Online Iterative Adaptive Dynamic Programming Approach for Solving the Zero-Sum Game for Nonlinear Continuous-Time Systems with Partially Unknown Dynamics

Fu, Bin; Sun, Bo; Guo, Hang; Yang, Tao; Fu, Wenxing

DOI
10.1007/978-981-99-0479-2_262

Publication date
2023

Document Version
Final published version

Published in
Proceedings of 2022 International Conference on Autonomous Unmanned Systems, ICAUS 2022

Online Iterative Adaptive Dynamic Programming Approach for Solving the Zero-Sum Game for Nonlinear Continuous-Time Systems with Partially Unknown Dynamics

Bin Fu[1], Bo Sun[2], Hang Guo[1(✉)], Tao Yang[1], and Wenxing Fu[1]

[1] Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, Shanxi, China
jsguoh@nwpu.edu.cn

[2] Faculty of Aerospace Engineering, Delft University of Technology, 2629HS Delft, The Netherlands

Abstract. The current study presents an online iterative adaptive dynamic programming approach to resolve the zero-sum game (ZSG) for nonlinear continuous-time (CT) systems containing a partially unknown dynamic. The Hamilton-Jacobian-Issacs (HJI) equation is solved along the state trajectory according to the value function approximation and the policy improvement online. Relaxed dynamic programming is utilized to ensure the algorithm's convergence. Model and costate networks were established to conduct the method. Computational simulations are performed to present the efficiency of the algorithm.

Keywords: Approximation dynamic programming · Zero-sum game · Integral reinforcement learning · Online learning · Value iteration

1 Introduction

Reinforcement Learning (RL) is an iterative process to improve the action based on the interactions with the environment [1]. In a typical RL application, each decision will be valued according to the current state. Based on this value, the superior decision will be remembered, and the inferior one will be eliminated. This learning process will finally provide a satisfactory solution [2]. Adaptive Dynamic Programming (ADP) is a typical application of RL. To interact with the environment, the ADP method constructs an Actor-Critic structure for approximating the value function and a proper system controller [3]. This iterative approximation method helps researchers solve complex problems that may not be worked out easily by traditional analytical approaches [4].

Several scholars have employed the ADP technique in the zero-sum game (ZSG) problem. When a control system is defined as containing two players, the problem dealing with optimal control is converted into a game problem, and the HJB equation is

W. Fu et al. (Eds.): ICAUS 2022, LNEE 1010, pp. 2833–2842, 2023.
https://doi.org/10.1007/978-981-99-0479-2_262

converted to a Hamilton-Jacobi-Isaacs (HJI) system that includes a group of first-order Partial Differential Equations (PDEs) coping with nonlinear equations. Traditional analytical solutions cannot be easily reached. The ADP approaches can be employed to find approximate solutions. In [5], the ADP method was presented for the online determination of the solution of the Nash equilibrium when the ZSDG (Zero Sum Differential Game) containing two players has both linear dynamics and unbounded horizon quadratic cost. In [6], an iterative ADP methodology was proposed for solving a category of CT nonlinear ZSDGs with an entire system understanding. A concurrent learning-based ADP technique was presented in [7] for a nonlinear CT system to ensure the uniform ultimate boundedness of the approximate HJI equation.

In this paper, expanding the result of [8], a game control algorithm of a CT nonlinear system is proposed by resolving the HJI system with the approach called ADP/RL. The convergence proof of the iteration process was based on the approach utilizing the relaxation of the dynamic programming [9, 10], which ensures the iterative value function's convergence to the actual value. Moreover, inspired by the idea of [11–13], a dual heuristic programming (DHP) iteration method is presented for approximating the costate function. The convergence proof is also provided.

The essential novelties of the current study are:

(1) An online ADP/RL using the value iteration method is investigated for the ZSG of a CT nonlinear system containing a partially unknown dynamic.
(2) Compared to [14], the current study establishes a direct approximation of the costate function in the iteration. The DHP approach avoids the inaccuracy caused by the derivation of the approximated value function in the HDP approach.
(3) The presented method employs the Bellman equation and iteration to attain an exact convergence to the actual value. In contrast, in [15], only uniform ultimate boundedness (UUB) can be guaranteed.

The remaining parts of the current study are arranged as the following. The fundamental problem description is given in Sect. 2. The convergence of the iteration HDP and DHP approach is given in Sect. 3. Section 4 constructs model and critic networks to perform the presented methodology. In the end, the numerical simulations show the efficiency of the presented approach.

2 Preliminaries and the Statement of the Problem

The subsequent nonlinear CT system is assumed to be expressed by

$$\begin{cases} \dot{x} = f(x) + g(x)u + h(x)w \\ y = z(x) \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^{\mathrm{m}}$, and $w \in \mathbb{R}^{\mathrm{p}}$ describe the state, control, and disturbance vectors, respectively. $f(x) \in \mathbb{R}^n$, $g(x) \in \mathbb{R}^{\mathrm{n}\times\mathrm{m}}$, and $h(x) \in \mathbb{R}^{n\times p}$ stand for both differentiable and smooth mappings. The subsequent presumptions that are satisfied through this article are assumed. The unknown inner dynamic of the system, $f(x)$, is considered in the study.

However, subsequent paragraphs are indented.

Assumption 1: $f(0) = 0$, and $x = 0$ denote the equilibrium standing of the system.

Assumption 2: The controllable system.

The index performance is assumed by (2):

$$\begin{aligned} J(x,u,w) &= \int_0^\infty \left(x^{\mathrm{T}}Qx + u^{\mathrm{T}}R_1u - w^{\mathrm{T}}R_2w\right)d\tau \\ &= \int_0^\infty U(x,u,w)d\tau \end{aligned} \tag{2}$$

where $Q \in \mathbb{R}^{n\times n}$, $R_1 \in \mathbb{R}^{m\times m}$, and $R_2 \in \mathbb{R}^{p\times p}$ are positive definite. For fixed control and disturbances policies represented by $u(x)$ and $w(x)$, we can define the value function as

$$V(x(t)) = \int_t^\infty U(x,u,w)d\tau \tag{3}$$

The value function (3) is differentiated as follows:

$$U(x,u,w) + (\nabla V(x))^{\mathrm{T}}(f(x) + g(x)u + h(x)w) = 0\,, \quad V(0) = 0 \tag{4}$$

where $\nabla V = \partial V/\partial x$.

$$H(x,u,w) = U(x,u,w) + \nabla V^{\mathrm{T}}(x)(f(x) + g(x)u + h(x)w) \tag{5}$$

A unique result of a saddle point concerning the game problem in differential form is assumed. Thus, the Nash condition expressed below holds:

$$V^*(x_0) = \min_u \max_w V(x,u,w) = \max_w \min_u V(x,u,w) \tag{6}$$

From Bellman's optimality principle, we have:

$$0 = \min_u \max_w H\left(x, \nabla V^*, u, w\right) \tag{7}$$

u^* and w^* should satisfy $\partial H(x,u,w)/\partial u = 0$ and $\partial H(x,u,w)/\partial w = 0$. Then, the optimum solution is given as

$$u^* = -\frac{1}{2}R_1^{-1}g^{\mathrm{T}}(x)\nabla V^*(x) \tag{8}$$

$$w^* = \frac{1}{2}R_2^{-1}h^{\mathrm{T}}(x)\nabla V^*(x) \tag{9}$$

Inserting Eq. (8) and Eq. (9) into Eq. (5) gives

$$\begin{aligned} 0 &= x^{\mathrm{T}}Qx + \frac{1}{4}\left(\nabla V^*(x)\right)^{\mathrm{T}}g(x)R_1^{-1}g^{\mathrm{T}}(x)\nabla V^*(x) \\ &\quad + \frac{1}{4}\left(\nabla V^*(x)\right)^{\mathrm{T}}h(x)R_2^{-1}\nabla V^*(x) + \left(\nabla V^*(x)\right)^{\mathrm{T}}f(x) \end{aligned} \tag{10}$$

Solving this nonlinear HJI equation is challenging. Since the system's inner dynamic is unknown for partially unknown nonlinear CT systems, it is challenging to solve this HJI equation. To resolve this issue, the cost Eq. (3) in the following interval reinforcement form can be rewritten [16]:

$$V(x(t)) = \int_t^{t+T} U(x, u, w)d\tau + V(x(t+T)) \tag{11}$$

For every $T > 0$, it is shown in [16] that Eq. (4) is entirely equivalent to an interval reinforcement form as Eq. (11). Assume the differential game problem has a unique result of a saddle point. Hence, the Nash expression presented below is valid:

$$\begin{aligned} V^*(x(t)) &= \min_u \max_w \left(\int_t^{t+T} U(x, u, w)d\tau + V^*(x(t+T)) \right) \\ &= \max_w \min_u \left(\int_t^{t+T} U(x, u, w)d\tau + V^*(x(t+T)) \right) \end{aligned} \tag{12}$$

The inner dynamic of the system is not involved in this form.

3 Main Results

This section presents a value iteration-based interval reinforcement learning method to resolve the ZSGs with two players concerning partially unknown nonlinear CT systems. The basic ideas are illustrated in the following.

Value Function Update

$V_0(x) = 0$ and $i = 0$ are considered primal values, and V_1 can be solved from the subsequent equation:

$$V_{i+1}(x(t)) = \int_t^{t+T} U(x(t), u_i, w_i)d\tau + V_i\left(x^{u_i, w_i}(t+T)\right) \tag{13}$$

Policy Improvement

According to Eq. (8) and Eq. (9), u_i and w_i are defined as

$$u_{i+1}(x) = -\frac{1}{2}R_1^{-1}g^{\mathrm{T}}(x)\nabla V_i(x) \tag{14}$$

$$w_{i+1}(x) = \frac{1}{2}R_2^{-1}h^{\mathrm{T}}(x)\nabla V_i(x) \tag{15}$$

Since the above mechanism is based on VI, the presented approach should be initialized with the initial value $V_0(x)$. Therefore, the control signal u_i does not need to be admissible.

Lemma 1. The Hamilton mapping, $H(x, u_i, w_i, \nabla V_i)$, is minimized by the signal u_i in Eq. (14), namely, an iterative control one and the value function in Eq. (13). Meanwhile, the iterative control policy w_i in Eq. (15) maximizes the Hamilton mapping $H(x, u_i, w_i, \nabla V_i)$, and so does the value function in Eq. (13).

Proof: The detailed proof is presented in [17].

Lemma 2. Both positive mappings $Y_a(x)$, and $Y_b(x)$ meet the inequality expressed by $Y_a(x) \leq Y_b(x)$. Now, we have

$$\begin{aligned}\min_u \max_w \left(\int_t^{t+T} U(x,u,w)d\tau + Y_a\left(x^{u,w}(t+T)\right)\right) \\ \leq \min_u \max_w \left(\int_t^{t+T} U(x,u,w)d\tau + Y_b\left(x^{u,w}(t+T)\right)\right)\end{aligned} \tag{16}$$

Proof: The detailed proof is presented in [14].

Now, we give the value iteration approach's convergence proof for the CT nonlinear zero-sum game. The above proof's fundamental idea is expanded from the idea of relaxing dynamic programming first proposed in [9] and [10].

Theorem 1 Suppose the following inequality conditions.

$$0 \leq V^*(x(t+T)) \leq \theta \int_t^{t+T} U(x,u,w)d\tau \tag{17}$$

$$\alpha V^* \leq V_0 \leq \beta V^* \tag{18}$$

hold uniformly for some $0 < \theta < \infty$, $0 \leq \alpha \leq 1 \leq \beta < \infty$. The control policy u_i, disturbance policy w_i, and the value function V_i are iterated through Eq. (13) to Eq. (15). Now, the value function V_i tends to V^* based on the following inequalities:

$$\left(1 + \frac{\alpha - 1}{\left(1+\theta^{-1}\right)^i}\right)V^* \leq V_i \leq \left(1 + \frac{\beta - 1}{\left(1+\theta^{-1}\right)^i}\right)V^* \tag{19}$$

Proof: The proof will be given in two parts.

1. The left-hand side in Eq. (19) is proven by the mathematical induction, i.e.,
   When $i = 1$, considering condition Eq. (17), we have:

$$\Gamma_1 \triangleq \frac{\alpha - 1}{1+\theta}\left(\theta \int_t^{t+T} U(x,u,w)d\tau - V^*(x(t+T))\right) \leq 0 \tag{20}$$

   Considering $aV^* \leq V_0$, and Lemma 2, we have

$$\begin{aligned} V_1(x(t)) &\geq \min_u \max_w \left(\int_t^{t+T} U(x,u,w)d\tau + \alpha V^*(x(t+T)) + \Gamma_1 \right) \\ &= \min_u \max_w \left\{ \left(1 + \theta\frac{\alpha-1}{1+\theta}\right) \int_t^{t+T} U(x,u,w)d\tau + \left(\alpha - \frac{\alpha-1}{1+\theta}\right)V^*(x(t+T)) \right\} \\ &= \left(1 + \frac{\alpha-1}{1+\theta^{-1}}\right)V^*(x(t)) \end{aligned} \tag{21}$$

   Thus, the proof is provided when $i = 1$.

Next, we assume that when $i-1$ is taken, the left-hand side in Eq. (19) becomes valid, and we will have

$$\begin{aligned} V_i(x(t)) &= \min_u \max_w \left(\int_t^{t+T} U(x,u,w)d\tau + V_{i-1}(x(t+T)) \right) \\ &\geq \min_u \max_w \left\{ \left(1 + \frac{\alpha-1}{\left(1+\theta^{-1}\right)^{i-1}} \right) V^*(x(t+T)) + \int_t^{t+T} U(x,u,w)d\tau \right\} \\ &= \min_u \max_w \left\{ \left(1 + \frac{\alpha-1}{\left(1+\theta^{-1}\right)^{i-1}} - \frac{(\alpha-1)\theta^{i-1}}{(1+\theta)^i} \right) V^*(x(t+T)) \right. \\ &\left. + \left(1 + \frac{(\alpha-1)\theta^i}{(1+\theta)^i} \right) \int_t^{t+T} U(x,u,w)d\tau \right\} \\ &= \left(1 + \frac{(\alpha-1)\theta^i}{(1+\theta)^i} \right) V^*(x(t)) \end{aligned} \tag{22}$$

This proves the left part in Eq. (19).

2. Similarly, the right part in Eq. (17) could also be proven.

When $i=1$ considering the condition in Eq. (17), we have:

$$\Gamma_2 \triangleq \frac{\beta-1}{1+\theta} \left(\theta \int_t^{t+T} U(x,u,w)d\tau - V^*(x(t+T)) \right) \geq 0 \tag{23}$$

Considering $V_0 \leq \beta V^*$ Lemma 2, we have

$$\begin{aligned} V_1(x(t)) &\leq \min_u \max_w \left(\int_t^{t+T} U(x,u,w)d\tau + \beta V^*(x(t+T)) + \Gamma_2 \right) \\ &= \left(1 + \theta \frac{\beta-1}{1+\theta} \right) V^*(x(t)) \end{aligned} \tag{24}$$

Thus, the case $i=1$ is proved.

Next, the right part in the inequality Eq. (19) is valid when $i-1$ applied is assumed, and we will have

$$\begin{aligned} V_i(x(t)) &= \min_u \max_w \left(\int_t^{t+T} U(x,u,w)d\tau + V_{i-1}(x(t+T)) \right) \\ &\leq \min_u \max_w \left\{ \left(1 + \frac{\beta-1}{\left(1+\theta^{-1}\right)^{i-1}} \right) V^*(x(t+T)) + \int_t^{t+T} U(x,u,w)d\tau \right\} \\ &= \left(1 + \frac{(\beta-1)\theta^i}{(1+\theta)^i} \right) V^*(x(t)) \end{aligned} \tag{25}$$

Accordingly, the proof is finished.

According to Theorem 1, further results can be made.

Theorem 2. u_i, w_i, and V_i, called control, disturbance policies, and value mapping, respectively are defined as the iterative processes represented by Eqs. (13) to (15) (u_i, w_i), called the control pair will converge to the saddle point denoted by $\left(u_i^*, w_i^*\right)$ when $i \to \infty$.

Proof: The detailed proof can be found in [18].

The DHP Iteration

To make the algorithm works more efficiently, we introduce the idea of DHP for solving the HJI equation of the ZSG for the nonlinear CT system.

The costate is defined as

$$\lambda_i(x(t)) = \frac{\partial V_i(x(t))}{\partial x(t)} \tag{26}$$

$$\lambda_{i+1}(x(t)) = 2Qx(t)T + \left(\frac{\partial x(t+T)}{\partial x(t)}\right)^{\mathrm{T}} \lambda_i(x(t+T)) \tag{27}$$

By introducing the costate λ_i, the policy update algorithms in Eq. (14) and Eq. (15) are expressed as:

$$u_{i+1}(x) = -\frac{1}{2}R_1^{-1}g^{\mathrm{T}}(x)\lambda_i(x(t)) \tag{28}$$

$$w_{i+1}(x) = \frac{1}{2}R_2^{-1}h^{\mathrm{T}}(x)\lambda_i(x(t)) \tag{29}$$

Due to the unknown model of the system and the costate of the HJI equation, the three-layer NN and the single-layer NN are established separately.

$$\hat{x}(t+T) = W_{mi}^{\mathrm{T}}\sigma\left(Y_{mi}^{\mathrm{T}}x(t)\right) + g(x_t)u_i + h(x_t)w_i \tag{30}$$

$$\hat{\lambda}_i(x(t)) = W_{\lambda i}^{\mathrm{T}}\Psi(x(t)) \tag{31}$$

4 Implementation of the Iterative DHP Approach

In the proposed approach, each iteration step will need to calculate the control pair and the value function. However, solving the equations for the nonlinear system with unknown dynamics is challenging. Thus, the Neural Network (NN) is utilized to implement this algorithm in the current part.

In this paper, since the system dynamic is partially unknown, we will need to construct a Model NN for approximating the system dynamics. Now, the Critic NN will be constructed for approximating the system's costate. Consequently, the control policy is calculated using the Critic NN. The framework diagram is presented in Fig. 1.

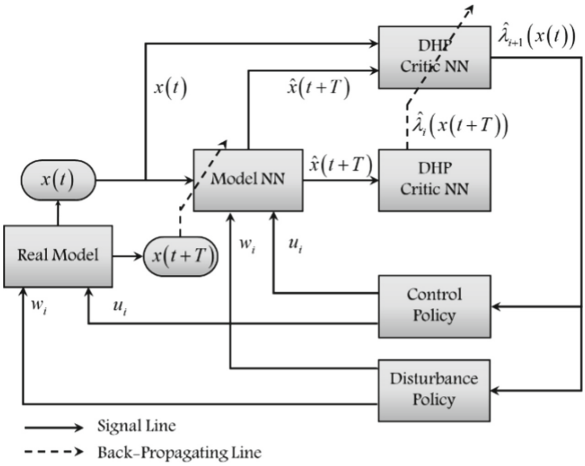

Fig. 1. The algorithm's signal flow.

5 Simulation

To demonstrate the presented approach's efficiency, the nonlinear system (32) is chosen.

$$\dot{x} = f(x) + g(x)u + h(x)w \tag{32}$$

$f(x)$, $g(x)$, and $h(x)$ are described as the inner dynamic, the matrices of the control coefficients, respectively as follows:

$$\begin{aligned} f(x) &= \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2\left(1 - (\cos(2x_1) + 2)^2\right) \end{bmatrix} \\ g(x) &= \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}, h(x) = \begin{bmatrix} 0 \\ \sin(4x_1) + 2 \end{bmatrix} \end{aligned} \tag{33}$$

The coefficients of the performance index are $Q = I^{2\times 2}$, $R_1 = 0.1I$, $R_2 = 2.5I$, and $x_0 = [1, -1]$. The activation mapping of the Critic Network was selected as follows:

$$\Psi(x(t)) = \left[x_1^2\ x_1x_2\ x_2^2\ x_1^4\ x_1^3x_2\ x_1^2x_2^2\ x_1x_2^3\ x_2^4\right]^{\mathrm{T}} \tag{34}$$

The interval $[-0.01, 0.01]$ is utilized to determine primal weights randomly. The number of neurons $L = 8$ is chosen by numerical simulation. It is found that a proper number of neurons will achieve a satisfying convergence precision with acceptable time consumption. The learning rates are $\alpha_M = 0.01$ and $\alpha_C = 0.02$. We set the simulation time as 15 s. The simulation time step is set as 0.001s. The residual error is set as 0.001. Since the proposed method is online, the tuning law of the weights vector runs literately in every time step until the costate vector NN converges. Then, the converged weights will be translated to the next step as an initial set.

Figure 2 depicts the simulation results. Figure 2(a) reveals that the weights vector is converged. The costate vector is presented in Fig. 2(b). Figure 2(c) depicts the optimal control pairs u, w. Figure 2(d) presents the trajectory states of the system.

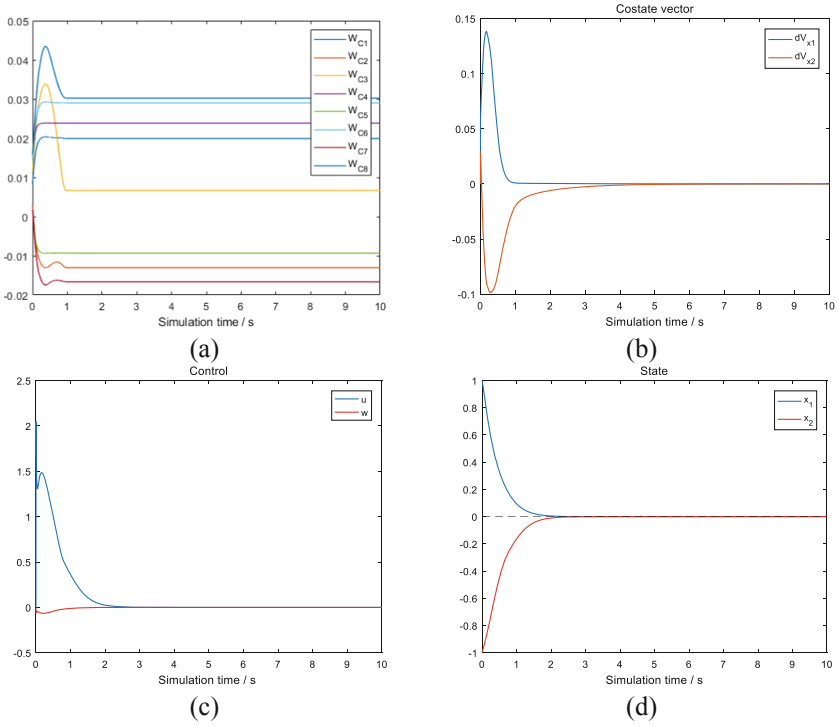

Fig. 2. Simulation results. (a) The convergence of the weight vector. (b) The costate trajectory. (c) The control signal. (d) The state trajectory.

6 Conclusion

The current study presents an NN-based online value iteration method to resolve the HJI system of the ZSG issue concerning the nonlinear CT systems containing a partially unknown dynamic. The proposed method provides a Nash equilibrium solution by iteration in every time step. The convergence of this method is provided by showing the convergence of the iterative value mapping to the HJI result. Moreover, the iteration approach of the DHP is proposed to approximate the system's costate vector. Then, model NN and costate NN are established to implement the presented approach. In the end, the simulation result reflects the presented approach's efficiency.

References

1. Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction. IEEE Trans. Neural Networks **9**(5), 1054 (1998)
2. Lewis, F., Vrabie, D., Vamvoudakis, K.: Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. IEEE Control Syst. **32**(6), 76–105 (2012)
3. Song, R., Lewis, F.L., Wei, Q., Zhang, H.: Off-policy actor-critic structure for optimal control of unknown systems with disturbances. IEEE Trans. Cybern. **46**(5), 1041–1050 (2016)

4. Khan, S.G., Herrmann, G., Lewis, F.L., Pipe, T., Melhuish, C.: Reinforcement learning and optimal adaptive control: an overview and implementation examples. Annu. Rev. Control. **36**(1), 42–59 (2012)
5. Vrabie, D., Lewis, F.: Adaptive Dynamic Programming algorithm for finding online the equilibrium solution of the two-player zero-sum differential game. In: Proceedings of the International Joint Conference on Neural Networks (2010)
6. Zhang, H., Wei, Q., Liu, D.: An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games. Automatica **47**(1), 207–214 (2011)
7. Yasini, S., Sistani, M.B.N., Karimpour, A.: Approximate dynamic programming for two-player zero-sum game related to H_∞ control of unknown nonlinear continuous-time systems. Int. J. Control. Autom. Syst. **13**(1), 99–109 (2014)
8. Su, H., Zhang, H., Zhang, K., Gao, W.: Online reinforcement learning for a class of partially unknown continuous-time non-linear systems via value iteration. Optim. Control Appl. Methods **39**(2), 1011–1028 (2018)
9. Rantzer, A.: Relaxed dynamic programming in switching systems. IEE Proc. - Control Theory Appl. **153**(5), 567–574 (2006)
10. Lincoln, B., Rantzer, A.: Relaxing dynamic programming. IEEE Trans. Automat. Contr. **51**(8), 1249–1260 (2006)
11. Wang, D., Liu, D.: Neuro-optimal control for a class of unknown nonlinear dynamic systems using SN-DHP technique. Neurocomputing **121**, 218–225 (2013)
12. Zhang, H., Qin, C., Luo, Y.: Neural-network-based constrained optimal control scheme for a discrete-time switched nonlinear system using dual heuristic programming, IEEE Trans. Autom. Sci. Eng. **11**(3), 839–849, 2014
13. Andrade, G.A., Da Fonseca Neto, J.V., Helena, P., Márcio, M.R., Gonsalves, E.: RLS estimator state space basis for the solution of HJB-Riccati Equation in approximate dynamic programming. In: IFAC Proceedings Volumes (IFAC-PapersOnline), vol. 3, no. PART 1, pp. 1153–1160 (2014)
14. Xiao, G., Zhang, H., Zhang, K., Wen, Y.: Value iteration based integral reinforcement learning approach for H∞ controller design of continuous-time nonlinear systems. Neurocomputing **285**, 51–59 (2018)
15. Sun, J., Liu, C.: Finite-horizon differential games for missile–target interception system using adaptive dynamic programming with input constraints. Int. J. Syst. Sci. **49**(2), 264–283 (2018)
16. Lewis, F.L., Vrabie, D.: Reinforcement learning and adaptive dynamic programming for feedback control. IEEE Circuits Syst. Mag. **9**(3), 32–50 (2009)
17. Zhu, Y., Zhao, D., Li, X.: Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data. IEEE Trans. Neural Networks Learn. Syst. **28**(3), 714–725 (2016)
18. Liu, D., Li, H., Wang, D.: Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm. Neurocomputing **110**, 92–100 (2013)