■

B. Actor Learning: Online Synchronous Policy Iteration

As shown in (22), the optimal control policy depends on the optimal value gradient $\frac{\partial V^*(s)}{\partial s}$. Therefore, consider the value gradient approximation using the critic weight W_c in (31), the policy can be determined as

$$u_c(s) = -\lambda \tanh(D_c) \tag{39}$$

$$D_c = \frac{1}{2\lambda} R^{-1} G^{\mathrm{T}} (\nabla\phi)^{\mathrm{T}} W_c \tag{40}$$

However, this policy improvement does not guarantee the stability of the system [8], [16]. Therefore, to ensure stability in a Lyaponuv sense (as discussed later), the policy applied to the system is implemented by the actor network as

$$u_a(s) = -\lambda \tanh(D_a) \tag{41}$$

$$D_a = \frac{1}{2\lambda} R^{-1} G^{\mathrm{T}} (\nabla\phi)^{\mathrm{T}} W_a \tag{42}$$

with the actor network weight W_a. The actor-critic adaptive learning is discussed as the following.

Theorem 2. *Consider the dynamical system (1). Let the control law* u_a *be given by* (41)*. Let the experience replay based critic learning be*

$$\dot{W}_c = -\alpha_c \frac{\sigma_a(t)}{[1+\sigma_a^{\mathrm{T}}(t)\sigma_a(t)]^2}\left[U_a(t) + W_c^{\mathrm{T}}(t)\sigma_a(t)\right], \tag{43}$$

where

$$\begin{aligned}\sigma_a(t) &:= \nabla\phi[F(s(t)) + G(s(t))u_a(t)], \\ U_a(t) &= U(s(t), u_a(t)),\end{aligned} \tag{44}$$

Let the gradient-descent based actor learning be

$$\dot{W}_a = -\alpha_a\left[\nabla\phi G e_u + \nabla\phi G \tanh^2(D_a) e_u + Y W_a\right], \tag{45}$$

which is designed to minimize the objective function

$$E_u = \frac{1}{2} e_u^{\mathrm{T}} R e_u \tag{46}$$

where

$$e_u = u_c - u_a = \lambda[\tanh(D_a) - \tanh(D_c)] \tag{47}$$

denotes the difference between the actor u_a (41) *applied to the system and the control input* u_c (39) *as an approximation*

of the optimal control policy (22) *with value function approximation by* (31)*. Then, under Assumptions 1 - 3 and suppose that the signal* $\frac{\sigma_a(t)}{1+\sigma_a^{\mathrm{T}}(t)\sigma_a(t)}$ *satisfies the PE condition, the closed-loop system states and the critic and actor NN errors are uniformly ultimately bounded for a sufficiently large number of NN basis, provided that*

$$Y > \frac{M_a M_a^{\mathrm{T}}}{2}$$

where $M_a = \nabla\phi G\lambda\left[\tanh\left(\kappa D_a\right) - \tanh\left(D_a\right)\right]$. □

Proof. When the policy u_a is applied to the system (12), we consider the following Lyapunov candidate

$$V = V^*\left(s\right) + \underbrace{\frac{1}{2}\tilde{W}_c^{\mathrm{T}}\alpha_1^{-1}\tilde{W}_c}_{V_c} + \underbrace{\frac{1}{2}\tilde{W}_a^{\mathrm{T}}\alpha_2^{-1}\tilde{W}_a}_{V_a} \quad (48)$$

where $V^*(s)$ is the optimal value function, $\tilde{W}_c$ and $\tilde{W}_a$ are the critic and actor weight errors, respectively. Differentiating V yields,

$$\dot{V} = \dot{V}^* + \dot{V}_c + \dot{V}_a. \quad (49)$$

The first term in (49) can be expressed as

$$\begin{aligned}\dot{V}^* &= \dot{V}^*\left(s\right) = \left\langle \nabla V^*\left(x\right), F + Gu_a\right\rangle \\ &= \left\langle \left(\nabla\phi\right)^{\mathrm{T}} W + \nabla\varepsilon\left(x\right), F - G\lambda\tanh\left(D_a\right)\right\rangle \\ &= W^{\mathrm{T}}\nabla\phi F - W^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right) + \varepsilon_1 \quad (50)\end{aligned}$$

where $\varepsilon_1 = \left(\nabla\varepsilon\right)^{\mathrm{T}}\left[F - G\lambda\tanh\left(D_a\right)\right]$ satisfying $\left\|\varepsilon_1\right\| \leqslant b_{d\varepsilon}b_f\left\|x\right\| + \lambda b_{d\varepsilon}b_g$. The HJB equation approximation (29) can be equivalently rewritten as

$$\varepsilon_{hjb} = W^{\mathrm{T}}\nabla\phi\left[F - G\lambda\tanh\left(D\right)\right] + Q\left(s\right) + \Theta, \quad (51)$$

where

$$\Theta = W^{\mathrm{T}}\nabla\phi g\lambda\tanh\left(D\right) + \lambda^2 R\log\left(1 - \tanh^2\left(D\right)\right), \quad (52)$$

with D as defined in (30). Then

$$W^{\mathrm{T}}\nabla\phi F = W^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D\right) - Q\left(s\right) - \Theta + \varepsilon_{hjb}, \quad (53)$$

Inserting (53) into (50) yields

$$\begin{aligned}\dot{V}^* = &-Q - \Theta + W^{\mathrm{T}}\nabla\phi\left(s\right)G\lambda\tanh\left(D\right) \\ &-W^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right) + \varepsilon_1 + \varepsilon_{hjb}. \quad (54)\end{aligned}$$

Note that Θ and Q are positive definite, there exists a positive constant q such that

$$-Q - \Theta \leqslant -Q \leqslant -qs^{\mathrm{T}}s. \quad (55)$$

The third term in (54) can be upper bounded by

$$W^{\mathrm{T}}\nabla\phi\left(x\right)G\lambda\tanh\left(D\right) \leqslant \lambda b_g b_{d\phi}\left\|W\right\| \quad (56)$$

Taking $W = W_a + \tilde{W}_a$ into (54) yields

$$\begin{aligned}&-W^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right) \\ &= -\tilde{W}_a^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right) - \underbrace{W_a^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right)}_{=2\lambda^2 R[D_a\tanh(D_a)]\geqslant 0} \\ &\leqslant -\tilde{W}_a^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right) \quad (57)\end{aligned}$$

where the above inequality results from the fact that $x^{\mathrm{T}}\tanh\left(x\right) \geqslant 0, \forall x$. Collecting (55) - (57), one has

$$\begin{aligned}\dot{V}^* &\leqslant -s^{\mathrm{T}}qs + b_{d\varepsilon}b_f\left\|s\right\| + \lambda b_g b_{d\phi}\left\|W\right\| + \lambda b_{d\varepsilon}b_g + b_h \\ &\quad -\tilde{W}_a^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right) \\ &= -s^{\mathrm{T}}qs + M_s\left\|s\right\| + N_s - \tilde{W}_a^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right), \quad (58)\end{aligned}$$

where $M_s = b_{d\varepsilon}b_f$ and $N_s = \lambda b_g b_{d\phi}\left\|W\right\| + \lambda b_{d\varepsilon}b_g + b_h$ are bounded.

Second, we analyze the term $\dot{V}_c$ in (49). From (35) and (43), one has

$$\dot{\tilde{W}}_c^{\mathrm{T}} = \alpha_c\frac{\sigma_a}{\left[1 + \sigma_a^{\mathrm{T}}\sigma_a\right]^2}e_c. \quad (59)$$

From (44), e_c in the above equation can be further written as

$$\begin{aligned}e_c &= Q + \Theta_a + W_c^{\mathrm{T}}\sigma_a \\ &= Q + \Theta_a + W_c^{\mathrm{T}}\sigma_a \\ &\quad -Q - \Theta - W^{\mathrm{T}}\nabla\phi\left[F - G\lambda\tanh\left(D\right)\right] + \varepsilon_{hjb},\end{aligned}$$

where the second equality results from (29) and Θ is defined in (52). Then, e_c can be equivalently written as

$$e_c = \Theta_a - \Theta + W_c^{\mathrm{T}}\sigma_a - W^{\mathrm{T}}\nabla\phi\left[F - G\lambda\tanh\left(D\right)\right] + \varepsilon_{hjb}, \quad (60)$$

where

$$\begin{aligned}&\Theta_a - \Theta \\ &= W_a^{\mathrm{T}}\nabla\phi\left(s\right)g\lambda\tanh\left(D_a\right) - W^{\mathrm{T}}\nabla\phi\left(s\right)g\lambda\tanh\left(D\right) \\ &+ \lambda^2 R\log\left(1 - \tanh^2\left(D_a\right)\right) - \lambda^2 R\log\left(1 - \tanh^2\left(D\right)\right). \quad (61)\end{aligned}$$

Note that the term $\lambda^2 R\log\left(1 - \tanh^2\left(D\right)\right)$ in (61) can be rewritten as

$$\begin{aligned}&\lambda^2 R\log\left(1 - \tanh^2\left(D\right)\right) \\ &= \lambda^2 R\left[\log 4 - 2D - 2\log\left(1 + e^{-2D}\right)\right],\end{aligned}$$

where $-2\log\left(1 + e^{-2D}\right)$ can be approximated as [16]

$$-2\log\left(1 + e^{-2D}\right) = 2D - 2D\mathrm{sgn}\left(D\right) + \varepsilon_D,$$

where $\left\|\varepsilon_D\right\| \leqslant \log 4$. Then,

$$\lambda^2 R\log\left(1 - \tanh^2\left(D\right)\right) = \lambda^2 R\left[\log 4 - 2D\mathrm{sgn}\left(D\right) + \varepsilon_D\right]. \quad (62)$$

Similarly,

$$\begin{aligned}&\lambda^2 R\log\left(1 - \tanh^2\left(D_a\right)\right) \\ &= \lambda^2 R\left[\log 4 - 2D_a\mathrm{sgn}\left(D_a\right) + \varepsilon_{D_a}\right], \quad (63)\end{aligned}$$

where $\left\|\varepsilon_{D_a}\right\| \leqslant \log 4$. Consider (61) - (63), one has

$$\begin{aligned}&\Theta_a - \Theta \\ &= W_a^{\mathrm{T}}\nabla\phi\left(s\right)g\lambda\tanh\left(D_a\right) - W^{\mathrm{T}}\nabla\phi\left(s\right)g\lambda\tanh\left(D\right) \\ &\quad +\lambda^2 R\left[2D\mathrm{sgn}\left(D\right) - 2D_a\mathrm{sgn}\left(D_a\right) + \varepsilon_{D_a} - \varepsilon_D\right] \quad (64)\end{aligned}$$

The approximation of function $x\mathrm{sgn}(x)$ is investigated in [18] with

$$0 \leqslant x\mathrm{sgn}\left(x\right) - x\tanh\left(\kappa x\right) \leqslant \frac{3.59}{\kappa} \quad (65)$$

Then, based on (30) and (41), one has

$$\begin{aligned}&\lambda^2 R\left[2D\mathrm{sgn}\left(D\right)-2D_a\mathrm{sgn}\left(D_a\right)\right]\\&=W^{\mathrm{T}}\nabla\phi\left(x\right)g\lambda\left[\tanh\left(\kappa D\right)-\tanh\left(\kappa D_a\right)\right]\\&+\tilde{W}_a^{\mathrm{T}}\nabla\phi\left(x\right)g\lambda\tanh\left(\kappa D_a\right)+\varepsilon_\kappa\end{aligned}\tag{66}$$

with approximation error satisfying $0\leqslant\varepsilon_\kappa\leqslant\frac{7.18}{\kappa}$. Based on (64), (66), by adding and subtracting $W^{\mathrm{T}}\nabla\phi G\lambda\tanh\left(D_a\right)$ in (60), one has

$$\begin{aligned}e_c\left(t\right)&=-\tilde{W}_c^{\mathrm{T}}\left(t\right)\sigma_a\left(t\right)+\tilde{W}_a^{\mathrm{T}}\left(t\right)M_a\left(t\right)+N_a\left(t\right)\\M_a\left(t\right)&=\nabla\phi\left(t\right)G\left(t\right)\lambda\left[\tanh\left(\kappa D_a\left(t\right)\right)-\tanh\left(D_a\left(t\right)\right)\right]\\N_a\left(t\right)&=W^{\mathrm{T}}\nabla\phi\left(t\right)G\left(t\right)\lambda\left[\tanh\left(\kappa D\left(t\right)\right)-\tanh\left(\kappa D_a\left(t\right)\right)\right]\\&\quad+\lambda^2R\left[\varepsilon_{D_a}\left(t\right)-\varepsilon_D\left(t\right)\right]+\varepsilon_{hjb}\left(t\right)+\varepsilon_\kappa\left(t\right)\end{aligned}\tag{67}$$

Based on Assumptions 1 and 3, both M_a and N_a are bounded. Substituting (67) into (59) yields,

$$\dot{\tilde{W}}_c=\alpha_c\frac{\sigma_a}{\left[1+\sigma_a^{\mathrm{T}}\sigma_a\right]^2}\left[-\tilde{W}_c^{\mathrm{T}}\left(t\right)\sigma_a+\tilde{W}_a^{\mathrm{T}}\left(t\right)M_a+N_a\right]$$

Therefore, $\dot{V}_c$ can be written as in (68), where the matrix $Q_c\succ0$ because the signal $\frac{\sigma_a(t)}{1+\sigma_a^{\mathrm{T}}(t)\sigma_a(t)}$ satisfies the PE condition. Also, note that N_a, M_{aj}, and N_{aj} are bounded. Then, M_1 in (68) is also bounded.

Next, we give the upper bound of $\dot{V}_a$. Based on (45), differentiating V_a yields (69). Based on Assumption 1 - 3 and the definition of the actor learning error e_u in (47), M_2 is also bounded.

Finally, collecting (58), (68) and (69), one can obtain

$$\begin{aligned}\dot{V}^*\leqslant&-s^{\mathrm{T}}qs+M_s\left\|s\right\|+N_s\\&-\tilde{W}_c^{\mathrm{T}}\left(t\right)Q_c\tilde{W}_c\left(t\right)+\tilde{W}_c^{\mathrm{T}}M_1\\&+\tilde{W}_a^{\mathrm{T}}\left(t\right)\left[\frac{M_aM_a^{\mathrm{T}}}{2}-Y\right]\tilde{W}_a\left(t\right)+\tilde{W}_a^{\mathrm{T}}M_2.\end{aligned}\tag{70}$$

Using the matrix theory, if the parameter Y is sufficiently large such that $Y-\frac{M_aM_a^{\mathrm{T}}}{2}>0$. Then, $\dot{V}$ is negative, provided that

$$\begin{aligned}\left\|s\right\|&\geqslant\frac{M_s}{2q}+\sqrt{\frac{M_s^2}{4q^2}+\frac{N_s}{q}}\\\left\|\tilde{W}_c\right\|&\geqslant\frac{M_1}{\lambda_{\min}(Q_c)}\\\left\|\tilde{W}_a\right\|&\geqslant\frac{M_2}{\lambda_{\min}\left(Y-\frac{M_aM_a^{\mathrm{T}}}{2}\right)}\end{aligned}$$

This completes the proof. ■

IV. Simulation Study

In order to validate the effectiveness of the presented actor-critic-barrier structure for solving Problem 1, the van der pol oscillator will be used in this section. Consider the controlled Van der Pol oscillator with the dynamics

$$\dot{x}=\left[\begin{array}{c}x_2\\-x_1+0.5\left(1-x_2^2\right)x_2\end{array}\right]+\left[\begin{array}{c}0\\x_1\end{array}\right]u.\tag{71}$$

In this scenario, it is desired that the state $x=\left[x_1\ x_2\right]^T$ satisfies the following constraints,

$$x_1\in\left(-0.6,0.2\right),x_2\in\left(-0.2,0.5\right)\tag{72}$$

According to the converse HJB method [19], when the performance parameters are selected as $Q=I_{2\times2}$ and $R=1$, the optimal controller would be $u^*(x)=-x_1x_2$. When applying this optimal control policy to the system (71), the evolution is shown in Figure 2, where the solid lines represent the state evolutions and the dashed lines denote the asymmetric bounds for the states. The system state trajectories in two-dimensional space is shown in Figure 4, where the black box denotes the safety region. It is desired to drive the system states to the origin without violating the safety constraints. Even the states are regulated to the origin asymptotically, the full state constraints can not be guaranteed.

Next, we apply the safe reinforcement learning algorithm with the actor-critic-barrier structure developed in Theorem 2, where the corresponding results are given in Figure 3. In this case, we start the system from the same initial condition. The system state trajectories in two-dimensional space is shown in Figure 4. In contrast to the previous case, one can observe that the state approach to the origin without violating the full state constraints. Both the full-state constraints and the closed-loop stability can be guaranteed when using the actor-critic-barrier learning algorithm.

V. Conclusion

This paper presents a novel barrier function based system transformation to deal with both symmetric and asymmetric constraints on full-state. With the barrier function, the regulation problem with full-state constraints is transformed into an unconstrained optimal control problem. It is shown that the solution of the unconstrained optimal control problem is guaranteed to tackle the full-state constrained regulation problem. Then, the barrier function is combined with the actor-critic structure to solve the unconstrained optimal control problem in an online fashion. It is shown that additional barrier function to the actor-critic structure guarantees the constraints would not be violated by the adaptive optimal controller design. Boundedness and stability of signals in the closed-loop system is analyzed, with the convergence to the optimal control policy. The efficacy of the presented approach is demonstrated using a simulation example in the end.

References

[1] W. He, Z. Li, and C. L. P. Chen, "A survey of human-centered intelligent robots: issues and challenges," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 602–609, 2017.

[2] 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.

[3] M. Chen, S. S. Ge, and B. Ren, "Adaptive tracking control of uncertain mimo nonlinear systems with input constraints," *Automatica*, vol. 47, no. 3, pp. 452 – 465, 2011.

[4] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal state feedback control of constrained nonlinear systems using a neural networks HJB approach," *Annual Reviews in Control*, vol. 28, no. 2, pp. 239 – 251, 2004.

[5] K. P. Tee, S. S. Ge, and E. H. Tay, "Barrier lyapunov functions for the control of output-constrained nonlinear systems," *Automatica*, vol. 45, no. 4, pp. 918 – 927, 2009.

$$\dot{V}_c = -\tilde{W}_c^{\mathrm{T}}(t)\frac{\sigma_a\sigma_a^{\mathrm{T}}}{(1+\sigma_a^{\mathrm{T}}\sigma_a)^2}\tilde{W}_c(t) + \tilde{W}_c^{\mathrm{T}}(t)\frac{\sigma_a}{[1+\sigma_a^{\mathrm{T}}\sigma_a]^2}M_a^{\mathrm{T}}\tilde{W}_a(t) + \tilde{W}_c^{\mathrm{T}}\frac{\sigma_a}{[1+\sigma_a^{\mathrm{T}}\sigma_a]^2}N_a$$
$$\leqslant -\tilde{W}_c^{\mathrm{T}}(t)\underbrace{\left[\frac{\sigma_a\sigma_a^{\mathrm{T}}}{(1+\sigma_a^{\mathrm{T}}\sigma_a)^2} - \frac{\sigma_a\sigma_a^{\mathrm{T}}}{2(1+\sigma_a^{\mathrm{T}}\sigma_a)^4}\right]}_{Q_c}\tilde{W}_c(t) + \frac{1}{2}\tilde{W}_a^{\mathrm{T}}(t)M_aM_a^{\mathrm{T}}\tilde{W}_a(t) + \tilde{W}_c^{\mathrm{T}}\underbrace{\frac{\sigma_a}{(1+\sigma_a^{\mathrm{T}}\sigma_a)^2}N_a}_{:=M_1}, \tag{68}$$

$$\dot{V}_a = \tilde{W}_a^{\mathrm{T}}\left[\nabla\phi G\lambda\tanh(D_a) - \nabla\phi G\lambda\tanh(D_c) + \nabla\phi G\tanh^2(D_a)e_u + YW - Y\tilde{W}_a\right]$$
$$= -\tilde{W}_a^{\mathrm{T}}Y\tilde{W}_a + \tilde{W}_a^{\mathrm{T}}\nabla\phi G\lambda\tanh(D_a) + \tilde{W}_a^{\mathrm{T}}\underbrace{\left[-\nabla\phi G\lambda\tanh(D_c) + \tilde{W}_a^{\mathrm{T}}\nabla\phi G\tanh^2(D_a)e_u + \tilde{W}_a^{\mathrm{T}}YW\right]}_{:=M_2}. \tag{69}$$

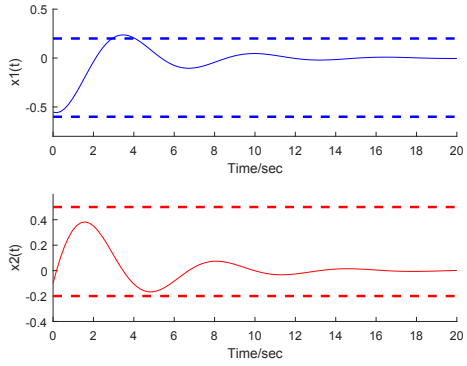

Fig. 2. State trajectories when using the converse HJB approach [19].

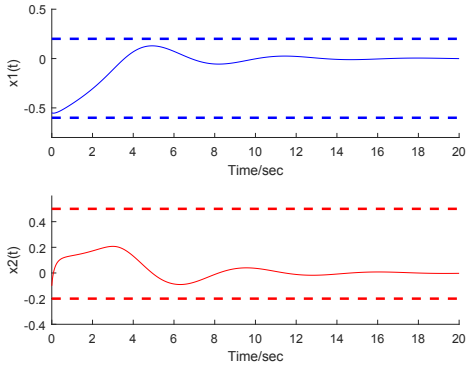

Fig. 3. State trajectories when applying actor-critic-barrier algorithm.

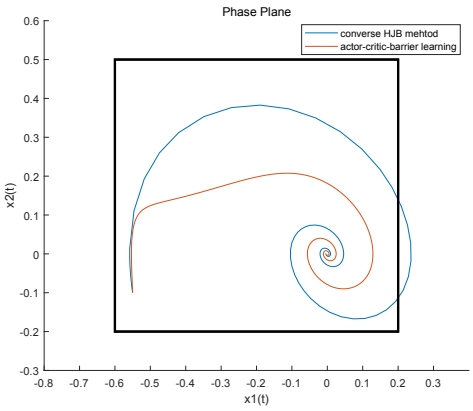

Fig. 4. Two-dimensional phase plot of state trajectories using the converse HJB approach [19] and the actor-critic-barrier algorithm. The black box denotes the safety region.

[6] W. He, Y. Chen, and Z. Yin, "Adaptive neural network control of an uncertain robot with full-state constraints," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 620–629, March 2016.

[7] Y. Yang, D. Wunsch, and Y. Yin, "Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 8, pp. 1929–1940, 2017.

[8] 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.

[9] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780 – 1792, 2014.

[10] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Leader-follower output synchronization of linear heterogeneous systems with active leader using reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2139–2153, June 2018.

[11] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Optimal containment control of unknown heterogeneous systems with active leaders," *IEEE Transactions on Control Systems Technology*, 2018, to Appear.

[12] Y. Yang, Z. Guo, H. Xiong, D. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2019, to Appear.

[13] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. IET Press, 2012.

[14] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Systems Magazine*, vol. 37, no. 1, pp. 33–52, Feb 2017.

[15] Y. Yang, K. G. Vamvoudakis, H. Ferraz, and H. Modares, "Dynamic intermittent Q-learning-based model-free suboptimal co-design of $\mathcal{L}_2$-stabilization," *International Journal of Robust and Nonlinear Control*, 2019, to Appear.

[16] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 10, pp. 1513–1525, 2013.

[17] K. G. Vamvoudakis, M. F. Miranda, and J. P. Hespanha, "Asymptotically stable adaptive-optimal control algorithm with saturating actuators and relaxed persistence of excitation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2386–2398, Nov 2016.

[18] M. M. Polycarpou and P. A. Ioannou, "A robust adaptive nonlinear control design," in *1993 American Control Conference*, June 1993, pp. 1365–1369.

[19] V. Nevisti and J. A. Primbs, "Constrained nonlinear optimal control: a converse HJB approach," *California Institute of Technology*, 1996.