Adaptive Optimal Control of Networked Nonlinear Systems With Stochastic Sensor and Actuator Dropouts Based on Reinforcement Learning

Jiang, Yi; Liu, Lu; Feng, Gang

Published in:
IEEE Transactions on Neural Networks and Learning Systems

Published: 01/03/2024

Document Version:
Post-print, also known as Accepted Author Manuscript, Peer-reviewed or Author Final version

Published version (DOI):
10.1109/TNNLS.2022.3183020

Publication details:
Jiang, Y., Liu, L., & Feng, G. (2024). Adaptive Optimal Control of Networked Nonlinear Systems With Stochastic Sensor and Actuator Dropouts Based on Reinforcement Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(3), 3107-3120. https://doi.org/10.1109/TNNLS.2022.3183020

Jiang, Y., Liu, L., & Feng, G. (2022). Adaptive Optimal Control of Networked Nonlinear Systems With Stochastic Sensor and Actuator Dropouts Based on Reinforcement Learning. *IEEE Transactions on Neural Networks and Learning Systems, 35*(3), 3197 – 3120. https://doi.org/10.1109/TNNLS.2022.3183020

Adaptive Optimal Control of Networked Nonlinear Systems with Stochastic Sensor and Actuator Dropouts based on Reinforcement Learning

Yi Jiang, *Member, IEEE*, Lu Liu, *Senior Member, IEEE*, Gang Feng, *Fellow, IEEE*

***Abstract*—This paper investigates the adaptive optimal control problem for networked discrete-time nonlinear systems with stochastic packet dropouts in both controller-to-actuator and sensor-to-controller channels. A Bernoulli model based Hamilton-Jacobi-Bellman (BHJB) equation is firstly developed to deal with the corresponding non-adaptive optimal control problem with known system dynamics and probability models of packet dropouts. The solvability of the non-adaptive optimal control problem is analyzed, the stability and optimality of the resulting closed-loop system are proven. Two reinforcement learning (RL) based policy iteration (PI) and value iteration (VI) algorithms are further developed to obtain the solution to the BHJB equation, and their convergence analysis is also provided. Furthermore, in the absence of *a priori* knowledge of partial system dynamics and probabilities of packet dropouts, two more online RL based PI and VI algorithms are developed by using critic-actor approximators and packet dropout probability estimator. It is shown that the concerned adaptive optimal control problem can be solved by the proposed online RL based PI and VI algorithms. Finally, simulation studies of a single-link manipulator are provided to illustrate the effectiveness of the proposed approaches.**

***Index Terms*—Adaptive optimal control, networked discrete-time nonlinear systems, reinforcement learning, Bernouli model based Hamilton-Jacobi-Bellman equation.**

I. Introduction

PRACTICAL systems in the real world are often nonlinear in nature [1], [2]. The optimal control problem for nonlinear systems has been a focus in the control community for a long time and many approaches have been proposed, such as calculus of variations [3], Pontryagin maximum principle [4], dynamic programming [5], [6], etc.. For finite horizon nonlinear optimal control problems, the calculus of variations and Pontryagin maximum principle are applicable but they produce open-loop control trajectories as a function of time as opposed to an optimal state feedback policy [7]. Moreover, the optimal control input obtained via a finite horizon performance index may lead to an unstable system. The stability will be ensured by designing a proper terminal cost function [8]. For infinite horizon nonlinear optimal control problems, dynamic programming offers a powerful tool and produces an optimal state feedback policy by establishing and solving a so-called Hamilton-Jacobi-Bellman (HJB) equation [6], [9].

This work was supported by the City University of Hong Kong under Grant CityU APRC 9610437 and 7005640. *(Corresponding author: Lu Liu.)*

The authors are with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong SAR (e-mail: yjian22@cityu.edu.hk, luliu45@cityu.edu.hk, megfeng@cityu.edu.hk).

However, the dynamic programming based optimal control approaches suffer from the inherent computational complexity issue, also known as the *curse of dimensionality*. Therefore, the accurate solution to the HJB equation remains intractable in all but the simplest cases. To obtain the solution to the HJB equation, as an approximative approach, the reinforcement leaning (RL) based iterative adaptive/approximate dynamic programming (ADP) algorithm was proposed in [10], [11] and gained much attention ever since [12]–[17]. The iterative ADP algorithm can be typically classified into two types, that is, policy iteration (PI) algorithms and value iteration (VI) algorithms. The main difference between these two types of algorithms is that the initial admissible feedback control policy is required in PI algorithms [18]. PI based algorithms were applied to solve the infinite horizon continuous-time affine nonlinear optimal control problem by using online data [19], [20]. The authors in [21]–[23] developed PI and VI based algorithms for solving the discrete-time (DT) affine nonlinear optimal control problem by using accurate system dynamics and critic-actor neural networks (NNs). In [24]–[26], completely model-free VI based algorithms were proposed for solving both affine and nonaffine nonlinear optimal control problems by using online data and model-critic-actor NNs.

It is noted that the online approaches developed for solving the nonlinear optimal control problem in the above-mentioned literature and references therein require that the online data are accurate, which implies that the communication links in the whole control system are reliable. However, in some practical applications, actuators, sensors and controllers of control systems may be distributed and connected by a communication network or multiple communication networks, such as industrial Ethernet, wireless network, Bluetooth, ZigBee. Such control systems are usually called networked control systems (NCSs) [27]. The network-induced transmission problems such as packet dropouts and time-delays, need to be taken into account in NCS analysis and synthesis. Packet dropouts are often described by stochastic processes [28]. A critical issue in solving nonlinear optimal control problems under packet dropouts is to design an optimal controller to minimize the performance index and tolerate the influence caused by the packet dropouts in both the controllenr-to-actuator (C/A) and sensor-to-controller (S/C) channels and at the same time to guarantee the stochastical stability of the resulting closed-loop system. The authors in [29]–[31] applied the model predictive control (MPC) framework to solve the finite horizon optimal control

problem for nonlinear systems in the presence of data loss in communication channels. However, the MPC scheme requires accurate mathematical model of the concerned nonlinear system and the finite horizon optimal performance index may lead to an optimal but unstable closed-loop system. Even though proper terminal cost functions or terminal constraints can be utilized to guarantee the stochastical stability of the closed-loop system, design of such proper terminal cost functions or terminal constraints requires the accurate mathematical model of the concerned nonlinear system [8], [32].

When the accurate mathematical model of the concerned system is not available and only S/C channel has packet dropouts, the Smith predictor based scheme can be applied to design direct adaptive optimal controllers for linear NCSs [33], [34] and indirect adaptive optimal controllers for nonlinear NCSs [35]. However, the C/A and S/C channels usually employ the same physical network link. This implies that the packet dropouts in both C/A and S/C channels often happen simultaneously. The packet dropouts in both channels should be considered simultaneously for analysis and synthesis of NCSs. In this work, we consider the case that the packet dropouts in both C/A and S/C channels are modeled by Bernoulli models and develop a novel RL based framework to solve the adaptive nonlinear optimal control problem in the absence of *a priori* knowledge of partial system dynamics and probabilities of packet dropouts. There are two major challenges in solving such a problem: 1) how to involve the influence of the packet dropouts on the traditional HJB equation; and 2) how to solve the resulting Bernoulli model based HJB (BHJB) equation adaptively. The authors in [36] introduced a nonlinear NCS representation incorporating the system uncertainties and network imperfections by using input and output measurements. An on-line neuro dynamic programming technique was developed to solve the stochastic optimal regulation problem of the concerned nonlinear NCS. It was shown that the closed-loop system is stochastically stable in the sense that all the closed-loop signals and neural network weights are uniformly ultimately bounded (UUB) in the mean while the approximated controller converges close to its target one in the sense of expectation with time. Contrary to the methodology in [36], we establish a deterministic formulation for the concerned problem via the newly developed BHJB equation, and propose novel deterministic approaches to solve this deterministic BHJB equation by using available measurements directly. Furthermore, we show that the learned optimal controller converges to the ideal one deterministically. The main contributions of this work can be summarized as follows.

1) We develop a BHJB equation by taking into consideration of stochastic packet dropouts for the infinite horizon nonlinear optimal control problem. The solvability of this problem is analyzed. It is shown that the solution to the BHJB results in an optimal feedback control policy and a globally stochastically asymptotically stable (GSAS) closed-loop system.
2) Two RL based algorithms, namely, the PI and VI algorithms, are proposed with the known system model to compute the optimal value function and feedback control policy iteratively. It is shown that the computed value function sequence is monotonically non-increasing in the proposed PI algorithm and monotonically non-decreasing in the proposed VI algorithm until convergence.
3) In the absence of *a priori* knowledge of partial system dynamics and probabilities of packet dropouts, two RL based online PI and VI algorithms are proposed to approximately compute the optimal value function and optimal feedback control policy by using critic-actor approximators and packet dropout probability estimator. In the online algorithms, reliable and unreliable data are utilized in different ways. The computed approximate controller is proven to be able to stochastically stabilize the nonlinear system with packet dropouts in both C/A and S/C channels.

The remainder of this paper is organized as follows: Section II formulates the adaptive nonlinear optimal control problem of networked nonlinear systems with packet dropouts in both S/C and C/A channels. In Section III, we develop the BHJB equation for solving the adaptive nonlinear optimal control problem and discuss its solvability. Then, the global stochastical asymptotical stability and optimality of the resulting closed-loop system are analyzed. In Section IV, two model-based RL algorithms for solving the BHJB equation and two online RL algorithms for solving the adaptive nonlinear optimal control problem are designed. In Section V, simulation studies of a single-link manipulator are provided to show the effectiveness of the proposed approaches. Section VI contains the concluding remarks and future work.

Notation: Throughout this paper, $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{N}$ denote the sets of real numbers, nonnegative real numbers and natural numbers excluding 0, respectively. For $n \in \mathbb{N}$ and $X \in \mathbb{R}^{n\times n}$, $X > 0$ ($X \geq 0$) means X is positive definite (positive semi-definite); X^{-1} denotes the inverse of nonsingular matrix X. For $m, n \in \mathbb{N}$ and $X \in \mathbb{R}^{m\times n}$, X^{T} denotes the transpose of X. $0_{a\times b}$ denotes a null matrix of dimension $a \times b$ and I_a denotes an identity matrix of dimension $a \times a$. For brevity, denote $\mathbb{R}^n := \mathbb{R}^{n\times 1}$. Moreover, $|\cdot|$ means the absolute value; $\|\cdot\|$ denotes Euclidean norm for vectors or the Frobenius norm for matrices; $\mathbb{P}(\cdot)$, $\mathbb{E}(\cdot)$ and $\mathbb{E}(\cdot|\cdot)$ denote the probability, expectation and conditional expectation, respectively; $\cap$ denotes the intersection of two sets; $\sup$ denotes the supremum; $\arg\min$ stands for the argument of the minimization operation. Definitions of $\mathcal{K}$, $\mathcal{K}_\infty$ and $\mathcal{KL}$ functions can be found in [2, Definitions 4.2-4.3]. $\mathcal{P}^n$ denotes the set of all bounded positive definite functions from $\mathbb{R}^n$ to $\mathbb{R}_+$.

II. Problem Formulation

Consider the following affine nonlinear DT system,

$$x(k+1) = f(x(k)) + g(x(k))u(k), \tag{1}$$

where $x \in \mathbb{R}^{n_x}$ and $u \in \mathbb{R}^{n_u}$ are the state and input, respectively. $f(\cdot) : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ and $g(\cdot) : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x\times n_u}$ are smooth and satisfy $f(0) = 0_{n_x\times 1}$ and $g(0) \neq 0_{n_x\times n_u}$. The following assumptions are made for $f(\cdot)$ and $g(\cdot)$ throughout this paper where Assumption 3 ensures that the concerned affine nonlinear DT system (1) is globally stablizable.

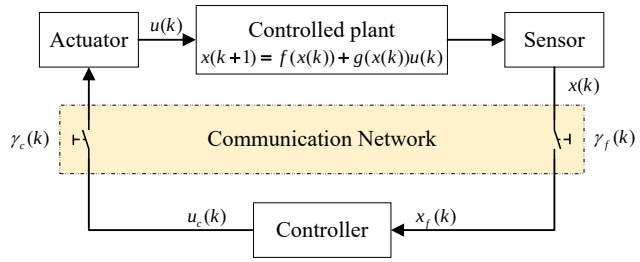


Fig. 1. Networked nonlinear control system with network-induced two channels dropouts.

Assumption 1: The function $f(\cdot)$ is unknown but the function $g(\cdot)$ is known.

Assumption 2: $f(\cdot)$ is globally Lipschitz, that is, for arbitrary $\mathbf{x}_1 \in \mathbb{R}^{n_x}$ and $\mathbf{x}_2 \in \mathbb{R}^{n_x}$, there exists a constant $L > 0$ such that $\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$.

Assumption 3: There exist a positive definite function $V^a(\cdot) \in \mathcal{P}^{n_x}$, a mapping $\kappa^a(\cdot) : \mathbb{R}^{n_x} \to \mathbb{R}^{n_u}$ with $\kappa^a(0) = 0$ such that for all $\mathbf{x} \in \mathbb{R}^{n_x}$, the following inequalities hold,

$$\begin{aligned}\underline{\alpha}^a(\|\mathbf{x}\|) \leq V^a(\mathbf{x}) \leq& \bar{\alpha}^a(\|\mathbf{x}\|),\\ V^a(f(\mathbf{x}) + g(\mathbf{x})\kappa^a(\mathbf{x})) - V^a(\mathbf{x}) \leq& -\alpha^a(\|\mathbf{x}\|),\end{aligned}$$

where $\underline{\alpha}^a$, $\bar{\alpha}^a$ and α^a are functions of class $\mathcal{K}_\infty$.

In this paper, we consider a NCS where the sensor signal and control signal are transmitted via a shared communication network. It is assumed that packet dropout phenomenon would occur and the shared communication network follows the Transmission Control Protocol. In other words, the S/C and C/A channel signals may be dropped through the shared communication network, as illustrated in Fig. 1. Two stochastic variables $\gamma_f(k), \gamma_c(k) \in \{0, 1\}$ are used to describe whether packet dropouts in S/C and C/A channels occur or not and they satisfy the Bernoulli model [37], that is, $\mathbb{P}(\gamma_f(k) = 1) = \bar{\gamma}_f$ and $\mathbb{P}(\gamma_c(k) = 1) = \bar{\gamma}_c$. Specifically, $\gamma_f(k) = 0$ and $\gamma_c(k) = 0$ mean the packet dropouts occur in S/C and C/A channels, respectively. The Bernoulli process is ideal to model packet dropouts and has been widely adopted in many works, see, for example, [31], [36], [37]. It is accurate enough to model the key properties of practical packet dropouts and simple enough for its feasible theoretical analysis. We thus adopt the Bernoulli processes for packet dropouts. Let $x_f(k)$ and $u_c(k)$ denote the sensor signal after network transmission and the controller signal before network transmission. As illustrated in Fig. 1, one has

$$x_f(k) = \gamma_f(k)x(k), \tag{2}$$

$$u_c(k) = \kappa(x_f(k)), \tag{3}$$

where $\kappa(\cdot)$ is the feedback control policy to be designed satisfying $\kappa(0) = 0$. In addition, the control input to the plant or the controller signal after network transmission can be expressed as follows,

$$u(k) = \gamma_c(k)\gamma_f(k)\kappa(x(k)). \tag{4}$$

The adaptive nonlinear optimal control problem under consideration can be formulated as follows: to design an adaptive optimal feedback control policy $\kappa^*(\cdot)$ by using the available information set $\mathbb{I}_{[1,k]} := \{g(\cdot), x_f(1), x_f(2), \ldots, x_f(k), \gamma_f(1), \gamma_f(2), \ldots, \gamma_f(k), u_c(1), u_c(2), \ldots, u_c(k), \gamma_c(1), \gamma_c(2), \ldots, \gamma_c(k-1)\}$ to minimize the following infinite horizon performance index,

$$J(x(k), \kappa) = \mathbb{E}\left\{\sum_{i=k}^{\infty}\left[Q(x(i)) + u^{\mathrm{T}}(i)Ru(i)\right]\right\}, \tag{5}$$

where $Q(\cdot) \in \mathcal{P}^{n_x}$ and $R > 0$. We will refer to the corresponding adaptive nonlinear optimal control problem as the nonlinear optimal control problem when the concerned nonlinear system dynamics and probabilities of packet dropouts in S/C and C/A channels are completely known, that is, function $f(\cdot)$, $\bar{\gamma}_f$ and $\bar{\gamma}_c$ are also known.

III. Solution of the Nonlinear Optimal Control Problem

In this section, we present a solution to the nonlinear optimal control problem by establishing the BHJB equation. Then, the solvability of the nonlinear optimal control problem, optimality and global stochastical asymptotical stability of the resulting closed-loop system are analyzed.

A. BHJB Equation

To address the nonlinear optimal control problem presented in the previous section, we develop a BHJB equation involving packet dropouts by using dynamic programming, principle of Bellman optimality and stochastic control theory [3], [5], [6], [9], [38].

It follows from the infinite horizon performance index (5) that for each feedback control policy $\kappa(\cdot)$, the Bellman equation of this problem can be expressed as follows,

$$\begin{aligned}V(x(k)) =& \mathbb{E}\left\{Q(x(k)) + u^{\mathrm{T}}(k)Ru(k) + V(x(k+1))|x(k)\right\}\\ =& Q(x(k)) + \bar{\gamma}\kappa^{\mathrm{T}}(x(k))R\kappa(x(k))\\ &+ (1-\bar{\gamma})V(f(x(k)))\\ &+ \bar{\gamma}V(f(x(k)) + g(x(k))\kappa(x(k))),\end{aligned} \tag{6}$$

where $V(x(k))$ is the value function and $\bar{\gamma} = \bar{\gamma}_c\bar{\gamma}_f$. Via the Bellman equation (6), the Hamiltonian function can be developed as follows,

$$\begin{aligned}H(x(k), \kappa, V) =& Q(x(k)) + \bar{\gamma}\kappa^{\mathrm{T}}(x(k))R\kappa(x(k))\\ &+ \bar{\gamma}V(f(x(k)) + g(x(k))\kappa(x(k)))\\ &+ (1-\bar{\gamma})V(f(x(k))) - V(x(k)).\end{aligned} \tag{7}$$

Based on the principle of Bellman optimality, the optimal value function $V^*(x(k))$ satisfies the following BHJB equation,

$$\begin{aligned}V^*(x(k)) = \min_{\kappa(x(k))}&\left[Q(x(k)) + \bar{\gamma}\kappa^{\mathrm{T}}(x(k))R\kappa(x(k))\right.\\ &+ (1-\bar{\gamma})V^*(f(x(k)))\\ &\left.+\bar{\gamma}V^*(f(x(k)) + g(x(k))\kappa(x(k)))\right].\end{aligned} \tag{8}$$

It follows from [3], [5], [6] that a necessary condition for the optimality is the stationarity condition given by

$$\left.\frac{\partial H(x(k), \kappa, V^*)}{\partial \kappa}\right|_{\kappa=\kappa^*} = 0. \tag{9}$$

Therefore, the optimal feedback control policy can be obtained as follows,

$$\kappa^*(x(k)) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x(k))\frac{\partial V^*(x^*(k+1))}{\partial x^*(k+1)}, \tag{10}$$

where $x^*(k+1) = f(x(k)) + g(x(k))\kappa^*(x(k))$. Substituting (10) to (8) yields the following BHJB equation in general form,

$$\begin{aligned} V^*(x(k)) =& Q(x(k)) + \frac{\bar{\gamma}}{4}\left(\frac{\partial V^*(x^*(k+1))}{\partial x^*(k+1)}\right)^{\mathrm{T}} g(x(k))R^{-1} \\ & \cdot g^{\mathrm{T}}(x(k))\frac{\partial V^*(x^*(k+1))}{\partial x^*(k+1)} + \bar{\gamma}V^*(x^*(k+1)) \\ & + (1-\bar{\gamma})V^*(f(x(k))). \end{aligned} \tag{11}$$

It can be summarized from the above discussions that the concerned problem is solved if the BHJB equation (11) can be solved and the optimal feedback control policy can be obtained from (10). However, it is worth noting that, even if the function $f(\cdot)$, $\bar{\gamma}_f$ and $\bar{\gamma}_c$ are completely known, solving the BHJB equation and thus the concerned problem is non-trivial in general and often quite challenging. In fact it is almost impossible to find the closed form solution to the BHJB equation except for the simplest cases due to its high nonlinearity and complexity.

B. Solvability of the Nonlinear Optimal Control Problem

Before showing how to solve the BHJB equation developed in the previous subsection, naturally, there arises a question: is this problem solvable for arbitrary $\bar{\gamma} \in [0,1]$? Consider the case with $\bar{\gamma} = 0$ and with system (1) being open-loop unstable. It is clear that this problem is not solvable in this case. This implies that the existence of the positive definite solution to the BHJB equation, and thus the solvability of this problem critically depends on the parameter $\bar{\gamma}$. In what follows, we will provide a way to reveal the relationship between $\bar{\gamma}$ and the solvability of the nonlinear optimal control problem.

To this end, we will show how the parameter $\bar{\gamma}$ influence the solution $V^*(\cdot)$ to the BHJB equation (8), which can guide us to explore the relationship between $\bar{\gamma}$ and the solvability of the nonlinear optimal control problem. For notation convenience, we use $V^*(x(k), a)$ to denote $V^*(x(k))$ with $\bar{\gamma} = a$. Notice that the partial derivative of $V^*(x(k), \bar{\gamma})$ with respect to $\bar{\gamma}$ can be computed as follows,

$$\begin{aligned} \frac{\partial V^*(x(k),\bar{\gamma})}{\partial \bar{\gamma}} =& \kappa^{*\mathrm{T}}(x(k))R\kappa^*(x(k)) - V^*(f(x(k)),\bar{\gamma}) \\ & + V^*(x^*(k+1),\bar{\gamma}) + \bar{\gamma}\frac{\partial V^*(x^*(k+1),\bar{\gamma})}{\partial \bar{\gamma}} \\ & + (1-\bar{\gamma})\frac{\partial V^*(f(x(k)),\bar{\gamma})}{\partial \bar{\gamma}} \\ =& \mathbb{E}\left\{\kappa^{*\mathrm{T}}(x(k))R\kappa^*(x(k)) - V^*(f(x(k),\bar{\gamma})\right. \\ & \left. + V^*(x^*(k+1),\bar{\gamma}) + \frac{\partial V^*(x(k+1),\bar{\gamma})}{\partial \bar{\gamma}}\right\} \\ =& \mathbb{E}\left\{\sum_{i=k}^{\infty}\left[\kappa^{*\mathrm{T}}(x(i))R\kappa^*(x(i))\right.\right. \\ & \left.\left. - V^*(f(x(i),\bar{\gamma}) + V^*(x^*(i+1),\bar{\gamma})\right]\right\}. \end{aligned}$$

Since the optimal feedback control policy is computed by using the condition (9), if $V^*(x(k),\bar{\gamma}) \in \mathcal{P}^{n_x}$, the optimal feedback control policy $\kappa^*(x(k))$ will minimize $\kappa^{*\mathrm{T}}(x(k))R\kappa^*(x(k)) + V^*(f(x(k)) + g(x(k))\kappa^*(x(k)),\bar{\gamma})$. As a result, for all $\kappa(\cdot) \neq \kappa^*(\cdot)$ and $x(k) \neq 0$, if $V^*(x(k),\bar{\gamma}) \in \mathcal{P}^{n_x}$, the following inequality holds,

$$\begin{aligned} & \frac{\partial V^*(x(k),\bar{\gamma})}{\partial \bar{\gamma}} - (1-\bar{\gamma})\frac{\partial V^*(f(x(k)),\bar{\gamma})}{\partial \bar{\gamma}} \\ & - \bar{\gamma}\frac{\partial V^*(x^*(k+1),\bar{\gamma})}{\partial \bar{\gamma}} \\ <& \kappa^{\mathrm{T}}(x(k))R\kappa(x(k)) - V^*(f(x(k)),\bar{\gamma}) + V^*(x^*(k+1),\bar{\gamma}). \end{aligned}$$

Note that the right hand side of the above inequality is 0 if $\kappa(\cdot) = 0$, which implies that the left hand side of the above inequality is negative definite and hence

$$\frac{\partial V^*(x(k),\bar{\gamma})}{\partial \bar{\gamma}} < 0, \forall V^*(x(k),\bar{\gamma}) \in \mathcal{P}^{n_x}, x(k) \neq 0 \tag{12}$$

holds strictly, and thus $V^*(x(k),\bar{\gamma})$ is monotonically decreasing with respect to $\bar{\gamma}$. Consider the case $\bar{\gamma} = 1$. It then follows from Assumption 3, $Q(\cdot) \in \mathcal{P}^{n_x}$ and $R > 0$ that for all $x(k) \neq 0$,

$$0 < V^*(x(k),1) = \min_{\kappa} J(x(k),\kappa) \leq J(x(k),\kappa^a) < +\infty.$$

Therefore, $V^*(x(k),1) \in \mathcal{P}^{n_x}$.

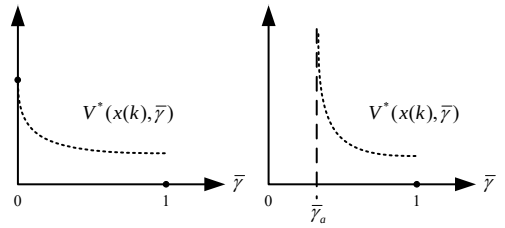
Now, let us discuss the solvability of the nonlinear optimal control problem by analyzing the existence of the positive definite solution to the BHJB equation in the following two scenarios.

Scenario 1): The system (1) *is open-loop asymptotically stable (OLAS).* In this scenario, the solution $V^*(x(k),0) = \mathbb{E}\{\sum_{i=k}^{\infty} Q(x(i))\}$ to the BHJB equation satisfies $V^*(x(k),0) \in \mathcal{P}^{n_x}$. It then follows from $V^*(x(k),0) \in \mathcal{P}^{n_x}$, $V^*(x(k),1) \in \mathcal{P}^{n_x}$ and the monotone property of $V^*(x(k),\bar{\gamma})$ with respect to $\bar{\gamma}$ that $V^*(x(k),0) \geq V^*(x(k),\bar{\gamma}) \geq V^*(x(k),1), \forall \bar{\gamma} \in [0,1]$, which implies that $V^*(x(k),\bar{\gamma}) \in \mathcal{P}^{n_x}, \forall \bar{\gamma} \in [0,1]$. Based on the above analysis, the relationship between $\bar{\gamma}$ and $V^*(x(k),\bar{\gamma})$ for all $x(k) \neq 0$ when the system (1) is OLAS can be illustrated in Fig. 2 (a). As a result, this problem is always solvable for all $\bar{\gamma} \in [0,1]$ when the system (1) is OLAS.

Scenario 2): The system (1) *is not OLAS.* To begin with, we define the following critical probability $\bar{\gamma}_a$ under which this problem is unsolvable when the system (1) is not OLAS,

$$\begin{aligned} \bar{\gamma}_a :=& \sup_{\bar{\gamma}}\{0 \leq \bar{\gamma} \leq 1 | V^*(x(k),\bar{\gamma}) \notin \mathcal{P}^{n_x} \text{ when the} \\ & \text{system (1) is not OLAS}\}. \end{aligned} \tag{13}$$

In this scenario, the solution $V^*(x(k),0)$ to the BHJB equation can be obtained by solving $V^*(x(k),0) = Q(x(k)) + V^*(f(x(k)),0)$. If $V^*(x(k),0) \in \mathcal{P}^{n_x}$, the global stochastical asymptotical stability of the open-loop system can be established by considering $V^*(x(k),0)$ as a Lyapunov function candidate, which contradicts the fact that the system (1) is not OLAS. This implies that $V^*(x(k),0) \notin \mathcal{P}^{n_x}$. It further follows from $V^*(x(k),0) \notin \mathcal{P}^{n_x}$, $V^*(x(k),1) \in \mathcal{P}^{n_x}$, and the monotone property of $V^*(x(k),\bar{\gamma})$ with respect to $\bar{\gamma}$ that there exists



(a) When the system (1) is OLAS. (b) When the system (1) is not OLAS.

Fig. 2. Relationship between $\bar{\gamma}$ and $V^*(x(k),\bar{\gamma})$ for $x(k)\neq 0$.

a point $\bar{\gamma}^*\geq 0$ such that $V^*(x(k),\bar{\gamma})\in\mathcal{P}^{n_x},\forall\bar{\gamma}\in(\bar{\gamma}^*,1]$. In other word, this problem is solvable for $\bar{\gamma}>\bar{\gamma}^*$ when the system (1) is not OLAS. Actually, based on the definition of the critical probability $\bar{\gamma}_a$ in (13), this point $\bar{\gamma}^*$ is the critical probability $\bar{\gamma}_a$. Based on the above analysis, the relationship between $\bar{\gamma}$ and $V^*(x(k),\bar{\gamma})$ for $x(k)\neq 0$ when the system (1) is not OLAS can be illustrated in Fig. 2 (b).

It follows from (12) that $V^*(x(k),\bar{\gamma})<\lim_{\bar{\gamma}\to\bar{\gamma}_a^+}V^*(x(k),\bar{\gamma}),\forall\bar{\gamma}\in(\bar{\gamma}_a,1],x(k)\neq 0$. As a result, the question posed in the beginning of this subsection can be answered by finding the critical probability $\bar{\gamma}_a$ such that $0<V^*(x(k),\bar{\gamma})<\infty,\forall\bar{\gamma}\in(\bar{\gamma}_a,1],x(k)\neq 0$ are satisfied. It follows from the Bellman equation (6) and the BHJB equation (8) that for all $\kappa(\cdot)$, $\gamma\in(\bar{\gamma}_a,1]$ and $x(k)\neq 0$,

$$0<V^*(x(k),\bar{\gamma})=\min_{\kappa}J(x(k),\kappa)\leq J(x(k),\kappa)=V(x(k)).$$

If the solution $V(x(k))$ to the Bellman equation (6) is positive definite and bounded for a fixed $\bar{\gamma}$, the boundedness of $V^*(x(k),\bar{\gamma})$ can be concluded. Thus the following assumption and lemma are provided to show when the Bellman equation (6) has a positive definite and bounded solution.

Assumption 4: There exist a positive definite function $V^0(\cdot)\in\mathcal{P}^{n_x}$ and a mapping $\kappa^0(\cdot):\mathbb{R}^{n_x}\to\mathbb{R}^{n_u}$ with $\kappa^0(0)=0$ such that

$$\begin{aligned}V^0(\mathbf{x})\geq&Q(\mathbf{x})+\bar{\gamma}(\kappa^0(\mathbf{x}))^{\mathrm{T}}R\kappa^0(\mathbf{x})+(1-\bar{\gamma})V^0(f(\mathbf{x}))\\&+\bar{\gamma}V^0(f(\mathbf{x})+g(\mathbf{x})\kappa^0(\mathbf{x})),\forall\mathbf{x}\in\mathbb{R}^{n_x},\end{aligned}\tag{14}$$

where $\kappa^0(\cdot)$ is the so-called admissible feedback control policy.

Lemma 1: Consider the system (1). If Assumption 4 is satisfied, the solution to the Bellman equation (6) is positive definite and bounded under $\kappa^0(x(k))$.

Proof: Consider $V^0(x(k))$ as a Lyapunov function candidate. Its difference can be calculated as

$$\begin{aligned}\Delta V^0(x(k))=&\mathbb{E}\left[V^0(x(k+1))-V^0(x(k))|x(k)\right]\\\leq&-Q(x(k))-\bar{\gamma}(\kappa^0(x(k)))^{\mathrm{T}}R\kappa^0(x(k))\\\leq&-\alpha^0(\|x(k)\|),\end{aligned}$$

where α^0 is a function of class $\mathcal{K}_\infty$. It then follows that the resulting closed-loop system is GSAS. Therefore, the solution $V(x(k))$ to the Bellman equation (6) under $\kappa^0(x(k))$ is thus positive definite and bounded, which completes the proof of Lemma 1. □

It is almost impossible to find the specific expression or the accurate value of $\bar{\gamma}_a$ except for the simplest cases. However, based on the above analysis in this subsection, a rough estimation for the range of the critical probability $\bar{\gamma}_a$ can be provided in the following theorem.

Theorem 1: Consider the system (1). Under Assumptions 2-3, the critical probability $\bar{\gamma}_a$ satisfies

$$\bar{\gamma}_a^{\min}\leq\bar{\gamma}_a\leq\bar{\gamma}_a^{\max},\tag{15}$$

where

$$\bar{\gamma}_a^{\max}:=\max_{\|x(k)\|>0}\frac{\bar{\alpha}^a(L\|x(k)\|)-\underline{\alpha}^a(\|x(k)\|)}{\bar{\alpha}^a(L\|x(k)\|)-\underline{\alpha}^a(\|x(k)\|)+\alpha^a(\|x(k)\|)}$$

and $\bar{\gamma}_a^{\min}$ is the largest value of $\bar{\gamma}$ such that the following equation does not have a positive definite solution,

$$V^{\triangle}(x(k))=Q(x(k))+(1-\bar{\gamma})V^{\triangle}(f(x(k))).\tag{16}$$

Proof: It follows from the statement of Lemma 1 and the analysis in this subsection that the range of the critical probability $\bar{\gamma}_a$ can be estimated by exploring the range of $\bar{\gamma}$ under which Assumption 4 is satisfied. Notice that (14) in Assumption 4 is equivalent to the following inequality,

$$\begin{aligned}&-Q(x(k))-\bar{\gamma}(\kappa^0(x(k)))^{\mathrm{T}}R\kappa^0(x(k))\\\geq&\bar{\gamma}V^0(f(x(k))+g(x(k))\kappa^0(x(k)))-V^0(x(k))\\&+(1-\bar{\gamma})V^0(f(x(k))),\forall x(k)\in\mathbb{R}^{n_x}.\end{aligned}\tag{17}$$

Let $V^0(\cdot)=NV^a(\cdot)$ and $\kappa^0(\cdot)=\kappa^a(\cdot)$ in (17), where N is a positive constant. When the system (1) is not OLAS, under Assumptions 2-3, one has that $L\geq 1$ and the right hand side of (17) satisfies the following inequality,

$$\begin{aligned}&(1-\bar{\gamma})NV^a(f(x(k)))-NV^a(x(k))\\&+\bar{\gamma}NV^a(f(x(k))+g(x(k))\kappa^a(x(k)))\\\leq&(1-\bar{\gamma})NV^a(f(x(k)))+(\bar{\gamma}-1)NV^a(x(k))\\&-\bar{\gamma}N\alpha^a(\|x(k)\|)\\\leq&N[-\bar{\gamma}(\bar{\alpha}^a(L\|x(k)\|)-\underline{\alpha}^a(\|x(k)\|)+\alpha^a(\|x(k)\|))\\&+(\bar{\alpha}^a(L\|x(k)\|)-\underline{\alpha}^a(\|x(k)\|))],\forall x(k)\in\mathbb{R}^{n_x}.\end{aligned}$$

If $\bar{\gamma}>\bar{\gamma}_a^{\max}$, one can always find a sufficiently large N such that the inequality (17) holds. This implies that if $\bar{\gamma}>\bar{\gamma}_a^{\max}$, Assumption 4 always holds and hence $V^*(x(k),\bar{\gamma})\in\mathcal{P}^{n_x}$. Besides, it can be easily verified that $0<\bar{\gamma}_a^{\max}<1$. As a result, based on the definition of the critical probability $\bar{\gamma}_a$ in (13), the critical probability $\bar{\gamma}_a$ satisfies $\bar{\gamma}_a\leq\bar{\gamma}_a^{\max}$. Furthermore, by comparing (16) and the BHJB function (8), one has that if $V^*(x(k))\in\mathcal{P}^{n_x}$, $0\leq V^{\triangle}(x(k))\leq V^*(x(k))$. By noting that $V^{\triangle}(x(k))\in\mathcal{P}^{n_x}$ when $\bar{\gamma}=1$, $V^{\triangle}(x(k))\notin\mathcal{P}^{n_x}$ when $\bar{\gamma}=0$ and

$$\begin{aligned}\frac{\partial V^{\triangle}(x(k))}{\partial\bar{\gamma}}=&-V^{\triangle}(f(x(k)))+(1-\bar{\gamma})\frac{\partial V^{\triangle}(f(x(k)))}{\partial\bar{\gamma}}\\=&\mathbb{E}\left\{\sum_{i=k}^{\infty}\left[-V^{\triangle}(\gamma_c(i)\gamma_f(i)f(x(i)))\right]\right\}\\\leq&0,\forall V^{\triangle}(x(k))\in\mathcal{P}^{n_x},\end{aligned}$$

one has that there exists a positive constant $0\leq\bar{\gamma}_a^{\min}<1$ such that $V^{\triangle}(x(k))\in\mathcal{P}^{n_x},\forall\bar{\gamma}>\bar{\gamma}_a^{\min}$. As a result, based on

the definition of the critical probability $\bar{\gamma}_a$ in (13), $\bar{\gamma}_a$ satisfies $\bar{\gamma}_a \geq \bar{\gamma}_a^{\min}$. The proof of Theorem 1 is thus completed. □

Remark 1: It is in general quite challenging to find $\bar{\gamma}_a^{\min}$ such that (16) does not have a positive definite solution except for the simplest cases due to its high nonlinearity. An alternative approach to finding $\bar{\gamma}_a^{\min}$ is shown as follows. Notice that $V^{\triangle}(x(k))$ can serve as the cost function of the performance index $\mathbb{E}\{\sum_{i=k}^{\infty} Q(x(i))\}$ for the following system,

$$x(k+1) = (1 - \gamma_c(k)\gamma_f(k))f(x(k)). \tag{18}$$

This implies that $\bar{\gamma}_a^{\min}$ can be obtained by analyzing when the system (18) is not GSAS. It follows from Assumption 2 that $\|\mathbb{E}\{x(k+1)\}\| \leq (1-\bar{\gamma})L\|x(k)\|$, $\forall x(k) \in \mathbb{R}^{n_x}$. This implies that if $\bar{\gamma} > 1 - 1/L$, the system (18) is GSAS and hence $\bar{\gamma}_a^{\min} \leq 1 - 1/L$.

Remark 2: It follows from the proof of Theorem 1 that if Assumptions 2-3 are satisfied, there always exists a proper $\bar{\gamma}$ such that the concerned system (1) with packet dropouts in both C/A and S/C channels can be stochastically stabilized. It is well known that the network bandwidth and transmission power determine the expectations $\bar{\gamma}_f$ and $\bar{\gamma}_c$ directly. Thus one just needs to design the network bandwidth and transmission power to achieve $\bar{\gamma} > \bar{\gamma}_{a\max}$.

C. Global Stochastical Asymptotical Stability and Optimality Analysis

In this subsection, we will present the global stochastical asymptotical stability and optimality of the resulting closed-loop system under the optimal feedback control policy obtained via (10).

Theorem 2: Consider the system (1). Under Assumption 4 and the optimal feedback control policy (10), the resulting closed-loop system is GSAS and the control policy minimizes the infinite horizon performance index (5).

Proof: See Appendix A. □

IV. RL based Algorithms for the Adaptive Nonlinear Optimal Control Problem

In this section, we develop two RL based algorithms, PI and VI algorithms, to approximate the optimal value function $V^*(\cdot)$ and the optimal feedback control policy $\kappa^*(\cdot)$. First, the PI and VI algorithms with the known system model are proposed. Then, the online realizations of the PI and VI algorithms are developed by using critic-actor approximators and packet dropout probability estimator. Finally, the performance analysis of the resulting closed-loop system is shown.

A. PI and VI Solutions to the BHJB Equation

The PI algorithm is shown as in Algorithm 1. By iteratively solving the value function $V_i(\cdot)$ in (19) and updating the feedback control policy $\kappa_i(\cdot)$ via (20)-(21), the ideal solution $V^*(\cdot)$ to the BHJB equation (11) and the optimal feedback control policy $\kappa^*(\cdot)$ in (10) are numerically approximated. The following theorem shows the convergence of the proposed PI Algorithm 1.

Theorem 3: Consider the PI Algorithm 1. Under Assumption 4, when the thresholds are set to zero and the step size is chosen small enough, the value function sequence $V_i(\cdot)$ and feedback control policy sequence $\kappa_i(\cdot)$ obtained via the algorithm have the following properties:

1) $V_i(x(k)) \geq V_{i+1}(x(k)) \geq V^*(x(k)), \forall i \geq 0, x(k) \in \mathbb{R}^{n_x}$;
2) $\kappa_i(\cdot), \forall i \geq 0$ are admissible feedback control policies;
3) $\lim_{i\to\infty} |V_i(x(k)) - V^*(x(k))| = 0, \forall x(k) \in \mathbb{R}^{n_x}$;
4) $\lim_{i\to\infty} \|\kappa_i(x(k)) - \kappa^*(x(k))\| = 0, \forall x(k) \in \mathbb{R}^{n_x}$;
5) $\lim_{j\to\infty} \|\kappa_{i,j}(x(k)) - \kappa_i(x(k))\| = 0, \forall x(k) \in \mathbb{R}^{n_x}$.

Proof: See Appendix B. □

The VI algorithm is shown as in Algorithm 2. By iteratively updating the value function $V_{i+1}(\cdot)$ and feedback control policy $\kappa_{i+1}(\cdot)$ via (22)-(24), the ideal solution $V^*(\cdot)$ to the BHJB equation (11) and the optimal feedback control policy $\kappa^*(\cdot)$ in (10) are numerically approximated. Different to the proposed PI Algorithm 1, the proposed VI Algorithm 2 does not require an initial admissible feedback control policy $\kappa_0(\cdot)$ to start the algorithm. However, compared with the PI algorithm, the VI algorithm usually needs more iterations to converge. The following theorem shows the convergence of the proposed VI Algorithm 2.

Theorem 4: Consider the VI Algorithm 2. Under Assumption 4, when the thresholds are set to zero and the step size

---

Algorithm 1 PI Algorithm for the BHJB Equation

Initiation: Choose two small error tolerance thresholds ε_p and ε_a and a small step size α. Start with an initial admissible feedback control policy $\kappa_0(\cdot)$.

Iteration: Iterate the following steps on i from $i = 0$ and j from $j = 0$ until convergence.

1) **Policy Evaluation**: Solve for the cost function $V_i(\cdot)$ using the following equation,

$$\begin{aligned} V_i(x(k)) =& Q(x(k)) + \bar{\gamma}\kappa_i^{\mathrm{T}}(x(k))R\kappa_i(x(k)) \\ &+ \bar{\gamma}V_i(f(x(k)) + g(x(k))\kappa_i(x(k))) \\ &+ (1-\bar{\gamma})V_i(f(x(k))); \end{aligned} \tag{19}$$

2) **Policy Improvement**: Update the feedback control policy $\kappa_{i+1}(\cdot)$ using the following equation,

$$\kappa_{i+1}(x(k)) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x(k))\frac{\partial V_i(x^{i+1}(k+1))}{\partial x^{i+1}(k+1)}, \tag{20}$$

where $x^{i+1}(k+1) = f(x(k)) + g(x(k))\kappa_{i+1}(x(k))$. Since $\kappa_{i+1}(\cdot)$ appears in both left and right hand sides of (20), the feedback control policy $\kappa_{i+1}(\cdot) = \kappa_{i+1,j}(\cdot)$ is computed by using the following equation until $\|\kappa_{i+1,j+1}(x(k)) - \kappa_{i+1,j}(x(k))\| \leq \varepsilon_a$,

$$\begin{aligned} \kappa_{i+1,j+1}(x(k)) =& \kappa_{i+1,j}(x(k)) \\ &- \alpha\frac{\partial \Xi(\kappa_{i+1,j}(x(k)), V_i)}{\partial \kappa_{i+1,j}(x(k))}, \end{aligned} \tag{21}$$

where $\Xi(\kappa_{i+1,j}(x(k)), V_i) := \kappa_{i+1,j}^{\mathrm{T}}(x(k))R\kappa_{i+1,j}(x(k)) + V_i(f(x(k)) + g(x(k))\kappa_{i+1,j}(x(k)))$ and $\kappa_{i+1,0}(\cdot) = \kappa_i(\cdot)$;

3) **Stop if**: $|V_i(x(k)) - V_{i-1}(x(k))| \leq \varepsilon_p$, otherwise set $i \leftarrow i+1$, $j \leftarrow 0$ and go to the policy evaluation.

Algorithm 2 VI Algorithm for the BHJB Equation

Initiation: Choose two small error tolerance thresholds ε_v and ε_a and a small step size α. Start with $\kappa_0(x(k)) = 0$ and set $V_0(x(k)) = 0$.

Iteration: Iterate the following steps on i form $i = 0$ and j from $j = 0$ until convergence.

1) **Value Update**: Update the cost function $V_{i+1}(\cdot)$ using the following equation,

$$\begin{aligned} V_{i+1}(x(k)) =& Q(x(k)) + \bar{\gamma}\kappa_i^{\mathrm{T}}(x(k))R\kappa_i(x(k)) \\ &+ \bar{\gamma}V_i(f(x(k)) + g(x(k))\kappa_i(x(k))) \\ &+ (1-\bar{\gamma})V_i(f(x(k))); \end{aligned} \tag{22}$$

2) **Policy Improvement**: Update the feedback control policy $\kappa_{i+1}(\cdot)$ using the following equation,

$$\begin{aligned} &\kappa_{i+1}(x(k)) \\ &= -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x(k))\frac{\partial V_{i+1}(x^{i+1}(k+1))}{\partial x^{i+1}(k+1)}. \end{aligned} \tag{23}$$

Similarly, the feedback control policy $\kappa_{i+1}(\cdot) = \kappa_{i+1,j}(\cdot)$ is computed by using the following equation until $\|\kappa_{i+1,j+1}(x(k)) - \kappa_{i+1,j}(x(k))\| \leq \varepsilon_a$,

$$\begin{aligned} \kappa_{i+1,j+1}(x(k)) =& \kappa_{i+1,j}(x(k)) \\ &- \alpha\frac{\partial \Xi(\kappa_{i+1,j}(x(k)), V_{i+1})}{\partial \kappa_{i+1,j}(x(k))}; \end{aligned} \tag{24}$$

3) **Stop if**: $|V_{i+1}(x(k)) - V_i(x(k))| \leq \varepsilon_v$, otherwise set $i \leftarrow i+1$, $j \leftarrow 0$ and go to the value update.

is chosen small enough, the value function sequence $V_i(\cdot)$ and feedback control policy sequence $\kappa_i(\cdot)$ obtained via the algorithm have the following properties:

1) $V_i(x(k)) \leq V_{i+1}(x(k)) \leq V^*(x(k)), \forall i \geq 0, x(k) \in \mathbb{R}^{n_x}$;
2) $\lim_{i\to\infty} |V_i(x(k)) - V^*(x(k))| = 0, \forall x(k) \in \mathbb{R}^{n_x}$;
3) $\lim_{i\to\infty} \|\kappa_i(x(k)) - \kappa^*(x(k))\| = 0, \forall x(k) \in \mathbb{R}^{n_x}$;
4) $\lim_{j\to\infty} \|\kappa_{i,j}(x(k)) - \kappa_i(x(k))\| = 0, \forall x(k) \in \mathbb{R}^{n_x}$.

Proof: See Appendix C. □

B. PI and VI Algorithms Online Realization and Performance Analysis of Resulting Closed-Loop System

In the previous subsection, the PI Algorithm 1 and VI Algorithm 2 are developed under the assumption that $f(\cdot)$ and $\bar{\gamma}$ are known, which contradicts the setting of the adaptive nonlinear optimal control problem, that is, both $f(\cdot)$ and $\bar{\gamma}$ are unknown. Besides, how to determine the appropriate form of the value function $V_i(\cdot)$ is still an open problem. To deal with these challenges, in this subsection, we propose the online realizations for both PI Algorithm 1 and VI Algorithm 2 by using critic-actor approximators, packet dropout probability estimator and online data.

1) Critic-Actor Approximators and Packet Dropout Probability Estimator: First, a critic approximator and an actor approximator are employed to approximate the value function $V_i(\cdot)$ and the feedback control policy $\kappa_i(\cdot)$ as follows,

$$\hat{V}_i(x(k)) = W_{ci}\sigma_c(x(k)), \tag{25}$$

$$\hat{\kappa}_i(x(k)) = W_{ai}\sigma_a(x(k)), \tag{26}$$

where $\hat{V}_i(\cdot)$ and $\hat{\kappa}_i(\cdot)$ are the approximations of the value function $V_i(\cdot)$ and the feedback control policy $\kappa_i(\cdot)$, respectively; $W_{ci} \in \mathbb{R}^{1\times n_c}$ and $W_{ai} \in \mathbb{R}^{n_u\times n_a}$ are the weights of critic and actor approximators with $n_c \in \mathbb{N}$ and $n_a \in \mathbb{N}$; $\sigma_c(\cdot) : \mathbb{R}^{n_x} \to \mathbb{R}^{n_c}$ and $\sigma_a(\cdot) : \mathbb{R}^{n_x} \to \mathbb{R}^{n_a}$ are the basis functions satisfying $\sigma_a(0) = 0$ and $\sigma_c(0) = 0$. It follows from approximation theory [39], that the approximation error can be made as small as desired as long as n_c and n_a are chosen large enough.

Then, based on the probability theory [40] and the available information set $\mathbb{I}_{[k_s,k_T+1]}$, where $[k_s, k_T+1]$ denotes the range of the collected data and $k_s, k_T \in \mathbb{N}$, the following packet dropout probability estimator can be employed to estimate the value of $\bar{\gamma}$,

$$\hat{\gamma} = \frac{\sum_{i=k_s}^{k_T}\gamma_f(i)\gamma_c(i)}{k_T - k_s + 1}, \tag{27}$$

where $\hat{\gamma}$ is the estimated value of $\bar{\gamma}$.

To remove the requirement of the knowledge of $f(x(k))$, notice that when $\gamma_f(k) = 1$ and $\gamma_f(k+1) = 1$, $x(k)$ and $x(k+1)$ are both available. Thus in this case, one can use the following equation to estimate the actual value of $f(k)$,

$$\hat{f}(x(k)) = x(k+1) - \gamma_c(k)g(x(k))u_c(k). \tag{28}$$

2) PI Algorithm Online Realization: Here, we will first introduce the online realization of the PI Algorithm 1. When $\gamma_f(k) = 1$ and $\gamma_f(k+1) = 1$, substituting (25)-(28) to (19)-(20) for the corresponding variables in the PI Algorithm 1 yields the following equations,

$$\begin{aligned} W_{ci}\sigma_c(x(k)) =& Q(x(k)) + \hat{\gamma}\hat{\kappa}_i^{\mathrm{T}}(x(k))R\hat{\kappa}_i(x(k)) \\ &+ \hat{\gamma}W_{ci}\sigma_c(\hat{x}^i(k+1)) \\ &+ (1-\hat{\gamma})W_{ci}\sigma_c(\hat{f}(x(k))), \end{aligned} \tag{29}$$

$$\begin{aligned} \kappa_{i+1,j+1}(x(k)) =& (I - 2\alpha R)\kappa_{i+1,j}(x(k)) - \alpha g^{\mathrm{T}}(x(k)) \\ &\cdot \frac{\partial \sigma_c^{\mathrm{T}}(\hat{x}^{i+1,j}(k+1))}{\partial \hat{x}^{i+1,j}(k+1)}W_{ci}^{\mathrm{T}}, \end{aligned} \tag{30}$$

$$W_{a(i+1)}\sigma_a(x(k)) = \kappa_{i+1}(x(k)), \tag{31}$$

where

$$\begin{aligned} \hat{x}^i(k+1) =& \hat{f}(x(k)) + g(x(k))\hat{\kappa}_i(x(k)), \\ \hat{x}^{i+1,j}(k+1) =& \hat{f}(x(k)) + g(x(k))\kappa_{i+1,j}(x(k)), \end{aligned}$$

and $\kappa_{i+1,0}(x(k)) = \hat{\kappa}_i(x(k))$. Based on the above equations, one can use data in the set $\mathbb{I}_k$ to define the following vectors and matrices,

$$\begin{aligned} \phi_{pci}(k_s, k_T, \hat{\gamma}) =& [f_{pci}(k_1), f_{pci}(k_2), \ldots, f_{pci}(k_n)], \\ \phi_{pai}(k_s, k_T) =& [\kappa_{i+1}(x(k_1)), \kappa_{i+1}(x(k_2)), \\ & \ldots, \kappa_{i+1}(x(k_n))], \\ \varphi_{pci}(k_s, k_T, \hat{\gamma}) =& [g_{pci}(k_1), g_{pci}(k_2), \ldots, g_{pci}(k_n)], \\ \varphi_{pa}(k_s, k_T) =& [\sigma_a(x(k_1)), \sigma_a(x(k_2)), \ldots, \sigma_a(x(k_n))], \end{aligned}$$

where

$$f_{pci}(l) = Q(x(l)) + \hat{\gamma}\hat{\kappa}_i^{\mathrm{T}}(x(l))R\hat{\kappa}_i(x(l)),$$

Algorithm 3 PI Algorithm Online Realization

Initiation: Choose an initial admissible feedback control policy $\kappa_0(\cdot)$, two small error tolerance thresholds $\bar{\varepsilon}_p$ and ε_a and a small step size α. Apply $u_c(k)$ to the actuator on $[k_s, k_T]$ to ensure that Assumption 5 holds. Set $i \leftarrow 0$ and $j \leftarrow 0$. Collect the data to obtain $\mathbb{I}_{k_T+1}$ and compute the estimated value $\hat{\gamma}$ via (27).

Online Iteration: Iterate the following steps on i until convergence.

1) **Policy Evaluation**: Solve for weights W_{ci} using the following equation,

$$\begin{aligned} W_{ci} =&\phi_{pci}(k_s, k_T, \hat{\gamma})\varphi_{pci}^{\mathrm{T}}(k_s, k_T, \hat{\gamma}) \\ &\cdot [\varphi_{pci}(k_s, k_T, \hat{\gamma})\varphi_{pci}^{\mathrm{T}}(k_s, k_T, \hat{\gamma})]^{-1}; \end{aligned} \tag{34}$$

2) **Policy Improvement**: Repeat computing the feedback control policy $\kappa_{i+1,j+1}(\cdot)$ using (30) until $\|\kappa_{i+1,j+1}(x(k)) - \kappa_{i+1,j}(x(k))\| \leq \varepsilon_a$ and solve for weights $W_{a(i+1)}$ using the following equation,

$$\begin{aligned} W_{a(i+1)} =&\phi_{pai}(k_s, k_T)\varphi_{pa}^{\mathrm{T}}(k_s, k_T) \\ &\cdot [\varphi_{pa}(k_s, k_T)\varphi_{pa}^{\mathrm{T}}(k_s, k_T)]^{-1}; \end{aligned} \tag{35}$$

3) **Stop if**: $\left\|W_{ci} - W_{c(i-1)}\right\| \leq \bar{\varepsilon}_p$, otherwise set $i \leftarrow i+1$, $j \leftarrow 0$ and go to step 1).

$$\begin{aligned} g_{pci}(l) =&\sigma_c(x(l)) - \hat{\gamma}\sigma_c(\hat{x}^i(l+1)) - (1-\hat{\gamma})\sigma_c(\hat{f}(x(l))), \\ k_{l+1} :=& \inf\{k \in [k_s, k_T] | k > k_l \cap \gamma_f(k) = 1 \\ &\cap \gamma_f(k+1) = 1\}, \\ n :=& \max\{l \in \mathbb{N} | k_l \in [k_s, k_T] \cap \gamma_f(k_l) = 1 \\ &\cap \gamma_f(k_l+1) = 1\}. \end{aligned}$$

Then, (29) and (31) can be rewritten as

$$W_{ci}\varphi_{pci}(k_s, k_T, \hat{\gamma}) = \phi_{pci}(k_s, k_T, \hat{\gamma}), \tag{32}$$

$$W_{a(i+1)}\varphi_{pa}(k_s, k_T) = \phi_{pai}(k_s, k_T). \tag{33}$$

Let $k_T - k_s$ be a sufficiently large positive integer, then the weights W_{ci} and $W_{a(i+1)}$ can be solved in terms of least squares solutions if the following assumption holds.

Assumption 5: There exist σ_c, σ_a and $\underline{k}$, such that for any $k_T - k_s \geq \underline{k}$ and $i > 0$, $\mathrm{rank}(\varphi_{pci}(k_s, k_T, \hat{\gamma})) = n_c$ and $\mathrm{rank}(\varphi_{pa}(k_s, k_T)) = n_a$.

Actually, Assumption 5 is like a condition for persistence of excitation in adaptive control theory [41]. Now, we are ready to present an online and data-driven algorithm, Algorithm 3, for online realization of Algorithm 1. Moreover, the convergence of Algorithm 3 is elucidated in the following theorem.

Theorem 5: Consider Algorithms 3. Under Assumptions 4-5 with $\hat{\gamma} = \bar{\gamma}$, when the thresholds are set to zero and the step size is chosen small enough, the approximate function sequence $\hat{V}_i(\cdot)$ and approximate feedback control policy sequence $\hat{\kappa}_i(\cdot)$ obtained via Algorithm 3 have the following properties:

$$\lim_{i, n_c, n_a \to \infty} |W_{ci}\sigma_c(x(k)) - V^*(x(k))| = 0, \tag{36}$$

$$\lim_{i, n_c, n_a \to \infty} \|W_{ai}\sigma_a(x(k)) - \kappa^*(x(k))\| = 0. \tag{37}$$

Proof: See Appendix D. □

3) VI Algorithm Online Realization: In this part, we will introduce the online realization of the VI Algorithm 2. When $\gamma_f(k) = 1$ and $\gamma_f(k+1) = 1$, substituting (25)-(28) to (22)-(24) for the corresponding variables in the VI Algorithm 2 yields the following equations and (31),

$$\begin{aligned} W_{c(i+1)}\sigma_c(x(k)) =&Q(x(k)) + \hat{\gamma}\hat{\kappa}_i^{\mathrm{T}}(x(k))R\hat{\kappa}_i(x(k)) \\ &+ \hat{\gamma}W_{ci}\sigma_c(\hat{x}^i(k+1)) \\ &+ (1-\hat{\gamma})W_{ci}\sigma_c(\hat{f}(x(k))), \end{aligned} \tag{38}$$

$$\begin{aligned} \kappa_{i+1,j+1}(x(k)) =&(I - 2\alpha R)\kappa_{i+1,j}(x(k)) - \alpha g^{\mathrm{T}}(x(k)) \\ &\cdot \frac{\partial \sigma_c^{\mathrm{T}}(\hat{x}^{i+1,j}(k+1))}{\partial \hat{x}^{i+1,j}(k+1)} W_{c(i+1)}^{\mathrm{T}}, \end{aligned} \tag{39}$$

with $W_{c0} = 0$ and $W_{a0} = 0$. Based on the above equations and (31), one can use data in the set $\mathbb{I}_k$ to define the following vectors and matrices,

$$\begin{aligned} \phi_{vci}(k_s, k_T, \hat{\gamma}) =&[f_{vci}(k_1), f_{vci}(k_2), \ldots, f_{vci}(k_n)], \\ \phi_{vai}(k_s, k_T) =&\phi_{pai}(k_s, k_T), \\ \varphi_{vc}(k_s, k_T) =&[\sigma_c(x(k_1)), \sigma_c(x(k_2)), \ldots, \sigma_c(x(k_n))], \\ \varphi_{va}(k_s, k_T) =&[\sigma_a(x(k_1)), \sigma_a(x(k_2)), \ldots, \sigma_a(x(k_n))], \end{aligned}$$

where

$$\begin{aligned} f_{vci}(l) =&Q(x(l)) + \hat{\gamma}\hat{\kappa}_i^{\mathrm{T}}(x(l))R\hat{\kappa}_i(x(l)) \\ &+ \hat{\gamma}W_{ci}\sigma_c(\hat{x}^i(l+1)) + (1-\hat{\gamma})W_{ci}\sigma_c(\hat{f}(x(l))). \end{aligned}$$

Then, (38) and (31) can be rewritten as

$$W_{c(i+1)}\varphi_{vc}(k_s, k_T, \hat{\gamma}) = \phi_{vci}(k_s, k_T, \hat{\gamma}), \tag{40}$$

$$W_{a(i+1)}\varphi_{va}(k_s, k_T) = \phi_{vai}(k_s, k_T). \tag{41}$$

The weights $W_{c(i+1)}$ and $W_{a(i+1)}$ can be solved in terms of least squares solutions if the following assumption holds.

Assumption 6: There exist σ_c, σ_a and $\underline{k}$, such that for any $k_T - k_s \geq \underline{k}$ and $i > 0$, $\mathrm{rank}(\varphi_{vc}(k_s, k_T, \hat{\gamma})) = n_c$ and $\mathrm{rank}(\varphi_{va}(k_s, k_T)) = n_a$.

Now, we are ready to present an online and data-driven algorithm, Algorithm 4, for online realization of Algorithm 2. Moreover, the convergence of Algorithm 4 is elucidated in the following theorem.

Theorem 6: Consider Algorithm 4. Under Assumptions 4 and 6 with $\hat{\gamma} = \bar{\gamma}$, when the thresholds are set to zero and the step size is chosen small enough, the approximate value function sequence $\hat{V}_i(\cdot)$ and approximate feedback control policy sequence $\hat{\kappa}_i(\cdot)$ obtained via Algorithm 4 satisfy (36)-(37), respectively.

Proof: The proof of Theorem 6 is similar to that of Theorem 5 and thus omitted. □

Remark 3: In online Algorithms 3-4, the control input signals applied to the actuator to generate data can be arbitrary but satisfy Assumptions 5-6. Besides, these algorithms only utilize the data satisfying $\gamma_f(k) = 1$ and $\gamma_f(k+1) = 1$ due to that the controller cannot receive the state $x(k)$ through the communication networks when $\gamma_f(k) = 0$. Therefore, the designed algorithms are off-policy algorithms and do not cause the incorrect solutions [25], [42]–[44].

Algorithm 4 VI Algorithm Online Realization

Initiation: Choose two small error tolerance thresholds $\bar{\varepsilon}_v$ and ε_a and a small step size α. Apply $u_c(k)$ to the actuator on $[k_s, k_T]$ to ensure that Assumption 6 holds. Set $i \leftarrow 0$, $j \leftarrow 0$, $W_{c0} = 0$ and $W_{a0} = 0$. Collect the data to obtain $\mathbb{I}_{k_T+1}$ and compute the estimated value $\hat{\gamma}$ via (27).

Online Iteration: Iterate the following steps on i until convergence.

1) **Value Evaluation**: Solve for weights $W_{c(i+1)}$ by using the following equation,

$$\begin{aligned} W_{c(i+1)} =&\phi_{vci}(k_s, k_T, \hat{\gamma})\varphi_{vc}^{\mathrm{T}}(k_s, k_T) \\ &\cdot [\varphi_{vc}(k_s, k_T)\varphi_{vc}^{\mathrm{T}}(k_s, k_T)]^{-1}; \end{aligned} \tag{42}$$

2) **Policy Improvement**: Repeat computing the feedback control policy $\kappa_{i+1,j+1}(\cdot)$ using (39) until $\|\kappa_{i+1,j+1}(x(k)) - \kappa_{i+1,j}(x(k))\| \leq \varepsilon_a$ and solve for weights $W_{a(i+1)}$ using the following equation,

$$\begin{aligned} W_{a(i+1)} =&\phi_{vai}(k_s, k_T)\varphi_{va}^{\mathrm{T}}(k_s, k_T) \\ &\cdot [\varphi_{va}(k_s, k_T)\varphi_{va}^{\mathrm{T}}(k_s, k_T)]^{-1}; \end{aligned} \tag{43}$$

3) **Stop if**: $\left\|W_{c(i+1)} - W_{ci}\right\| \leq \bar{\varepsilon}_v$, otherwise set $i \leftarrow i+1$, $j \leftarrow 0$ and go to step 1).

4) Performance Analysis of Resulting Closed-Loop System: The following theorem presents the stochastical stability of the resulting closed-loop system under the approximate feedback controller via Algorithms 3-4.

Theorem 7: Consider the system (1) and the approximate feedback controller obtained via Algorithms 3-4. Under $\hat{\gamma} = \bar{\gamma}$, the resulting closed-loop system is stochastically input-to-state stable (SISS) and the state is UUB. Furthermore, the resulting closed-loop system is GSAS if i, n_a and n_c go to infinity, the thresholds are set to zero and the step size is chosen small enough.

Proof: From Theorems 5-6, there exist a positive constant $\bar{\epsilon}$ and positive integers i^*, n_c^* and n_a^* such that for all $i \geq i^*$, $n_c \geq n_c^*$ and $n_a \geq n_a^*$,

$$\begin{aligned} \hat{V}_i(x(k)) \geq& Q(x(k)) + \bar{\gamma}\hat{\kappa}_i^{\mathrm{T}}(x(k))R\hat{\kappa}_i(x(k)) \\ &+ \bar{\gamma}\hat{V}_i(f(x(k)) + g(x(k))\hat{\kappa}_i(x(k))) \\ &+ (1-\bar{\gamma})\hat{V}_i(f(x(k))) - \bar{\epsilon}. \end{aligned}$$

Therefore, by considering $\hat{V}_i(x(k))$ as a Lyapunov function candidate, its difference satisfies

$$\begin{aligned} \Delta\hat{V}_i(x(k)) =&\mathbb{E}\left[\hat{V}_i(x(k+1)) - \hat{V}_i(x(k))\right] \\ \leq& -Q(x(k)) - \bar{\gamma}\hat{\kappa}_i^{\mathrm{T}}(x(k))R\hat{\kappa}_i(x(k)) + \bar{\epsilon} \\ \leq& -\alpha_i(\|x(k)\|) + \bar{\epsilon}, \end{aligned} \tag{44}$$

where α_i is a function of class $\mathcal{K}_\infty$. It follows from (44) that the resulting closed-loop system is SISS [45], [46]. Using the DT comparison principle used in the proof of [47, Lemma 1], it further follows from (44) that there exist a function β_i of class $\mathcal{KL}$ and a function α_σ of class $\mathcal{K}_\infty$ such that

$$\hat{V}_i(x(k)) \leq \beta_i(\|x(0)\|, k) + \alpha_\sigma(\bar{\epsilon}).$$

This implies that the state is UUB. Moreover, it is worth noting that, $\lim_{i,n_a,n_c\to\infty}\bar{\epsilon} = 0$ and hence $\lim_{i,n_a,n_c\to\infty}\alpha_\sigma(\bar{\epsilon}) = 0$. The proof of Theorem 7 is thus completed. □

Remark 4: Theorems 5-7 in this paper are established under the condition $\hat{\gamma} = \bar{\gamma}$. One can collect as many data as possible to facilitate it in the online realization of those algorithms since $\lim_{(k_T-k_s)\to\infty}\hat{\gamma} = \bar{\gamma}$.

Remark 5: It can be observed that the BHJB equation (11) is parameterized by $\mathbb{E}\{\gamma_f(k)\gamma_c(k)\}$ and the packet dropout probability estimator in (27) is used to estimate it. Therefore, when $\gamma_f(k)$ and $\gamma_c(k)$ are dependent, the proposed Algorithms 3-4 can still be applied.

C. Discussion on Extension

The proposed approaches in this paper can be extended to solve the adaptive nonlinear optimal control problem for nonaffine nonlinear DT systems with packet dropouts in both C/A and S/C channels, where the system model is described as follows,

$$x(k+1) = f_n(x(k), u(k)),$$

where $f_n(\cdot,\cdot) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$ satisfying $f_n(0,0) = 0$. In order to solve the adaptive nonlinear optimal control problem for nonaffine nonlinear DT systems, we need an additional model approximator given as follows,

$$\hat{f}_n(x(k), u(k)) = W_m\sigma_m(x(k), u(k)),$$

where $\hat{f}_n(\cdot,\cdot)$ is the approximation of the nonlinear function $f_n(\cdot,\cdot)$, $W_m \in \mathbb{R}^{n_x\times n_m}$ are the weights of the model approximator with $n_m \in \mathbb{N}$, $\sigma_m(\cdot,\cdot) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_m}$ is the basis function satisfying $\sigma_m(0,0) = 0$. Based on the reliable online data, the model approximator can be trained. Moreover, one can predict the current state value $x(k)$ when the dropout phenomenon occurs in the S/C channel by repeating $\bar{x}(i) = \hat{f}_n(\bar{x}(i-1), \gamma_c(i-1)u_c(i-1)), i = k_d+1, k_d+2, \ldots, k$, where $\gamma_f(k_d) = 1$ and $\bar{x}(k_d) = x_f(k_d) = x(k_d)$. In this scenario, a similar BHJB equation (11) for the optimal nonaffine nonlinear control problem is established with the parameter $\bar{\gamma}$ being replaced by $\bar{\gamma}_c$. In this case, the packet dropout probability estimator in (27) need to be replaced by the following one,

$$\hat{\gamma}_c = \frac{\sum_{i=k_s}^{k_T}\gamma_c(i)}{k_T - k_s + 1},$$

where $\hat{\gamma}_c$ is the estimated value of $\bar{\gamma}_c$. Then, Algorithms 3-4 can be applied to learn the optimal value function and optimal feedback control policy by using online data, model-critic-actor approximators and packet dropout probability estimator.

Similar to the BHJB equation (8), the BHJB equation resulting from the nonaffine nonlinear system has the same form as follows,

$$\begin{aligned} V^*(x(k)) = \min_{\kappa(x(k))}&\left[Q(x(k)) + \hat{\gamma}_c\kappa^{\mathrm{T}}(x(k))R\kappa(x(k))\right. \\ &+ (1-\hat{\gamma}_c)V^*(f_n(x(k),0)) \\ &\left.+\hat{\gamma}_cV^*(f_n(x(k),\kappa(x(k))))\right]. \end{aligned}$$

It then follows from the proofs of Theorems 2-7 that the conclusions in Theorems 2-7 still hold for the optimal nonaffine nonlinear control problem.

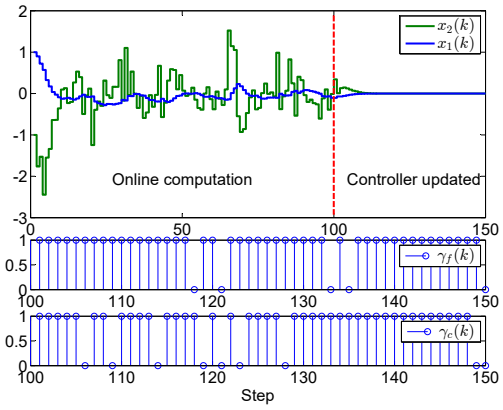


Fig. 3. Trajectories of $x(k)$, $\gamma_f(k)$ and $\gamma_c(k)$ via Algorithm 3.

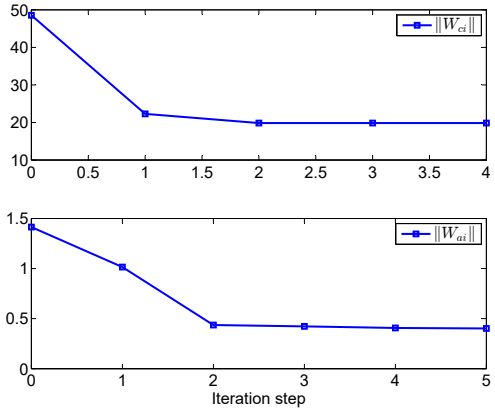


Fig. 4. Trajectories of $\|W_{ci}\|$ and $\|W_{ai}\|$ via Algorithm 3.

V. Numerical Example

In this section, we apply the proposed approaches to stabilize a single-link manipulator with packet dropouts in both C/A and S/C channels to illustrate their effectiveness.

A. Single-Link Manipulator Model and Control Parameter Design

The dynamic of a single-link manipulator can be found in [47]. Via the Euler method with the sampling interval $\Delta t = 0.1$s the single-link manipulator system can be discretized and rewritten in the form of (1) with

$$f(x) = \begin{bmatrix} x_1 + 0.1x_2 \\ -1.962\sin(x_1) + 0.2x_2 \end{bmatrix}, g(x) = \begin{bmatrix} 0 \\ 0.4 \end{bmatrix}.$$

We choose the parameters in the performance index (5) as $Q(x(k)) = x^{\mathrm{T}}(k)x(k)$, $R = I$. In practical NCSs, actuator, sensors and controllers are distributed and connected by communication networks. Due to the bandwidth limitation and/or environment disruptions, packet dropouts often occur in these communication networks. In the simulation experiments, the successful probabilities $\bar{\gamma}_f$ and $\bar{\gamma}_c$ of the S/C and C/A channels signal transmission are chosen as $\bar{\gamma}_f = \bar{\gamma}_c = \sqrt{0.8}$. Thus, $\bar{\gamma} = 0.8$.

B. PI Algorithm 3 Simulation Results

In this simulation, the following choices are made. $u_c(k)$ is random signals on the period [1,100], which yields $k_s = 1$ and $k_T = 100$. The initial value of the state is $x(1) = [1, -1]^{\mathrm{T}}$. The basis functions $\sigma_a(\cdot)$ and $\sigma_c(\cdot)$ for both the critic and actor approximators are $\sigma_c(x) = [x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, x_1^4, x_1^3x_2, x_1^2x_2^2, x_1x_2^3, x_2^4]^{\mathrm{T}}$ and $\sigma_a(x) = [\sigma_c^{\mathrm{T}}(x), x_1, x_2]^{\mathrm{T}}$, which yields $n_c = 12$ and $n_a = 14$. $W_{a0} = [0_{1\times 12}, 1, 1]$; $\bar{\varepsilon}_p = 0.0001$; $\varepsilon_a = 0.00001$ and the step size $\alpha = 0.01$. When $k \geq 100$, the control input signals are generated by the approximate optimal control policy obtained via PI Algorithm 3. $\hat{\gamma} = 0.8$. Fig. 3 shows the trajectories of $x(k)$, $\gamma_f(k)$ and $\gamma_c(k)$ via Algorithm 3. It can be observed from $x(k), k \geq 100$ that the controlled NCS can be stabilized by the approximate optimal controller via PI Algorithm 3 in the presence of network-induced two channels dropouts. Fig. 4 shows trajectories of $\|W_{ci}\|$ and $\|W_{ai}\|$ via Algorithm 3. The approximate cost function $V^*(x(k))$ and the cost functions $V_i(x(k))$ learned via Algorithm 3 are shown in Fig. 5. Clearly, the approximate cost function $V^*(x(k))$ is positive definite and the property 1) of Theorem 3 can be thus verified.

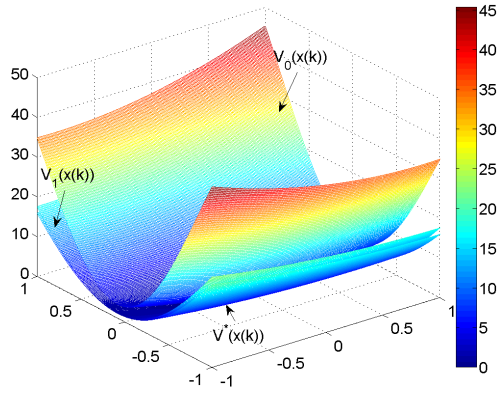


Fig. 5. Cost functions $V_i(x(k))$ learned via Algorithm 3.

C. VI Algorithm 4 Simulation Results

In this simulation, the following choices are made. $u_c(k)$ is random signals on [1,100]. The initial value of the state is $x(1) = [1, -1]^{\mathrm{T}}$. The basis functions $\sigma_a(\cdot)$ and $\sigma_c(\cdot)$ for both the critic and actor approximators are the same as them in the PI Algorithm 3 simulation experiment; $W_{c1} = 0_{1\times 12}$; $W_{a1} = 0_{1\times 14}$; $\bar{\varepsilon}_p = 0.0001$; $\varepsilon_a = 0.00001$ and the step size $\alpha = 0.01$. When $k \geq 100$, the control input signals are generated by the approximate optimal control policy obtained via VI Algorithm 4. $\hat{\gamma} = 0.8$. Fig. 6 shows the trajectories of $x(k)$, $\gamma_f(k)$ and $\gamma_c(k)$ via Algorithm 3. It can be observed from $x(k), k \geq 100$ that the controlled NCS can be stabilized by the approximate optimal controller via VI Algorithm 4

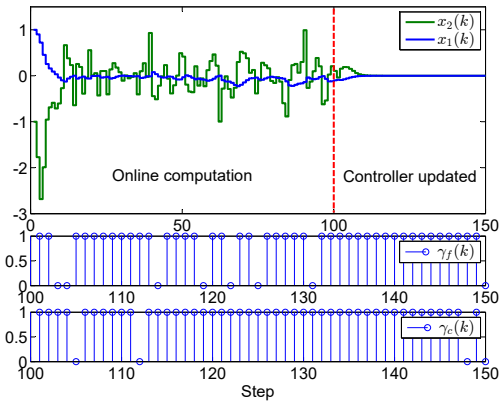


Fig. 6. Trajectories of $x(k)$, $\gamma_f(k)$ and $\gamma_c(k)$ via Algorithm 4.

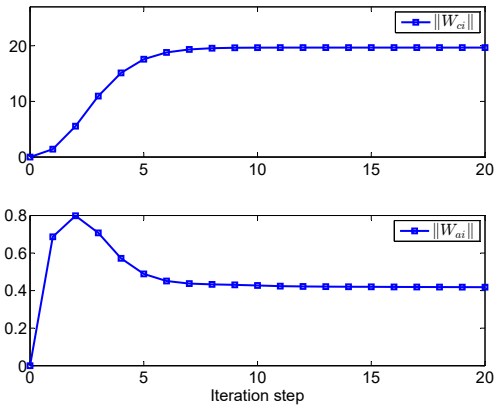


Fig. 7. Trajectories of $\|W_{ci}\|$ and $\|W_{ai}\|$ via Algorithm 4.

in the presence of network-induced two channels dropouts. Fig. 7 shows trajectories of $\|W_{ci}\|$ and $\|W_{ai}\|$ via Algorithm 4. The approximate cost function $V^*(x(k))$ and the cost functions $V_i(x(k))$ learned via Algorithm 4 are compared in Fig. 8. Clearly, the approximate cost function $V^*(x(k))$ is positive definite and the property 1) of Theorem 4 can be thus verified. Moreover, it can be observed that the approximate cost functions $V^*(x(k))$ learned via both Algorithms 3 and 4 are almost the same.

VI. Conclusions

In this work, we have studied the adaptive nonlinear optimal control problem of networked nonlinear DT systems with packet dropouts in both S/C and C/A channels. A BHJB equation is built to deal with the packet dropouts. RL based PI and VI algorithms are then developed to solve the obtained BHJB equation online in the absence of *a priori* knowledge of the partial system dynamics and probabilities of packet dropouts . It is shown that the optimal controller can be approximately obtained by the proposed online algorithms via using measurable data subject to packet dropouts. In our future work, the adaptive optimal output regulation problem for a class of networked DT strict feedback nonlinear systems will be investigated.

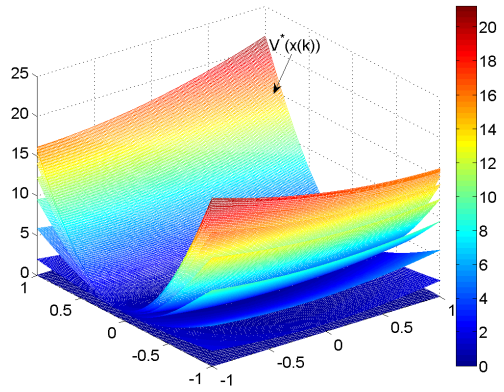


Fig. 8. Cost functions $V_i(x(k))$ learned via Algorithm 4.

Appendix A
Proof of Theorem 2

It follows from (8) that $V^*(\cdot)$ and $\kappa^*(\cdot)$ satisfy (14) in Assumption 4 and thus the global stochastical asymptotical stability can be established directly based on the proof of Lemma 1. For the optimality, let $V^{0*}(\cdot)$ be the solution to $H(x(k),\kappa^0,V^{0*})=0$. It then follows from (8) that

$$\begin{aligned}
0=&H(x(k),\kappa^0,V^{0*})-H(x(k),\kappa^*,V^*)\\
=&\bar{\gamma}\kappa^{0\mathrm{T}}(x(k))R\kappa^0(x(k))-\bar{\gamma}\kappa^{*\mathrm{T}}(x(k))R\kappa^*(x(k))\\
&+\bar{\gamma}V^{0*}(f(x(k))+g(x(k))\kappa^0(x(k)))-V^{0*}(x(k))\\
&-\bar{\gamma}V^*(f(x(k))+g(x(k))\kappa^*(x(k)))+V^*(x(k))\\
&+(1-\bar{\gamma})V^{0*}(f(x(k)))-(1-\bar{\gamma})V^*(f(x(k)))\\
=&\bar{\gamma}M(\kappa^0(x(k)),\kappa^*(x(k)),V^*)\\
&+\mathbb{E}\left[V^{0*}(f(x(k))+\gamma_f(k)\gamma_c(k)g(x(k))\kappa^0(x(k)))\right.\\
&-V^*(f(x(k))+\gamma_f(k)\gamma_c(k)g(x(k))\kappa^0(x(k)))\\
&\left.-V^{0*}(x(k))+V^*(x(k))|x(k)\right],
\end{aligned}\tag{45}$$

where $M(\kappa^0(x(k)),\kappa^*(x(k)),V^*) := \Xi(\kappa^0(x(k)),V^*)-\Xi(\kappa^*(x(k)),V^*)$. By summing (45) from k to ∞ and considering the fact that $V^*(x(\infty))=0$ and $V^{0*}(x(\infty))=0$, which is concluded by the global stochastical asymptotical stability, one has the following equation,

$$V^{0*}(x(k))=\sum_{i=k}^{\infty}\bar{\gamma}M(\kappa^0(x(i)),\kappa^*(x(i)),V^*)+V^*(x(k)).$$

To prove the optimality, it is indeed to prove that

$$\begin{cases} M(\kappa^0(x(k)),\kappa^*(x(k)),V^*)=0 & \kappa^0(x(k))=\kappa^*(x(k)),\\ M(\kappa^0(x(k)),\kappa^*(x(k)),V^*)>0 & \kappa^0(x(k))\neq\kappa^*(x(k)).\end{cases}$$

It follows from $R>0$ and $V^*(\cdot)\in\mathcal{P}^{n_x}$ that $M(\kappa^0(x(k)),\kappa^*(x(k)),V^*)$ is a convex function with respect to $\kappa^0(x(k))$ and minimized when the following equation holds,

$$0=\frac{\partial M(\kappa^0(x(k)),\kappa^*(x(k)),V^*)}{\partial\kappa^0(x(k))}$$

$$= -2R\kappa^0(x(k)) + \frac{\partial V^*(f(x(k)) + g(x(k))\kappa^0(x(k)))}{\partial \kappa^0(x(k))}.$$

This implies that $M(\kappa^0(x(k)), \kappa^*(x(k), V^*)$ is minimized when $\kappa^0(x(k)) = \kappa^*(x(k))$ with its minimal value being 0. The optimality is proven and the proof of Theorem 2 is thus completed. □

APPENDIX B
PROOF OF THEOREM 3

To begin with, we suppose that property 5) holds to prove properties 1)-4), which will be proven in the end of this proof. It follows from the policy evaluation in (19) that

$$\begin{aligned}0 =& H(x(k), \kappa_0, V_0) - H(x(k), \kappa_1, V_1)\\ =& \bar{\gamma} M(\kappa_0(x(k)), \kappa_1(x(k)), V_0)\\ &+ \mathbb{E}\left[V_0(f(x(k)) + \gamma_f(k)\gamma_c(k)g(x(k))\kappa_0(x(k)))\right.\\ &- V_1(f(x(k)) + \gamma_f(k)\gamma_c(k)g(x(k))\kappa_0(x(k)))\\ &\left.-V_0(x(k)) + V_1(x(k))\right]. \end{aligned} \tag{46}$$

Since $\kappa_0(\cdot)$ is an admissible feedback control policy, it can be concluded that $V_0(x(\infty)) = 0$ and $V_0(x(k)) \in \mathcal{P}^{n_x}$. Summing (46) from k to ∞ yields the following equation,

$$V_0(x(k)) = \sum_{i=k}^{\infty} \bar{\gamma} M(\kappa_0(x(i)), \kappa_1(x(i)), V_0) + V_1(x(k)).$$

Besides, it follows from $H(x(k), \kappa^*, V^*) - H(x(k), \kappa_1, V_1) = 0$ that

$$V_1(x(k)) = \sum_{i=k}^{\infty} \bar{\gamma} M(\kappa_1(x(i)), \kappa^*(x(i)), V^*) + V^*(x(k)).$$

By noting that $R > 0$, $V_0(x(k)) \in \mathcal{P}^{n_x}$ and $V^*(\cdot) \in \mathcal{P}^{n_x}$, one has that $M(\kappa_0(x(k)), \kappa_1(x(k)), V_0)$ and $M(\kappa_1(x(k)), \kappa^*(x(k)), V^*)$ are convex functions with respect to $\kappa_0(x(k))$ and $\kappa_1(x(k))$, respectively. This implies that they are minimized when the following equations hold,

$$\begin{aligned}0 =& \frac{\partial M(\kappa_0(x(k)), \kappa_1(x(k)), V_0)}{\partial \kappa_0(x(k))}\\ =& -2R\kappa_0(x(k)) + \frac{\partial V_0(f(x(k)) + g(x(k))\kappa_0(x(k)))}{\partial \kappa_0(x(k))},\\ 0 =& \frac{\partial M(\kappa_1(x(k)), \kappa^*(x(k)), V^*)}{\partial \kappa_1(x(k))}\\ =& -2R\kappa_1(x(k)) + \frac{\partial V^*(f(x(k)) + g(x(k))\kappa_1(x(k)))}{\partial \kappa_1(x(k))}.\end{aligned}$$

Therefore, the minimal values of $M(\kappa_0(x(k)), \kappa_1(x(k)), V_0)$ and $M(\kappa_1(x(k)), \kappa^*(x(k)), V^*)$ are 0. This implies that $V_0(x(k)) \geq V_1(x(k)) \geq V^*(x(k))$. It then follows from $V_0(x(k)) \geq 0$ and $V^*(x(k)) \geq 0$ that $V_1(x(k)) \geq 0$. Therefore, $\kappa_1(\cdot)$ is an admissible feedback control policy based on Lemma 1. Repeating above analysis for $i = 1, 2, \ldots$, one can conclude properties 1)-2). It then follows from the monotone and boundedness properties of $V_i(x(k))$ that $V_\infty(x(k))$ and $\kappa_\infty(x(k))$ exist and satisfy the following equations,

$$\begin{aligned}V_\infty(x(k)) =& Q(x(k)) + \bar{\gamma}\kappa_\infty^{\mathrm{T}}(x(k))R\kappa_\infty(x(k))\\ &+ \bar{\gamma} V_\infty(f(x(k)) + g(x(k))\kappa_\infty(x(k)))\\ &+ (1-\bar{\gamma})V_\infty(f(x(k))),\end{aligned} \tag{47}$$

$$\kappa_\infty(x(k)) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x(k))\frac{\partial V_\infty(x^\infty(k+1))}{\partial x^\infty(k+1)}. \tag{48}$$

It can be seen that $V_\infty(x(k))$ solves the BHJB equation (11). Therefore, $V_\infty(x(k)) = V^*(x(k))$ and $\kappa_\infty(x(k)) = \kappa^*(x(k))$. This implies that properties 3) and 4) hold. For the property 5), it follows from (20) and $V_i(x(k)) \in \mathcal{P}^{n_x}, \forall i \geq 0$ that $\kappa_{i+1}(x(k))$ is the optimal solution to minimize $\Xi(\kappa(x(k)), V_i)$. This implies that $\Xi(\kappa_{i+1,j}(x(k)), V_i) \geq \Xi(\kappa_{i+1}(x(k)), V_i)$. By noting that

$$\begin{aligned}&\Theta(\alpha)\\ :=& \frac{\partial \Xi(\kappa_{i+1,j+1}(x(k)), V_i)}{\partial \alpha}\\ =& -\left(\frac{\partial \Xi(\kappa_{i+1,j+1}(x(k)), V_i)}{\partial \kappa_{i+1,j+1}(x(k))}\right)^{\mathrm{T}} \frac{\partial \Xi(\kappa_{i+1,j}(x(k)), V_i)}{\partial \kappa_{i+1,j}(x(k))},\end{aligned}$$

one has that $\Theta(0) < 0$ when $\kappa_{i+1,j}(x(k)) \neq \kappa_{i+1}(x(k))$. Therefore, it follows from the continuity of $\Theta(\alpha)$ that there exists a sufficiently small positive constant α such that $\Theta(\alpha) < \Theta(0) < 0$. This implies that $\Xi(\kappa_{i+1,j+1}(x(k)), V_i) < \Xi(\kappa_{i+1,j}(x(k)), V_i)$. It then follows from the monotone and boundedness properties of $\Xi(\kappa_{i+1,j}(x(k)), V_i)$ that the property 5) holds and the proof of Theorem 3 is thus completed. □

APPENDIX C
PROOF OF THEOREM 4

To begin with, we suppose that property 4) holds to prove properties 1)-3), whose proof is similar to the proof of property 5) in Theorem 3 and thus omitted. It follows from (22) that $V_i(x(k)) \in \mathcal{P}^{n_x}, \forall i > 0$. This implies that the feedback control policy $\kappa_i(\cdot)$ obtained via (23) is the optimal solution to minimize $\Xi(\kappa(x(k)), V_i)$. Therefore, one has the following equation,

$$\begin{aligned}V_{i+1}(x(k)) = \min_{\kappa(x(k))} &\left[Q(x(k)) + \bar{\gamma}\kappa^{\mathrm{T}}(x(k))R\kappa(x(k))\right.\\ &+ \bar{\gamma} V_i(f(x(k)) + g(x(k))\kappa(x(k)))\\ &\left.+(1-\bar{\gamma})V_i(f(x(k)))\right].\end{aligned} \tag{49}$$

Let $\mu_i(\cdot)$ be any arbitrary sequence of feedback control policies and Λ_i be defined by

$$\begin{aligned}\Lambda_{i+1}(x(k)) =& Q(x(k)) + \bar{\gamma}\mu_i^{\mathrm{T}}(x(k))R\mu_i(x(k))\\ &+ \bar{\gamma}\Lambda_i(f(x(k)) + g(x(k))\mu_i(x(k)))\\ &+ (1-\bar{\gamma})\Lambda_i(f(x(k))).\end{aligned}$$

If $V_0(x(k)) = \Lambda_0(x(k)) = 0$, then it follows from (49) that $V_i(x(k)) \leq \Lambda_i(x(k))$, $\forall i \geq 0$. Now, assume that $\mu_i(x(k)) = \kappa_{i+1}(x(k))$ such that

$$\begin{aligned}\Lambda_{i+1}(x(k)) =& Q(x(k)) + \bar{\gamma}\kappa_{i+1}^{\mathrm{T}}(x(k))R\kappa_{i+1}(x(k))\\ &+ \bar{\gamma}\Lambda_i(f(x(k)) + g(x(k))\kappa_{i+1}(x(k)))\\ &+ (1-\bar{\gamma})\Lambda_i(f(x(k))),\end{aligned} \tag{50}$$

and consider (23). It will next be proven by induction that if $V_0(x(k)) = \Lambda_0(x(k)) = 0$, then $V_{i+1}(x(k)) \geq \Lambda_i(x(k))$. Induction is initialized by letting $V_0(x(k)) = \Lambda_0(x(k)) = 0$

and hence $V_1(x(k))-\Lambda_0(x(k))=Q(x(k))\geq 0$. Now, assume that $V_i(x(k))\geq \Lambda_{i-1}(x(k))$. Then, by subtracting (49) from (23), it follows that

$$\begin{aligned}&V_{i+1}(x(k))-\Lambda_i(x(k))\\=&\bar{\gamma}V_i(f(x(k))+g(x(k))\kappa_i(x(k)))+(1-\bar{\gamma})V_i(f(x(k)))\\&-\bar{\gamma}\Lambda_{i-1}(f(x(k))+g(x(k))\kappa_i(x(k)))\\&-(1-\bar{\gamma})\Lambda_{i-1}(f(x(k)))\geq 0,\end{aligned}$$

which completes the proof that $V_{i+1}(x(k))\geq \Lambda_i(x(k))$. Based on $V_{i+1}(x(k))\geq \Lambda_i(x(k))$ and $V_i(x(k))\leq \Lambda_i(x(k))$, it is direct to conclude that $V_i(x(k))\leq V_{i+1}(x(k)), \forall i\geq 0$. Clearly, $V_i(x(k))$ is a nondecreasing sequence. It will next be proven that if $V_i(x(k))$ is bounded, then $V_\infty(x(k))$ exists. Now, let $\eta(x(k))$ be any admissible feedback control policy and $V_0(x(k))=Z_0(x(k))=0$, where $Z_i(x(k))$ is updated as

$$\begin{aligned}Z_{i+1}(x(k))=&Q(x(k))+\bar{\gamma}\eta^{\mathrm{T}}(x(k))R\eta(x(k))\\&+\bar{\gamma}Z_i(f(x(k))+g(x(k))\eta(x(k)))\\&+(1-\bar{\gamma})Z_i(f(x(k))).\end{aligned}$$

Since $\eta(x(k))$ is an admissible feedback control policy, based on the statement in Lemma 1, one has that

$$\begin{aligned}Z_{i+1}(x(k))=&\sum_{n=0}^{i}\left[Q(x(k+n))+\bar{\gamma}\eta^{\mathrm{T}}(k+n)R\eta(k+n)\right]\\\leq&\sum_{n=0}^{\infty}\left[Q(x(k+n))+\bar{\gamma}\eta^{\mathrm{T}}(k+n)R\eta(k+n)\right]\\=&Z^*(x(k)),\end{aligned}$$

where $Z^*(\cdot)\in\mathcal{P}^{n_x}$. Clearly, $0\leq Z^*(x(k))\leq\infty$. From (49), it follows that $V_i(x(k))\leq Z_i(x(k))\leq Z^*(x(k))\leq\infty$, which implies that $V_i(x(k))$ is bounded. It then follows from the monotone and boundedness properties of $V_i(x(k))$ that $V_\infty(x(k))$ and $\kappa_\infty(x(k))$ exist and satisfy (47)-(48). Furthermore, $V_\infty(x(k))$ solves the BHJB equation (11). Therefore, $V_\infty(x(k))=V^*(x(k))$ and $\kappa_\infty(x(k))=\kappa^*(x(k))$. This implies that properties 1)-3) hold and the proof of Theorem 4 is thus completed. □

APPENDIX D
PROOF OF THEOREM 5

It follows from properties 3)-4) in Theorem 3 that (36)-(37) can be established by proving the following equations,

$$\lim_{n_c,n_a\to\infty}|W_{ci}\sigma_c(x(k))-V_i(x(k))|=0, \tag{51}$$

$$\lim_{n_c,n_a\to\infty}\|W_{ai}\sigma_a(x(k))-\kappa_i(x(k))\|=0. \tag{52}$$

Let W_{ci}^* and W_{ai}^* be the ideal weights in (32)-(33) such that

$$\begin{aligned}W_{ci}^*\varphi_{pci}(k_s,k_T,\hat{\gamma})=&\phi_{pci}(k_s,k_T,\hat{\gamma})+\epsilon_{pc},\\W_{a(i+1)}^*\varphi_{pa}(k_s,k_T)=&\phi_{pai}(k_s,k_T)+\epsilon_{pa},\end{aligned}$$

where ϵ_{pc} and ϵ_{pa} are reconstruction errors. It follows from approximation theory [39] that ϵ_{pc} and ϵ_{pa} satisfy $\lim_{n_c,n_a\to\infty}\epsilon_{pc}=0$ and $\lim_{n_c,n_a\to\infty}\epsilon_{pa}=0$. Denoting the approximate errors in (32)-(33) as e_{pc} and e_{pa}, one has the following equations,

$$\begin{aligned}e_{pc}=&W_{ci}\varphi_{pci}(k_s,k_T,\hat{\gamma})-\phi_{pci}(k_s,k_T,\hat{\gamma})\\=&\bar{W}_{ci}\varphi_{pci}(k_s,k_T,\hat{\gamma})+\epsilon_{pc},\end{aligned} \tag{53}$$

$$\begin{aligned}e_{pa}=&W_{a(i+1)}\varphi_{pa}(k_s,k_T)-\phi_{pai}(k_s,k_T)\\=&\bar{W}_{a(i+1)}\varphi_{pa}(k_s,k_T)+\epsilon_{pa},\end{aligned} \tag{54}$$

where $\bar{W}_{ci}=W_{ci}-W_{ci}^*$ and $\bar{W}_{a(i+1)}=W_{a(i+1)}-W_{a(i+1)}^*$. Since the weights are obtained via using the least-square method, W_{ci} and $W_{a(i+1)}$ computed via (34)-(35) minimize $e_{pc}^{\mathrm{T}}e_{pc}$ and $e_{pa}^{\mathrm{T}}e_{pa}$. It follows from $e_{pc}^{\mathrm{T}}e_{pc}=\epsilon_{pc}^{\mathrm{T}}\epsilon_{pc}$ when $W_{ci}=W_{ci}^*$ and $e_{pa}^{\mathrm{T}}e_{pa}=\epsilon_{pa}^{\mathrm{T}}\epsilon_{pa}$ when $W_{a(i+1)}=W_{a(i+1)}^*$ that $e_{pc}^{\mathrm{T}}e_{pc}\leq\epsilon_{pc}^{\mathrm{T}}\epsilon_{pc}$ and $e_{pa}^{\mathrm{T}}e_{pa}\leq\epsilon_{pa}^{\mathrm{T}}\epsilon_{pa}$. Furthermore, by considering Assumption 5 and (53)-(54), one has that there exist two positive constants δ_{pc} and δ_{pa} such that

$$\begin{aligned}\delta_{pc}\bar{W}_{ci}\bar{W}_{ci}^{\mathrm{T}}\leq&\bar{W}_{ci}\varphi_{pci}(k_s,k_T,\hat{\gamma})\varphi_{pci}^{\mathrm{T}}(k_s,k_T,\hat{\gamma})\bar{W}_{ci}^{\mathrm{T}}\\=&(e_{pc}-\epsilon_{pc})(e_{pc}-\epsilon_{pc})^{\mathrm{T}}\\\leq&2e_{pc}e_{pc}^{\mathrm{T}}+2\epsilon_{pc}\epsilon_{pc}^{\mathrm{T}},\\\delta_{pa}\bar{W}_{a(i+1)}\bar{W}_{a(i+1)}^{\mathrm{T}}\leq&\bar{W}_{a(i+1)}\varphi_{pa}(k_s,k_T)\varphi_{pa}^{\mathrm{T}}(k_s,k_T)\\&\cdot\bar{W}_{a(i+1)}^{\mathrm{T}}\\=&(e_{pa}-\epsilon_{pa})(e_{pa}-\epsilon_{pa})^{\mathrm{T}}\\\leq&2e_{pa}e_{pa}^{\mathrm{T}}+2\epsilon_{pa}\epsilon_{pa}^{\mathrm{T}}.\end{aligned}$$

It then follows from $e_{pc}^{\mathrm{T}}e_{pc}\leq\epsilon_{pc}^{\mathrm{T}}\epsilon_{pc}$ and $e_{pa}^{\mathrm{T}}e_{pa}\leq\epsilon_{pa}^{\mathrm{T}}\epsilon_{pa}$ that there exist two positive constants $\bar{\delta}_{pc}$ and $\bar{\delta}_{pa}$ such that $\|\bar{W}_{ci}\|\leq\bar{\delta}_{pc}\|\epsilon_{pc}\|$ and $\|\bar{W}_{a(i+1)}\|\leq\bar{\delta}_{pa}\|\epsilon_{pa}\|$. By considering the fact that $\lim_{n_c,n_a\to\infty}\epsilon_{pc}=0$ and $\lim_{n_c,n_a\to\infty}\epsilon_{pa}=0$, one has that $\lim_{n_c,n_a\to\infty}\bar{W}_{ci}=0$ and $\lim_{n_c,n_a\to\infty}\bar{W}_{a(i+1)}=0$. It then follows from $\lim_{n_c,n_a\to\infty}W_{ci}^*\sigma_c(x(k))=V_i(x(k))$ and $\lim_{n_c,n_a\to\infty}W_{ai}^*\sigma_a(x(k))=\kappa_i(x(k))$ that the following equations hold,

$$\begin{aligned}&\lim_{n_c,n_a\to\infty}|W_{ci}\sigma_c(x(k))-V_i(x(k))|\\\leq&\lim_{n_c,n_a\to\infty}\|\bar{W}_{ci}\|\|\sigma_c(x(k))\|=0,\\&\lim_{n_c,n_a\to\infty}\|W_{ai}\sigma_a(x(k))-\kappa_i(x(k))\|\\\leq&\lim_{n_c,n_a\to\infty}\|\bar{W}_{ai}\|\|\sigma_a(x(k))\|=0,\end{aligned}$$

which thus completes the proof of Theorem 5. □

REFERENCES

[1] A. Isidori, *Nonlinear Control Systems*. Springer Science & Business Media, 1985.

[2] H. K. Khalil, *Noninear Systems*. Prentice-Hall, New Jersey, USA, 1996.

[3] D. Liberzon, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2011.

[4] L. S. Pontryagin, *Mathematical Theory of Optimal Processes*. Routledge, 1987.

[5] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[6] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific Belmont, 2000.

[7] J. A. Primbs, "Nonlinear optimal control: A receding horizon appoach," Ph.D. dissertation, California Institute of Technology, 1999.

[8] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, Jun. 2000.

[9] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. John Wiley & Sons, 2012.

[10] P. J. Werbos, W. Miller, and R. Sutton, "A menu of designs for reinforcement learning over time," in *Neural Networks for Control*. MIT Press Cambridge, MA, 1990, pp. 67–95.

[11] P. J. Werbos, "Approximate dynamic programming for realtime control and neural modelling," in *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. Van Nostrand, 1992, pp. 493–525.

[12] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, May 2014.

[13] W. Gao and Z.-P. Jiang, "Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2614–2624, Jun. 2017.

[14] Y. Jiang, J. Fan, T. Chai, J. Li, and F. L. Lewis, "Data-driven flotation industrial process operational optimal control based on reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 1974–1989, May 2018.

[15] X. Lu, B. Kiumarsi, T. Chai, Y. Jiang, and F. L. Lewis, "Operational control of mineral grinding processes using adaptive dynamic programming and reference governor," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2210–2221, Apr. 2019.

[16] Y. Jiang, J. Fan, T. Chai, and F. L. Lewis, "Dual-rate operational optimal control for flotation industrial process with unknown operational model," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4587–4599, Jun. 2019.

[17] X. Yang and Q. Wei, "Adaptive critic learning for constrained optimal event-triggered control with discounted cost," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 91–104, Jan. 2021.

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

[19] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. John Wiley & Sons, 2017.

[20] ——, "Global adaptive dynamic programming for continuous-time nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2917–2929, Nov. 2015.

[21] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 621–634, Mar. 2013.

[22] Q. Wei, D. Liu, and H. Lin, "Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 840–853, 2015.

[23] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[24] D. Wang, D. Liu, Q. Wei, D. Zhao, and N. Jin, "Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming," *Automatica*, vol. 48, no. 8, pp. 1825–1832, Aug. 2012.

[25] J. Li, T. Chai, F. L. Lewis, Z. Ding, and Y. Jiang, "Off-policy interleaved Q-learning: Optimal control for affine nonlinear discrete-time systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1308–1320, May 2018.

[26] C. Mu, D. Wang, and H. He, "Novel iterative neural dynamic programming for data-based approximate optimal control design," *Automatica*, vol. 81, pp. 240–252, Jul. 2017.

[27] R. A. Gupta and M.-Y. Chow, "Networked control system: Overview and research trends," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 7, pp. 2527–2535, Jul. 2009.

[28] J. P. Hespanha, P. Naghshtabrizi, and Y. G. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, Jan. 2007.

[29] D. E. Quevedo and D. Nešić, "Robust stability of packetized predictive control of nonlinear systems with disturbances and markovian packet losses," *Automatica*, vol. 48, no. 8, pp. 1803–1811, Aug. 2012.

[30] D. E. Quevedo and I. Jurado, "Stability of sequence-based control with random delays and dropouts," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1296–1302, May 2013.

[31] H. Li and Y. Shi, "Network-based predictive control for constrained nonlinear systems with two-channel packet dropouts," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 3, pp. 1574–1582, Mar. 2014.

[32] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, Jul. 2003.

[33] Y. Jiang, J. Fan, T. Chai, F. L. Lewis, and J. Li, "Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4607–4620, Oct. 2018.

[34] J. Fan, Q. Wu, Y. Jiang, T. Chai, and F. L. Lewis, "Model-free optimal output regulation for linear discrete-time lossy networked control systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4033–4042, Nov. 2020.

[35] J. Qiu, T. Wang, S. Yin, and H. Gao, "Data-based optimal control for networked double-layer industrial processes," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 4179–4186, May 2017.

[36] H. Xu and S. Jagannathan, "Stochastic optimal controller design for uncertain nonlinear networked control system via neuro dynamic programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 3, pp. 471–484, Mar. 2013.

[37] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, "Foundations of control and estimation over lossy networks," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 163–187, Jan. 2007.

[38] K. J. Åström, *Introduction to Stochastic Control Theory*. Courier Corporation, 2012.

[39] M. J. D. Powell, *Approximation Theory and Methods*. Cambridge University Press, 1981.

[40] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[41] K. J. Åström and B. Wittenmark, *Adaptive Control*. Courier Corporation, 2013.

[42] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, Oct. 2012.

[43] Y. Jiang, B. Kiumarsi, J. Fan, T. Chai, and F. L. Lewis, "Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3147–3156, Jul. 2020.

[44] Y. Jiang, W. Gao, J. Na, D. Zhang, T. T. Hämäläinen, V. Stojanovic, and F. L. Lewis, "Value iteration for adaptive optimal output regulation of linear continuous-time systems with assured convergence rate," *Control Engineering Practice*, vol. 121, p. 105042, Apr. 2022.

[45] E. D. Sontag, "Input to state stability: Basic concepts and results," in *Nonlinear and Optimal Control Theory*. Springer, 2008, pp. 163–220.

[46] Z.-P. Jiang and Y. Wang, "Input-to-state stability for discrete-time nonlinear systems," *Automatica*, vol. 37, no. 6, pp. 857–869, Jun. 2001.

[47] Y. Jiang, J. Fan, W. Gao, T. Chai, and F. L. Lewis, "Cooperative adaptive optimal output regulation of discrete-time nonlinear multi-agent systems," *Automatica*, vol. 121, p. 109149, Nov. 2020.