Policy Iteration for H_∞ Optimal Control of Polynomial Nonlinear Systems via Sum of Squares Programming

Yuanheng Zhu, *Member, IEEE*, Dongbin Zhao, *Senior Member, IEEE*, Xiong Yang, and Qichao Zhang

***Abstract*—Sum of squares (SOS) polynomials have provided a computationally tractable way to deal with inequality constraints appearing in many control problems. It can also act as an approximator in the framework of adaptive dynamic programming. In this paper, an approximate solution to the H_∞ optimal control of polynomial nonlinear systems is proposed. Under a given attenuation coefficient, the Hamilton–Jacobi–Isaacs equation is relaxed to an optimization problem with a set of inequalities. After applying the policy iteration technique and constraining inequalities to SOS, the optimization problem is divided into a sequence of feasible semidefinite programming problems. With the converged solution, the attenuation coefficient is further minimized to a lower value. After iterations, approximate solutions to the smallest L_2-gain and the associated H_∞ optimal controller are obtained. Four examples are employed to verify the effectiveness of the proposed algorithm.**

***Index Terms*—Adaptive dynamic programming (ADP), H_∞ optimal control, policy iteration (PI), polynomial nonlinear systems, sum of squares (SOS).**

I. Introduction

IN THE field of robust control, H_∞ control [1], [2] is a powerful tool in solving the disturbance attenuation problem occurred in many practical systems. In these systems, an exogenous disturbance is mixed into the dynamics. An H_∞ controller renders the L_2-gain, or H_∞ norm of the system from disturbance to output within a prescribed constant in the time domain. One approach to synthesize such a controller is by solving the Hamilton–Jacobi–Isaacs (HJI) equation [3], [4], which is a first-order, nonlinear partial differential equation. For general cases, it is difficult to give a universal analytic solution to the HJI equation, not to mention finding the H_∞ optimal controller with the smallest L_2-gain. When considering systems with linear dynamics, the HJI equation is reduced to an algebraic Riccati equation (ARE) [5], and the linear matrix inequality (LMI) toolbox [6], [7] provides a tractable approach to find its solution. To deal with nonlinear dynamics, numerical methods are proposed to find the approximate solutions to the HJI equation [8], [9].

Adaptive dynamic programming (ADP) exhibits promising performance in solving many control problems, including H_∞ control (see [10]–[12] and references quoted therein). Among the existing literature, policy iteration (PI) is an effective approach to solve nonlinear partial differential equations, including the Hamilton–Jacobi–Bellman (HJB) equation in the optimal control [13]–[16], the HJI equation in H_∞ control [17]–[19], and the Hamilton–Jacobi (HJ) equation in the nonzero sum game problem [20], [21]. PI mainly includes two steps. One calculates a given policy's value function, and the other updates the policy based on the result of the first step. Since the policy is monotonically improved after each iteration, it will finally converge to the optimal solution.

In practical applications, approximation techniques have to be used to efficiently describe the complicated value functions and policy functions appearing in PI, and neural network (NN) technique [22]–[24] is the most common one. A group of basis functions together with the corresponding weights define a network. The weights are determined on the basis of system data that are collected offline or online. Our previous work [25] has proved the convergence of ADP under the approximation of NNs in H_∞ control. Unfortunately, the universal approximation property of NNs is held only in a compact set, not over the entire state space. More importantly, NNs rarely consider the non-negativity of target functions, which is an essential requirement for value functions in PI and many other similar functions appearing in control theory.

Recent exciting developments on positive polynomial theory, especially the sum of squares (SOS) theory [26], [27] have provided a feasible solution to ensure the global non-negativity for polynomial functions. A sufficient condition for positivity of a polynomial is to be expressible as a sum of squared polynomials. The existence of an SOS decomposition is equivalent to a semidefinite programming (SDP) feasible problem. Many toolboxes such as SOSTOOLS [28] and GloptiPoly [29], have been fully developed to solve the problem. With the development of SOS techniques, a new avenue is opened

Manuscript received July 1, 2016; revised October 20, 2016; accepted December 19, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61603382, Grant 61273136, Grant 61573353, Grant 61533017, and Grant 61503379, and in part by the Early Career Development Award of SKLMCCS. This paper was recommended by Associate Editor H. Zhang.

Y. Zhu, D. Zhao, and Q. Zhang are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yuanheng.zhu@ia.ac.cn; dongbin.zhao@ia.ac.cn; zhangqichao2014@ia.ac.cn).

X. Yang is with the School of Electrical Engineering and Automation, Tianjin University, Tianjin 300072, China (e-mail: xiong.yang@tju.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2643687

to analyze and synthesize controllers for nonlinear systems with polynomial dynamics. In [30]–[32], state dependent LMIs are constructed to design controllers via control Lyapunov functions. The controllers are computed after relaxation of inequalities to SOS constraints. In [33], the Bellman equation is relaxed to an inequality, and an approximate value function is computed offline by solving a semidefinite program. Stochastic optimal control is also studied in [34] and [35].

The success of SOS polynomials attracts interests of researchers from the ADP community. Jiang and Jiang [36] proposed an SOS-based global ADP algorithm to solve the HJB equation using online data in the absence of system dynamics information. The learned policy is globally stabilizing for a general class of polynomial nonlinear systems. Another combination of ADP and SOS is seen in [37] via the backstepping technique to control a class of block strict-feedback nonlinear systems. It is worth noting that in the past few years, polynomials have been widely used in ADP algorithms as basis functions for NN approximation. The polynomial coefficients therein are determined in the principle of minimizing approximation errors. While in SOS programs, the coefficients are constrained in the convex SOS feasible set, which guarantees non-negativity of the resultant approximation. However, for the H_∞ optimal control problem, SOS-based ADP is rarely considered, which motivates this paper.

In this paper, we consider polynomial nonlinear systems and deal with the H_∞ optimal control problem. The contribution is threefold. First, given a prescribed attenuation coefficient, the H_∞ control is relaxed to an optimization problem by treating the disturbance as independent variables. Second, an SOS-based PI is proposed to solve the relaxed L_2-gain optimization problem. By adding SOS constraints to the set of inequality conditions, each iteration becomes a semidefinite program. Third, once we obtain the approximate solution to the H_∞ control, we further minimize the L_2-gain to a lower value using SDP. A new H_∞ control problem is formulated on the basis of the new L_2-gain. After iterations, the H_∞ optimal control problem in finding the smallest L_2-gain and the associated controller is approximately solved in a numerical way. It is the first time that ADP successfully solves this problem. Examples on a scalar system and a linear system verify the effectiveness of finding the optimal solutions. A nonlinear example shows our algorithm is capable of finding an H_∞ controller with a smaller L_2-gain than other methods. A comparison with the NN-based ADP is also presented. In addition, an application to the active suspension problem reveals our algorithm is capable to solve realistic problems.

The remainder of this paper is organized as follows. Section II describes the basic knowledge of the H_∞ control and the HJI equation. Section III relaxes the HJI equation to an optimization problem. Section IV gives a PI approach and introduces SOS constraints to solve the optimization problem. Section V proposes an approximate solution to the H_∞ optimal control of polynomial nonlinear systems. Section VI presents four examples to test the performance of our algorithm. Section VII gives the remarkable conclusion in the end.

II. Preliminary of H_∞ Control and HJI Equation

We consider the continuous-time nonlinear system in the form

$$\begin{aligned} \dot{x} &= f(x) + g(x)u + k(x)d \\ z &= h(x) \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ is the control input, $d(t) \in \mathbb{R}^p$ is the external disturbance, and $z(t) \in \mathbb{R}^q$ is the output. The system is assumed to be zero-state observable. Dynamics functions f, g, and k are assumed to be Lipschitz continuous and $f(0) = 0$. Then $x = 0$ is the equilibrium. h is the output function. If there is no confusion in the context, argument x is omitted in functions.

For the H_∞ control, it is desired to find a controller that stabilizes the system at $d(t) = 0$ and renders the cost

$$J = \int_0^T \left(\|h\|^2 + u^T R u - \gamma_0^2 \|d\|^2\right) dt \tag{2}$$

nonpositive for $x(0) = 0$ and $\forall d \in L_2(0, T)$. R is a positive symmetric matrix. γ_0 is a positive constant, which is also known as *attenuation coefficient*. If such a controller exists, the closed-loop system is said to have L_2-gain $\leq \gamma_0$. H_∞ *optimal control* is to find the smallest γ^* and the associated controller such that the above problem is still solvable.

Assumption 1: For the system (1), suppose there exists a controller u such that the closed-loop system has L_2-gain $\leq \gamma_0$ and it globally stabilizes the system when $d(t) = 0$.

As shown in [4], given a fixed γ_0, a sufficient condition for solvability of the L_2-gain problem is that there exists a smooth positive semidefinite solution V^* to the HJI equation

$$\begin{aligned} &\left(\nabla V^*\right)^T f - \frac{1}{4}\left(\nabla V^*\right)^T g R^{-1} g^T \nabla V^* \\ &\quad + \frac{1}{4\gamma_0^2}\left(\nabla V^*\right)^T k k^T \nabla V^* + \|h\|^2 = 0,\ V^*(0) = 0. \end{aligned} \tag{3}$$

If such a solution exists, the controller

$$u^*(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V^*(x)$$

is the H_∞ controller with L_2-gain $\leq \gamma_0$.

Among the various approaches that have been proposed to solve the nonlinear HJI equation, PI is the most commonly used one. Van der Schaft [4] and Abu-Khalaf *et al.* [17] divided the HJI equation into an infinite sequence of PIs on the control input following:

1) *Policy Evaluation:*

$$\begin{aligned} &(\nabla V_i)^T (f + g u_i) + \frac{1}{4\gamma_0^2} (\nabla V_i)^T k k^T \nabla V_i \\ &\quad + \|h\|^2 + u_i^T R u_i = 0,\ V_i(0) = 0. \end{aligned} \tag{4}$$

2) *Policy Improvement:*

$$u_{i+1}(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V_i(x). \tag{5}$$

V_i is also the *available storage* function for u_i. Some useful facts are listed in Theorem 1 that can be seen as an extension of [17]. The previous work considers input-saturated systems. We generalize the results to system (1) without control constraints. Similar research is also seen in [9], but it fails in analyzing the L_2-gain property of their policies. The proof of the theorem is given in the Appendix.

Theorem 1: Let u_1 satisfy Assumption 1. Suppose for every i, there exists a smooth solution $V_i \geq 0$ to (4). Following the PI of (4) and (5), the following properties hold $\forall i$:

1) u_{i+1} is a globally stabilizing controller and makes the closed-loop system L_2-gain $\leq \gamma_0$;
2) $V_i(x) \geq V_{i+1}(x) \geq 0$, $\forall x \in \mathbb{R}^n$;
3) the sequence $\{V_i\}$ is convergent and the limitation $V^o = \lim_{i\to\infty} V_i$ satisfies the HJI equation

$$\left(\nabla V^o\right)^T f - \frac{1}{4}\left(\nabla V^o\right)^T gR^{-1}g^T\nabla V^o + \frac{1}{4\gamma_0^2}\left(\nabla V^o\right)^T kk^T\nabla V^o + \|h\|^2 = 0. \quad (6)$$

Notice that (4) is also a nonlinear partial differential equation. Reference [9] further introduces an additional PI on disturbance to iteratively solve (4) by a sequence of linear partial differential equations

$$\nabla V_{i,j}^T\left(f + gu_i + kd_{i,j}\right) + \|h\|^2 + u_i^T Ru_i - \gamma_0^2\|d_{i,j}\|^2 = 0,\ V_{i,j}(0) = 0 \quad (7)$$

and the disturbance update law

$$d_{i,j+1}(x) = \frac{1}{2\gamma_0^2}k^T(x)\nabla V_{i,j}(x).$$

Subscript i indicates the outer iteration on policies, while j indicates the inner iteration on disturbance. When updating disturbance, i keeps constant. The complete analysis is available in [9] and [17].

Note that even for linear partial differential equations, it is still difficult to find analytical solutions. So approximation techniques like NNs have to be used. In order to achieve small approximation errors, a large number of basis functions are necessary for NNs and the computation is heavy. In addition, disturbance d is treated as a state-dependent function, just like the control policy u. It naturally increases the computational burden.

III. Relaxed L_2-Gain Optimization Problem

In this section, we introduce a relaxation to the HJI equation. By completing the square, (3) becomes

$$\left(\nabla V^*\right)^T f - \frac{1}{4}\left(\nabla V^*\right)^T gR^{-1}g^T\nabla V^* + \left(\nabla V^*\right)^T kd + \|h\|^2 - \gamma_0^2\|d\|^2 = -\gamma_0^2\left\|d - \frac{1}{2\gamma_0^2}k^T\nabla V^*\right\|^2 \leq 0.$$

Based on that, a relaxed L_2-gain optimization problem is formulated.

Definition 1: The *relaxed L_2-gain optimization problem* is defined as

$$\min_V \quad \int_\Omega V dx \quad (8)$$

$$\text{s.t.} \quad -\nabla V^T f + \frac{1}{4}\nabla V^T gR^{-1}g^T\nabla V - \nabla V^T kd - \|h\|^2 + \gamma_0^2\|d\|^2 \geq 0 \quad (9)$$

$$V \geq 0 \quad (10)$$

where $\Omega \in \mathbb{R}^n$ is an arbitrary compact set containing the origin. It represents the area where the H_∞ control is mostly interested, or in other words, the area where disturbance attenuation is mostly expected.

In difference to the HJI equation, the equality constraint is relaxed to an inequality in (9). The inequality condition should be held for arbitrary disturbance, so d acts as an independent variable just like x. In what follows, a useful lemma is presented.

Lemma 1: Given a control policy $u(x)$, if there exists $V \geq 0$ such that $\mathcal{L}(V, u, \gamma_0) \geq 0$, $\forall x$ and $\forall d$, then u is globally stabilizing and the associated closed-loop system has L_2-gain $\leq \gamma_0$. $\mathcal{L}(V, u, \gamma_0)$ is defined as

$$\mathcal{L}(V, u, \gamma_0) = -\nabla V^T(x)(f(x) + g(x)u(x) + k(x)d) - \|h(x)\|^2 - u^T(x)Ru(x) + \gamma_0^2\|d\|^2.$$

Proof: According to the definition of $\mathcal{L}$, if $d = 0$, it has

$$\nabla V^T(f + gu) = -\mathcal{L}(V, u, \gamma_0) - \|h\|^2 - u^T Ru.$$

Under the zero-state observability, if $\mathcal{L}(V, u, \gamma_0) \geq 0$, the closed-loop system is stabilizable under the well-defined Lyapunov function V. When considering the disturbance

$$\nabla V^T(f + gu + kd) = -\mathcal{L}(V, u, \gamma_0) - \|h\|^2 - u^T Ru + \gamma_0^2\|d\|^2.$$

Integrate both sides over the interval $[0, T]$

$$V(x(T)) - V(x(0)) \leq \int_0^T \left(-\|h\|^2 - u^T Ru + \gamma_0^2\|d\|^2\right)dt.$$

Under the positive assumption of V, u renders the system L_2-gain $\leq \gamma_0$. ■

Theorem 2: Suppose for arbitrary initial policy u_0 satisfying Assumption 1, the unique limit solution V^o to PI in (4) and (5) exits and is positive smooth. The following facts hold for the relaxed L_2-gain optimization problem.

1) The problem has a nonempty feasible set.
2) Let V be a feasible solution. Then $u' = -(1/2)R^{-1}g^T\nabla V$ is globally stabilizing and the closed-loop system has L_2-gain $\leq \gamma_0$.
3) V^o is the optimal solution to the problem.

Proof:

1) Obviously, V^o is a feasible solution to (8)–(10), so the feasible set is nonempty.
2) For a feasible solution V, the inequality condition (9) can be rewritten as

$$\mathcal{L}\left(V, u', \gamma_0\right) \geq 0.$$

By Lemma 1, the stabilizing and L_2-gain properties of the control policy u' are ensured.

3) From 2), for any feasible solution V, $u' = -(1/2)R^{-1}g^T\nabla V$ is an L_2-gain $\leq \gamma_0$ controller. In

addition, the following inequality holds for arbitrary d according to (9):

$$\nabla V^T(f+gu') + \frac{1}{4\gamma_0^2}\nabla V^T kk^T \nabla V + \|h\|^2 + u'^T R u' \le \gamma_0^2 \left\| d - \frac{1}{2\gamma_0^2} k^T \nabla V \right\|^2.$$

If we let $d = (1/2\gamma_0^2)k^T \nabla V$

$$\nabla V^T(f+gu') + \frac{1}{4\gamma_0^2}\nabla V^T kk^T \nabla V + \|h\|^2 + u'^T R u' \le 0.$$

Thus, V is a possible storage function for u'. Following the results of Theorem 1, if we start the PI with $u_0 = u'$ and compute the value sequence $\{V_i\}$ using (4) and (5), we have $V^o \le \cdots \le V_{i+1} \le V_i \le \cdots \le V_0 \le V$. Since the inequality $V^o \le V$ holds for arbitrary feasible solutions, V^o is the optimal solution of the problem. ■

Notice that (9) is nonlinear in ∇V. Based on the PI technique, the nonlinear inequality constraint can be divided into a sequence of linear inequalities. But testing the non-negativity of a function is still NP-hard. With the development of the polynomial theory, the positivity of a polynomial can be ensured by testing for an SOS decomposition, which is a tractable SDP problem. In what follows, an SOS-based PI for polynomial nonlinear systems is proposed to give an approximate solution to the relaxed L_2-gain optimization problem.

IV. Policy Iteration for Relaxed L_2-Gain Optimization Using SOS

A multivariable polynomial $p(x)$ is called an SOS if there exist polynomials $f_1(x), \ldots, f_m(x)$ such that

$$p(x) = \sum_{i=1}^{M} f_i^2(x).$$

It is clear that an SOS polynomial is *globally non-negative*. But the converse is not true. An SOS decomposition of a polynomial is equivalent to an SDP feasible problem. SOS polynomials provide a computationally tractable way to deal with non-negativity constraints. More detailed descriptions about SOS can refer to [26] and [27].

Now, we suppose that the dynamics and output functions in (1) are all polynomials. We also define the value function in the polynomial form

$$V(x) = \sum_{j=1}^{N} c_j m_j(x)$$

where $m_j(x)$ are predefined monomials in terms of x and c_j are coefficients to be determined. Using more monomials is helpful to reduce errors when approximating true functions, but more coefficients are to be determined, increasing the computational cost. In practice, designers should make a balance between the computational burden and the control performance.

Based on the SOS theory, a new assumption is made.

Assumption 2: For system (1), there exist polynomial functions $V_0(x)$ and $u_1(x)$ such that V_0 is SOS and $\mathcal{L}(V_0, u_1, \gamma_0)$ is SOS.

Prajna *et al.* [30] provided a computational method to find such V_0, u_1, and γ_0 that satisfy Assumption 2 if they exist. Interested readers may refer to their work.

From Lemma 1, Assumption 2 implies u_1 is an L_2-gain $\le \gamma_0$ controller. The SOS-based PI for the relaxed L_2-gain optimization problem with the initial u_1 is given as follows.

Algorithm 1 (Relaxed L_2-Gain Optimization Problem):

1) *Policy Evaluation:* For $i = 1, 2, \ldots,$ solve the SOS program for the optimal solution V_i

$$\min_V \int_\Omega V dx \tag{11}$$

$$\text{s.t.} \quad \mathcal{L}(V, u_i, \gamma_0) \text{ is SOS} \tag{12}$$

$$V_{i-1} - V \text{ is SOS} \tag{13}$$

$$V \text{ is SOS.} \tag{14}$$

2) *Policy Improvement:* Update the control policy by

$$u_{i+1}(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla V_i(x). \tag{15}$$

Go back to 1) with $i = i + 1$.

The terminal condition for the algorithm is that the difference of monomial coefficients $\{c_j\}$ of V_i between two successive iterations is less than a prescribed threshold. In comparison with the original PI in solving the HJI equation in Section II, the above SOS-based method replaces the HJ equation (4) with a relaxed optimization problem (11)–(14). The inequalities are restricted to SOS constraints so that the problem is computationally solvable by SDP.

Theorem 3: Suppose there exists a positive smooth solution V^o to (6). Under Assumption 2, for $i = 1, 2, \ldots$:

1) the SOS program (11)–(14) has a nonempty feasible set;
2) the controller u_{i+1} is globally stabilizing and the closed-loop system has L_2-gain $\le \gamma_0$;
3) the optimal solution V_i at each iteration always has $V_i(x) \ge V_{i+1}(x) \ge 0, \forall x \in \mathbb{R}^n$;
4) the sequence $\{V_i\}$ is convergent and V^o provides the lower bound for the limitation, i.e., $V_\infty(x) = \lim_{i\to\infty} V_i(x) \ge V^o(x), \forall x \in \mathbb{R}^n$.

Proof:

1) For $i = 1$, under Assumption 2, V_0 is a feasible solution to (11)–(14). So the feasible set is nonempty. When $i \ge 2$, suppose at the ith iteration the SOS program still has a nonempty feasible set and V_i is the optimal solution. After substituting $u_{i+1} = -(1/2)R^{-1}g^T\nabla V_i$ into (12), it has

$$\mathcal{L}(V_i, u_{i+1}, \gamma_0) = \mathcal{L}(V_i, u_i, \gamma_0) + (u_i - u_{i+1})^T R (u_i - u_{i+1}). \tag{16}$$

So $\mathcal{L}(V_i, u_{i+1}, \gamma_0)$ is also an SOS. By the definition of the SOS program, V_i is a feasible solution to (11)–(14) at the $(i+1)$th iteration. By induction, the first statement is true.

2) From the analysis of 1), $\mathcal{L}(V_i, u_{i+1}, \gamma_0) \ge 0$. The conclusion is derived directly from Lemma 1.

3) Constraints (13) and (14) imply $V_{i-1} \geq V_i \geq 0, \forall i$.
4) As shown in the proof of 3), the sequence $\{V_i\}$ is decreasing and lower bounded by 0. So there exists $V_\infty(x) = \lim_{i\to\infty} V_i(x)$. Furthermore, V_∞ satisfies

$$\begin{aligned}&\mathcal{L}\left(V_\infty, -\frac{1}{2}R^{-1}g^T\nabla V_\infty, \gamma_0\right) \text{ is SOS}\\ &V_\infty \text{ is SOS}\end{aligned}$$

indicating V_∞ is a feasible solution to (8)–(10). According to 3) of Theorem 2, $V_\infty \geq V^o$. ■

Now, the HJI equation is first relaxed to an optimization problem, and then relaxed to a sequence of SOS programs that are computationally solvable by SDP. The objective function (11) restricts the algorithm to find the closest solution to the HJI solution V^o in the feasible set at each iteration. After iterating, the result is continuously decreased and it approaches more and more closely to V^o. Note that the final converged solution V_∞ to (11)–(14) is not necessarily equal to V^o. But the associated controller $u_\infty = -(1/2)R^{-1}g^T\nabla V_\infty$ has the closest possible storage function to V^o than any other controllers generated during the PI. So u_∞ possesses promising performance in the H_∞ control.

The SOS constraint on $\mathcal{L}(V, u_i, \gamma_0)$ in (12) includes two variables: x and d. The relaxation of disturbance being independent variables is first presented in [4]. But the author fails in finding an effective way to deal with the inequality. In this paper, we successfully convert the problem to a feasible SDP problem.

In the existing literature, polynomial NNs are widely used as approximation of the value function in the form

$$\hat{V}(x) = \sum_{j=1}^{N} c_j\phi_j(x)$$

where $\phi_j(x)$ is the polynomial basis function and c_j is the corresponding weight. The approximation $\hat{V}$ is also a polynomial in x. But the weights c_j are determined mostly in the principle of minimizing the approximation error, which occurs when substituting $\hat{V}$ into the HJI equation or other relevant equations. The resultant weights cannot ensure the non-negativity of $\hat{V}$.

V. Approximate Solution to H_∞ Optimal Control

In the previous section, we propose an approximate solution to the HJI equation under a given attenuation coefficient. The value of γ_0 needs to be prescribed and an initial L_2-gain $\leq \gamma_0$ controller is required. In addition, the H_∞ optimal control problem in finding the smallest γ^* is still unsolved.

From the analysis in the proof of Theorem 3, (16) implies

$$\mathcal{L}(V_i, u_{i+1}, \gamma_0) \geq \mathcal{L}(V_i, u_i, \gamma_0).$$

According to the definition of $\mathcal{L}$, the above inequality implies that it is possible to find a smaller $\gamma' \leq \gamma_0$ to ensure

$$\mathcal{L}\left(V_i, u_{i+1}, \gamma'\right) \geq 0.$$

Thus, the conclusion of Lemma 1 holds for u_{i+1} with γ'. In other words, the improved policy after the PI in (11)–(15) allows a smaller attenuation coefficient to achieve the L_2-gain performance. Based on that fact, we use the final converged u_∞ to construct an SOS program in finding a smaller attenuation coefficient for the H_∞ control. The problem is formulated as

$$\begin{aligned}\min_{\gamma}\quad & \gamma\\ \text{s.t.}\quad & \mathcal{L}(V, u_\infty, \gamma) \text{ is SOS}\\ & V \text{ is SOS.}\end{aligned}$$

Since γ_0 is a feasible solution, the problem has a nonempty feasible set. We denote the optimal solution found by SDP as γ_1, and it is no greater than γ_0. With the new γ_1, we can say the closed-loop system with u_∞ has L_2-gain $\leq \gamma_1$.

Inspired by the idea of PI, we propose the following iterative SOS-based algorithm to obtain the approximate solution of the H_∞ optimal control. A two-loop iteration is included, where the inner loop searches the HJI solution under a given attenuation coefficient, and the outer loop minimizes the coefficient under the result of the inner loop. The whole process is listed in Algorithm 2. The initial controller $u^{(0)}$ is assumed to be globally stabilizing and has a finite L_2-gain.

Algorithm 2 (H_∞ Optimal Control):

1) For $l = 0, 1, \ldots,$ find the optimal solution γ to the SOS program

$$\min_{\gamma}\quad \gamma \tag{17}$$

$$\text{s.t.}\quad \mathcal{L}\left(V, u^{(l)}, \gamma\right) \text{ is SOS} \tag{18}$$

$$V \text{ is SOS.} \tag{19}$$

Denote $\gamma^{(l)} = \gamma$, and $V_0^{(l)} = V$. Let $u_1^{(l)} = u^{(l)}$.

a) For $i = 1, 2, \ldots,$ find the optimal solution V to the SOS program

$$\min_{V}\quad \int_\Omega V dx \tag{20}$$

$$\text{s.t.}\quad \mathcal{L}\left(V, u_i^{(l)}, \gamma\right) \text{ is SOS} \tag{21}$$

$$V_{i-1}^{(l)} - V \text{ is SOS} \tag{22}$$

$$V \text{ is SOS.} \tag{23}$$

Then, denote $V_i^{(l)} = V$.

b) Update the control policy by

$$u_{i+1}^{(l)}(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla V_i^{(l)}(x). \tag{24}$$

If $\|V_i^{(l)} - V_{i-1}^{(l)}\| \geq \varepsilon$, go back to step a) and let $i = i + 1$. Otherwise, stop the inner iteration and let $u^{(l+1)} = u_{i+1}^{(l)}$.

2) If $\|\gamma^{(l)} - \gamma^{(l-1)}\| \leq \varepsilon$, the convergence is reached and the algorithm is terminated. Output the approximate optimal solution $\gamma^{(l)}$ and $u^{(l+1)}$. Otherwise, replace l by $(l + 1)$ and go back to step 1).

In the above algorithm, an initial globally stabilizing $u^{(0)}$ is required, and we assume it has finite L_2-gain. But the exact L_2-gain value is not needed and is computed using SOS program (17)–(19). After the learning process, we let the converged $\gamma^{(l)}$ be the near-optimal L_2-gain and let $u^{(l+1)}$ be the associated H_∞ near-optimal controller. It should be mentioned that our algorithm solves the H_∞ optimal control in a numerical way with the help of SOS programming. In the conventional ADP literature, the optimal γ^* is suggested to be searched using the bisection method [9]. It has to test the validity of the L_2-gain problem for every candidate γ, which inevitably results in heavy computational burden.

VI. Experimental Study

In this part, we use four polynomial examples to test the performance of our algorithm. The first is a scalar nonlinear system whose exact optimal γ^* can be analytically solved. The second is a 3-D F-16 system with the linear dynamics, so the HJI equation is reduced to an ARE. By using the LMI toolbox, the optimal solution is easily obtained. The third is a 2-D nonlinear system. We compare the results with another SOS-based algorithm [30] and with the conventional NN-based ADP [17]. The last is an application to the active suspension problem. SOSTOOLS [28] is used throughout our experiments to solve SOS programs.

A. Scalar Nonlinear Example

Consider the system with dynamics

$$\dot{x} = -x^3 + u + d$$
$$z = x^3$$

where $x \in \mathbb{R}$ is the state variable, $u \in \mathbb{R}$ is the control input, $d \in \mathbb{R}$ is the disturbance signal, and z denotes the output signal. Let $R = I$. The corresponding HJI equation is

$$-\nabla V \cdot x^3 - \frac{1}{4}\nabla V^2 + \frac{1}{4\gamma^2}\nabla V^2 + x^6 = 0.$$

By solving the nonlinear equation, the solution has

$$V^* = \frac{\gamma}{2\left(\sqrt{2\gamma^2 - 1} + \gamma\right)} x^4.$$

Note that the solution ceases to be valid for $\gamma < (1/\sqrt{2})$. So the optimal γ is equal to $(1/\sqrt{2})$.

Now we apply our algorithm to the problem. The initial globally stabilizing controller selects $u^{(0)}(x) = 0$. The polynomial form of V is set to $V(x) = c_1x^2 + c_2x^3 + c_3x^4$, where c_1, c_2, and c_3 are unknown coefficients and are to be determined by SDP. The optimizing area selects $\Omega = \{x| -1 \le x \le 1\}$. The optimal solution $\gamma^{(0)}$ for $u^{(0)}$ after solving (17)–(19) is equal to 2. After ten iterations, the outer loop converges and outputs $\gamma^{(10)} = 0.70715$, which is nearly equal to $(1/\sqrt{2})$. The final converged $V^{(10)}(x)$ is $0.49258x^4$, in comparison with the exact optimal solution $V^*(x) = (1/2)x^4$ at $\gamma^* = (1/\sqrt{2})$. The results of value functions after inner loop iterations are shown in Fig. 1. The H_∞ near-optimal controller obtained by our algorithm is $u^{(11)}(x) = -0.98515x^3$.

B. 3-D Linear Example

The second example considers the F-16 aircraft plant, whose finite L_2-gain problem has been solved in [38] and [39] using NNs with polynomial basis functions. Here we consider its H_∞ optimal control problem. The system dynamics is described by

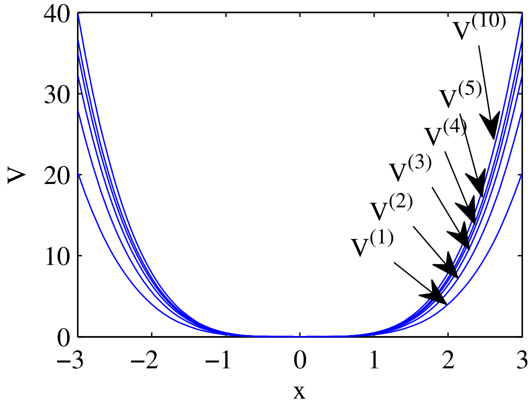


Fig. 1. Comparison of value functions after inner loop iterations in Example 1.

$$\dot{x} = Ax + Bu + Cd$$
$$z = x$$

where

$$A = \begin{bmatrix} -1.01887 & 0.90506 & -0.00215 \\ 0.82225 & -1.07741 & -0.17555 \\ 0 & 0 & -1 \end{bmatrix}$$
$$B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

The state vector is composed of $x = [\alpha, q, \delta_e]^T$, where α denotes the angle of attack, q is the pitch rate, and δ_e is the elevator deflection angle. The control input u is the elevator command and the disturbance d is a wind gust into the angle of attack. Still let $R = I$. Since the dynamics is linear and the cost is quadratic, the HJI equation is reduced to the ARE

$$AQ + QA^T + \frac{1}{\gamma^2}CC^T - BB^T + QQ^T = 0.$$

Finding the minimum γ and positive matrix Q is equivalent to the minimization problem expressed in the form of LMIs

$$\min_{\gamma} \quad \gamma$$
$$\text{s.t.} \quad \begin{bmatrix} AQ + QA^T + \frac{1}{\gamma^2}CC^T - BB^T & Q \\ Q^T & -I \end{bmatrix} \le 0.$$

By MATLAB LMI toolbox, the optimal solution to the above problem is $\gamma^* = 3.46469$.

Now apply our H_∞ optimal control algorithm with the initial stabilizing controller $u^{(0)}(x) = -x_1 - 0.1x_2 + 0.1x_3$. V is defined in the form of $V(x) = c_1x_1^2 + c_2x_1x_2 + c_3x_1x_3 + c_4x_2^2 + c_5x_2x_3 + c_6x_3^2$. Ω selects $\{x| -1 \le x_i \le 1, i = 1, 2, 3\}$. After eight iterations, the algorithm reaches the convergence. The value of γ is significantly reduced after the first outer-loop iteration, as shown in Fig. 2. The converged $\gamma^{(8)}$ is equal to 3.46470. The final H_∞ optimal controller is $u^{(9)}(x) = 0.34324x_1 + 0.35648x_2 - 0.46166x_3$. To test the attenuation effect of $u^{(9)}$, we select the disturbance signal $d(t) = 8t\cos(t/5)\exp(-t/3)/(t+1)$ and initialize the system

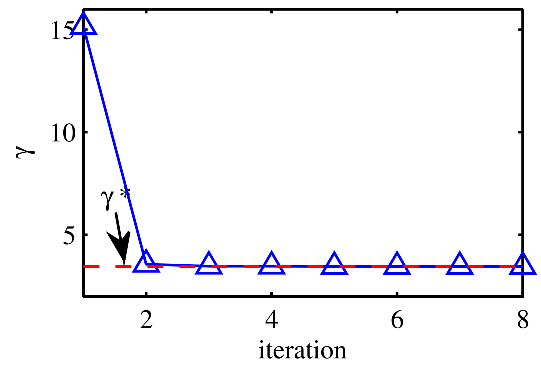


Fig. 2. Values of γ during the learning process in Example 2.

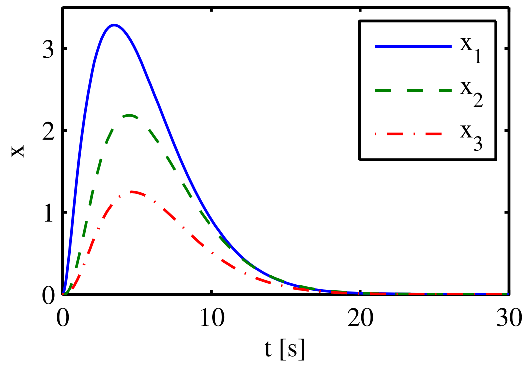


Fig. 3. State trajectories with $u^{(9)}$ in Example 2.

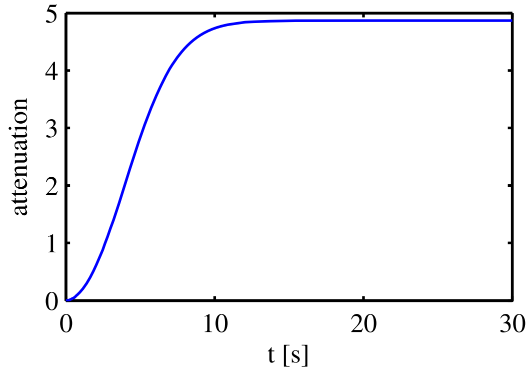


Fig. 4. Disturbance attenuation with $u^{(9)}$ in Example 2.

at $x(0) = 0$. The trajectories of state and disturbance attenuation $(\int_0^T (\|h\|^2 + u^T Ru)dt / \int_0^T \|d\|^2 dt)$ are plotted in Figs. 3 and 4. It is obvious that the attenuation keeps less than the square of $\gamma^{(8)}$.

C. 2-D Nonlinear Example

Consider a nonlinear system with dynamics [30]

$$\dot{x} = \begin{bmatrix} -x_1 + \frac{1}{4}x_2 + x_1^2 - \frac{3}{2}x_1^3 - x_1^2 x_2 - \frac{3}{4}x_1 x_2^2 - \frac{1}{2}x_2^3 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 1 \\ 0 \end{bmatrix} d$$

$$z = [1 \quad 0]x.$$

To apply the proposed algorithm, we select the initial controller $u^{(0)}(x) = 2x_1 - x_2$ and let $V(x) = c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_2^2 + c_4 x_1^3 + c_5 x_1^2 x_2 + c_6 x_1 x_2^2 + c_7 x_2^3 + c_8 x_1^4 + c_9 x_1^3 x_2 + c_{10} x_1^2 x_2^2 + c_{11} x_1 x_2^3 + c_{12} x_2^4$. Ω selects $\{x| - 1 \leq x_i \leq 1, i = 1, 2\}$. $R = I$. After six iterations, γ converges to 1.07529. The curve of $\gamma^{(i)}$ along the outer-loop iteration is given in Fig. 5. The final H_∞ near-optimal controller is expressed as $u^{(7)}(x) = -0.46766x_1 - 0.66409x_2 - 0.76782x_1^2 - 0.74141x_1 x_2 - 0.42354x_2^2 - 0.02336x_1^3 - 0.57793x_1^2 x_2 - 0.32172x_1 x_2^2 - 0.47494x_2^3$.

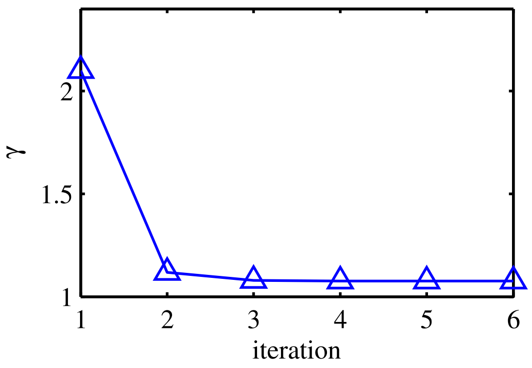


Fig. 5. Values of γ during the learning process in Example 3.

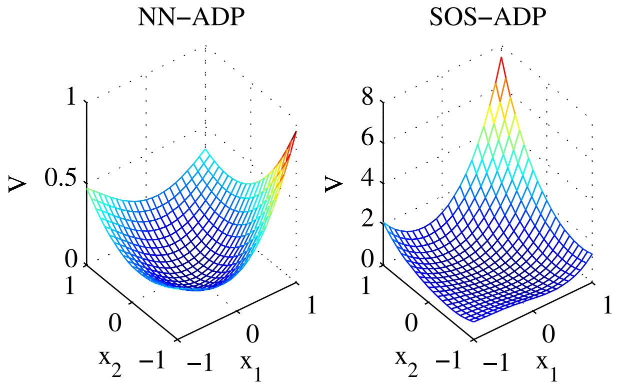


Fig. 6. Comparison of learned value functions by [17] (left) and our method (right) over the area $[-1, 1] \times [-1, 1]$ in Example 3.

The same H_∞ optimal control problem has been studied in [30]. The authors convert the problem to a set of state-dependent LMIs and use SOS optimization to find the minimum γ. Their final optimal value of γ is equal to 1.15, which is larger than our result. It means our algorithm is capable of finding a more accurate solution to the H_∞ optimal control. Another drawback appeared in the method of [30] is that it requires the system input gain matrix $g(x)$ must have at least one zero entry row, while our algorithm has no such limitation.

We further compare our results with the NN-based ADP algorithm proposed in [17]. We let the network basis functions for the value function use the same group of monomials in our experiment, and specify $\gamma = 1.07529$. After substituting the network into the PI formula (7), the NN weights are determined by minimizing the least-squared error. We sample the state space in the same area $\Omega = \{x|-1 \leq x_i \leq 1, i = 1, 2\}$ for 441 points. After convergence, the learned value function by NN-based ADP algorithm is $V = 0.3401x_1^2 - 0.1115x_1 x_2 + 0.2435x_2^2 + 0.0782x_1^3 - 0.1009x_1^2 x_2 + 0.0488x_1 x_2^2 - 0.0005x_2^3 - 0.0396x_1^4 - 0.0493x_1^3 x_2 - 0.0068x_1^2 x_2^2 - 0.0206x_1 x_2^3 - 0.0167x_2^4$.

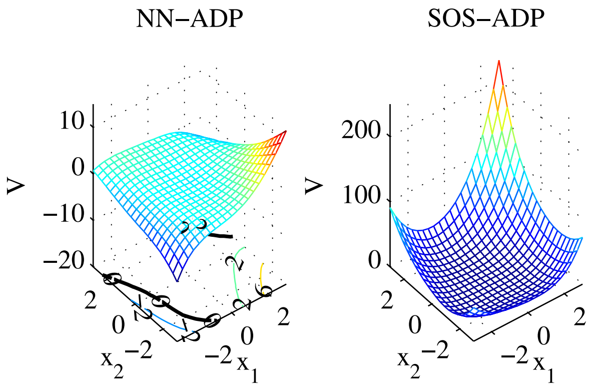


Fig. 7. Comparison of learned value functions by [17] (left) and our method (right) over the area $[-3, 3] \times [-3, 3]$ in Example 3.

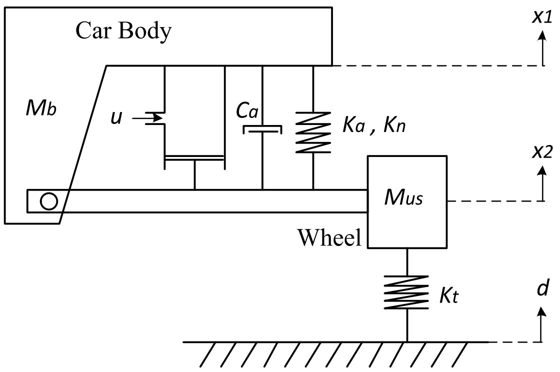


Fig. 8. Quater-car model.

For comparison, we plot the two value functions of NN-based and our SOS-based ADP algorithms over Ω in Fig. 6. Both values are positive in the area. However, when we extend the plotting area, the value function by NN-based ADP is no longer positive in some outside regions, as shown in Fig. 7. It can be explained by the fact that NN only approximates a complex function in a compact set. For regions outside the sampled area, positivity is not guaranteed and neither is the H_∞ control of the learned controller. To solve the problem, designers have to extend the sampled area, which inevitable leads to more computation. In our SOS-based ADP algorithm, the value function is globally positive supported by SOS theory. So the H_∞ control is valid over the whole state space.

D. Active Suspension Problem

The last experiment considers the application of our algorithm to the active suspension control of a quarter-car system [36], [40], whose model is depicted in Fig. 8. Its nonlinear dynamics is described by

$$\begin{aligned}
\dot{x}_1 &= x_2\\
\dot{x}_2 &= -\frac{1}{M_b}\Big[K_a(x_1 - x_3) + K_n(x_1 - x_3)^3\\
&\qquad\qquad + C_a(x_2 - x_4) - u\Big]\\
\dot{x}_3 &= x_4\\
\dot{x}_4 &= \frac{1}{M_{us}}\Big[K_a(x_1 - x_3) + K_n(x_1 - x_3)^3\\
&\qquad\qquad + C_a(x_2 - x_4) - K_t(x_3 - d) - u\Big]
\end{aligned}$$

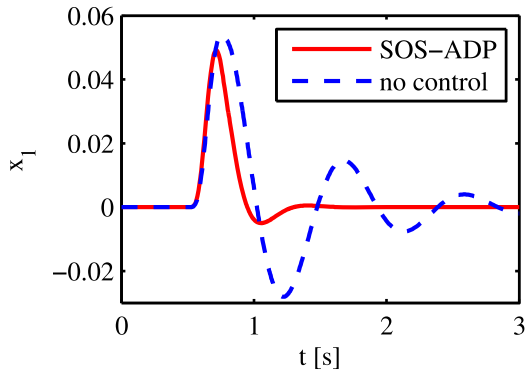


Fig. 9. Comparison of performance with the learned controller and with no control input in Example 4.

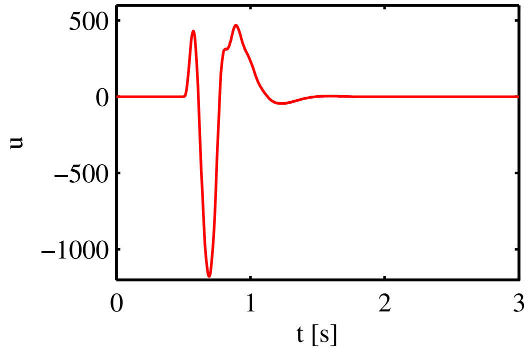


Fig. 10. Control signal with the learned controller in Example 4.

where x_1, x_2, and M_b denote the position, velocity, and mass of the car body. x_3, x_4, and M_{us} are the position, velocity, and mass of the wheel. K_t, K_a, K_n, and C_a are the tyre stiffness, the linear suspension stiffness, the nonlinear suspension stiffness, and the damping rate of the suspension. u is the control force from the hydraulic actuator and d is the road disturbance. In the experiment, dynamics parameters are set to standard values

$$\begin{aligned}
M_b &= 300 \text{ Kg} \quad M_{us} = 60 \text{ Kg}\\
K_t &= 190000 \text{ N/m} \quad K_a = 16000 \text{ N/m}\\
K_n &= K_a/10 \quad C_a = 1000 \text{ N/(m/sec)}.
\end{aligned}$$

Our main consideration is to maintain position of the car body x_1 in the presence of road disturbance, so the cost for the H_∞ control selects

$$J = \int_{t=0}^{\infty} \left(10^9 x_1^2 + u^2 - \gamma^2 d^2\right) dt.$$

To apply our algorithm, monomials for the value function select products of the set $\{x_1, x_2, x_3, x_4, x_1^2, x_2^2, x_3^2, x_4^2\}$ with itself. So the polynomial degree is 4. The system is desired to attenuate disturbance effects in the area $\Omega = \{x | x \in \mathbb{R}^4, |x_1| \leq 0.05, |x_2| \leq 0.5, |x_3| \leq 0.05, |x_4| \leq 1\}$. Since the system is self-stable, the initial controller is $u^{(0)}(x) = 0$. After five outer iterations, the algorithm converges to $\gamma^{(5)} = 94465.1$ and the approximate H_∞ optimal controller is expressed by $u^{(5)}(x) = -24473.2x_1 - 0.00110x_2^3 - 0.00020x_2^2 x_4 + 0.00004x_2 x_4^2 - 3602.06x_2 + 21584.4x_3 + 0.00592x_4^3 + 343.62x_4$.

To test the learned controller, we set the system at rest and choose the road disturbance as a single bump in the form

$$d(t) = \begin{cases} 0.038(1-\cos(8\pi t)), & 0.5 \le t \le 0.75 \\ 0, & \text{otherwise.} \end{cases}$$

After applying $u^{(5)}$, the trajectories of x_1 and u are plotted in Figs. 9 and 10. For comparison, the trajectory of x_1 at the same environment but with no control input is also plotted in Fig. 9. It is obvious that the learned controller exhibits satisfying performance in disturbance attenuation.

VII. Conclusion

The H_∞ optimal control problem is solved based on PI and SOS programming. A two-loop iteration is proposed, where the inner loop calculates the approximate HJI solution under a given attenuation coefficient, while the outer loop further minimizes the coefficient value. However, our algorithm supposes the stabilizing and L_2-gain region is global. But in many cases, the valid state space is bounded. $\mathcal{S}$-procedure [30], [33] provides a relaxation to deal with bounded state space. Besides, this paper only concentrates on polynomial nonlinear systems. In practice there exist numerous systems with polynomial fractions and other complicated dynamics. Further efforts are needed to extend this paper to these systems.

The proposed algorithm relies on the SOS theory, which has its own challenges. Since it is only an approximation to the optimal solution, the approximability needs to be addressed. In addition, the computation and numerical efficiency is another issue that limits its application. The development of SOS theory is the key to overcome the challenges.

Appendix
Proof of Theorem 1

1) The statement is true for $i = 1$ under Assumption 1. Suppose at the ith iteration, we have obtained the globally stabilizing u_i and the smooth solution V_i to (4). Since $V_i \ge 0$, it is a well-defined Lyapunov function. Following (4) and (5), u_{i+1} and V_i satisfy:

$$(\nabla V_i)^T(f+gu_{i+1}) + \frac{1}{4\gamma_0^2}(\nabla V_i)^T kk^T \nabla V_i + \|h\|^2 + u_{i+1}^T R u_{i+1} = -(u_i - u_{i+1})^T R(u_i - u_{i+1}) \le 0. \tag{25}$$

Consider the closed-loop system $\dot{x} = f + gu_{i+1}$, the time derivative of V_i satisfies

$$\begin{aligned}\dot{V}_i &= (\nabla V_i)^T(f+gu_{i+1}) \\ &\le -\frac{1}{4\gamma_0^2}(\nabla V_i)^T kk^T \nabla V_i - \|h\|^2 - u_{i+1}^T R u_{i+1} \\ &\le 0.\end{aligned}$$

So u_{i+1} is globally stabilizing. When the exogenous disturbance exists

$$\begin{aligned}&(\nabla V_i)^T(f+gu_{i+1}+kd) \\ &\le -\frac{1}{4\gamma_0^2}(\nabla V_i)^T kk^T \nabla V_i - \|h\|^2 \\ &\qquad - u_{i+1}^T R u_{i+1} + (\nabla V_i)^T kd \\ &\le -\|h\|^2 - u_{i+1}^T R u_{i+1} + \gamma_0^2\|d\|^2.\end{aligned}$$

After integrating both sides over the interval $[0, T]$, the following integral dissipation inequality exists:

$$\begin{aligned}&V_i(x(T)) - V_i(x(0)) \\ &\le \int_0^T \Big(-\|h\|^2 - u_{i+1}^T R u_{i+1} + \gamma_0^2\|d\|^2\Big)dt.\end{aligned}$$

Taking $x(0) = 0$ and $d \in L_2(0, T)$, and using $V_i \ge 0$, u_{i+1} is an L_2-gain $\le \gamma_0$ controller.

2) Reviewing the inequality given in (25), V_i and u_{i+1} satisfy the HJ inequality function

$$(\nabla V_i)^T(f+gu_{i+1}) + \frac{1}{4\gamma_0^2}(\nabla V_i)^T kk^T \nabla V_i + \|h\|^2 + u_{i+1}^T R u_{i+1} \le 0 \tag{26}$$

which means V_i is a possible storage function for u_{i+1}. Since we assume V_{i+1} is the smooth solution to (26) when equality holds, then V_{i+1} is the available storage function for u_{i+1} and has $0 \le V_{i+1}(x) \le V_i(x)$, $\forall x \in \mathbb{R}^n$.

3) After the above analysis, it is inferred that $\{V_i\}$ is a decreasing sequence and has the lower bound 0. The sequence is convergent and its limitation V^o satisfies the HJI equation.

The proof is complete. ■

References

[1] T. Başar and P. Bernhard, *H_∞ Optimal Control and Related Minimax Design Problems*. Boston, MA, USA: Birkhäuser, 1995.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Boston, MA, USA: Birkhäuser, 2008.

[3] A. Isidori and A. Astolfi, "Disturbance attenuation and H_∞-control via measurement feedback in nonlinear systems," *IEEE Trans. Autom. Control*, vol. 37, no. 9, pp. 1283–1293, Sep. 1992.

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

[5] J. B. Burl, *Linear Optimal Control: H_2 and H_∞ Methods*. Menlo Park, CA, USA: Addison-Wesley, 1998.

[6] P. Gahinet and P. Apkarian, "A linear matrix inequality approach to H_∞ control," *Int. J. Robust Nonlin. Control*, vol. 4, no. 4, pp. 421–448, 1994.

[7] P. Gahinet, A. Nemirovskii, A. J. Laub, and M. Chilali, "The LMI control toolbox," in *Proc. IEEE Conf. Decis. Control*, vol. 3. Lake Buena Vista, FL, USA, 1994, pp. 2038–2041.

[8] J. Huang and C.-F. Lin, "Numerical approach to computing nonlinear H_∞ control laws," *J. Guid. Control Dyn.*, vol. 18, no. 5, pp. 989–994, 1995.

[9] R. Beard and T. McLain, "Successive Galerkin approximation algorithms for nonlinear optimal and robust control," *Int. J. Control*, vol. 71, no. 5, pp. 717–743, 1998.

[10] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ, USA: Wiley, 2013.

[11] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.

[12] D. Zhao, Y. Zhu, and H. He, "Neural and fuzzy dynamic programming for under-actuated systems," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Brisbane, QLD, Australia, 2012, pp. 1–7.

[13] 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, Sep. 2009.

[14] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 882–893, May 2014.

[15] Y. Zhu, D. Zhao, H. He, and J. Ji, "Event-triggered optimal control for partially-unknown constrained-input systems via adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, doi: 10.1109/TIE.2016.2597763, to be published.

[16] Y. Zhu, D. Zhao, and D. Liu, “Convergence analysis and application of fuzzy-HDP for nonlinear discrete-time HJB systems,” *Neurocomputing*, vol. 149, pp. 124–131, Feb. 2015.

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

[18] H. Zhang, C. Qin, B. Jiang, and Y. Luo, “Online adaptive policy learning algorithm for H_∞ state feedback control of unknown affine nonlinear discrete-time systems,” *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2706–2718, Dec. 2014.

[19] Q. Zhang, D. Zhao, and Y. Zhu, “Event-triggered H_∞ control for continuous-time nonlinear system via concurrent learning,” *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2016.2531680.

[20] H. Zhang, L. Cui, and Y. Luo, “Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP,” *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 206–216, Feb. 2013.

[21] D. Zhao, Q. Zhang, D. Wang, and Y. Zhu, “Experience replay for optimal control of nonzero-sum game systems with unknown dynamics,” *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 854–865, Mar. 2016.

[22] 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.

[23] T. Dierks and S. Jagannathan, “Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1118–1129, Jul. 2012.

[24] R. Song, F. L. Lewis, Q. Wei, and H. Zhang, “Off-policy actor-critic structure for optimal control of unknown systems with disturbances,” *IEEE Trans. Cybern.*, vol. 46, no. 5, pp. 1041–1050, May 2016.

[25] Y. Zhu, D. Zhao, and X. Li, “Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data,” *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2561300.

[26] P. A. Parrilo, “Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization,” Ph.D. dissertation, Dept. Control Dyn. Syst., California Inst. Technol., Pasadena, CA, USA, 2000.

[27] G. Blekherman, P. A. Parrilo, and R. R. Thomas, *Semidefinite Optimization and Convex Algebraic Geometry*. Philadelphia, PA, USA: SIAM, 2013.

[28] A. Papachristodoulou *et al.* (2013). *SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB*. [Online]. Available: http://arxiv.org/abs/1310.4716

[29] D. Henrion and J.-B. Lasserre, “Gloptipoly: Global optimization over polynomials with MATLAB and SeDuMi,” *ACM Trans. Math. Softw.*, vol. 29, no. 2, pp. 165–194, 2003.

[30] S. Prajna, A. Papachristodoulou, and F. Wu, “Nonlinear control synthesis by sum of squares optimization: A Lyapunov-based approach,” in *Proc. IEEE 5th Asian Control Conf.*, vol. 1. 2004, pp. 157–165.

[31] J. Xu, L. Xie, and Y. Wang, “Simultaneous stabilization and robust control of polynomial nonlinear systems using SOS techniques,” *IEEE Trans. Autom. Control*, vol. 54, no. 8, pp. 1892–1897, Aug. 2009.

[32] W.-C. Huang, H.-F. Sun, and J.-P. Zeng, “Robust control synthesis of polynomial nonlinear systems using sum of squares technique,” *Acta Automatica Sinica*, vol. 39, no. 6, pp. 799–805, 2013.

[33] T. H. Summers *et al.*, “Approximate dynamic programming via sum of squares programming,” in *Proc. IEEE Eur. Control Conf. (ECC)*, 2013, pp. 191–197.

[34] M. B. Horowitz and J. W. Burdick, “Semidefinite relaxations for stochastic optimal control policies,” in *Proc. Amer. Control Conf. (ACC)*, Portland, OR, USA, 2014, pp. 3006–3012.

[35] Y. P. Leong, M. B. Horowitz, and J. W. Burdick, “Linearly solvable stochastic control Lyapunov functions,” *SIAM J. Control Optim.*, vol. 54, no. 6, pp. 3106–3125, 2016.

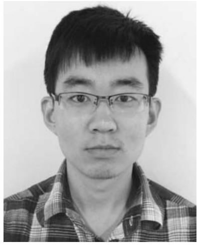
[36] Y. Jiang and Z.-P. Jiang, “Global adaptive dynamic programming for continuous-time nonlinear systems,” *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2917–2929, Nov. 2015.

[37] Z. Wang, X. Liu, K. Liu, S. Li, and H. Wang, “Backstepping-based Lyapunov function construction using approximate dynamic programming and sum of square techniques,” *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2574747.

[38] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, “Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control,” *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[39] H.-N. Wu and B. Luo, “Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear H_∞ control,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1884–1895, Dec. 2012.

[40] J.-S. Lin and I. Kanellakopoulos, “Nonlinear design of active suspensions,” *IEEE Control Syst.*, vol. 17, no. 3, pp. 45–59, Jun. 1997.



Yuanheng Zhu (M’15) received the B.S. degree from Nanjing University, Nanjing, China, in 2010, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015.

He is currently an Assistant Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His current research interests include optimal control, adaptive dynamic programming, and reinforcement learning.

Dongbin Zhao (M’06–SM’10) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1994, 1996, and 2000, respectively.

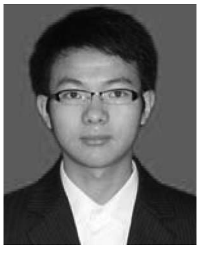
He was a Post-Doctoral Fellow with Tsinghua University, Beijing, China, from 2000 to 2002. He has been a Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, since 2002, and a Professor with the University of Chinese Academy of Sciences, Beijing. From 2007 to 2008, he was a Visiting Scholar with the University of Arizona, Tucson, AZ, USA. He has published four books, and over 50 international journal papers. His current research interests include computational intelligence, adaptive dynamic programming, deep reinforcement learning, robotics, intelligent transportation systems, and smart grids.

Dr. Zhao has been an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS since 2012 and the IEEE COMPUTATION INTELLIGENCE MAGAZINE since 2014. He has been serving as the Chair of the Adaptive Dynamic Programming and Reinforcement Learning Technical Committee since 2015 and the Multimedia Subcommittee since 2015 of the IEEE Computational Intelligence Society. He serves as a guest editors for several international journals. He is involved in organizing several international conferences.

Xiong Yang received the B.S. degree in mathematics and applied mathematics from Central China Normal University, Wuhan, China, in 2008, the M.S. degree in pure mathematics from Shandong University, Jinan, China, in 2011, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014.

He was an Assistant Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, from 2014 to 2016. He is currently an Associate Professor with the School of Electrical Engineering and Automation, Tianjin University, Tianjin, China. His current research interests include adaptive dynamic programming, reinforcement learning, data-driven control, robust control, and neural networks.

Dr. Yang was a recipient of the Excellent Award of Presidential Scholarship of the Chinese Academy of Sciences in 2014.



Qichao Zhang received the B.S. degree in automation from Northeastern Electric Power University, Jilin City, China, in 2012, and the M.S. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2014. He is currently pursuing the Ph.D. degree in control theory and control engineering with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His current research interests include reinforcement learning, game theory, and multiagent systems.