IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 December 2022, accepted 29 December 2022, date of publication 2 January 2023, date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3233567

RESEARCH ARTICLE

Optimal Drug Dosage Control Strategy of Immune Systems Using Reinforcement Learning

LIN CHEN[1], **YONG-WEI ZHANG**[2], **AND SHUN-CHAO ZHANG**[3]

[1]Scientific Research Center, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518107, China
[2]School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[3]School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou 510521, China

Corresponding author: Shun-Chao Zhang (47-319@gduf.edu.cn)

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110022.

ABSTRACT In this article, a reinforcement learning-based drug dosage control strategy is developed for immune systems with input constraints and dynamic uncertainties to sustain the number of tumor and immune cells in an acceptable level. First of all, the state of the immune system and the desired number of tumor and immune cells are constructed into an augmented state to derive an augmented immune system. By designing a discounted non-quadratic performance index function, the robust tracking control problem of immune systems with uncertainties is transformed into an optimal tracking control problem of nominal immune systems and the drug dosage can be limited within the specified range. Hereafter, a reinforcement learning algorithm and a critic-only structure are adopted to acquire the approximate optimal drug dosage control strategy. Furthermore, theoretical proof reveals that the proposed reinforcement learning-based drug dosage control strategy ensures the number of tumor and immune cells reaches the preset level under limited drug dosages and model uncertainties. Finally, simulation study verifies the availability of the developed drug dosage control strategy in different growth models of tumor cell.

INDEX TERMS Reinforcement learning, immune systems, immunotherapy, drug dosage control, robust control, neural networks.

I. INTRODUCTION

Cancer is a leading cause of death worldwide in recent decades, accounting for nearly 10 million deaths in 2020. Its morbidity expects up to 29 million cases by 2040 [1]. Cancer development is a multistep process. The risk factors of tumorigenesis are highly diverse, including genetic alterations, poor diet, physical inactivity, chronic infections and so on [2], [3]. Normal cells grow out of control when harmful changes interfere with orderly cellular biological process, forming precancerous lesions. Further, precancerous lesions develop into tumors. Cancer is characterized as malignant tumor. Traditional treatments of cancer mainly include surgery, radiotherapy, chemotherapy. Treatment options depends on the type and stage of cancer and the individual status of patients. Most types of cancer are separated by tumor-node-metastasis classification system including stage I to stage IV [4]. Stage I cancer is limited to primary location and can be removed through surgery. Stage II-III cancers have spread deeply into nearby tissues and even lymph nodes. Stage IV cancer that has spread to remote organs of the body is called advanced or metastatic cancer. Widespread metastases are the leading causes of cancer death. Once the cancer is diagnosed at stage II-IV, it should be treated with radiotherapy, chemotherapy or combined chemo-radiation therapy.

Along with the cancer progression, abnormal cells can be recognized and eliminated by the immune system inside the body due to the differences in cancer cells and normal cells. Immune cells are the main components of immune system and it can be divided into innate immune cells and adaptive immune cells. Activated innate immune cells could eliminate cancer cells through extensive phagocytosis and further activate adaptive immunity [5], [6]. Adaptive immune cells like cytotoxic CD8+ T cells directly target cancer cells through recognizing corresponding antigens [7], [8] and it is different

The associate editor coordinating the review of this manuscript and approving it for publication was Jianxiang Xi.

from radiotherapy and chemotherapy eliminating both cancer cells and normal cells. In addition, immunological memory, a significant characteristic of adaptive immunity, favors to consistent antitumor effects [9]. Thus, immunotherapy was proposed to prevent and treat cancer through reconstruction and enhancement of immune ability [8], [10], [11]. However, tumor cells could employ many strategies to escape immune surveillance and elimination, such as avoiding the immune recognition and recruiting of immunosuppressive immune cells [12]. Development and application of combined chemo and immunotherapies have been regarded as promising strategy to fight against cancer [13], [14]. The balance between tumor cells and immune cells determines tumor fate.

For the sake of describing the correlation between tumor cells and immune cells in human body, many scholars have established appropriate mathematical models for them, among which the most classic one is Stepanova's model. This model uses two differential equations to describe the changes of tumor cells and immune cells in immune systems. Based on it, many researchers have proposed different treatment plans based on control theory. The core idea is to design an appropriate control scheme for immune systems based on control theory, namely drug dosage control strategy, to ensure the level of tumor cells and immune cells in immune systems is maintained at a desired level. In [15], an adaptive robust control scheme was developed for cancer tumor-immune systems with model uncertainties. By designing a sliding-mode observer and a pair of adaptive control laws, the level of tumor and immune cells can be maintained on a preset value. In [16], the tracking control problem of cancer tumor-immune systems was addressed by proposing an adaptive control approach. However, these methods does not consider the drug dosage during treatment. Since drugs have side effects on the human body, we hope that the drug dosage should be as small as possible while ensuring the treatment effect. Fortunately, this requirement can be achieved by using the optimal control approach. In recent years, several researchers have proposed tumor treatment protocols based on optimal control theory. In [17], the chemotherapy administration problem was investigated by developing state dependent riccati equation based optimal control scheme. In [18], the initial malignant state of tumor was transferred to the benign region by adopting optimal control method. On the whole, a performance index function that contains drug dosages, tumor cells, and immune cells is defined, and then an optimal control strategy is developed to minimize the performance index function while ensuring that the desired level of tumor cells and immune cells. Although optimal control methods have been adopted to develop appropriate tumor treatment regimens, this research is still in its infancy and requires further investigated.

As is known to all, reinforcement learning (RL) is widely employed on control systems to handle various control problems, such as optimal regulation, trajectory tracking control, fault-tolerant control, robust control, differential game, and so on [19]. For the optimal regulation problem, Tamimi et al. [20] and Liu et al. [21] addressed it by proposing classical RL algorithms, namely value iteration (VI) and policy iteration (PI). Furthermore, the convergence and optimality of both algorithms were strictly analyzed. In recent years, several improved iterative RL algorithms have been proposed to overcome the shortcomings of traditional algorithms. Ha et al. [22] proposed a novel VI algorithm to speed up the convergence rate of the iterative value function and ensure the admissibility of the iterative control law. Jiang et al. [23] developed a bias PI algorithm to remove the initial admissible control law in traditional PI. For the trajectory tracking control problem, Modares et al. [24] designed a data-based integral RL algorithm to address the linear quadratic trajectory tracking control problem. Later, an off-policy integral RL algorithm was proposed to cope with the optimal exponential tracking control of unknown linear systems [25]. Lu et al. [26] addressed the optimal parallel tracking control problem under event-triggered mechanism. For the fault-tolerant control problem, Zhao et al. [27] developed an RL-based fault-tolerant controller by adding fault information into the performance index function. Subsequently, Zhang et al. [28] developed a fuzzy RL scheme to deal with the fault-tolerant tracking control problem. For the robust control problem, Liu et al. [32] shown that the robust guaranteed cost control of nonlinear systems with mismatched uncertainties can be transformed to an optimal control problem through designing appropriate value function and developed an RL-based optimal robust controller. After that, Wang et al. [33] addressed the same issue under event-triggered framework to save the computing resource. For the differential game problem, many scholars have proposed RL-based methods to acquire Nash equilibrium solutions of zero-sum games [29], nonzero sum games [30], and Stackerberg games [31]. In addition, due to the limited executive capacity of the actuator, the control input cannot exceed the prescribed range. To overcome this problem, researchers in RL community usually designed a non-quadratic performance index function to ensure the control input satisfies the specified range. This method was first proposed by Abu-Khalaf et al. [34] and has been widely employed to obtain the constrained optimal regulation controller, optimal tracking controller or robust controller for discrete-time or continuous-time nonlinear systems with input constraints. In discrete-time systems, Su et al. [35] developed event-triggered constrained optimal controller for sensor-actuator network systems via RL technique. Wei et al. [36] investigated event-triggered near-optimal tracking control of boiler-turbine systems with asymmetric input constraints. In continuous-time systems, Yang et al. [37] addressed the event-triggered constrained robust control problem for nonlinear systems with mismatched uncertainties via single network adaptive critic design. Xue et al. [38] proposed event-triggered integral RL scheme to cope with the constrained H_∞ tracking control problem.

In practical application, RL is also adopted to deal with the control problem of different practical systems including energy systems [39], stirred tank reactor systems [40], spring-mass-damper systems [41], modular reconfigurable robots [42], residential energy scheduling systems [43], boiler-turbine systems [44], hypersonic vehicles [45], and hypersonic flight vehicles [46]. In recent years, several scholars developed RL-based optimal regulation approaches for immune systems. In [47], the optimal regulation problem of immune systems was modeled as nonzero-sum games and the optimal drug dosage policies were deduced to form Nash equilibrium. In [48], the mathematical model of immune systems was established in differential equation form and the optimal dosage of chemotherapeutic and immunotherapeutic drugs was obtained by adopting RL technique. Nevertheless, the above-mentioned results consider ideal situations and there are exist many issues that need to further investigate. For example, 1) the human immune system is complicated, it is intractable to build an accurately mathematical model to describe the relationship between immune cells and tumor cells. Moreover, different environments and ages will affect the model parameters. Therefore, model uncertainty should be considered when designing drug dosage strategy. 2) drugs have side effects on the human body and people in different ages can tolerate different dosages. It is necessary to develop a constrained drug dosage strategy which can be obtained by addressing the input constraint problem in control community. 3) most of existing results investigate the optimal regulation problem. However, the number of immune and tumor cells requires to maintain at a certain level and the trajectory tracking control problem needs to be considered. According to the aforementioned statement, it is imperative to study the robust tracking control of immune systems subject to input constraints and model uncertainties. To our best knowledge, it has not been studied yet and inspires our research.

In this article, an RL-based drug dosage control strategy is presented for immune systems to guarantee the number of tumor and immune cells reaches a specified level. The characteristics of this research are summarized as two aspects.

1) Compared with existing approaches [15], [16] which developed robust control schemes for uncertain immune systems to maintain the number of immune cells and tumor cells at a appropriate level only, this paper further considers the drug dosage optimization problem. By employing RL technique, the drug dosage is reduced as much as possible while ensuring the treatment effect. Therefore, it is salutary to human body.
2) Unlike existing immune optimization regulation approaches [47], [48] that considered idea model only, this paper considers model uncertainties and input constraints simultaneously, which is more appropriate in actual scenario. By designing a discounted non-quadratic performance index function, the developed RL-based drug dosage control strategy guarantees the number of immune cells and tumor cells maintain at the desired level under model uncertainties and limited drug dosages.

The arrangement of this article is given as follows. In Section II, the mathematical model and the augmented form of immune systems are formulated, and the control problem is described. In Section III, RL-based drug dosage control strategy is developed for augmented immune systems. Moreover, the NN implementation and the stability analysis are given. Section IV verifies the effectiveness of the proposed RL-based drug dosage control strategy on two different growth models of tumor cells. In Section V, we provide the conclusion of this paper.

II. PROBLEM STATEMENT

According to [15] and [16], the mathematical model of immune systems is described by

$$\begin{aligned}\dot{P}_T &= \nu_c P_T F(P_T) - \gamma P_T P_I - \kappa_T P_T \mu_T + \Psi_1(P_T),\\ \dot{P}_I &= \nu_I (P_T - \beta P_T^2) P_I - \delta P_I + \alpha + \kappa_I P_I \mu_I + \Psi_2(P_I),\end{aligned}$$

where $P_T \in \mathbb{R}$ is the tumor volume, $P_I \in \mathbb{R}$ is the immune cell density, $\nu_c \in \mathbb{R}$ is the tumor growth rate, $\gamma \in \mathbb{R}$ is the elimination rate of tumor cells under the action of immune cells, $F(\cdot)$ denotes the tumor cell growth model, $\alpha \in \mathbb{R}$ is the T-cells' generation rate, $\delta \in \mathbb{R}$ is the natural death rate of immune cells, $\mu_T \in \mathbb{R}$ is the drug dosage of the tumor cells, $\mu_I \in \mathbb{R}$ is the booster drug for immune cells, $\kappa_I \in \mathbb{R}$ and $\kappa_T \in \mathbb{R}$ are corresponding control activities, respectively. $\Psi_1(P_T) \in \mathbb{R}$ and $\Psi_2(P_T) \in \mathbb{R}$ denote the model uncertainties, $\beta \in \mathbb{R}$ is the stimulation way that tumor cells influence immune cells, and $\nu_I \in \mathbb{R}$ is used to calibrate interaction between tumor cells and immune cells. Consider the drug dosage should not exceed the specified range, we assume that μ_T and μ_I satisfy $|\mu_T| \leq \bar{\mu}_T$ and $|\mu_I| \leq \bar{\mu}_I$, where $\bar{\mu}_T$ and $\bar{\mu}_I$ are upper bounds.

In practice, external environment or patient age will affect the mathematical model of immune systems, and it can be considered as model uncertainties. By using mathematical transformation, the immune system is reformulated as

$$\dot{\mathcal{X}}(t) = \mathcal{F}\big(\mathcal{X}(t)\big) + \mathcal{G}\big(\mathcal{X}(t)\big)\mathcal{S}(t) + \Psi\big(\mathcal{X}(t)\big), \tag{1}$$

where $\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2]^\mathsf{T} = [P_T, P_I]^\mathsf{T} \in \mathbb{R}^2$ is the immune system state, $\mathcal{S} = [\mathcal{S}_1, \mathcal{S}_2]^\mathsf{T} = [\mu_T, \mu_I]^\mathsf{T} \in \mathbb{R}^2$ is the control input, $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ are given by

$$\mathcal{F}(\mathcal{X}) = \begin{bmatrix} \nu_c \mathcal{X}_1 F(\mathcal{X}_1) - \gamma \mathcal{X}_1 \mathcal{X}_2 \\ \nu_I(\mathcal{X}_1 - \beta \mathcal{X}_1^2)\mathcal{X}_2 - \delta \mathcal{X}_2 + \alpha \end{bmatrix}$$

and

$$\mathcal{G}(\mathcal{X}) = \begin{bmatrix} -\kappa_T \mathcal{X}_1 \\ \kappa_I \mathcal{X}_2 \end{bmatrix},$$

respectively, $\Psi(\mathcal{X}) = [\Psi_1(\mathcal{X}_1), \Psi_2(\mathcal{X}_2)]^\mathsf{T} \in \mathbb{R}^2$ is the uncertain term. In this paper, we consider matched uncertainties, that is $\Psi(\mathcal{X}) = \mathcal{G}(\mathcal{X})\mathcal{D}(\mathcal{X})$ with $\mathcal{D}(\mathcal{X})$ is a uncertain function and satisfies $\|\mathcal{D}(\mathcal{X})\| \leq \Psi_M(\mathcal{X})$, where $\Psi_M(\mathcal{X})$ is the upper bound function of uncertain term and $\Psi_M(0) = 0$.

Assumption 1: The system functions $\mathcal{F}(\mathcal{X})$ and $\mathcal{G}(\mathcal{X})$ are Lipschitz continuous on a compact set Ω and the system (1) is controllable on Ω.

In immune systems, the number of tumor cells and immune cells needs to maintain at an appropriate level such that the tumor growth can be inhibited or even stopped. To accomplish this goal, we need to develop drug dosage control approach to guarantee the number of tumor cells and immune cells follows the predesigned indicator $\pi(t)$ under model uncertainties and input constraints. In this paper, we will demonstrate that it can be achieved by designing a constrained optimal drug dosage control strategy for its nominal system.

Assume that the dynamics of predesigned indicator satisfies

$$\dot{\pi}(t) = \psi_d\big(\pi(t)\big), \tag{2}$$

where $\psi_d(\cdot)$ is a C^∞ function with $\psi_d(0) = 0$. Therefore, the indicator tacking error is defined as

$$e(t) = \pi(t) - \mathcal{X}(t). \tag{3}$$

In order to deal with the constrained robust tracking control problem, an augmented immune systems is established as

$$\dot{\mathcal{K}}(t) = \mathcal{F}_{\mathcal{K}}\big(\mathcal{K}(t)\big) + \mathcal{G}_{\mathcal{K}}\big(\mathcal{K}(t)\big)\Big(\mathcal{S}(t) + \mathcal{D}\big(\mathcal{X}(t)\big)\Big), \tag{4}$$

where $\mathcal{K}(t) = [e(t), \pi(t)]^{\mathsf{T}}$ is the augmented state and

$$\mathcal{F}_{\mathcal{K}}(\mathcal{K}) = \begin{bmatrix} \mathcal{F}\big(e(t) + \pi(t)\big) - \psi_d\big(\pi(t)\big) \\ \psi_d\big(\pi(t)\big) \end{bmatrix},$$
$$\mathcal{G}_{\mathcal{K}}(\mathcal{K}) = \begin{bmatrix} \mathcal{G}\big(e(t) + \pi(t)\big) \\ 0 \end{bmatrix}.$$

The nominal form of system (4) is provided as

$$\dot{\mathcal{K}}(t) = \mathcal{F}_{\mathcal{K}}\big(\mathcal{K}(t)\big) + \mathcal{G}_{\mathcal{K}}\big(\mathcal{K}(t)\big)\mathcal{S}(t). \tag{5}$$

The discounted performance index function of system (5) is defined as

$$\begin{aligned} \mathcal{P}\big(\mathcal{K}(t)\big) = &\int_t^\infty e^{-\eta(\nu - t)}\Big(\xi\Psi_M^2\big(\mathcal{X}(\nu)\big) \\ &+ \mathcal{C}\big(\mathcal{K}(\nu), \mathcal{S}(\nu)\big)\Big)d\nu, \end{aligned} \tag{6}$$

where ξ is a positive constant, $0 < \eta < 1$ is a discount factor, $\mathcal{C}(\mathcal{K}, \mathcal{S}) = \mathcal{K}^{\mathsf{T}}\mathcal{Q}\mathcal{K} + \mathcal{Y}(\mathcal{S})$ is the utility function and $\mathcal{Q} = \mathrm{diag}\{\bar{\mathcal{Q}}, 0_{2\times 2}\}$ with a positive definite matrix $\bar{\mathcal{Q}} \in \mathbb{R}^{2\times 2}$. In order to ensure the drug dosage within a limited range, $\mathcal{Y}(\mathcal{S})$ is designed as

$$\mathcal{Y}(\mathcal{S}) = 2\int_0^{\mathcal{S}} \zeta^{-\mathsf{T}}(\bar{\mathcal{S}}^{-1}\nu)d\nu, \tag{7}$$

where $\zeta(\cdot) \in \mathbb{R}$ is a monotonic odd function satisfying $|\zeta(\cdot)| < 1$, $\bar{\mathcal{S}} = \mathrm{diag}\{\bar{\mu}_T, \bar{\mu}_I\}$ is the upper bound of the drug dosage. The Hamiltonian is defined as

$$\begin{aligned} \mathcal{H}\big(\mathcal{K}, \mathcal{S}, \nabla\mathcal{P}(\mathcal{K})\big) = &\ \mathcal{C}(\mathcal{K}, \mathcal{S}) + \xi\Psi_M^2(\mathcal{X}) - \eta\mathcal{P}(\mathcal{K}) \\ &+ \nabla\mathcal{P}^{\mathsf{T}}(\mathcal{K})\Big(\mathcal{F}_{\mathcal{K}}\big(\mathcal{K}\big) + \mathcal{G}_{\mathcal{K}}\big(\mathcal{K}\big)\mathcal{S}\Big). \end{aligned} \tag{8}$$

The optimal performance index function satisfies

$$\begin{aligned} \mathcal{P}^*(\mathcal{K}) = \min_{\mathcal{S}\in\Re(\Omega)} &\int_t^\infty e^{-\eta(\nu - t)}\Big(\xi\Psi_M^2\big(\mathcal{X}(\nu)\big) \\ &+ \mathcal{C}\big(\mathcal{K}(\nu), \mathcal{S}(\nu)\big)\Big)d\nu, \end{aligned} \tag{9}$$

where $\Re(\Omega)$ denotes the admissible control set. Therefore, the constrained optimal drug dosage control strategy is obtained by

$$\mathcal{S}^*(\mathcal{K}) = -\bar{\mathcal{S}}\zeta\Big(\frac{1}{2}\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\mathcal{P}^*(\mathcal{K})\Big). \tag{10}$$

According to (8) and (10), the Hamilton Jacobi-Bellman equation is provided as

$$\begin{aligned} 0 = &\ \mathcal{H}\big(\mathcal{K}, \mathcal{S}^*, \nabla\mathcal{P}^*(\mathcal{K})\big) \\ = &\ \nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\big(\mathcal{F}_{\mathcal{K}}\big(\mathcal{K}\big) + \mathcal{G}_{\mathcal{K}}\big(\mathcal{K}\big)\mathcal{S}^*\big) + \mathcal{C}\big(\mathcal{K}, \mathcal{S}^*\big). \end{aligned} \tag{11}$$

Obviously, the optimal performance index function is required to construct the optimal drug dosage control strategy. Unfortunately, it is scarcely possible to acquire it by solving (11) directly. In the following, we will employ RL algorithm to overcome this difficulty.

Remark 1: In optimal control field, the performance index function includes process cost and control cost. It reflects the rapidity of the system response and the energy saving of the system. For traditional optimal regulation problem, the performance index function is defined as a quadratic form with respect to system state and control input. However, in trajectory tracking scenario, the augmented immune system state and the control input will not converge to zero when the system state follows the desired trajectory. Therefore, in order to guarantee the boundedness of the performance index function in infinite horizon, a discounted factor is added in performance index function. Moreover, a non-quadratic term is adopted to ensure the control input stays within the specified range. In addition, the upper bound function of the uncertain term is added in performance index function. The purpose of this is to consider the influence of model uncertainties when designing the controller and ensure the obtained controller is robust.

Remark 2: The assumption of "$\psi_d(\cdot)$ is a C^∞ function" is used to guarantee the augmented system functions $\mathcal{F}_{\mathcal{K}}$ and $\mathcal{G}_{\mathcal{K}}$ are Lipschitz continuous. It is noted that the Lipschitz continuous of system functions is basic and general for nonlinear systems in control community, which guarantees the solution of differential equation is unique [20], [21]. Moreover, a lot of common trajectories satisfy this condition, such as step functions with any magnitudes, ramp functions with any slopes, and sinusoidal functions functions with any and initial phases [54], [56]. Therefore, this assumption is necessary and reasonable.

III. ROBUST DRUG DOSAGE CONTROL STRATEGY DESIGN VIA REINFORCEMENT LEARNING

A. ROBUST DRUG DOSAGE CONTROL STRATEGY DESIGN

In this section, we will provide rigorous mathematical theorem to exhibit that the constrained robust tracking control problem is addressed by developing a constrained optimal drug dosage control strategy for the nominal system (5). Before that, we provide some general Assumptions that have been given in [20], [21], [29], [32], [37], [49], [50], and [51].

Assumption 2: The optimal performance index function $\mathcal{P}^*(\mathcal{K})$ and it's partial derivative with respect to $\mathcal{K}$ satisfy,

$$\|\mathcal{P}^*(\mathcal{K})\| \leq c_1\|\mathcal{K}\|^2, \|\nabla\mathcal{P}^*(\mathcal{K})\| \leq c_2\|\mathcal{K}\|, \tag{12}$$

where c_1 and c_2 are positive constants.

Assumption 3: The system function $\mathcal{G}_{\mathcal{K}}(\mathcal{K})$ is norm-bounded, that is,

$$0 < \|\mathcal{G}_{\mathcal{K}}(\mathcal{K})\| \leq \bar{\mathcal{G}}_{\mathcal{K}}, \tag{13}$$

where $\bar{\mathcal{G}}_{\mathcal{K}}$ is a positive constant.

Theorem 1: Consider the uncertain immune system (1) and its augmented system (4), the constrained optimal drug dosage control strategy given by (10), and Assumptions 2 and 3, if the following inequalities

$$\lambda_{\min}(\mathcal{Q}) > \eta + \frac{1}{2}c_2^2\bar{\mathcal{G}}_{\mathcal{K}}^2, \tag{14}$$

$$\xi > \frac{1}{2} \tag{15}$$

hold, then the tracking error is guaranteed to be asymptotically stable. It implies that the number of tumor cells and immune cells can be maintained at the desired level.

Proof. The Lyapunov function candidate is constructed as

$$\mathcal{L}_{T1} = \mathcal{P}^*(\mathcal{K}). \tag{16}$$

Based on (4), the time derivative of (16) is calculated by

$$\dot{\mathcal{L}}_{T1} = \nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\big(\mathcal{F}_{\mathcal{K}}(\mathcal{K}) + \mathcal{G}_{\mathcal{K}}(\mathcal{K})\mathcal{S}^* + \mathcal{G}_{\mathcal{K}}(\mathcal{K})\mathcal{D}\big). \tag{17}$$

According to (17) and Assumptions 2 and 3, it holds that

$$\begin{aligned}\dot{\mathcal{L}}_{T1} =& -\mathcal{C}(\mathcal{K}, \mathcal{S}^*) - \xi\Psi_M^2(\mathcal{X}) + \eta\mathcal{P}^*(\mathcal{K}) \\ &+\nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\mathcal{G}_{\mathcal{K}}(\mathcal{K})\mathcal{D} \\ \leq& -\mathcal{K}^{\mathsf{T}}\mathcal{Q}\mathcal{K} + \eta\|\mathcal{K}\|^2 + \frac{1}{2}\|\nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\mathcal{G}_{\mathcal{K}}(\mathcal{K})\|^2 \\ &-\xi\Psi_M^2(\mathcal{X}) + \frac{1}{2}\|\mathcal{D}\|^2 \\ \leq& -\lambda_{\min}(\mathcal{Q})\|\mathcal{K}\|^2 + \eta\|\mathcal{K}\|^2 \\ &+\frac{1}{2}c_2^2\|\mathcal{K}\|^2\bar{\mathcal{G}}_{\mathcal{K}}^2 - \left(\xi - \frac{1}{2}\right)\Psi_M^2(\mathcal{X}). \end{aligned} \tag{18}$$

Therefore, if (14) and (15) hold, we derive that $\dot{\mathcal{L}}_{T1} < 0$. It indicates that the level of tumor cells and immune cells are reached to expected value.

B. NEURAL NETWORK IMPLEMENTATION

In this section, the critic neural network (NN) is introduced to formulate the optimal performance index function $\mathcal{P}^*(\mathcal{K})$ as

$$\mathcal{P}^*(\mathcal{K}) = \varphi_c^{*\mathsf{T}}\chi_c(\mathcal{K}) + \epsilon_c(\mathcal{K}), \tag{19}$$

where $\varphi_c^* \in \mathbb{R}^{h_c}$ is the optimal weight vector, $\chi_c(\mathcal{K}) \in \mathbb{R}^{h_c}$ is the activation function, h_c is the number of hidden layer neurons, and $\epsilon_c(\mathcal{K}) \in \mathbb{R}$ is the approximation error. Consider the optimal weight vector is unknown, we provide the approximate performance index function as

$$\hat{\mathcal{P}}(\mathcal{K}) = \hat{\varphi}_c^{\mathsf{T}}\chi_c(\mathcal{K}), \tag{20}$$

where $\hat{\varphi}_c$ is the estimate of φ_c^*.

Consequently, the constrained optimal drug dosage control strategy and its approximate version are given as

$$\mathcal{S}^*(\mathcal{K}) = -\bar{\mathcal{S}}\zeta\left(\frac{1}{2}\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\big(\nabla\chi_c^{\mathsf{T}}(\mathcal{K})\varphi_c^* + \nabla\epsilon_c(\mathcal{K})\big)\right), \tag{21}$$

$$\hat{\mathcal{S}}(\mathcal{K}) = -\bar{\mathcal{S}}\zeta\left(\frac{1}{2}\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\chi_c^{\mathsf{T}}(\mathcal{K})\hat{\varphi}_c\right). \tag{22}$$

Based on (20) and (22), the approximate Hamiltonian is provided as

$$\begin{aligned}\mathcal{H}\big(\mathcal{K}, \hat{\mathcal{S}}, \nabla\hat{\mathcal{P}}(\mathcal{K})\big) =& \ \hat{\varphi}_c^{\mathsf{T}}\nabla\chi_c(\mathcal{K})\big(\mathcal{F}_{\mathcal{K}}(\mathcal{K}) + \mathcal{G}_{\mathcal{K}}(\mathcal{K})\hat{\mathcal{S}}\big) \\ &+\mathcal{C}(\mathcal{K}, \hat{\mathcal{S}}) + \xi\Psi_M^2(\mathcal{X}) - \eta\hat{\varphi}_c^{\mathsf{T}}\chi_c(\mathcal{K}) \\ \triangleq& \ e_c. \end{aligned} \tag{23}$$

To ensure the approximate weight approach the optimal weight, we need to minimize the object function $\mathcal{E} = \frac{1}{2}e_c^2$. By adopting the gradient descent approach, the critic NN weight is renovated by

$$\begin{aligned}\dot{\hat{\varphi}}_c =& -\frac{\alpha_c\Upsilon}{(1 + \Upsilon^{\mathsf{T}}\Upsilon)^2}\Big(\hat{\varphi}_c^{\mathsf{T}}\Upsilon + \mathcal{C}(\mathcal{K}, \hat{\mathcal{S}}) + \xi\Psi_M^2(\mathcal{X}) \\ &-\eta\hat{\varphi}_c^{\mathsf{T}}\chi_c(\mathcal{K})\Big), \end{aligned} \tag{24}$$

where $\alpha_c > 0$ is the learning rate and

$$\Upsilon = \nabla\chi_c(\mathcal{K})\big(\mathcal{F}_{\mathcal{K}}(\mathcal{K}) + \mathcal{G}_{\mathcal{K}}(\mathcal{K})\hat{\mathcal{S}}\big).$$

Lemma 1: Consider the nominal nonlinear system (5), the critic NN weight estimation error is guaranteed to be uniformly ultimately bounded (UUB) with the critic NN weight tuning rule (24).

Proof. The proof of Lemma 1 has been provided in [29], [32], and [37], so the detail is omitted here.

C. STABILITY ANALYSIS

Assumption 4: $\tilde{\varphi}_c$, $\nabla\chi_c(\mathcal{K})$, and $\nabla\epsilon_c(\mathcal{K})$ satisfy

$$\|\tilde{\varphi}_c\| \leq \bar{\varphi}_c, \ \|\nabla\chi_c(\mathcal{K})\| \leq \bar{\chi}_c, \ \|\nabla\epsilon_c(\mathcal{K})\| \leq \bar{\epsilon}_c,$$

where $\tilde{\varphi}_c = \varphi_c^* - \hat{\varphi}_c$, $\bar{\varphi}_c$, $\bar{\chi}_c$, and $\bar{\epsilon}_c$ are positive constants.

Assumption 5: $\zeta(\cdot)$ is Lipschitz continuous and satisfies

$$\|\zeta(x) - \zeta(y)\| \leq \mathcal{L}_\zeta\|x - y\|, \tag{25}$$

where $\mathcal{L}_\zeta$ is a positive constant, x and y are vectors with appropriate dimensions.

Theorem 2: For the nominal immune system (5), the approximate optimal drug dosage control strategy given by (22), and Assumptions 2–5, if $\mathcal{Q}$ is selected to satisfies

$$\lambda_{\min}(\mathcal{Q}) > \frac{\frac{1}{2}\bar{\mathcal{G}}_{\mathcal{K}}^2 + \eta c_1}{\mathcal{A}_2}, \tag{26}$$

where $0 < \mathcal{A}_2 < 1$, then the tracking error is guaranteed to be UUB.

Proof. The Lyapunov function candidate is established as

$$\mathcal{L}_{T2} = \mathcal{P}^*(\mathcal{K}). \tag{27}$$

The time derivative of (27) is calculated by

$$\begin{aligned}\dot{\mathcal{L}}_{T2} =&\ \nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\big(\mathcal{F}_{\mathcal{K}}(\mathcal{K}) + \mathcal{G}_{\mathcal{K}}(\mathcal{K})\hat{\mathcal{S}}\big)\\ =&\ -\mathcal{C}(\mathcal{K},\hat{\mathcal{S}}) + \nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\mathcal{G}_{\mathcal{K}}(\mathcal{K})\big(\hat{\mathcal{S}} - \mathcal{S}^*\big)\\ &-\xi\Psi_M^2(\mathcal{X}) + \eta\mathcal{P}(\mathcal{K})\\ \le&\ -\mathcal{K}^{\mathsf{T}}\mathcal{Q}\mathcal{K} + \frac{1}{2}\|\nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\mathcal{G}_{\mathcal{K}}(\mathcal{K})\|^2 - \xi\Psi_M^2(\mathcal{X})\\ &+\eta\mathcal{P}(\mathcal{K}) + \frac{1}{2}\|\hat{\mathcal{S}} - \mathcal{S}^*\|^2. \end{aligned} \tag{28}$$

According to Assumptions 2–5, (21) and (22), the last part of (28) is expanded as

$$\begin{aligned}\frac{1}{2}\|\hat{\mathcal{S}} - \mathcal{S}^*\|^2 \le&\ \frac{1}{2}\Big\| - \bar{\mathcal{S}}\zeta\Big(\frac{1}{2}\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\chi_c^{\mathsf{T}}(\mathcal{K})\hat{\varphi}_c\Big)\\ &+\bar{\mathcal{S}}\zeta\Big(\frac{1}{2}\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\big(\nabla\chi_c^{\mathsf{T}}(\mathcal{K})\varphi_c^*\\ &+\nabla\epsilon_c(\mathcal{K})\big)\Big)\Big\|^2\\ \le&\ \frac{1}{8}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\big\|\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\chi_c^{\mathsf{T}}(\mathcal{K})\tilde{\varphi}_c\\ &+\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\epsilon_c(\mathcal{K})\big\|^2\\ \le&\ \frac{1}{4}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\big\|\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\chi_c^{\mathsf{T}}(\mathcal{K})\tilde{\varphi}_c\big\|^2\\ &+\frac{1}{4}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\|\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\epsilon_c(\mathcal{K})\|^2\\ \le&\ \frac{1}{4}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\bar{\mathcal{G}}_{\mathcal{K}}^2(\bar{\chi}_c^2\bar{\varphi}_c^2 + \bar{\epsilon}_c^2). \end{aligned} \tag{29}$$

Therefore, we further have

$$\begin{aligned}\dot{\mathcal{L}}_{T2} \le&\ -\mathcal{K}^{\mathsf{T}}\mathcal{Q}\mathcal{K} + \frac{1}{2}\|\nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\mathcal{G}_{\mathcal{K}}(\mathcal{K})\|^2 + \eta c_1\|\mathcal{K}\|^2\\ &+\frac{1}{4}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\bar{\mathcal{G}}_{\mathcal{K}}^2(\bar{\chi}_c^2\bar{\varphi}_c^2 + \bar{\epsilon}_c^2)\\ \le&\ -\mathcal{A}_1\mathcal{K}^{\mathsf{T}}\mathcal{Q}\mathcal{K} - \mathcal{A}_2\mathcal{K}^{\mathsf{T}}\mathcal{Q}\mathcal{K} + \frac{1}{2}\|\nabla\mathcal{P}^{*\mathsf{T}}(\mathcal{K})\mathcal{G}_{\mathcal{K}}(\mathcal{K})\|^2\\ &+\eta c_1\|\mathcal{K}\|^2 + \frac{1}{4}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\bar{\mathcal{G}}_{\mathcal{K}}^2(\bar{\chi}_c^2\bar{\varphi}_c^2 + \bar{\epsilon}_c^2)\\ \le&\ -\beta_1^2\mathcal{A}_1\lambda_{\min}(\bar{\mathcal{Q}})\|e\|^2 + (\beta_1^2 - \mathcal{A}_1)\lambda_{\min}(\bar{\mathcal{Q}})\|e\|^2\\ &-\lambda_{\min}(\mathcal{Q})\mathcal{A}_2\|\mathcal{K}\|^2 + \frac{1}{2}\bar{\mathcal{G}}_{\mathcal{K}}^2c_2^2\|\mathcal{K}\|^2 + \eta c_1\|\mathcal{K}\|^2\\ &+\frac{1}{4}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\bar{\mathcal{G}}_{\mathcal{K}}^2(\bar{\chi}_c^2\bar{\varphi}_c^2 + \bar{\epsilon}_c^2), \end{aligned} \tag{30}$$

where $0 < \mathcal{A}_1 < 1$ and satisfies $\mathcal{A}_1 + \mathcal{A}_2 = 1$. Letting

$$\Theta_1 = \frac{1}{4}\|\bar{\mathcal{S}}\|^2\mathcal{L}_\zeta^2\bar{\mathcal{G}}_{\mathcal{K}}^2(\bar{\chi}_c^2\bar{\varphi}_c^2 + \bar{\epsilon}_c^2). \tag{31}$$

Therefore, $\dot{\mathcal{L}}_{T2} < 0$ if the tracking error e is outside the following set

$$\Omega_e = \left\{e\colon \|e\| \le \sqrt{\frac{\Theta_1}{\mathcal{A}_1 - \beta_1^2}}\right\}. \tag{32}$$

The proof is finished.

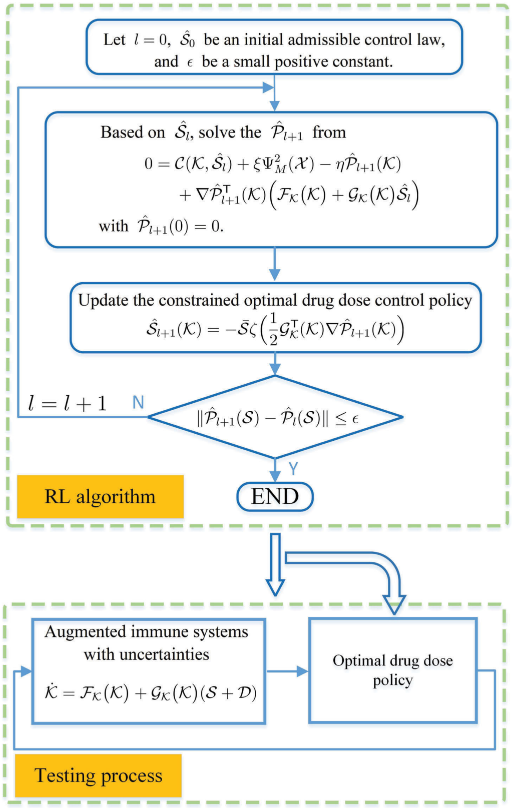
Let $l = 0$, $\hat{\mathcal{S}}_0$ be an initial admissible control law, and ϵ be a small positive constant.
Based on $\hat{\mathcal{S}}_l$, solve the $\hat{\mathcal{P}}_{l+1}$ from
$0 = \mathcal{C}(\mathcal{K},\hat{\mathcal{S}}_l) + \xi\Psi_M^2(\mathcal{X}) - \eta\hat{\mathcal{P}}_{l+1}(\mathcal{K}) + \nabla\hat{\mathcal{P}}_{l+1}^{\mathsf{T}}(\mathcal{K})\big(\mathcal{F}_{\mathcal{K}}(\mathcal{K}) + \mathcal{G}_{\mathcal{K}}(\mathcal{K})\hat{\mathcal{S}}_l\big)$
with $\hat{\mathcal{P}}_{l+1}(0) = 0$.
Update the constrained optimal drug dose control policy
$\hat{\mathcal{S}}_{l+1}(\mathcal{K}) = -\bar{\mathcal{S}}\zeta\big(\frac{1}{2}\mathcal{G}_{\mathcal{K}}^{\mathsf{T}}(\mathcal{K})\nabla\hat{\mathcal{P}}_{l+1}(\mathcal{K})\big)$
$l = l + 1$
N
$\|\hat{\mathcal{P}}_{l+1}(\mathcal{S}) - \hat{\mathcal{P}}_l(\mathcal{S})\| \le \epsilon$
Y
RL algorithm
END
Augmented immune systems with uncertainties
$\dot{\mathcal{K}} = \mathcal{F}_{\mathcal{K}}(\mathcal{K}) + \mathcal{G}_{\mathcal{K}}(\mathcal{K})(\mathcal{S} + \mathcal{D})$
Optimal drug dose policy
Testing process

FIGURE 1. **The implementation process of RL-based drug dosage control scheme.**

The overall structure of the RL-based drug dosage control scheme is shown in Fig. 1, where we find that it can be divided into two parts, that is, 1) obtain a optimal drug dosage control strategy by using RL algorithm and 2) employ this control strategy on augmented immune system with model uncertainties and input constraints.

Remark 3: For Assumption 2, $\mathcal{P}^*(\mathcal{K})$ represents the optimal performance index function and is continuously differentiable on a compact set. Thus, this assumption is reasonable and widely used in the existing results. For Assumption 3, consider the augmented immune system is controllable, it is reasonable to assume the system input function $\mathcal{G}_{\mathcal{K}}$ is norm-bounded as a positive constant. For Assumption 4, $\tilde{\varphi}_c$ is the critic NN estimation error. From Lemma 1, we can know that it is guaranteed to be UUB with the critic NN weight updating law (24). Therefore, it is reasonable to assume that $\tilde{\varphi}_c$ is norm-bounded. Moreover, $\nabla\epsilon_c(\mathcal{K})$ is the NN reconstruction error. Since it can not be infinite in practice, so the norm-bounded assumption is reasonable. For Assumption 5, $\zeta(\cdot)$ is a monotonic odd function and can be selected as *tanh*$(\cdot)$. Hence, it is reasonable to assume to be Lipschitz continuous.

Remark 4: 1) Different from the existing results [45] and [46] which handled input constraints by using prescribed performance control approach, this paper adopts RL technique to design a constrained robust tracking controller in a nearly optimal manner. Therefore, the developed controller can guarantee the tracking performance and reduce the control cost as much as possible. 2) Unlike the existing RL-based tracking control approaches [52] and [53] for ideal system model, this paper considers both input constraints and model uncertainties. By establishing an augmented system and designing a novel discounted non-quadratic performance index function, an RL-based constrained robust tracking controller is developed and it is more appropriate in practice.

Remark 5: According to the properties of performance index function, minimum tracking error and energy can be achieved by minimizing the performance index function. To achieve this objective, a Hamiltonian function is defined and its partial derivative with respect to control input is calculated, thus the equation of the optimal control input is obtained. Consider the optimal control input relies on optimal performance index function, RL algorithm and critic-only structure are adopted to obtain its approximate value by an iterative manner. Consequently, the minimum tracking error and the minimum control energy can be guaranteed by using the developed constrained optimal controller.

IV. SIMULATION

In this part, an immune system is adopted to confirm the availability of the RL-based optimal drug dosage control strategy. The system parameters of the immune system are selected as $\nu_c = 0.56$, $\nu_I = 0.005$, $\gamma = 1$, $\beta = 0.0026$, $\delta = 0.375$, $\alpha = 0.118$. The model uncertainties are chosen as

$$\Psi_1(P_T) = r_1 P_T \sin^5(P_I)\cos^2(P_T),$$
$$\Psi_2(P_I) = r_2 P_I \cos^5(P_T)\sin^2(P_I),$$

where $r_1 = r_2 = 10$.

Case 1: The growth model of tumor cell is exponential.

In case 1, the growth model of tumor cell is chosen as $F(P_T) = 1$. It means that tumor cells grow exponentially. The predesigned indicators of tumor cells and immune cells are set as $\pi = [P_T, P_I]^{\mathsf{T}} = [2, 1]^{\mathsf{T}}$, the parameters in (6) are chosen as $\mathcal{Q} = [1\ 0\ 0\ 0, 0\ 1\ 0\ 0, 0\ 0\ 0\ 0, 0\ 0\ 0\ 0]^{\mathsf{T}}$ and $\zeta(\cdot) = \tanh(\cdot)$, the activation function of the critic NN is selected as $\chi_c(\mathcal{K}) = [\mathcal{K}_1^2, \mathcal{K}_1\mathcal{K}_2, \mathcal{K}_2^2]^{\mathsf{T}}$, and the upper bounds of drug dosages are picked as $\bar{\mu}_T = 10$ and $\bar{\mu}_I = 10$.

The simulation verification consequences are given in Figs. 2–7. Fig. 2 provides the evolution curves of critic NN weights, where we can conclude that the critic NN weight vectors will reach to $\hat{\varphi}_c = [81.57, 30.59, 12.95]^{\mathsf{T}}$. The tracking error of the immune system is given in Fig. 3. We find that the tracking error will reach to zero. To demonstrate the effectiveness of the RL-based optimal drug dosage control approach, we compared it with the existing ADP-based constrained tracking method in literature [38]. It can be seen from Fig. 4 that our method can guarantee the tracking performance when there exist model uncertainties. Fig. 5 reveals the curves of drug dosage control strategies under different schemes. It is pretty clear that the drug dosage control strategies developed by this paper are bounded and the drug dosage will not exceed the predetermined range. However, drug control strategies designed in [55] will exceed the predetermined values. Fig. 6 provides the curves of performance index function for different methods. Compared with the existing methods in [15] and [16], the convergence value of the performance index in this paper is smaller, it means that the developed RL-based drug dosage control method can guarantee the tracking performance with lower control cost. Moreover, a time-varying expectation indicator $\pi(t) = [\cos(0.5t), \cos(0.2t)]^{\mathsf{T}}$ is selected to further verify the effectiveness of the RL-based optimal drug dosage control method. According to Fig. 7, we can see that the number of tumor cells and immune cells still catch up with time-varying expected values.

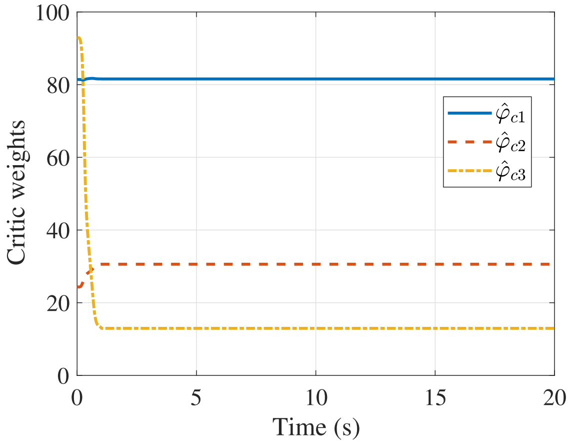

FIGURE 2. Critic NN weights of case 1.

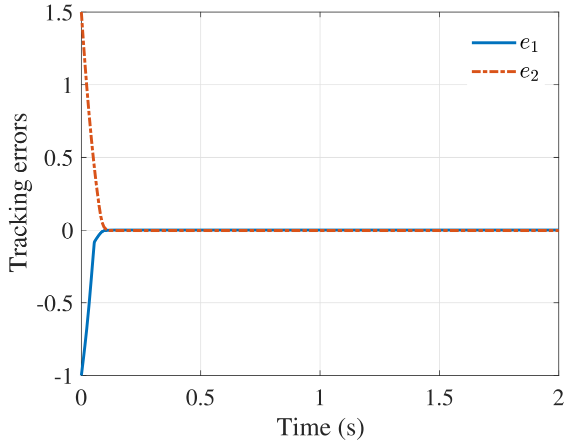

FIGURE 3. Tracking errors of case 1.

Case 2: The tumor cell is Gompertzian growth.

In this case, we consider that the tumor cell is Gompertzian growth, that is, $F(P_T) = -\ln(\frac{P_T}{P_\infty})$, where $P_\infty = 10$ denotes the fixed carrying capacity of the tumor. Let the predesigned indicators of tumor cells and immune cells be $\pi = [P_T, P_I]^{\mathsf{T}} = [5, 3]^{\mathsf{T}}$ and the upper bounds of the

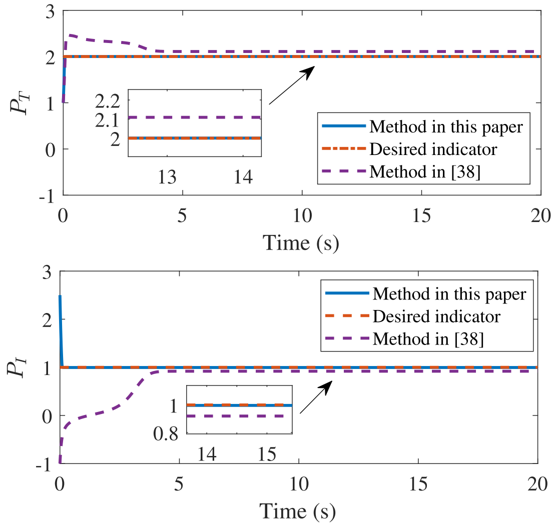

FIGURE 4. The number of tumor cells and immune cells in case 1.

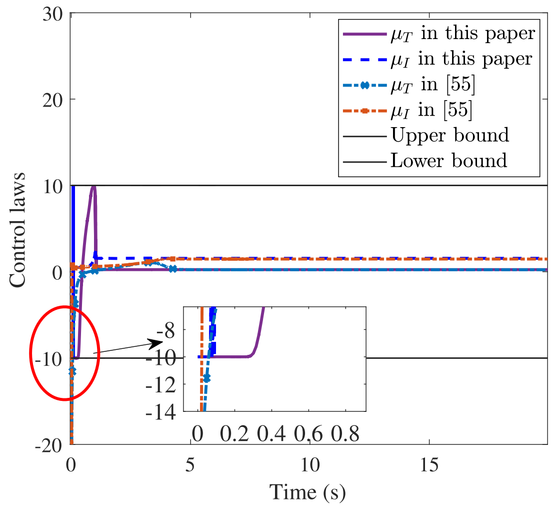

FIGURE 5. Tracking control laws of case 1.

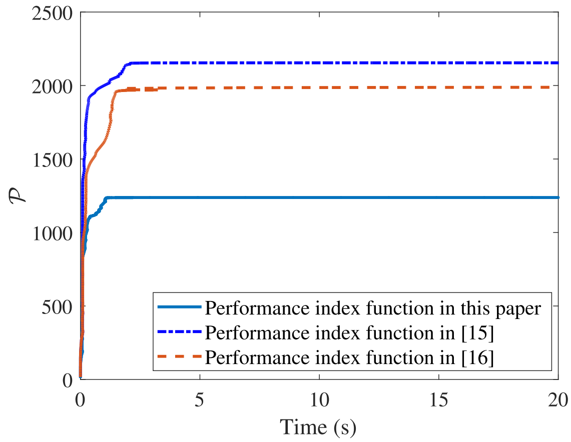

FIGURE 6. Performance index functions of different method in case 1.

drug dosage be $\bar{\mu}_T = 5$, $\bar{\mu}_I = 5$. Simulation results are shown in Figs. 8–11. From Fig. 8 we can conclude that compared with the tracking control approach in [38], the developed RL-based optimal drug dosage control scheme guarantees the number of tumor cells and immune cells track the predesigned indicator precisely. The drug dosage curves obtained by different methods are given in Fig. 9. We can find that the developed RL-based optimal drug dosage curves can stay within the specified range, whereas the conventional approach will exceed. The curves of performance index function are displayed in Fig. 10. We can conclude that the RL-based optimal drug dosage control method maintains the number of tumor cells and immune cells at expected level with less drug dosage. Fig. 11 shows that the number of tumor cells and immune cells can track the time-varying expectation indicator $\pi(t) = [\cos(0.5t), \cos(t)]^{\mathsf{T}}$.

According to the simulation results from cases 1 and 2, we can conclude that the proposed RL-based drug dosage control strategy is available, that is, tumor cells and immune cells can be maintained at desired levels by using limited drug dosages.

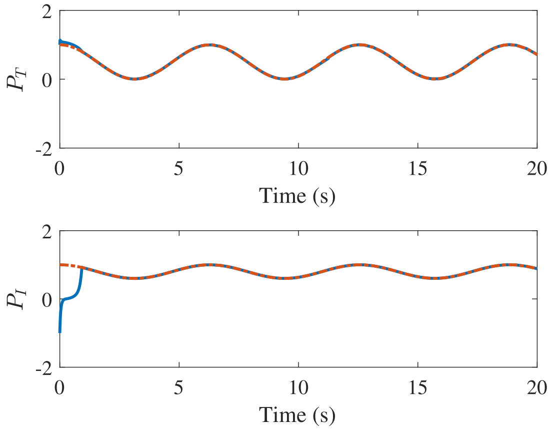

FIGURE 7. The number of tumor cells and immune cells under time-varying expectation indicator in case 1.

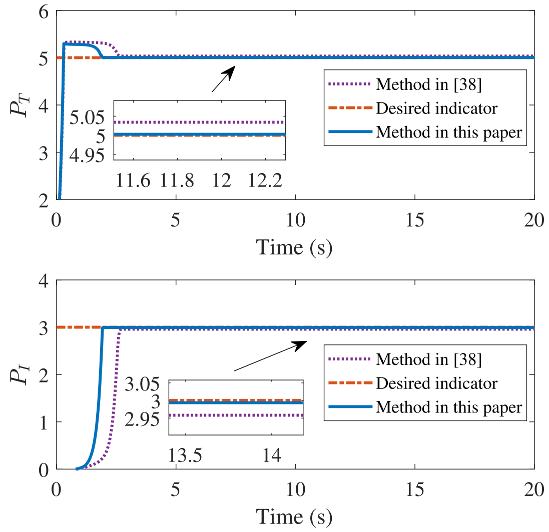

FIGURE 8. The number of tumor cells and immune cells of case 2.

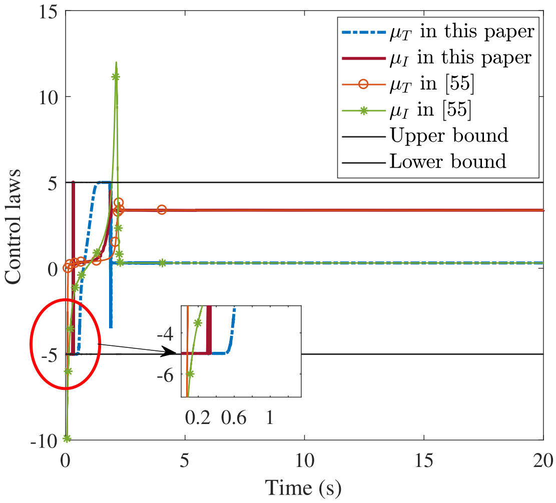

FIGURE 9. Tracking control laws of case 2.

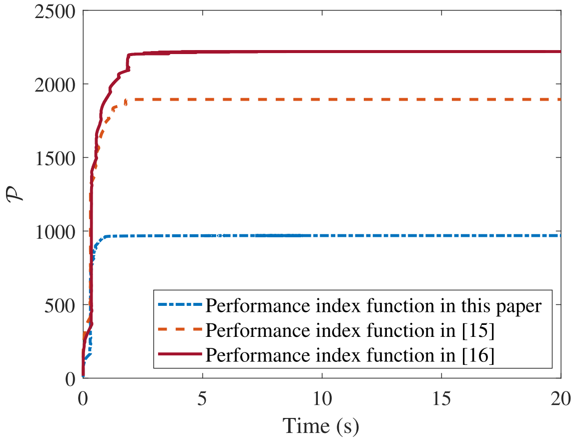

FIGURE 10. Performance index functions of different methods in case 2.

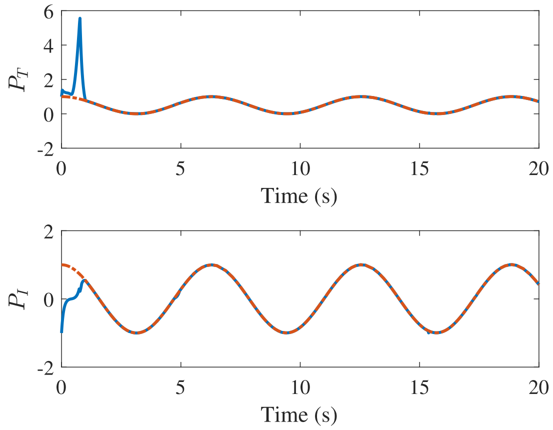

FIGURE 11. The number of tumor cells and immune cells under time-varying expectation indicator in case 2.

V. CONCLUSION

This paper provides an immunotherapy regimen for cancer via RL technique. We show that it can be obtained by addressing the robust tracking control problem of immune systems subject to input constraints and dynamic uncertainties in control community. To accomplish this goal, an augmented immune system and a discounted non-quadratic performance index function are established such that the robust tracking control problem of uncertain immune systems is converted to an optimal tracking control problem of its nominal plant. Subsequently, we develop constrained drug dosage control strategy by using RL algorithm and critic-only structure. According to the Lyapunov theory, we proof that the developed RL-based drug dosage control strategy ensures the number of tumor and immune cells reaches to the preset level with limited drug dosages. At last, simulation results display that the developed immunotherapy regimen is feasible.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] D. Hanahan, "Hallmarks of cancer: New dimensions," *Cancer Discovery*, vol. 12, no. 1, pp. 31–46, Jan. 2022.

[3] D. Hanahan and R. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–774, Mar. 2011.

[4] F. Greene and L. Sobin, "The staging of cancer: A retrospective and prospective appraisal," *CA, Cancer J. Clinicians*, vol. 58, no. 3, pp. 180–190, May 2008.

[5] R. Nowarski, N. Gagliani, S. Huber, and R. A. Flavell, "Innate immune cells in inflammation and cancer," *Cancer Immunol. Res.*, vol. 1, no. 2, pp. 77–84, Aug. 2013.

[6] S. Woo, L. Corrales, and T. Gajewski, "Innate immune recognition of cancer," *Annu. Rev. Immunol.*, vol. 33, pp. 445–474, Jan. 2015.

[7] M. St. Paul and P. S. Ohashi, "The roles of CD8$^+$ T cell subsets in antitumor immunity," *Trends Cell Biol.*, vol. 30, no. 9, pp. 695–704, Sep. 2020.

[8] T. K. Kim, E. N. Vandsemb, R. S. Herbst, and L. Chen, "Adaptive immune resistance at the tumour site: Mechanisms and therapeutic opportunities," *Nature Rev. Drug Discovery*, vol. 21, no. 7, pp. 529–540, Jul. 2022.

[9] T. Wang, Y. Shen, S. Luyten, Y. Yang, and X. Jiang, "Tissue-resident memory CD8$^+$ T cells in cancer immunology and immunotherapy," *Pharmacolog. Res.*, vol. 159, Sep. 2020, Art. no. 104876.

[10] D. S. Chen and I. Mellman, "Elements of cancer immunity and the cancer–immune set point," *Nature*, vol. 541, no. 7637, pp. 321–330, Jan. 2017.

[11] M. Nishino, N. H. Ramaiya, H. Hatabu, and F. S. Hodi, "Monitoring immune-checkpoint blockade: Response evaluation and biomarker development," *Nature Rev. Clin. Oncol.*, vol. 14, no. 11, pp. 655–668, Nov. 2017.

[12] J. Cao and Q. Yan, "Cancer epigenetics, tumor immunity, and immunotherapy," *Trends Cancer*, vol. 6, no. 7, pp. 580–592, Jul. 2020.

[13] L. Zitvogel, L. Apetoh, F. Ghiringhelli, and G. Kroemer, "Immunological aspects of cancer chemotherapy," *Nature Rev. Immunol.*, vol. 8, no. 1, pp. 59–73, Jan. 2008.

[14] P. Gotwals, S. Cameron, D. Cipolletta, V. Cremasco, A. Crystal, B. Hewes, B. Müeller, S. Quaratino, C. Sabatos-Peyton, L. Petruzzelli, J. A. Engelman, and G. Dranoff, "Prospects for combining targeted and conventional cancer therapy with immunotherapy," *Nature Rev. Cancer*, vol. 17, no. 5, pp. 286–301, May 2017.

[15] M. Sharifi, A. A. Jamshidi, and N. N. Sarvestani, "An adaptive robust control strategy in a cancer tumor-immune system under uncertainties," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 865–873, May 2019.

[16] H. Jiao, Q. Shen, Y. Shi, and P. Shi, "Adaptive tracking control for uncertain cancer-tumor-immune systems," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2753–2758, Nov. 2021.

[17] M. Itik, M. U. Salamci, and S. P. Banks, "SDRE optimal control of drug administration in cancer treatment," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 18, pp. 715–729, Jan. 2010.

[18] U. Ledzewicz, M. Naghnaeian, and H. Schattler, "Bifurcation of singular arcs in an optimal control problem for cancer immune system interactions under treatment," in *Proc. 49th IEEE Conf. Decis. Control (CDC)*, Dec. 2010, pp. 7039–7044.

[19] D. Liu, S. Xue, B. Zhao, B. Luo, and Q. Wei, "Adaptive dynamic programming for control: A survey and recent advances," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 1, pp. 142–160, Jan. 2021.

[20] 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, Jun. 2008.

[21] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Mar. 2014.

[22] M. Ha, D. Wang, and D. Liu, "A novel value iteration scheme with adjustable convergence rate," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 28, 2022, doi: 10.1109/TNNLS.2022.3143527.

[23] H. Jiang and B. Zhou, "Bias-policy iteration based adaptive dynamic programming for unknown continuous-time linear systems," *Automatica*, vol. 136, Feb. 2022, Art. no. 110058.

[24] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 3051–3056, Nov. 2014.

[25] C. Chen, H. Modares, K. Xie, F. L. Lewis, Y. Wan, and S. Xie, "Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4423–4438, Nov. 2019.

[26] J. Lu, Q. Wei, Y. Liu, T. Zhou, and F.-Y. Wang, "Event-triggered optimal parallel tracking control for discrete-time nonlinear systems," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 6, pp. 3772–3784, Jun. 2022.

[27] B. Zhao, D. Liu, and Y. Li, "Observer based adaptive dynamic programming for fault tolerant control of a class of nonlinear systems," *Inf. Sci.*, vol. 384, pp. 21–33, Apr. 2017.

[28] H. G. Zhang, K. Zhang, Y. Cai, and H. Jian, "Adaptive fuzzy fault-tolerant tracking control for partially unknown systems with actuator faults via integral reinforcement learning method," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 10, pp. 1986–1998, Oct. 2019.

[29] X. Shan, L. Biao, and L. Derong, "Event-triggered adaptive dynamic programming for zero-sum game of partially unknown continuous-time nonlinear systems," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 50, no. 9, pp. 3189–3199, Sep. 2020.

[30] K. G. Vamvoudakis and F. L. Lewis, "Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations," *Automatica*, vol. 47, no. 8, pp. 1556–1569, Aug. 2011.

[31] M. Li, J. Qin, N. M. Freris, and D. W. C. Ho, "Multiplayer Stackelberg–nash game for nonlinear system via value iteration-based integral reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1429–1440, Apr. 2022.

[32] D. Liu, D. Wang, F.-Y. Wang, H. Li, and X. Yang, "Neural-network-based online HJB solution for optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2834–2847, Dec. 2014.

[33] D. Wang and D. Liu, "Learning and guaranteed cost control with event-based adaptive critic implementation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6004–6014, Dec. 2018.

[34] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, May 2005.

[35] H. Su, H. Zhang, H. Jiang, and Y. Wen, "Decentralized event-triggered adaptive control of discrete-time nonzero-sum games over wireless sensor-actuator networks with input constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4254–4266, Oct. 2020.

[36] Q. Wei, J. Lu, T. Zhou, X. Cheng, and F.-Y. Wang, "Event-triggered near-optimal control of discrete-time constrained nonlinear systems with application to a boiler-turbine system," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 3926–3935, Jun. 2022.

[37] X. Yang and H. B. He, "Event-triggered robust stabilization of nonlinear input-constrained systems using single network adaptive critic designs," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 9, pp. 3145–3157, Sep. 2020.

[38] S. Xue, B. Luo, D. Liu, and Y. Gao, "Neural network-based event-triggered integral reinforcement learning for constrained H_∞ tracking control with experience replay," *Neurocomputing*, vol. 513, pp. 25–35, Nov. 2022.

[39] Z. Chen, S.-Z. Chen, K. Chen, and Y. Zhang, "Constrained decoupling adaptive dynamic programming for a partially uncontrollable time-delayed model of energy systems," *Inf. Sci.*, vol. 608, pp. 1352–1374, Aug. 2022.

[40] X. Yang and Q. Wei, "Adaptive critic designs for optimal event-driven control of a CSTR system," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 484–493, Jan. 2021.

[41] D. Wang and C. Mu, "Adaptive-critic-based robust trajectory tracking of uncertain dynamics and its application to a spring–mass–damper system," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 654–663, Jan. 2018.

[42] B. Zhao and D. Liu, "Event-triggered decentralized tracking control of modular reconfigurable robots through adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 67, no. 4, pp. 3054–3064, Apr. 2020.

[43] D. Liu, Y. Xu, Q. Wei, and X. Liu, "Residential energy scheduling for variable weather solar energy based on adaptive dynamic programming," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 36–46, Jan. 2018.

[44] Q. Wei, J. Lu, T. Zhou, X. Cheng, and F.-Y. Wang, "Event-triggered near-optimal control of discrete-time constrained nonlinear systems with application to a boiler-turbine system," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 3926–3935, Jun. 2022.

[45] X. Bu, Y. Xiao, and H. Lei, "An adaptive critic design-based fuzzy neural controller for hypersonic vehicles: Predefined behavioral nonaffine control," *IEEE/ASME Trans. Mechatron.*, vol. 24, no. 4, pp. 1871–1881, May 2019.

[46] X. Bu and Q. Qi, "Fuzzy optimal tracking control of hypersonic flight vehicles via single-network adaptive critic design," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 1, pp. 270–278, Jan. 2022.

[47] J. Sun, H. Zhang, Y. Yan, S. Xu, and X. Fan, "Optimal regulation strategy for nonzero-sum games of the immune system using adaptive dynamic programming," *IEEE Trans. Cybern.*, early access, Aug. 31, 2021, doi: 10.1109/TCYB.2021.3103820.

[48] J. Sun, J. Dai, H. Zhang, S. Yu, S. Xu, and J. Wang, "Neural-network-based immune optimization regulation using adaptive dynamic programming," *IEEE Trans. Cybern.*, early access, Jun. 29, 2022, doi: 10.1109/TCYB.2022.3179302.

[49] Q. Li, L. Xia, and R. Song, "Output resilient containment control of heterogeneous systems with active leaders using reinforcement learning under attack inputs," *IEEE Access*, vol. 7, pp. 162219–162228, 2019.

[50] J. Li, J. Xi, M. He, and B. Li, "Formation control for networked multiagent systems with a minimum energy constraint," *Chin. J. Aeronaut.*, vol. 36, no. 1, pp. 349–362, 2023.

[51] X. Yang, L. Liao, Q. Yang, B. Sun, and J. Xi, "Limited-energy output formation for multiagent systems with intermittent interactions," *J. Franklin Inst.*, vol. 358, no. 13, pp. 6462–6489, Sep. 2021.

[52] X. Bu, B. Jiang, and H. Lei, "Low-complexity fuzzy neural control of constrained waverider vehicles via fragility-free prescribed performance approach," *IEEE Trans. Fuzzy Syst.*, early access, Oct. 26, 2022, doi: 10.1109/TFUZZ.2022.3217378.

[53] X. Bu, B. Jiang, and H. Lei, "Nonfragile quantitative prescribed performance control of waverider vehicles with actuator saturation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 3538–3548, Aug. 2022.

[54] J. Huang, *Nonlinear Output Regulation: Theory and Applications*. Philadelphia, PA, USA: SIAM, 2004.

[55] D. Wang, D. Liu, Y. Zhang, and H. Li, "Neural network robust tracking control with adaptive critic framework for uncertain nonlinear systems," *Neural Netw.*, vol. 97, pp. 11–18, Jan. 2018.

[56] J. Xi, L. Wang, J. Zheng, and X. Yang, "Energy-constraint formation for multiagent systems with switching interaction topologies," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 7, pp. 2442–2454, Jul. 2020.

LIN CHEN received the B.S. degree in clinical medicine from Gannan Medical University, Jiangxi, China, in 2017, and the M.S. degree in oncology from the Sun Yat-Sen University Cancer Center, Guangzhou, China, in 2020. She is currently pursuing the Ph.D. degree in molecular medicine with The Seventh Affiliated Hospital, Sun Yat-sen University. Her current research interests include tumor immunity and regulation of tumor micro environment.

YONG-WEI ZHANG received the B.S. degree in automation from the School of Electronic and Information Engineering, Jiaying University, Meizhou, China, in 2016, and the Ph.D. degree in control science and engineering from the School of Automation, Guangdong University of Technology, Guangzhou, China, in 2021. He is currently a Postdoctoral Fellow with the Guangdong University of Technology. His current research interests include adaptive dynamic programming, optimal control, and multi-agent systems.

SHUN-CHAO ZHANG received the B.S. degree in measurement and control technology and instrument from the School of Electrical and Information Engineering, Hunan Institute of Engineering, Xiangtan, China, in 2016, and the M.S. degree in control engineering and the Ph.D. degree in control science and engineering from the School of Automation, Guangdong University of Technology, Guangzhou, China, in 2019 and 2022, respectively. He is currently a Lecturer with the School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou. His current research interests include optimal control and adaptive dynamic programming.