WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia IJCNN

H_∞ Control of Unknown Discrete-Time Nonlinear Systems with Control Constraints Using Adaptive Dynamic Programming

Derong Liu, *Fellow, IEEE*, Hongliang Li, and Ding Wang

***Abstract*—In this paper, we solve the H_∞ robust optimal control problem for discrete-time nonlinear systems with control saturation constraints using the iterative adaptive dynamic programming algorithm. First, a heuristic dynamic programming algorithm is derived to solve the Hamilton-Jacobi-Isaacs equation associated with the H_∞ control problem, and a convergence analysis is provided. Then, a dual heuristic dynamic programming algorithm with nonquadratic performance functional is developed to overcome the control saturation constraints. Finally, to facilitate the implementation of the algorithm, four neural networks are used to approximate the unknown nonlinear system, the control policy, the disturbance policy, and the value function.**

I. Introduction

DURING the last decades, adaptive dynamic programming (ADP) [1], [2] has received much attention as an intelligent scheme for solving the optimal control problems by an online data-based procedure, and the exact knowledge of the system is not required. Existing ADP approaches can be classified into several main schemes [3]: heuristic dynamic programming (HDP), dual heuristic dynamic programming (DHP), globalized dual heuristic dynamic programming (GDHP), and their action-dependent (AD) versions, ADHDP, ADDHP, ADGDHP. The optimal state feedback control policy for nonlinear systems can be found by solving the Hamilton-Jacobi-Bellman (HJB) [4] equation, while it reduces to Riccati equation for linear quadratic regulator (LQR) problem. However, the theoretical solution of the HJB equation is difficult to obtain due to its inherently nonlinear nature. Many efforts using ADP have been made to solve the HJB equation [5]–[7]. Reinforcement learning (RL) [8] is a machine learning method for an agent or controller to learn the optimal control policies based on the observed responses from the environment or system. In recent years, RL has been applied to feedback control [9]. The main algorithms of RL, i.e., policy iteration (PI) and value iteration (VI) have been developed to solve the HJB equation of the optimal control problems. PI algorithms contain policy evaluation and policy improvement [10]–[12], where an initial stabilizing control policy is required. VI algorithms solve the optimal control problem without requirement of an initial stabilizing control policy [13]–[17]. However, most of the previous researches on ADP algorithms provide an online or offline approach to the solution of optimal control problems assuming that the system is not affected by disturbance. But the disturbance exists in reality and affects the control performance. The ADP algorithm considering the disturbance is the interest of our paper.

As a kind of robust optimal control methods, the H_∞ optimal control seeks to not only minimize a cost function, but also attenuates a worst-case disturbance [18]. The H_∞ control problem was converted into an L_2-gain optimal control problem [19] using the concept of dissipative system [20]. It relies on solving the Hamilton-Jacobi-Isaacs (HJI) equation which reduces to the game algebraic Riccati equation (GARE). The HJI equation is more difficult to solve than the HJB equation for the nonlinear dynamical systems. Furthermore, the H_∞ control has a strong connection with zero-sum game [21], where the controller is a minimizing player and the disturbance is a maximizing player. The Nash equilibrium solution is usually obtained by means of offline iterative computation, and the exact knowledge of the system dynamics is required.

In [22], Vrabie and Lewis presented an ADP algorithm for determining online the Nash equilibrium soulution for the two-player zero-sum differential game with linear continuous-time dynamics. For continuous-time nonlinear systems, Abu-Khalaf et al. [23], [24] derived an H_∞ suboptimal state feedback controller for constrained input systems. This method was offline and there existed two iterative loops. In [25], Zhang et al. used four action networks and two critic networks to obtain the saddle point solution of the game, and the full knowledge of the system dynamics was required. In [26], Vamvoudakis and Lewis presented an online adaptive learning algorithm based on PI to solve the continuous-time two-player zero-sum game for nonlinear systems with known dynamics. In [27], Dierks and Jagannathan solved the HJI equation online and forward-in-time using a novel single online approximator-based scheme to achieve optimal regulation and tracking control of affine nonlinear continuous-time systems. In [28], Al-Tamimi et al. solved online the zero-sum game of linear discrete-time (DT) system using HDP and DHP. In [29], Mehraeen et al. developed an iterative approach to solve offline the approximate HJI equation by using the Taylor series expansion of the value function and derived sufficient conditions for the convergence of the approximate HJI solution

The authors are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China (phone: +86-10-62557379; fax: +86-10-62650912; e-mail: derong.liu@ia.ac.cn; Hongliang.Li625@gmail.com; ding.wang@ia.ac.cn).

This work was supported in part by the National Natural Science Foundation of China under Grants 60904037, 60921061, and 61034002, and by the Beijing Natural Science Foundation under Grant 4102061.

to the saddle point.

To our knowledge, there still exist no results to solve the HJI equation for unknown discrete-time nonlinear systems with control saturation constraints. In this paper, we propose two value iteration methods based on HDP and DHP to solve the HJI equation for discrete nonlinear systems, in which the knowledge of the internal system dynamics is not needed. The method in [29] has two iterative loops, i.e., the control and disturbance policies are asynchronously updated. In our scheme, only one iterative loop is used, and the initial stabilizing control policy is not required. To prove the convergence of this scheme, we use relaxed dynamic programming method introduced in [30], [31]. To facilitate the implementation of the algorithm, four neural networks are used to approximate the unknown nonlinear system, the control policy, the disturbance policy, and the value function.

The rest of the paper is organized as follows. Section II provides the problem formulation and DT HJI equation for nonlinear systems. In Section III, we derive the value iteration algorithm, give the convergence analysis, and then solve the control constraints problem. Section IV discusses the NN implementation of the iterative ADP algorithm and is followed by concluding remarks in Section V.

II. Problem Formulation

Consider the discrete-time affine nonlinear dynamical systems described by

$$x_{k+1} = f(x_k) + g(x_k)u_k + h(x_k)w_k, \tag{1}$$

where $x_k \in \Omega \subseteq \mathbb{R}^n$ is the state vector, $u_k = u(x_k) \in \Omega_u \subseteq \mathbb{R}^m$ is the control input, and $w_k = w(x_k) \in \mathbb{R}^q$ is the disturbance input. $f(x_k) \in \mathbb{R}^n$, $g(x_k) \in \mathbb{R}^{n\times m}$ and $h(x_k) \in \mathbb{R}^{n\times q}$ are smooth and differentiable functions. We denote $\Omega_u = \{u(x_k)|\ |u_i(x_k)| \le \bar{u}_i,\ i = 1,\dots,m\}$, and let $\bar{U} = \text{diag}\{\bar{u}_1,\dots,\bar{u}_m\}$ be the constant diagonal matrix. We assume that the following assumptions hold throughout the paper.

Assumption 1: $f(0) = 0$, and $x_k = 0$ is an equilibrium state of the system.

Assumption 2: $f + gu + hw$ is Lipschitz continuous on a compact set $\Omega \subseteq \mathbb{R}^n$ containing the origin.

Assumption 3: The system (1) is controllable in the sense that there exists a continuous control policy on Ω that asymptotically stabilizes the system.

In this paper, we define the infinite horizon cost function as follows:

$$\begin{aligned} J(x_0) &= \sum_{k=0}^{\infty} \left\{ x_k^T Q x_k + P(u_k) - \gamma^2 w_k^T w_k \right\} \\ &= \sum_{k=0}^{\infty} l(x_k, u_k, w_k), \end{aligned} \tag{2}$$

where Q is positive definite matrix, $P(u_k) \in \mathbb{R}$ is also positive definite, and γ is a prescribed positive constant. For unconstrained control problem, $P(u_k)$ can be chosen as quadratic form. To overcome control saturation constraints, we employ a non-quadratic functional [16]

$$P(u_k) = 2\int_0^{u_k} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds, \tag{3}$$

where $\boldsymbol{\psi}^{-1}(u_k) = [\psi^{-1}(u_{1k}),\dots,\psi^{-1}(u_{mk})]^T$, R is positive definite diagonal matrix, $s \in \mathbb{R}^m$, $\boldsymbol{\psi} \in \mathbb{R}^m$, $\boldsymbol{\psi}^{-T}$ denotes $(\boldsymbol{\psi}^{-1})^T$. $\psi(\cdot)$ is a bounded one-to-one function satisfying $|\psi(\cdot)| \le 1$ and belonging to $C^p(p \ge 1)$ and $L_2(\Omega)$, and it is a monotonic odd function with its first derivative bounded by a constant M. The hyperbolic tangent function $\psi(\cdot) = \tanh(\cdot)$ is one example satisfying these conditions. Besides, it is important to note that $P(u_k)$ is positive definite since $\psi^{-1}(\cdot)$ is a monotonic odd function and R is positive definite.

Note that the control policy $u(x_k)$ must not only stabilize the system on Ω but also guarantee that (2) is finite, i.e., the control policy must be admissible [7].

Definition 1: (Admissible Control Policy) A control policy $u(x)$ is said to be admissible with respect to (2) on Ω, denoted by $u(x) \in \Psi(\Omega)$, if $u(x)$ is continuous on a compact set $\Omega \subseteq \mathbb{R}^n$, $u(0) = 0$, $u(x)$ stabilizes (1) on Ω and for $\forall x_0 \in \Omega$, $J(x_0)$ is finite.

For the admissible control policy $u(x_k)$ and disturbance policy $w(x_k)$, define the value function as

$$\begin{aligned} V(x_k, u_k, w_k) = \sum_{i=k}^{\infty} \Bigg\{ & x_i^T Q x_i + 2\int_0^{u_i} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds \\ & - \gamma^2 w_i^T w_i \Bigg\}. \end{aligned} \tag{4}$$

The Hamilton function can be defined as

$$\begin{aligned} H(x_k, u_k, w_k) = V(f + gu_k + hw_k) - V(x_k) + x_k^T Q x_k \\ + 2\int_0^{u_k} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds - \gamma^2 w_k^T w_k. \end{aligned} \tag{5}$$

According to [21], this control problem can be referred to a two-player zero-sum differential game, where the infinite-horizon value function is to be minimized by the control policy player $u(x_k)$ and maximized by the disturbance policy player $w(x_k)$. Our goal is to find the the feedback saddle point solution (u_k^*, w_k^*) or the Nash equilibrium such that

$$V^*(x_0) = \min_{u_k} \max_{w_k} \{V(x_0, u_k, w_k)\} \tag{6}$$

or $V(u_k^*, w_k) \le V(u_k^*, w_k^*) \le V(u_k, w_k^*)$ for all u_k and w_k. The sufficient condition for the existence of a saddle point is

$$\min_{u_k} \max_{w_k} \{V(x_0, u_k, w_k)\} = \max_{w_k} \min_{u_k} \{V(x_0, u_k, w_k)\}. \tag{7}$$

According to Bellman's optimality principle, the optimal value function $V^*(x_k)$ satisfies the DT HJI equation [28]

$$V^*(x_k) = \min_{u_k} \max_{w_k} \{l(x_k, u_k, w_k) + V^*(x_{k+1})\}. \tag{8}$$

The optimal control policy $u^*(x_k)$ and the worst case disturbance $w^*(x_k)$ should satisfy $\partial H(x_k, u_k, w_k)/\partial u_k = 0$ and $\partial H(x_k, u_k, w_k)/\partial w_k = 0$. Therefore, we obtain

$$u^*(x_k) = \bar{U}\boldsymbol{\psi}\left(-\frac{1}{2}(\bar{U}R)^{-1} g^T(x_k) \frac{\partial V^*(x_{k+1})}{\partial x_{k+1}} \right), \tag{9}$$

and

$$w^*(x_k) = \frac{1}{2}\gamma^{-2}h^T(x_k)\frac{\partial V^*(x_{k+1})}{\partial x_{k+1}}. \tag{10}$$

Then, the DT HJI equation becomes

$$V^*(x_k) = x_k^T Q x_k + 2\int_0^{u_k^*} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds - \gamma^2 w_k^{*T} w_k^* + V^*(x_{k+1}). \tag{11}$$

This equation reduces to GARE in the zero-sum linear quadratic case. However, in the general nonlinear case, the value function of the optimal control problem cannot be obtained.

For the problem of disturbance attenuation, we need the definition of the L_2-gain for DT nonlinear system.

Definition 2: (L_2-gain) The nonlinear system (1) with state feedback control policy u_k and disturbance policy $w_k \in L_2$ is said to have an L_2-gain less than or equal to γ if

$$\sum_{k=0}^{\infty}\left\{x_k^T Q x_k + 2\int_0^{u_k} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds\right\} \le \sum_{k=0}^{\infty}\gamma^2 w_k^T w_k. \tag{12}$$

If there exists a neighborhood around the origin such that $\forall w_k \in L_2$ the trajectories of the closed-loop system (1) starting from the origin remain in the same neighborhood, and (12) satisfies, the disturbance w_k is locally attenuated by a real value $\gamma > 0$. Let γ^* stands for the smallest γ for which the system is stabilized. Then, we can find a suboptimal H_∞ state feedback controller for any $\gamma > \gamma^*$.

III. Iterative Adaptive Dynamic Programming Algorithm for H_∞ Control

This section consists of three subsections. The iterative HDP algorithm is developed to solve the H_∞ control problem for DT nonlinear system in the first subsection. The corresponding convergence proof is presented in the second subsection, and the iterative DHP algorithm is given in the third subsection.

A. Derivation of Iterative HDP Algorithm for H_∞ Control

Since direct solution of the HJI equation is computationally intensive, we present an iterative HDP algorithm based on Bellman's principle of optimality.

First, we start with an initial value function $V_0(\cdot) = 0$ which is not necessarily optimal and set $\gamma > 0$. Then, we find $V_1(x_k)$ by solving

$$V_{i+1}(x_k) = \min_{u_k}\max_{w_k}\left\{x_k^T Q x_k + 2\int_0^{u_k} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds - \gamma^2 w_k^T w_k + V_i(f(x_k) + g(x_k)u_k + h(x_k)w_k)\right\} \tag{13}$$

with $i = 0$. The greedy policies $u_i(x_k)$ and $w_i(x_k)$ are updated by

$$u_i(x_k) = \bar{U}\boldsymbol{\psi}\left(-\frac{1}{2}(\bar{U}R)^{-1}g^T(x_k)\frac{\partial V_i(x_{k+1})}{\partial x_{k+1}}\right), \tag{14}$$

and

$$w_i(x_k) = \frac{1}{2}\gamma^{-2}h^T(x_k)\frac{\partial V_i(x_{k+1})}{\partial x_{k+1}}. \tag{15}$$

Therefore, $V_1(x_k)$ is calculated by

$$V_{i+1}(x_k) = x_k^T Q x_k + 2\int_0^{u_i(x_k)} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds - \gamma^2 w_i^T w_i + V_i(f(x_k) + g(x_k)u_i(x_k) + h(x_k)w_i(x_k)) \tag{16}$$

with $i = 0$. After $V_1(x_k)$ is found, we repeat the same value iteration process for $i = 1, 2, \ldots$. Furthermore, it should be satisfied that $V_i(0) = 0$, $\forall i \ge 0$. Note that i is the iteration index and k is the time index. As a value iteration algorithm, this iterative ADP algorithm does not require an initial stabilizing controller. In the next section we will prove the convergence, i.e., $V_i \to V^*$, $u_i \to u^*$ and $w_i \to w^*$ as $i \to \infty$.

B. Convergence Analysis of Iterative HDP Algorithm for H_∞ Control

Theorem 1: (Monotonicity Property) Define the control policy sequence $\{u_i\}$ as in (14), the disturbance policy sequence $\{w_i\}$ as in (15), and the value function sequence $\{V_i\}$ as in (16) with $V_0(\cdot) = 0$. Then $V_{i+1}(x_k) \ge V_i(x_k)$, $\forall i$ and x_k.

Proof: It is easy to see that $V_1(x_k) \ge V_0(x_k)$ as $u_0(x_k) = w_0(x_k) = 0$. Assume that $V_i(x_k) \ge V_{i-1}(x_k)$, $\forall i$ and x_k.

$$\begin{aligned} V_{i+1}(x_k) &= \min_{u_k}\max_{w_k}\left\{x_k^T Q x_k + 2\int_0^{u_k} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds - \gamma^2 w_k^T w_k + V_i(f(x_k) + g(x_k)u_k + h(x_k)w_k)\right\} \\ &\ge \min_{u_k}\max_{w_k}\left\{x_k^T Q x_k + 2\int_0^{u_k} \boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds - \gamma^2 w_k^T w_k + V_{i-1}(f(x_k) + g(x_k)u_k + h(x_k)w_k)\right\} \\ &= V_i(x_k). \end{aligned} \tag{17}$$

Therefore, we complete the proof by mathematical induction. ■

Next, we will demonstrate the convergence of iterative HDP algorithm for H_∞ control according to the work of [30] and [31].

Theorem 2: (Convergence Property) Suppose the condition $0 \le V^*(f(x) + g(x)u(x) + h(x)w(x)) \le \theta l(x, u, w)$ holds uniformly for some $0 < \theta < \infty$ and that $0 \le \alpha V^* \le V_0 \le \beta V^*$, $0 \le \alpha \le 1$ and $1 \le \beta < \infty$. The control policy sequence $\{u_i\}$, the disturbance policy sequence $\{w_i\}$ and the value function sequence $\{V_i\}$ are iteratively updated by (14), (15) and (16). Then the value function V_i approaches V^* according to the inequalities

$$\left[1 + \frac{\alpha - 1}{(1 + \theta^{-1})^i}\right]V^*(x) \le V_i(x) \le \left[1 + \frac{\beta - 1}{(1 + \theta^{-1})^i}\right]V^*(x). \tag{18}$$

Define $V_\infty(x_k) = \lim_{i\to\infty} V_i(x_k)$, then

$$V_\infty(x_k) = V^*(x_k). \tag{19}$$

Proof: First, we will demonstrate that the system defined in this paper satisfies the conditions of Theorem 2. According to Assumption 2 that $f+gu+hw$ is Lipschitz continuous, the system state cannot jump to infinity by any one step of finite control input, i.e., $f(x)+g(x)u(x)+h(x)w(x)$ is finite. Considering that $V^*(x,u,w)$ is finite for any finite state and control and that $l(x,u,w)$ is positive definite function, there exists some $0<\theta<\infty$ that makes $0\le V^*(f(x)+g(x)u(x)+h(x)w(x))\le \theta l(x,u)$ hold uniformly. Besides, for any finite initial value function V_0, there exist α and β such that $0\le \alpha V^*\le V_0\le \beta V^*$ is satisfied, $0\le\alpha\le 1$ and $1\le\beta<\infty$.

Next, we will demonstrate the left hand side of the inequality (18) by mathematical induction, i.e.,

$$\left[1+\frac{\alpha-1}{(1+\theta^{-1})^i}\right]V^*(x)\le V_i(x). \tag{20}$$

When $i=1$, since

$$\frac{\alpha-1}{1+\theta}(\theta l(x_k,u(x_k),w(x_k))-V^*(x_{k+1}))\le 0, 0\le\alpha\le 1, \tag{21}$$

and $\alpha V^*\le V_0$, $\forall x_k$, we have

$$\begin{aligned}
V_1(x_k)&=\min_{u_k}\max_{w_k}\left\{l(x_k,u_k,w_k)+V_0(x_{k+1})\right\}\\
&\ge\min_{u_k}\max_{w_k}\left\{l(x_k,u_k,w_k)+\alpha V^*(x_{k+1})\right\}\\
&\ge\min_{u_k}\max_{w_k}\left\{\left(1+\theta\frac{\alpha-1}{1+\theta}\right)l(x_k,u_k,w_k)\right.\\
&\quad\left.+\left(\alpha-\frac{\alpha-1}{1+\theta}\right)V^*(x_{k+1})\right\}\\
&=\left[1+\frac{\alpha-1}{(1+\theta^{-1})}\right]\min_{u_k}\max_{w_k}\left\{l(x_k,u_k,w_k)\right.\\
&\quad\left.+V^*(x_{k+1})\right\}\\
&=\left[1+\frac{\alpha-1}{(1+\theta^{-1})}\right]V^*(x_k).
\end{aligned} \tag{22}$$

Assume that the inequality (20) holds for $i-1$. Then, we have

$$\begin{aligned}
V_i(x_k)&=\min_{u_k}\max_{w_k}\left\{l(x_k,u_k,w_k)+V_{i-1}(x_{k+1})\right\}\\
&\ge\min_{u_k}\max_{w_k}\left\{\left[1+\frac{\alpha-1}{(1+\theta^{-1})^{i-1}}\right]V^*(x_{k+1})\right.\\
&\quad\left.+l(x_k,u_k,w_k)\right\}\\
&\ge\min_{u_k}\max_{w_k}\left\{\left[1+\frac{(\alpha-1)\theta^i}{(\theta+1)^i}\right]l(x_k,u_k,w_k)+\right.\\
&\quad\left.\left[1+\frac{\alpha-1}{(1+\theta^{-1})^{i-1}}-\frac{(\alpha-1)\theta^{i-1}}{(\theta+1)^i}\right]V^*(x_{k+1})\right\}\\
&=\left[1+\frac{(\alpha-1)\theta^i}{(\theta+1)^i}\right]\min_{u_k}\max_{w_k}\left\{l(x_k,u_k,w_k)\right.\\
&\quad\left.+V^*(x_{k+1})\right\}\\
&=\left[1+\frac{(\alpha-1)}{(1+\theta^{-1})^i}\right]V^*(x_k).
\end{aligned} \tag{23}$$

Thus, the left hand side of the inequality (18) is proved and the right hand side can be shown by the same way.

Lastly, we will demonstrate the convergence of value function as the iteration index i goes to infinity. When $i\to\infty$, for $0<\theta<\infty$, we have

$$\lim_{i\to\infty}\left[1+\frac{\alpha-1}{(1+\theta^{-1})^i}\right]V^*(x_k)=V^*(x_k) \tag{24}$$

and

$$\lim_{i\to\infty}\left[1+\frac{\beta-1}{(1+\theta^{-1})^i}\right]V^*(x_k)=V^*(x_k). \tag{25}$$

Therefore, we can get

$$V_\infty(x_k)=V^*(x_k). \tag{26}$$

The proof is completed. ■

Remark 1: From the above demonstration, we know that we can find upper and lower bounds for every iterative value function based on the optimal value function. As the iterative index i increases, the upper bound will exponentially approach the lower bound. When the iterative index i goes to infinity, the upper bound will be nearly equal to the lower bound, which is just the optimal value function. From Theorem 2, we can also find the convergence speed of the value function. According to the inequality (18), smaller θ will lead to faster convergence speed of the value function. Moreover, it should be mentioned that conditions in Theorem 2 can be satisfied according to Assumptions 1−3, which are some mild assumptions for general control problems. Specially, when $V_0(\cdot)=0$, we can have $\alpha=0,\beta=1$. From the inequality (18), we have

$$\left[1-\frac{1}{(1+\theta^{-1})^i}\right]V^*(x)\le V_i(x)\le V^*(x). \tag{27}$$

According to the results of Theorem 2, we can derive the following corollary.

Theorem 3: Define the control policy sequence $\{u_i\}$ as in (14), the disturbance policy sequence $\{w_i\}$ as in (15), and the value function $\{V_i\}$ as in (16) with $V_0(\cdot)=0$. If the system state x_k is controllable, then the control pair (u_i,w_i) converges to the saddle point (u^*,w^*) as $i\to\infty$.

Proof: According to Theorem 2, we have proved that $\lim_{i\to\infty}V_i(x_k)=V_\infty(x_k)=V^*(x_k)$, so

$$V_\infty(x_k)=\min_{u_k}\max_{w_k}\left\{l(x_k,u_k,w_k)+V_\infty(x_{k+1})\right\}. \tag{28}$$

That is to say the value function sequence $\{V_i\}$ converges to the optimal value function of the DT HJI equation. Considering (9) and (14), (10) and (15), the corresponding control pair (u_i,w_i) converges to the saddle point (u^*,w^*) as $i\to\infty$. ■

C. Derivation of Iterative DHP Algorithm for H_∞ Control

We find that there exists an integral term to compute in (16), i.e., $2\int_0^{u_i(x_k)}\boldsymbol{\psi}^{-T}(\bar{U}^{-1}s)\bar{U}Rds$. To reduce the computing burden, we develop iterative DHP algorithm to solve H_∞ control for DT nonlinear systems with control saturation

constraints. The costate function [16], [28] is denoted as

$$\begin{aligned}\lambda_{i+1}(x_k) &= \frac{\partial V_{i+1}(x_k)}{\partial x_k} \\ &= \frac{\partial l(x_k, u_i(x_k), w_i(x_k))}{\partial x_k} \\ &\quad + \left(\frac{\partial u_i(x_k)}{\partial x_k}\right)^T \frac{\partial l(x_k, u_i(x_k), w_i(x_k))}{\partial u_i(x_k)} \\ &\quad + \left(\frac{\partial w_i(x_k)}{\partial x_k}\right)^T \frac{\partial l(x_k, u_i(x_k), w_i(x_k))}{\partial w_i(x_k)} \\ &\quad + \left(\frac{\partial x_{k+1}}{\partial x_k}\right)^T \frac{\partial V_i(x_{k+1})}{\partial x_{k+1}} \\ &\quad + \left(\frac{\partial u_i(x_k)}{\partial x_k}\right)^T \left(\frac{\partial x_{k+1}}{\partial u_i(x_k)}\right)^T \frac{\partial V_i(x_{k+1})}{\partial x_{k+1}} \\ &\quad + \left(\frac{\partial w_i(x_k)}{\partial x_k}\right)^T \left(\frac{\partial x_{k+1}}{\partial w_i(x_k)}\right)^T \frac{\partial V_i(x_{k+1})}{\partial x_{k+1}}.\end{aligned} \tag{29}$$

Considering $u_i(x_k)$ and $w_i(x_k)$ in (14) and (15), we have

$$\lambda_{i+1}(x_k) = 2Qx_k + \left(\frac{\partial x_{k+1}}{\partial x_k}\right)^T \lambda_i(x_{k+1}), \tag{30}$$

where the integral term in (16) is removed. The greedy policies $u_i(x_k)$ and $w_i(x_k)$ are updated by

$$u_i(x_k) = \bar{U}\psi\left(-\frac{1}{2}(\bar{U}R)^{-1}g^T(x_k)\lambda_i(x_{k+1})\right), \tag{31}$$

and

$$w_i(x_k) = \frac{1}{2}\gamma^{-2}h^T(x_k)\lambda_i(x_{k+1}). \tag{32}$$

The initial costate function is chosen as $\lambda_0(\cdot) = 0$, and the iterative process of DHP algorithm is as the same as HDP algorithm.

Remark 2: In this paper, we assume that the value function $V(x)$ is smooth so that the costate function $\lambda(x)$ exists. Furthermore, the costate function sequence is also convergent, i.e., $\lambda_i \to \lambda^*$, $u_i \to u^*$ and $w_i \to w^*$ as $i \to \infty$, which is not included due to the space limitations.

IV. Implementation of the Iterative Adaptive Dynamic Programming Algorithm

In Section III, we have demonstrated the convergence of the iterative ADP algorithm under the assumption that the costate function (30), the control policy (31), and the disturbance policy (32) update equations can exactly be solved at each iteration. However, it is difficult to solve these equations for unknown nonlinear systems. In this section, we will use neural networks (NN) to implement the iterative ADP algorithm.

The structure diagram of the iterative DHP algorithm is given in Fig. 1. In the DHP algorithm, there are four neural networks, which are model network, critic network, action network, and disturbance network. The model network approximates the unknown nonlinear system, the critic network approximates the costate function $\lambda_i(x_k)$, the action network approximates the control policy $u_i(x_k)$, and the disturbance network approximates the disturbance policy $w_i(x_k)$.

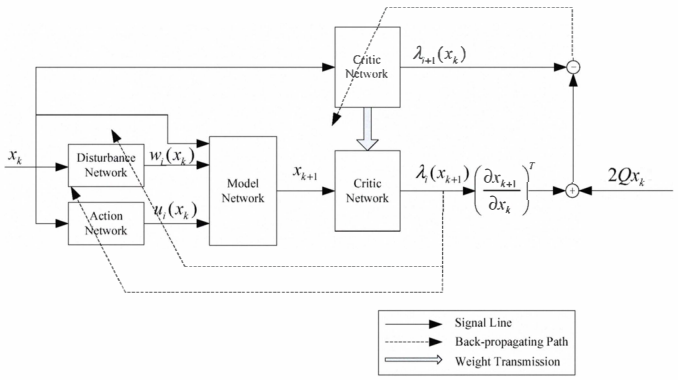

Fig. 1. The structure diagram of DHP algorithm

We chose the two-layer feed-forward NN as our function approximation scheme. The first step is to train the model network. The output of model network is denoted as

$$\tilde{x}_{k+1} = \omega_m^T\sigma(\bar{z}_k) = \omega_m^T\sigma(\nu_m^T z_k), \tag{33}$$

where $z_k = [x_k\ u_i(x_k)\ w_i(x_k)]^T$ is the input vector of model network. The error function for training model network is defined as

$$e_m(x_k) = \tilde{x}_{k+1} - x_{k+1}, \tag{34}$$

and the objective function to be minimized is defined as

$$E_m(x_k) = \frac{1}{2}e_m^T(x_k)e_m(x_k). \tag{35}$$

When the weights of model network converge, these weights are kept unchanged. Then, the estimated value of the control coefficient matrix $\hat{g}(x_k)$ is given by

$$\hat{g}(x_k) = \frac{\partial(\omega_m^T(k)\sigma(\bar{z}_k))}{\partial u_k}, \tag{36}$$

and the estimated value of the disturbance coefficient matrix $\hat{h}(x_k)$ is given by

$$\hat{h}(x_k) = \frac{\partial(\omega_m^T(k)\sigma(\bar{z}_k))}{\partial w_k}. \tag{37}$$

The output of the critic network is denoted as

$$\tilde{\lambda}_{i+1}(x_k) = \omega_{c(i+1)}^T\sigma(\nu_{c(i+1)}^T x_k). \tag{38}$$

The target costate function is given in (30), where $\lambda_i(x_{k+1}) = \omega_{c(i)}^T\sigma(\nu_{c(i)}^T x_{k+1})$. Then, the error function for training critic network is defined as

$$e_{c(i+1)}(x_k) = \tilde{\lambda}_{i+1}(x_k) - \lambda_{i+1}(x_k), \tag{39}$$

and the objective function to be minimized is defined as

$$E_{c(i+1)}(x_k) = \frac{1}{2}e_{c(i+1)}^T(x_k)e_{c(i+1)}(x_k). \tag{40}$$

In the action network, the state x_k is used as input to obtain the optimal control. The output can be formulated as

$$\tilde{u}_i(x_k) = \omega_{a(i)}^T\sigma\left(\nu_{a(i)}^T x_k\right). \tag{41}$$

The target of control input is calculated in (31). The error function of the action network can be defined as

$$e_{a(i)}(x_k) = \tilde{u}_i(x_k) - u_i(x_k). \tag{42}$$

The weights of the action network are updated to minimize the following objective function:

$$E_{a(i)}(x_k) = \frac{1}{2} e_{a(i)}^T(x_k) e_{a(i)}(x_k). \tag{43}$$

In the disturbance network, the state x_k is used as input to obtain the worst case disturbance policy. The output can be formulated as

$$\tilde{w}_i(x_k) = \omega_{d(i)}^T \sigma\left(\nu_{d(i)}^T x_k\right). \tag{44}$$

The target of disturbance input is calculated in (32). The error function of the disturbance network can be defined as

$$e_{d(i)}(x_k) = \tilde{w}_i(x_k) - w_i(x_k). \tag{45}$$

The weights of the disturbance network are updated to minimize the following objective function:

$$E_{d(i)}(x_k) = \frac{1}{2} e_{d(i)}^T(x_k) e_{d(i)}(x_k). \tag{46}$$

With these objective functions, many methods like gradient descent algorithm and Levenberg-Marquardt algorithm can be used to tune the weights of NN.

V. Conclusions

In this paper, the H_∞ control for discrete-time nonlinear systems using iterative adaptive dynamic programming algorithm is developed. The heuristic dynamic programming algorithm is derived to solve the Hamilton-Jacobi-Isaacs equation, and the convergence analysis is rigorously proved. The dual heuristic dynamic programming algorithm with non-quadratic performance functional is given to overcome the control saturation constraints. Four neural networks are used to approximate the unknown nonlinear system, the control policy, the disturbance policy, and the value function.

References

[1] F. Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: an introduction," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 39–47, May 2009.

[2] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, July 2009.

[3] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, Sept. 1997.

[4] F. L. Lewis and V. L. Syrmos, *Optimal Control*. New York: Wiley, 1995.

[5] R. Beard, G. Saridis, and J. Wen, "Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation," *Automatica*, vol. 33, no. 12, pp. 2158–2177, Aug. 1997.

[6] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 32, no. 2, pp. 140–153, May 2002.

[7] 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, 779–791, May 2005.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning : An Introduction*. Cambridge, MA: MIT Press, 1998.

[9] F. L. Lewis, G. Lendaris, and D. Liu, "Special issue on approximate dynamic programming and reinforcement learning for feedback control," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 896–897, Aug. 2008.

[10] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, Feb. 2009.

[11] D. Vrabie and F. L. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, Apr. 2009.

[12] 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, May 2010.

[13] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[14] T. Dierks, B. T. Thumati, and S. Jagannathan, "Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence," *Neural Networks*, vol. 22, no. 5–6, pp. 851–860, July–Aug. 2009.

[15] H. Zhang, Q. Wei, and Y. Luo, "A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 937–942, Aug. 2008.

[16] H. Zhang, Y. Luo, and D. Liu, "Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints," *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1490–1503, Sept. 2009.

[17] F. Wang, N. Jin, D. Liu, and Q. Wei, "Adaptive dynamic programming for finite-horizon optimal control of discrete-time nonlinear sytems with ε-error bound," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 24–36, Jan. 2011.

[18] W. Lin and C. I. Byrnes, "H_∞ control of discrete-time nonlinear systems," *IEEE Trans. Autom. Control*, vol. 41, no. 4, pp. 494–510, Apr. 1996.

[19] A. J. van der Shaft, "L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control," *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 770–784, June 1992.

[20] J. C. Willems, "Dissipative dynamical systems part 1: general theory," *Archive for ratioinal mechanics and analysis*, vol. 45, no. 1, pp. 321–351, 1972.

[21] T. Basar and P. Bernhard, *H_∞ Optimal Conrol and Related Minimax Design Problems: A Dynamic Game Approach*, Second Edition, Boston: 1995.

[22] D. Varbie and F. L. Lewis, "Adaptive dynamic programming algorithm for finding online the equilibrium solution of the two-player zero-sum differential game," in *Proceedings of International Joint Conference on Neural Networks*, Barcelona, Spain, July 2010, pp. 1–8.

[23] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Policy iterations and the Hamilton-Jacobi-Isaacs equation for H_∞ state feedback control with input saturation," *IEEE Trans. Autom. Control*, vol. 51, no. 12, pp. 1989–1995, Dec. 2006.

[24] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Neurodynamic progarmming and zero-sum games for constrained control systems," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1243–1252, July 2008.

[25] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, no. 1, pp. 207–214, Jan. 2011.

[26] K. G. Vamvoudakis and F. L. Lewis, "Online solution of nonlinear two-player zero-sum games using synchronous policy iteration," *International Journal of Robust and Nonlinear Control*, 2011. (in press, DOI: 10.1002/rnc.1760)

[27] T. Dierks and S. Jagannathan, "Optimal control of affine nonlinear continuous-time systems using an online Hamilton-Jacobi-Isaacs formulation," in *IEEE Conference on Decision and Control*, Atlanta, GA, Dec. 2010, pp. 3048–3053.

[28] A. Al-Tamimi, M. Abu-Khalaf, and F. L. Lewis, "Adaptive critic designs for discretet-time zero-sum games with application to H_∞ control," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 1, pp. 240–247, Feb. 2007.

[29] S. Mehraeen, T. Dierks, S. Jagannathan, and M. L. Crow, "Zero-sum two-player game theoretic formulation of affine nonlinear discrete-time systems using neural networks," in *Proceedings of International Joint Conference on Neural Networks*, Barcelona, Spain, pp. 1–8, July 2010.

[30] B. Lincoln and A. Rantzer, "Relaxing dynamic programming," *IEEE Trans. Autom. Control*, vol. 51, no. 8, pp. 1249–1260, Aug. 2006.

[31] A. Rantzer, "Relaxed dynamic programming in switching systems," *Proc. Inst. Elect. Eng.*, vol. 153, pp. 567–574, 2006.