Tracking Control for Linear Discrete-Time Networked Control Systems With Unknown Dynamics and Dropout

Yi Jiang, Student Member, IEEE, Jialu Fan, Member, IEEE, Tianyou Chai, Fellow, IEEE, Frank L. Lewis, Fellow, IEEE, and Jinna Li, Member, IEEE

Abstract—This paper develops a new method for solving the optimal control tracking problem for networked control systems (NCSs), where network-induced dropout can occur and the system dynamics are unknown. First, a novel dropout Smith predictor is designed to predict the current state based on historical data measurements over the communication network. Then, it is shown that the quadratic form of the performance index is preserved even with dropout, and the optimal tracker solution with dropout is given based on a novel dropout generalized algebraic Riccati equation. New algorithms for off-line policy iteration (PI), online PI, and Q-learning PI are presented for NCS with dropout. The Q-learning algorithm adaptively learns the optimal control online using data measured over the communication network based on reinforcement learning, including dropout, without requiring any knowledge of the system dynamics. Simulation results are provided to show that the proposed approaches give proper optimal tracking performance for the NCS with unknown dynamics and dropout.

Index Terms—Dropout, networked control system (NCS), Q-learning, reinforcement learning (RL).

I. INTRODUCTION

In networked control systems (NCSs), communication networks are employed to exchange the information and control signals (reference input, feedback state, and control input signals) between the sensors, the controllers, and the actuators. With the introduction of the communication networks, the advantages of NCSs are low cost, reduced weight, simple installation and maintenance, and high reliability; therefore, NCSs are widely applied to robot control, motor control, industrial manufacture, vehicles, aircraft, and spacecraft [1]–[7].

Despite their advantages and wide applications, communication networks will introduce errors through unreliable transmission, such as network-induced dropout [5]–[8], network-induced time delay [8]–[10], network noise [6], [11], [12], and quantization problem [13]–[17], which are usually caused by limited bit rate of the communication channels, by the sharing of limited network bandwidth, or by a node waiting to send out a packet via a busy channel. It is known that the occurrence of dropout, time delay, noise, and quantization degrade the stability and the control performance of the closed-loop system. Many researchers have paid attention to the study of the controller design for stabilization and proper performance for the NCSs.

For linear NCS with unreliable transmission, [4]–[9] and [13] use Lyapunov functions to obtain a group of linear matrix inequalities, whose solution gives stabilizing feedback gains. Papers [18]–[21] discuss the transmission control protocol (TCP) and the user datagram protocol (UDP), combine the Kalman filter and stochastic optimal control, give the relationship between the stability and the probability of the dropout, and prove that the optimal performance index has a quadratic form. Paper [22] considers the delay compensations in both the feedback and forward channels in NCSs for networked output tracking control problem. Reference [23] proposes the networked-predictive-control scheme to compensate for the network-induced delay. Paper [17] investigates the problem of event-triggered $H_{\infty}$ control for a networked singular system with both state and input subject to quantizations. For the nonlinear case, paper [24] proposes a model predictive control framework to compensate for the two-channel packet dropout. All these approaches require the knowledge of the system dynamics, and some of these approaches even require the statistical characteristics of the communication networks.

Reinforcement learning (RL) [25]–[34] is a class of machine learning method widely used in optimal control to find an optimal policy for unknown system dynamics. Papers [31] and [35] use Q-learning to solve the tracking and zero-sum game problem with unknown system dynamics, [36] proves the convergence of the value iteration (VI), and [37] utilizes...
the measured input and output data to implement policy iteration (PI) and VI. Reference [38] uses actor-critic neural networks to learn the optimal control online for uncertain nonlinear discrete-time (DT) systems.

To address unknown system dynamics NCS, an optimal control approach for linear NCS and a network scheduling protocol in a distributed framework is presented in [39]. Stochastic Q-learning method is used to design the optimal controller in an event-sampled framework. This design considers an ideal communication network without any network imperfections. Paper [40] proposes the stochastic optimal control method using an adaptive estimator and ideas from Q-learning to solve the infinite horizon optimal regulation problem for unknown NCS with time-varying system matrices. Then, paper [41] extends this approach to the nonlinear case.

However, these papers do not consider the influence of network-induced feedback dropout. When a dropout occurs in the feedback loop, no current state can be obtained for PI, then, the optimal control cannot be calculated, and even the unstable control input will be calculated. Thus, this paper develops a new Q-learning approach for the tracking problem in linear DT NCSs with unknown dynamics and dropout. A novel Smith predictor is developed to take advantage of the historical data to estimate the current state of the system. This information is used to achieve Q-learning online when dropout occurs.

The contributions of this paper are as follows.
1) A novel network-induced dropout Smith predictor is developed to predict the current state from historical measured data when network-induced dropout occurs.
2) It is proven that the NCS performance index is still quadratic even with dropout, and the optimal tracking control solution is provided by solving an NCS generalized algebraic Riccati equation (GARE).
3) A Q-learning method is given for using PIs to solve for the Q-function online using measured data with completely unknown system dynamics and network-induced dropout.

This paper is organized as follows. In Section II, we formulate the NCS optimal control problem with network-induced dropout. In Section III, we establish a novel network-induced dropout Smith predictor, prove the quadratic form of the NCS optimal control problem, and solve this problem by solving a GARE. In Section IV, we provide an off-line PI and an online iteration to solve the GARE off-line and online. In Section V, we provide the PI to solve Q-function online using measured data with network-induced dropout. Simulation experiments are used to verify the performance of the algorithms.

II. PROBLEM FORMULATION

The objective of this paper is to make the output of a dynamical system track a reference signal. This is known as the tracking problem. In this section, we introduce the linear quadratic tracker (LQT) problem and formulate the network-induced dropout problem, where some portions of signals can be lost due to failures in a communication system.

We formulate the system of interest as the DT linear system having dynamics

\[
\begin{align*}
x(k+1) &= Ax(k) + Bu(k) \\
y(k) &= Cx(k)
\end{align*}
\]

where \(x \in \mathbb{R}^{nx}\) is the state, \(u \in \mathbb{R}^{nu}\) is the input, and \(y \in \mathbb{R}^{ny}\) is the output, respectively. \(A \in \mathbb{R}^{nx \times nx}\), \(B \in \mathbb{R}^{nx \times nu}\), and \(C \in \mathbb{R}^{ny \times nx}\) are constant matrices.

This is a standard form used to model a system that develops through time in response to control \(u(k)\) and generates an output signal \(y(k)\). Our objective in the tracker problem is to design a control input \(u(k)\) that makes the output \(y(k)\) follow a desired reference signal \(r(k)\). We take the reference signal \(r(k)\) generated by the dynamical system

\[
r(k+1) = Fr(k)
\]

where \(r \in \mathbb{R}^{nr}\) and \(F \in \mathbb{R}^{nr \times nr}\). In the tracker problem, it is desired for the output \(y(k)\) of system (1) to track the reference signal \(r(k)\).

A networked control system is shown in Fig. 1. Therefore, the state \(x(k)\) is measured and passes through a communication network to a controller, which computes \(u(k)\). It is assumed that the state information is transmitted with a single packet and that some packets are lost in transmission, known as network-induced dropout. Therefore, the state \(x_f(k)\) available to the controller with network-induced dropout is

\[
x_f(k) = \delta(k)x(k) + (1 - \delta(k))x_f(k - 1)
\]

where \(\delta(k) \sim (0, 1)\), \(\delta(k) = 0\) means dropout occurs and \(\delta(k) = 1\) means a successful transmission. Then, the following assumptions are made.

Assumption 1: The pair \((A, B)\) is stabilizable and the pair \((A, C)\) is observable.

Assumption 2: The maximum number of contiguous feedback packet dropout \(\delta_f(k)\) is \(\delta_f^{max}\), equivalently

\[
\delta_f^{max} \sum_{i=0}^{\delta_f^{max}} \delta(k - i) > 0.
\]

This paper assumes that the number of contiguous feedback packet dropout \(\delta(k)\) is bounded; when the contiguous feedback packet dropout number is infinite, the control system is equivalent to an open-loop control system.
To solve the tracker problem, construct the augmented system as

\[
X(k + 1) = \begin{bmatrix} A & 0 \\ F & 0 \end{bmatrix} x(k) + \begin{bmatrix} B \\ 0 \end{bmatrix} u(k) \\
A_2 X(k) + B_2 u(k) \\
y(k) = [C \ 0] X(k) \equiv C_2 X(k) \tag{5}
\]

where the augmented state is \( X(k) = \begin{bmatrix} x(k) \\ r(k) \end{bmatrix} \). In the tracker problem, it is desired for the system output \( y(k) \) in (1) to follow the reference signal \( r(k) \) generated by (2). This is accomplished by selecting a suitable control signal \( u(k) \) in (1). To formulate a design problem for choosing \( u(k) \) to make \((r(k) - y(k))\) go to zero, thus achieving tracking, define an LQT problem performance index as

\[
J(k) = \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} [(r(i) - y(i))^T Q (r(i) - y(i)) + u^T (i) R u(i)]
\]

\[
= \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} U(i)
\]

\[
= \frac{1}{2} U(k) + \gamma J(k + 1) \tag{6}
\]

where \( 0 < \gamma \leq 1 \) is a discount factor. The tracker problem is solved by selecting control \( u(k) \) to minimize this performance index. If the reference generator (2) is asymptotically stable, one may select \( \gamma = 1 \). If (2) is critical stable, e.g., in tracking a unit step, then it is required that \( \gamma < 1 \). In fact, one selects the discount factor so that \((\gamma^{0.5} F)\) is stable, which can be proven in Section III.

Remark 1: The reference generator (2) is a standard form used in many works. The system is said to be stable if signal \( r(k) \) goes to zero with time. In general, we assume that \( F \) is not stable. If \( F \) is stable, then a trivial control guarantees trajectory following by making \( y(k) \) go to zero. The model (2) can describe many sorts of signals. By selecting \( F \) to have one pole at the origin, we can generate a whole family of unit step reference signals. If \( F \) has nonrepeated poles on the imaginary axis, we can generate a family of sinusoidal reference signals.

III. SOLUTION TO THE DROPOUT LQT PROBLEM

In this section, we introduce a dropout Smith predictor to predict the current state when network-induced dropout occurs. Using this Smith predictor, we establish that the performance index (6) has a quadratic form. The solution to the dropout LQT problem is given in terms of a GARE. This section develops a dropout LQT Bellman equation, which is used in Section V to use RL to solve the dropout LQT problem online when the dynamics (1) and (2) are unknown.

A. Network-Induced Dropout Smith Predictor

The solution to the LQT problem requires knowledge of the states \( x(k) \) and \( r(k) \). In this section, we design a dropout Smith predictor that uses the past values of the available measurements with dropout (3) to predict the current state \( X(k) \) in (5).

The feedback information \( x(k) \) is transmitted through the network using a single packet that may succumb to network-induced dropout as in (3). Assume that the network control system communication protocol is the TCP or the UDP. In the TCP case, there is acknowledgment of received packets, while in the UDP case, no feedback is provided on the communication link. But, using either TCP or UDP, the dropout history is known and can be used to improve the controller design.

Now, define the number of contiguous packet dropouts that has occurred prior to step \( k \) as \( \delta_f(k) \). If \( x(k) \) is transmitted successfully according to (3), we define \( \delta_f(k) = 0 \). Obviously, \( 0 \leq \delta_f(k) \leq \delta_f \max \). Then, the most recent data available through the network at time \( k \) are

\[
x_f(k) = x(k - \delta_f(k)). \tag{7}
\]

With this definition of contiguous feedback packet dropout number, the network-induced dropout can be considered as a stochastic bounded time delay. Then, the current state can be predicted by using historical state and control input information as

\[
x(k) = A^{\delta_f(k)} x(k - \delta_f(k)) + \sum_{i=1}^{\delta_f(k)} A^{i-1} B u(k - i). \tag{8}
\]

Using TCP or UDP, \( \delta_f(k) \) is known. Equation (7) gives the most recent data available through the network at time \( k \) as \( x(k - \delta_f(k)) \). Equation (8) explicitly computes current state \( x(k) \) in terms of this most recent data and the controls \( u(k - i) \) that have been applied, since time \( k - \delta_f(k) \).

Now, a dropout Smith predictor [42] using the measurable state can be constructed. Define the available measurement vector at time \( k \) as

\[
z(k) = \begin{bmatrix} x_f^T(k) & \cdots & x_f^T(k) \\ \delta_f \max + 1 & \cdots & \delta_f \max \end{bmatrix} \begin{bmatrix} u^T(k - 1) \\ \cdots \\ u^T(k - \delta_f \max) \\ r^T(k) \end{bmatrix}^T. \tag{9}
\]

Thus, the predictive state is

\[
X(k) = MN(\delta_f(k)) z(k) \tag{10}
\]

where

\[
M = \begin{bmatrix} I & A & \cdots & A^{\delta_f \max} & B & AB & \cdots & A^{\delta_f \max - 1} B & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & I \end{bmatrix} \tag{11}
\]

and \( N(\delta_f(k)) \) is a \([n_x \times (\delta_f \max + 1) + n_u \times \delta_f \max + n_r] \times [n_x \times (\delta_f \max + 1) + n_u \times \delta_f \max + n_r] \) matrix depending on \( \delta_f(k) \) defined as follows. Define \( \tilde{z}(k) = N(\delta_f(k)) z(k) \) that when \( \delta_f(k) = 0,1,\cdots,\delta_f \max \), \( \tilde{z}(k) \) is calculated,
respectively, as
\[
\bar{z}(k) = \begin{bmatrix} x_T^T(k) & 0 & \ldots & 0 & 0 & \ldots & 0 & r^T(k) \end{bmatrix}^T
\]
where \( \delta_{f_n}(k) = 0 \)
(12)
\[
\bar{z}(k) = \begin{bmatrix} 0 & x_T^T(k) & \ldots & 0 \\
\delta_{f_n}(k) + 1 & u^T(k - 1) & \ldots & 0 \delta_{f_n}(k) + 1 \\
\end{bmatrix} \begin{bmatrix} r^T(k) \end{bmatrix}, \quad \delta_{f_n}(k) = 1
\]
(13)
\[
\bar{z}(k) = \begin{bmatrix} 0 & 0 & \ldots & x_T^T(k) & \delta_{f_n}(k) = \delta_{f_{\max}} \\
\delta_{f_n}(k) + 1 & u^T(k - 1) & \ldots & u^T(k - \delta_{f_{\max}}) \end{bmatrix} \begin{bmatrix} r^T(k) \end{bmatrix}^T.
\]
(14)
Note that \( \bar{z}(k) \) is known at time \( k \). Then, the predictive state (5) can be described as
\[
X(k) = M \bar{z}(k).
\]
(15)

B. Quadratic Form of the Dropout LQT Problem

With augmented system (5), the network-induced dropout LQT problem performance index is
\[
J(k) = \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} \left[ X^T(i) Q_1 X(i) + u^T(i) R u(i) \right]
\]
where \( Q_1 = C_1^T Q C_1 \) with \( C_1 = [C, -I] \). Define the control input as
\[
u(k) = K_1 x(k) + K_2 r(k)
\]
\[= K X(k) = K M \bar{z}(k) \equiv \bar{K} \bar{z}(k). \]
(17)

**Theorem 1:** Take any stabilizing input (17) and select discount factor \( \gamma \), such that \((\gamma^{-0.5} F)\) is stable. Then, the dropout LQT problem performance index has a quadratic form
\[
J(k) = \frac{1}{2} X^T(k) P X(k) - \frac{1}{2} \bar{z}^T(k) \bar{P} \bar{z}(k)
\]
for some matrix \( P = \bar{P}^T > 0 \), where \( \bar{P} = \bar{P}^T = M^T P M > 0 \).

**Proof:** Refer to [35], for any fixed stabilizing policy (17); then, the dropout LQT problem performance index has a quadratic form depending on \( X(k) \) as
\[
J(k) = \frac{1}{2} X^T(k) P X(k)
\]
where \( P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \), the definitions of \( P_{11}, P_{12}, P_{21}, \) and \( P_{22} \) are [35]
\[
P_{11} = \sum_{i=0}^{\infty} \gamma^{i} \left[ (A_1^T)^i (C^T Q C + K_1^T R K_1) A_1^i \right]
\]
\[
P_{12} = \sum_{i=0}^{\infty} \gamma^{i} \left[ (A_1^T)^i (-C^T Q + K_1^T R K_2) F^i \right. \]
\[+ (A_1^T)^i (C^T Q C + K_1^T R K_1) T \]
\[
P_{21} = P_{12}^T
\]
\[
P_{22} = \sum_{i=0}^{\infty} \gamma^{i} \left[ (T^T \left( -C^T Q + K_1^T R K_2 \right) F^i \right. \]
\[+ T^T \left( C^T Q C + K_1^T R K_1 \right) T \]
\[+ (F^i)^T \left( -Q C + K_2^T R K_1 \right) T \]
\[+ (F^i)^T \left( Q + K_2^T R K_2 \right) T \]
\]
where \( A_\epsilon = A + B K_1 \) and \( T = \sum_{n=0}^{\infty} A_1^{T-n} B K_2 F^n \), when \((\gamma^{-0.5} F)\) is stable, \( P_{22} \) is bounded. Replace \( X(k) \) by \( \bar{z}(k) \), then one obtains
\[
J(k) = \frac{1}{2} \bar{z}^T(k) \bar{P} \bar{z}(k) = \frac{1}{2} \bar{z}^T(k) \bar{P} \bar{z}(k) + \gamma \bar{z}^T(k + 1) \bar{P} \bar{z}(k + 1)
\]
(20)
where \( \bar{P} = M^T P M \).
The proof is completed.

\[\square\]

C. Solution to the Dropout LQT Problem

In this section, we derive the Bellman equation for the dropout LQT, and a dropout LQT GARE is given to obtain the solution of the dropout LQT.

Introducing the quadratic form of the performance index (18) into (6), we obtain the LQT Bellman equation
\[
X^T(k) P X(k) = X^T(k) Q_1 X(k) + u^T(k) R u(k)
\]
\[+ \gamma X^T(k + 1) P X(k + 1). \]
(21)
Replacing \( X(k) \) by \( \bar{z}(k) \), the dropout LQT Bellman equation is
\[
\bar{z}^T(k) \bar{P} \bar{z}(k) = \bar{z}^T(k) M \bar{P} \bar{z}(k) + u^T(k) R u(k)
\]
\[+ \gamma \bar{z}^T(k + 1) \bar{P} \bar{z}(k + 1). \]
(22)
Define the LQT Hamiltonian as
\[
H(k) = \bar{z}^T(k) M \bar{P} \bar{z}(k) + u^T(k) R u(k)
\]
\[+ \gamma \bar{z}^T(k + 1) \bar{P} \bar{z}(k + 1) - \bar{z}^T(k) \bar{P} \bar{z}(k). \]
(23)

or
\[
H(k) = \bar{z}^T(k) M \bar{P} \bar{z}(k) + u^T(k) R u(k)
\]
\[+ \gamma J(k + 1) - J(k). \]
(24)

Theorem 2 provides the solution to the dropout LQT problem.

**Theorem 2:** The optimal control input of the dropout LQT problem is
\[
u(k) = K M \bar{z}(k) = \bar{K} \bar{z}(k)
\]
(25)
where
\[
\bar{K} = (R + \gamma B_1^2 P B_2) -1 \gamma B_2^T P A_2 M
\]
(26)
or
\[\bar{K} = (R + \gamma B_2^T M^T \bar{P} M^* B_2)^{-1} \gamma B_2^T M^T \bar{P} M^* A_2 M \]
(27)
where $M^* = M^T(MM^T)^{-1}$ and $\tilde{P}$ satisfies the dropout LQT GARE

$$
M^TQ_1M - \tilde{P} + \gamma M^T A_2^T M^* \tilde{P} M^* A_2M - \gamma^2 M^T A_2^T M^{*T} M^* B_2 \left(R + \gamma B_2^T M^* \tilde{P} M^* B_2\right)^{-1} B_2^T M^{*T} \tilde{P} M^* A_2M = 0
$$

(28)
or equivalent to the following LQT ARE:

$$
Q_1 - P + \gamma A_2^T P A_2 - \gamma^2 A_2^T P B_2 \times \left(R + \gamma B_2^T P B_2\right)^{-1} B_2^T P A_2 = 0.
$$

(29)

Proof: The stationary condition of the LQT Bellman equation is

$$
\frac{\partial H(k)}{\partial u(k)} = 2Ru(k) + \gamma \frac{\partial \tilde{z}(k + 1)^T}{\partial u(k)} \frac{\partial J(\tilde{z}(k + 1))}{\partial \tilde{z}(k + 1)} = 0.
$$

(30)

Using (5) and (15), one has

$$
X(k + 1) = M \tilde{z}(k + 1) = A_2 X(k) + B_2 u(k).
$$

(31)

Therefore

$$
\tilde{z}(k + 1) = M^* A_2 \tilde{z}(k) + M^* B_2 u(k).
$$

(32)

where $M^* = M^T(MM^T)^{-1}$ is the reverse of $M$, and thus, $P = M^{*T} \tilde{P} M^*$.

The stationary condition of the dropout LQT Bellman equation is

$$
\frac{\partial H(k)}{\partial u(k)} = 2Ru(k) + \gamma \frac{\partial \tilde{z}(k + 1)^T}{\partial u(k)} \frac{\partial J(\tilde{z}(k + 1))}{\partial \tilde{z}(k + 1)} = 2Ru(k) + 2\gamma B_2^T M^{*T} \tilde{P} \tilde{z}(k + 1) = 0.
$$

(33)

Then, the optimal input is

$$
u(k) = -(R + \gamma B_2^T M^{*T} \tilde{P} M^* B_2)^{-1} \gamma B_2^T M^{*T} \tilde{P} M^* A_2 \tilde{z}(k).
$$

(34)

Substituting (5) and (34) in the Bellman equation (22) results in the dropout LQT GARE (28)

$$
M^TQ_1M - \tilde{P} + \gamma M^T A_2^T M^{*T} \tilde{P} M^* A_2M - \gamma^2 M^T A_2^T M^{*T} M^* B_2 \left(R + \gamma B_2^T M^{*T} \tilde{P} M^* B_2\right)^{-1} B_2^T M^{*T} \tilde{P} M^* A_2M = 0.
$$

(35)

It is direct to show that this is equivalent to (29).

The proof is completed. \hfill \square

Theorem 3 shows the stability and the optimality of the solution by solving dropout LQT GARE.

Theorem 3: Apply the control (25) to the system (5) with network-induced dropout. Select discount factor $\gamma$, such that $(\gamma^{0.5} F)$ is stable, and

$$
0 < (P_{11} - C^T QC)(P_{11} + G)^{-1} < \gamma^2 I
$$

(36)

where $G = P_{11} B (R + B^T P_{11} B)^{-1} R (R + B^T P_{11} B)^{-1} B^T P_{11}$, and then, the closed-loop system (5) is stable. Moreover, this control is optimal and it minimizes the performance index (16).

Proof: The closed-loop system with any fixed stabilizing policy (17) will be

$$
x(k + 1) = (A + BK_1)x(k) + BK_2 r(k) = A_c x(k) + B_c r(k).
$$

(37)

If the eigenvalues of the closed-loop system $A_c$ are in the unit circle, then the stability of closed-loop system holds.

Inspired by [43] and [44], assume that $\lambda$ is an eigenvalue of the closed-loop system dynamics $A_c$, so that $A_c x_\lambda = \lambda x_\lambda$, where $x_\lambda$ is the feature vector. Then, multiplying the left-hand and right-hand sides of the following closed-loop dynamics LQT ARE by $x_\lambda^T$ and $x_\lambda$:

$$
C^T QC - P_{11} + \gamma A_2^T P_{11} A_c + K_1^T R K_1 = 0
$$

(38)

where $K_1 = -(R + \gamma B^T P_{11} B)^{-1} \gamma B^T P_{11} A$, one has

$$
x_\lambda^T(k) C^T QC x_\lambda - x_\lambda^T(k) P_{11} x_\lambda + \gamma |\lambda|^2 x_\lambda^T(k) P_{11} x_\lambda
$$

$$
+ \gamma^2 |\lambda|^2 x_\lambda^T(k) P_{11} B (R + \gamma B^T P_{11} B)^{-1}
$$

$$
R (R + \gamma B^T P_{11} B)^{-1} B^T P_{11} x_\lambda = 0.
$$

(39)

Since $R$ and $Q$ are positive definite, and $(R + \gamma B^T P_{11} B)^{-1} > (R + B^T P_{11} B)^{-1}$, which yields

$$
(1 - \gamma |\lambda|^2) P_{11} + \gamma^2 |\lambda|^2 P_{11} B (R + B^T P_{11} B)^{-1}
$$

$$
R (R + B^T P_{11} B)^{-1} B^T P_{11}.
$$

(40)

Define $G = P_{11} B (R + B^T P_{11} B)^{-1} R (R + B^T P_{11} B)^{-1} B^T P_{11}$, then

$$
P_{11} - C^T QC > \gamma^2 |\lambda|^2 (P_{11} + G)
$$

(41)

or equivalent

$$
\frac{1}{\gamma^2} (P_{11} - C^T QC)(P_{11} + G)^{-1} > |\lambda|^2 I.
$$

(42)

If the eigenvalues of the closed-loop system $A_c$ are in the unit circle, which means $|\lambda| \leq 1$, then the stability of closed-loop system holds. Therefore

$$
0 < (P_{11} - C^T QC)(P_{11} + G)^{-1} < \gamma^2 I.
$$

(43)

Therefore, stability is proven.

To prove optimality, multiplying both sides of the performance index (6) by $\gamma^k$ gives

$$
\gamma^k J(k) = \gamma^k U(k) + \gamma^{k+1} J(k + 1).
$$

(44)

Subtracting $\gamma^{k+1} J(k + 1)$ from both sides of this equation yields

$$
\gamma^{k+1} J(k + 1) - \gamma^k J(k) = -\gamma^k U(k).
$$

(45)

Sum both sides of (45) from $k$ to $\infty$ to obtain

$$
\gamma^{\infty} J(\infty) - \gamma^k J(k) = -\sum_{i=k}^{\infty} \gamma^i U(i).
$$

(46)

Since $\gamma^{\infty} J(\infty) = 0$, thus

$$
\gamma^k J(k) = \sum_{i=k}^{\infty} \gamma^i U(i).
$$

(47)
Sum the left-hand side of (45) from \( k \) to \( \infty \) to obtain

\[
\gamma^k J(k) = \sum_{i=k}^{\infty} \left[ \gamma^i J(i) - \gamma^{i+1} J(i+1) \right]
\]

\[
= \frac{1}{2} \sum_{i=k}^{\infty} \left[ \gamma^i \tilde{z}^T(i) \tilde{P} \tilde{z}(i) - \gamma^{i+1} \tilde{z}^T(i+1) \tilde{P} \tilde{z}(i+1) \right].
\]

Introduce the dropout LQT GARE (35) to (48), and combine it with (47) to yield

\[
\gamma^k J(k) = \frac{1}{2} \sum_{i=k}^{\infty} \left[ u(i) + (R + \gamma B_2^T M^* T \tilde{P} M^* B_2)^{-1} \right]
\]

\[
\times (R + \gamma B_2^T M^* T \tilde{P} M^* B_2) \left[ u(i) + (R + \gamma B_2^T M^* T \tilde{P} M^* B_2)^{-1} \right]
\]

\[
\gamma B_2^T M^* T \tilde{P} M^* A_2 M \tilde{z}(i)).
\]

Since \((R + \gamma B_2^T M^* T \tilde{P} M^* B_2)\) is positive definite, to minimize the value of (49), the optimal control input satisfies the control policy (25) and (27), the proof of the optimality is complete.

The proof is completed.

\( \Box \)

Remark 2: In this paper, we found a bound for \( \gamma \) to guarantee stability. The bound for \( \gamma \) just shows that we need to pick \( \gamma \) close enough to 1.

Remark 3: It is shown that the tracking error is not asymptotically stable, but the stability can be guaranteed; it means that the tracking error is not zero because of using the discount factor in the performance index. However, by choosing a proper discount factor and a large enough positive definite matrix \( \tilde{Q} \) in the performance index, one can obtain the steady tracking error as small as desired.

IV. REINFORCEMENT LEARNING TO SOLVE THE DROPOUT LQT PROBLEM ONLINE

In Section III, it was seen that the optimal tracker control can be computed by finding the solution to the LQT (29) and then computing the optimal control gain feedback using (25) and (26). In this section, we utilize the dropout LQT Bellman equation (22) to design algorithms based on RL that solve the dropout LQT problem without solving the dropout LQT GARE in Theorem 2. This is accomplished in two steps. First, Bellman equation (22) is used to directly develop an off-line PI algorithm (Algorithm 1) that does not require solution of the LQT GARE (29). However, Algorithm 1 still requires knowledge of the system dynamics (1). As such, it is still an offline algorithm that requires repeated solutions of the Lyapunov equation (50). Therefore, we next derive Algorithm 2, which is an online data-based algorithm that finds the optimal control using data measured in real time.

Using the feedback control (17) and (27) introduces them into the Bellman equation (22). Then, it is seen that the dropout LQT Bellman equation is equivalent to the dropout LQT Lyapunov equation

\[
\tilde{P} = M^T Q_1 M + \tilde{K}^T R \tilde{K}
\]

\[
+ \gamma (M^* A_2 M + M^* B_2 \tilde{K})^T \tilde{P} (M^* A_2 M + M^* B_2 \tilde{K}).
\]

The following offline PI algorithm calculates the optimal control by repeatedly solving this dropout LQT Lyapunov equation instead of solving the dropout LQT GARE.

Algorithm 1 Off-Line PI for Dropout LQT Solution

Initiation: Start with a stabilizing control policy \( \tilde{K}^1 \). Iterate the following two steps on \( j \) until convergence.

1) Policy Evaluation: solve for \( \tilde{P}^{j+1} \) using the dropout LQT Lyapunov equation

\[
\tilde{P}^{j+1} = M^T Q_1 M + (\tilde{K}^j)^T R \tilde{K}^j
\]

\[
+ \gamma (M^* A_2 M + M^* B_2 \tilde{K}^j)^T \tilde{P}^{j+1} (M^* A_2 M + M^* B_2 \tilde{K}^j)
\]

2) Policy Improvement: Update the policy using (27), that is

\[
\tilde{K}^{j+1} = -(R + \gamma B_2^T M^* T \tilde{P}^{j+1} M^* B_2)^{-1}
\]

\[
\gamma B_2^T M^* T \tilde{P}^{j+1} M^* A_2 M
\]

The convergence proof of Algorithm 1 is similar to the convergence proof in [26], and therefore, it is shown that \( \tilde{P}^{j+1} \) is convergent to the solution \( \tilde{P}^* \) of dropout LQT GARE.

Instead of using dropout LQT Lyapunov equation, the next algorithm shows how to utilize data measured online in real time to implement the policy evaluation.

Algorithm 2 On-Line PI for Dropout LQT Solution

Initiation: Start with a stabilizing control policy \( K^1 \). Iterate the following three steps on \( j \) until convergence.

1) Policy Evaluation: solve for \( \tilde{P}^{j+1} \) using the dropout LQT Bellman equation (22)

\[
\tilde{z}^T(k) M^T P^{j+1} M \tilde{z}(k)
\]

\[
= \tilde{z}^T(k) M^T (Q_1 + (K^j)^T R \tilde{K}^j) M \tilde{z}(k)
\]

\[
+ \gamma \tilde{z}^T(k+1) M^T P^{j+1} M \tilde{z}(k+1)
\]

(53)

2) Policy Improvement: Update the policy using (26), that is

\[
K^{j+1} = -(R + \gamma B_2^T P^{j+1} B_2)^{-1} \gamma B_2^T P^{j+1} A_2
\]

(54)

3) Variable Transformation:

\[
\tilde{P}^{j+1} = M^T P^{j+1} M
\]

(55)

\[
\tilde{K}^{j+1} = K^{j+1} M
\]

(56)

Policy evaluation (55) is equivalent to

\[
X^T(k) P^{j+1} X(k) = X^T(k) (Q_1 + (K^j)^T R K^j) X(k)
\]

\[
+ \gamma X^T(k+1) P^{j+1} X(k+1).
\]

(57)
To solve (53) or (57), standard least square (LS) or the recursive LS can be utilized as follows. Matrix \( P_{j+1} \) is a symmetric \((n_x + n_r) \times (n_x + n_r)\) matrix with \((n_x + n_r) \times (n_x + n_r + 1)/2\) independent elements. Therefore, at least \((n_x + n_r) \times (n_x + n_r + 1)/2\) data tuples are required before (53) or (57) can be solved using LS. Using the Kronecker product, one has

\[
\begin{bmatrix}
\bar{z}(k) \\
\bar{z}(k+1) \\
\vdots \\
\bar{z}(k+s) \\
\end{bmatrix} \\ = \\
\begin{bmatrix}
X(k)(Q_1 + (K')^T R K') X(k) \\
X(k+s)(Q_1 + (K')^T R K') X(k+s) \\
\vdots \\
\end{bmatrix}
\]

where \( s \geq (n_x + n_r) \times (n_x + n_r + 1)/2 - 1 \). Then, \( P_{j+1} \) can be solved using LS as the equation

\[
\text{vec}(P_{j+1}) = (\eta^T(k)\eta(k))^{-1} \eta^T(k)\chi(k). \tag{58}
\]

Both the off-line PI algorithm (Algorithm 1) and the online PI algorithm (Algorithm 2) require the knowledge of the system dynamics. In Section V, we introduce a Q-learning algorithm to avoid the requirement of the knowledge of the system dynamics.

**Remark 4:** Using the VI algorithm [26] instead of using the PI algorithm, an initial stabilizing policy is not needed.

**V. Q-LEARNING TO SOLVE DROPOUT LQT BASED ON RL**

In Section IV, it was seen that the optimal tracker control can be computed by Algorithms 1 and 2. However, this requires knowing the system dynamics (5), or equivalently (1). Unfortunately, in many cases of interest in process control and elsewhere, the dynamics are unknown. In practice, time signals, such as \( u(k) \), \( x_f(k) \), and \( r(k) \), can be measured in the closed-loop system. Therefore, in this section, we introduce a Q-learning algorithm based on RL to solve the dropout LQT online, and this algorithm avoids the requirement of the knowledge of the system dynamics.

**A. Q-Function of the Dropout LQT**

In this section, we establish the Q-function of the dropout LQT.

Based on the dropout LQT Bellman equation (22), the DT dropout LQT Q-function is defined as

\[
Q(k) = \frac{1}{2} \bar{z}^T(k)M^T Q_1 M \bar{z}(k) + \frac{1}{2} u^T(k)R u(k) \\
+ \frac{1}{2}\gamma \bar{z}^T(k+1) \bar{P} \bar{z}(k+1). \tag{59}
\]

By using augmented system dynamics (5) and (32), (59) becomes

\[
Q(k) = \frac{1}{2} \bar{z}^T(k)M^T Q_1 M \bar{z}(k) \\
+ \frac{1}{2}\gamma (M^* A_2 \bar{M} \bar{z}(k) + M^* B_2 u(k))^T \\
\bar{P}(M^* A_2 \bar{M} \bar{z}(k) + M^* B_2 u(k)) \\
= \frac{1}{2} \begin{bmatrix} \bar{z}(k) \\ u(k) \end{bmatrix}^T \begin{bmatrix} H & \bar{u}(k) \\ \bar{u}(k) & \bar{u}(k) \end{bmatrix} \begin{bmatrix} \bar{z}(k) \\ u(k) \end{bmatrix} \tag{60}
\]

where \( H \) refers to as shown at the bottom of this page. Therefore, define

\[
Q(k) = \frac{1}{2} \begin{bmatrix} \bar{z}(k) \\ u(k) \end{bmatrix}^T \begin{bmatrix} H & \bar{u}(k) \\ \bar{u}(k) & \bar{u}(k) \end{bmatrix} \begin{bmatrix} \bar{z}(k) \\ u(k) \end{bmatrix} \\
= \begin{bmatrix} \bar{z}(k) \\ u(k) \end{bmatrix}^T \begin{bmatrix} H \bar{z} \bar{u} & \bar{z} \bar{u} \\ \bar{u} \bar{u} & \bar{u} \bar{u} \end{bmatrix} \begin{bmatrix} \bar{z}(k) \\ u(k) \end{bmatrix}. \tag{61}
\]

Applying \( \partial Q(k)/\partial u(k) = 0 \) to (61) yields the optimal control input

\[
\begin{align*}
\bar{u}(k) &= -H_{uu}^{-1} H_{uz} \bar{z}(k) \\
&= -(R + \gamma B_2^T M^* \bar{P} M^* B_2)^{-1} \gamma B_2^T M^* \bar{P} M^* A_2 \bar{z}(k).
\end{align*} \tag{62}
\]

**B. Reinforcement Learning to Solve Q-Function Online Using Measure Data**

To solve for the Q-function online, define the state

\[
Z(k) = \begin{bmatrix} \bar{z}(k) \\ u(k) \end{bmatrix}. \tag{63}
\]

The Q-function satisfies the dropout LQT Bellman equation

\[
Q(k) = \frac{1}{2} \bar{z}^T(k)M^T Q_1 M \bar{z}(k) + \frac{1}{2} u^T(k)R u(k) + \gamma Q(k+1). \tag{64}
\]

With state (63), the Q-function Bellman equation is

\[
Z^T(k) H Z(k) = \bar{z}^T(k)M^T Q_1 M \bar{z}(k) + u^T(k)R u(k) \\
+ \gamma Z^T(k+1) H Z(k+1). \tag{65}
\]

Define

\[
\begin{bmatrix} \bar{z}(k) \\ r(k) \end{bmatrix} = \begin{bmatrix} \bar{z}_1(k) \\ \bar{r}(k) \end{bmatrix} \tag{66}
\]

\[
M = \begin{bmatrix} I & M \bar{M} & 0 \\ 0 & 0 & I \end{bmatrix}. \tag{67}
\]

where \( \bar{z}_1(k) \) is the element of \( \bar{z}(k) \) form row 1 to \( n_z \). Clearly, \( \bar{z}_1(k) = x(k) \) when \( \delta(k) = 1 \) and \( \bar{z}_1(k) = 0 \) when \( \delta(k) = 0 \). Define \( \bar{r}(k) \) as the elements of \( \bar{r}(k) \) without \( \bar{z}_1(k) \) and \( r(k) \).
Thus, one obtains
\[
\dot{z}^T(k)M^TQ_1Mz(k) = z^T_1(k)CTQC\bar{z}_1(k) + r^T(k)Qr(k)
\]
\[
+ z^T_2(k)\tilde{M}^TCTQC\tilde{z}_2(k) - 2r^T(k)QC\bar{z}_1(k) - 2r^T(k)QC\tilde{z}_2(k).
\]
(68)

Using the Kronecker product, define \( \phi(k) = Z^T(k) \otimes Z^T(k) \), \( \sigma(k) = z^T_1(k) \otimes z^T_2(k) \), and \( \vartheta(k) = z^T_2(k) \otimes r^T(k) \). Then, the Q-function Bellman equation is
\[
\phi(k)\text{vec}(H) = z^T_1(k)CTQC\bar{z}_1(k) + r^T(k)Qr(k)
\]
\[
- 2r^T(k)QC\bar{z}_1(k) + u^T(k)Ru(k)
\]
\[
+ \sigma(k)\text{vec}(\tilde{M}^TCTQM)
\]
\[
- 2\vartheta(k)\text{vec}(QCM) + \gamma \phi(k + 1)\text{vec}(H).
\]
(69)

or equivalently
\[
[\phi(k) - \gamma \phi(k + 1), -\sigma(k), 2\vartheta(k)]
\]
\[
\begin{bmatrix}
\text{vec}(H) \\
\text{vec}(\tilde{M}^TCTQM) \\
\text{vec}(QCM)
\end{bmatrix}
\]
\[
= u^T(k)Ru(k) + z^T_1(k)CTQC\bar{z}_1(k) + r^T(k)Qr(k)
\]
\[
- 2r^T(k)QC\bar{z}_1(k).
\]
(70)

Define
\[
\zeta(k) = \begin{bmatrix}
\phi(k) - \gamma \phi(k + 1), -\sigma(k), 2\vartheta(k) \\
\phi(k + s) - \gamma \phi(k + s + 1), -\sigma(k + s), 2\vartheta(k + s)
\end{bmatrix}
\]
\[
\equiv [\xi(k)]_N
\]
(71)

\[
\zeta(k) = \begin{bmatrix}
u^T(k)Ru(k) + z^T_1(k)CTQC\bar{z}_1(k) + r^T(k)Qr(k) - 2r^T(k)QC\bar{z}_1(k)
\end{bmatrix}
\]
(72)

where \( N \) is the elementary column transformation matrix. Scalar \( s \) is an integral number that depends on the maximum number \( \delta_f \) of contiguous feedback packet dropout \( \delta(k) \).

Then, one has the rank condition
\[
\text{rank}(\zeta^T(k)\zeta(k)) = S_r
\]
(73)

where \( \text{rank}[\cdot] \) respects the rank of the matrix, and
\[
S_r = \sum_{i=0}^{\delta_f \text{max}} (n_x + n_r + n_u + i + n_u)
\times (n_x + n_r + n_u + i + n_u + 1)/2
\]
\[
- \delta_f \text{max} \times (n_x + n_u) \times (n_x + n_u + 1)/2
\]
\[
+ (n_x + \delta_f \text{max} \times n_u) \times (n_x + \delta_f \text{max} \times n_u + 1)/2
\]
\[
+ (n_x + \delta_f \text{max} \times n_u + n_r).
\]

The Q-function Bellman equation becomes
\[
\xi(k) \begin{bmatrix}
\text{vec}(H) \\
\text{vec}(\tilde{M}^TCTQM) \\
\text{vec}(QCM)
\end{bmatrix} = \zeta(k)
\]
(74)

or equivalently
\[
\xi(k)\text{vec}(\bar{H}) = \zeta(k)
\]
(75)

where
\[
\text{vec}(\bar{H}) = N^{-1} \begin{bmatrix}
\text{vec}(\tilde{M}^TCTQM) \\
\text{vec}(QCM)
\end{bmatrix}
\]

Finally, one obtains
\[
\text{vec}(\bar{H}) = (\xi^T(k)\xi(k))^{-1}\xi^T(k)\xi(k).
\]
(76)

Now, the following algorithm can be written.

**Algorithm 3** PI to Solve Q-Function Online Using Measure Data

**Initiation:** Start with a stabilizing control policy \( \tilde{K}^1 \). Iterate the following two steps on \( j \) until convergence.

1) **Policy Evaluation:** solve for \( \bar{H}^{j+1} \) using the LS equation (70)
\[
[u^T(k)Ru(k) + z^T_1(k)CTQC\bar{z}_1(k) + r^T(k)Qr(k) - 2r^T(k)QC\bar{z}_1(k)]^{j+1}
\]
\[
\times \begin{bmatrix}
\text{vec}(H) \\
\text{vec}(\tilde{M}^TCTQM) \\
\text{vec}(QCM)
\end{bmatrix}
\]
(77)

2) **Policy Improvement:** Update the policy using (62), that is
\[
u^{j+1}(k) = - (H_{uu}^{-1})^{j+1}H_{u'z}^{j+1}\zeta(k)
\]
(78)

The convergence of Algorithm 3 can be proven as in [45]. Note that, PI using Q-function is performed online and can be implemented without requiring any knowledge of the augmented system dynamics.

**Remark 5:** PI-based adaptive optimal control schemes require a persistent excitation (PE) condition [26] to ensure the sufficient exploration of the state space. If the state almost converges to the desired position and becomes stationary, the PE is no longer satisfied. A white noise signal can be added to the control input to ensure PE qualitatively.

**VI. SIMULATION EXPERIMENTS**

In this section, some simulation experiments are used to verify the effectiveness of the proposed method. We verify that Algorithms 1–3 give the same results.

Consider the following linear DT system:
\[
x(k + 1) = \begin{bmatrix}
-1 & 2 \\
2.2 & 1.7
\end{bmatrix} x(k) + \begin{bmatrix}
2 \\
1.6
\end{bmatrix} u(k)
\]
\[
y(k) = \begin{bmatrix} 1 & 2 \end{bmatrix} x(k).
\]
(79)

Then, \( n_x = 2, n_u = 1, \) and \( n_y = n_r = 1, \) respectively. The open-loop poles are \(-2.1445\) and \(2.8445\), so the system is unstable. The parameters in performance index (6) are selected as \( Q = 6, R = 1, \) and \( \gamma = 0.8, \) and the reference trajectory is given by
\[
r(k + 1) = -r(k)
\]
with \( r(1) = 5 \) and \( F = -1. \)
The maximum number of contiguous feedback packet dropout is \( \delta_{j_{\text{max}}} = 1 \). Therefore, the dropout Smith predictor matrix is
\[
M = \begin{bmatrix} I & A & B & 0 \\ 0 & 0 & 0 & I \end{bmatrix}.
\] (81)

The optimal matrices \( P^* \) and \( \tilde{P}^* \) by solving the dropout LQT AREs (28) and (29) are as follows, (82) and (83) shown at the bottom of this page.

Thus, the optimal feedback gains \( K^* \) and \( \tilde{K}^* \) are
\[
K^* = [0.1898 \quad -1.0085 \quad 0.0640]
\] (84)
\[
\tilde{K}^* = K^*M = [0.1898 \quad -1.0085 \quad -2.4086 \\
-1.3348 \quad -1.2340 \quad 0.0640].
\] (85)

**A. Off-Line Policy Iteration Simulation Experiments**

In this section, Algorithm 1 is used to calculate the optimal matrix \( \tilde{P} \) and the optimal feedback gain \( \tilde{K} \). The initial stabilizing control policy \( \tilde{K}^1 \) is selected as
\[
\tilde{K}^1 = [0.2 \quad -1.1 \quad -2.62 \quad -1.47 \quad -1.36 \quad 0.05].
\] (86)

All the initial elements of the \( \tilde{P}^{j+1} \) are 0. Fig. 2 shows the error between \( \tilde{P}^{j+1} \) and the optimal matrix \( P \) during learning. At two iterations, Algorithm 1 is applied, and after four iterations, \( \tilde{P}^{j+1} \) is calculated as (87), shown at the bottom of this page, and the feedback control gain is
\[
\tilde{K}^{j+1} = [0.1898 \quad -1.0085 \quad -2.4086 \\
-1.3348 \quad -1.2340 \quad 0.0640].
\] (87)

Fig. 3 shows the convergence of the \( \tilde{P} \) matrix parameters to their optimal values for off-line PI algorithm (Algorithm 1).

**B. Online Policy Iteration Simulation Experiments**

In this section, the online PI algorithm (Algorithm 2) is applied to system (79). The initial stabilizing control policy \( \tilde{K}^1 \) is also selected as (86), all the initial elements of \( P^{j+1} \) are 0, and the initial state is \( x(1) = [5, -5]^T \). Note that, Algorithm 2 starts at control step 10, and after each 6 control steps, \( P^{j+1} \) updates one time. After three iterations, \( P^{j+1} \) is calculated as
\[
P^{j+1} = \begin{bmatrix} 133.3864 & 16.0506 & 31.1407 \\
16.0506 & 25.1590 & -10.8271 \\
31.1407 & -10.8271 & 18.4868 \end{bmatrix}
\] (89)
Fig. 4. Evaluation of the output, the input, and the reference trajectory for off-line PI algorithm (Algorithm 1).

Fig. 5. Error between $P_j^{+1}$ and the optimal matrix $P^*$ during learning.

Fig. 6. Convergence of the $P$ matrix parameters to their optimal values for online PI algorithm (Algorithm 2).

and the feedback control gain is calculated as

$$K_{j+1} = [0.1899 \ -1.0085 \ 0.0640]. \quad (90)$$

Fig. 5 shows the error norm between $P_j^{+1}$ and the optimal matrix $P^*$ during learning. Fig. 6 shows the convergence of the $P$ matrix parameters to its optimal values for online PI algorithm (Algorithm 2). Fig. 7 shows the evaluation of the output, the input, and the reference trajectory for online PI algorithm (Algorithm 2). It is seen that the online algorithm (Algorithm 2) converges very close to the optimal controller and the results given by Algorithm 1.

C. Policy Iteration to Solve Q-Function Online Simulation Experiments

In this section, the online PI to solve Q-function in Algorithm 3 is applied to system (79).

In Q-function (60), the optimal matrix $\tilde{H}^*$ is as (91), shown at the bottom of this page, thus

$$\tilde{H}_{uz}^* = 10^3 \times [-0.1066 \ 0.5663 \ 1.3525 \ 0.7496 \ 0.6929 \ -0.0360] \quad (92)$$

$$\tilde{H}_{uu}^* = 561.5492. \quad (93)$$

To guarantee the excitation condition (PE), a white noise signal is added to the control input. The initial stabilizing control policy $\tilde{K}_1$ is also selected as (86). All the initial

$$\tilde{H}^* = 10^3 \times \begin{bmatrix} 0.1536 & -0.0915 & -0.3548 & 0.1518 & 0.1609 & 0.0380 & -0.1066 \\ -0.0915 & 0.5963 & 1.4034 & 0.8308 & 0.7712 & -0.0471 & 0.5663 \\ -0.3548 & 1.4034 & 3.4423 & 1.6761 & 1.5357 & -0.1416 & 1.3525 \\ 0.1518 & 0.8308 & 1.6761 & 1.7159 & 1.6329 & -0.0041 & 0.7496 \\ 0.1609 & 0.7712 & 1.5357 & 1.6329 & 1.5557 & 0.0006 & 0.6929 \\ 0.0380 & -0.0471 & -0.1416 & -0.0041 & 0.0006 & 0.0208 & -0.0360 \\ -0.1066 & 0.5663 & 1.3525 & 0.7496 & 0.6929 & -0.0360 & 0.5615 \end{bmatrix} \quad (91)$$
The elements of $\tilde{H}^{j+1}_{u\dot{z}}$ and $\tilde{H}^{j+1}_{uu}$ are 0, and the initial state is $x(1) = [5, -5]^T$. $\tilde{H}^{j+1}_{u\dot{z}}$ and $\tilde{H}^{j+1}_{uu}$ in Algorithm 3 are updated, which depend on rank condition (73). After three iterations, $\tilde{H}^{j+1}_{u\dot{z}}$ and $\tilde{H}^{j+1}_{uu}$ are calculated as

$$\tilde{H}^{j+1}_{u\dot{z}} = 10^3 \times \begin{bmatrix} 0.1066 & 0.5663 & 1.3525 \\ 0.7496 & 0.6929 & -0.0360 \end{bmatrix}$$

(94)

and the feedback control gain is

$$\tilde{K}^{j+1} = \begin{bmatrix} 0.1898 & -1.0085 & -2.4086 \\ -1.3349 & -1.2340 & 0.0640 \end{bmatrix}$$

(96)

Fig. 8 shows the evaluation of the output, the input, and the reference trajectory for Q-learning PI algorithm (Algorithm 3). Fig. 9 shows the convergence of $\tilde{H}^{j+1}_{u\dot{z}}$ and $\tilde{H}^{j+1}_{uu}$ to its optimal values $\tilde{H}^{*}_{u\dot{z}}$ and $\tilde{H}^{*}_{uu}$ during the learning process. Fig. 10 shows the convergence of $\tilde{H}$ to its optimal values $\tilde{H}^{*}$ during the learning process.

D. Comparison Simulation Experiment

In this section, a comparison simulation experiment is presented. In Algorithm 3, the feedback control gain $K_1$ and the feedforward control gain $K_2$ with Smith predictor are calculated as (96). In the comparison simulation experiment, the feedback control gain $K_1$ is calculated by Algorithm 1, the Smith predictor is the same as in (81). Then, refer to [22], the feedforward control gain $K_2$ is

$$K_2 = \Pi_{22} - K_1 \Pi_{12}$$

(97)

where \( \Pi_{11} \Pi_{12} \Pi_{21} \Pi_{22} = \Psi^T (\Psi \Psi^T)^{-1} \), and $\Psi = \begin{bmatrix} A - I & B \\ C & 0 \end{bmatrix}$.

The results of applying the control input calculated by Algorithm 3 and compared approach to the system (1) are in Fig. 14. To evaluate the control performance, the integral absolute error and the mean square error [46] are used, and
the evaluation equation is given by

\[ IAE_Y = \frac{1}{k^*} \sum_{k=1}^{k^*} |r(k) - y(k)| \]  
(98)

\[ MSE_Y = \frac{1}{k^*} \sum_{k=1}^{k^*} (r(k) - y(k))^2 \]  
(99)

The evaluation indices are shown in Table I. Clearly, the performance in this paper is better than the compared approach; moreover, the approach of this paper does not require the knowledge of system dynamics.

Remark 6: The reference trajectory of the recomparison simulation experiment is saltus step, if the reference trajectory is constant, and the evaluation indices of the compared approach are better than Algorithm 3, but the advantage of Algorithm 3 is a model-free approach.

VII. CONCLUSION

This paper learns the NCS tracker problem with unknown dynamics and network-induced dropout. The Q-learning PI are proposed, which can learn the optimal control online using measurable data with network dropout without the requirement of knowledge of system dynamics. The simulation results show that the proposed approach gives good tracking performance for the network-based control system with unknown dynamics and dropout.

The future work will focus on how to utilize the input, output, and reference data to implement the Q-learning PI.

REFERENCES


**Yi Jiang** (S’14) was born in Hubei, China. He received the B.Eng. degree in automation and the M.S. degree in control theory and control engineering from the State Key Laboratory of Synthetical Automation for Process Industries, Information Science and Engineering College, Northeastern University, Shenyang, China, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree.

In 2017, he was a Visiting Scholar with the UTA Research Institute, The University of Texas at Arlington, Arlington, TX, USA. His current research interests include networked control systems, industrial process operational control, and reinforcement learning.

**Jialu Fan** (M’12) received the B.Eng. degree in automation from Northeastern University, Shenyang, China, in 2006, and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2011.

She was a Visiting Scholar with Pennsylvania State University, State College, PA, USA, from 2009 to 2010. She is currently an Associate Professor with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. Her current research interests include networked control systems, delay-tolerant networks, and mobile social networks.

**Tianyou Chai** (M’90–SM’97–F’08) received the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 1985.

He has been with the Research Center of Automation, Northeastern University, since 1985, where he became a Professor in 1988 and a Chair Professor in 2004. He is currently the Founder and the Director of the National Engineering and Technology Research Center, Shenyang. He has made a number of important contributions in control technologies and applications. He has authored and co-authored two monographs, 84 peer-reviewed international journal papers, and around 219 international conference papers. His current research interests include adaptive control, intelligent decoupling control, integrated plant control and systems, and the development of control technologies with applications to various industrial processes.

Dr. Chai is a member of the Chinese Academy of Engineering, an Academician of the International Eurasian Academy of Sciences, and a fellow of the International Federation of Automatic Control (IFAC). He is a Distinguished Visiting Fellow of the Royal Academy of Engineering (U.K.) and an Invitation Fellow of the Japan Society for the Promotion of Science. For his contributions, he received three prestigious awards of the National Science and Technology Progress, the 2002 Technological Science Progress Award from the Ho Leung Ho Lee Foundation, the 2007 Industry Award for Excellence in Transitional Control Research from the IEEE Control Systems Society, and the 2010 Yang Jia-Chi Science and Technology Award from the Chinese Association of Automation. He has been invited to deliver more than 20 plenary speeches in international conferences of the IFAC and the IEEE.
Frank L. Lewis (S’70–M’81–SM’86–F’94) received the bachelor’s degree in physics and electrical engineering and the M.S.E.E. degree from Rice University, Houston, TX, USA, the M.S. degree in aeronautical engineering from the University of West Florida, Pensacola, FL, USA, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently a Distinguished Scholar Professor and a Distinguished Teaching Professor with The University of Texas at Arlington, Fort Worth, TX, USA, the Moncrief-ODonnell Chair of the University of Texas at Arlington Research Institute, Fort Worth, TX, USA, the Qian Ren Thousand Talents Professor and the Project 111 Professor with Northeastern University, Shenyang, China, and a Distinguished Visiting Professor with the Nanjing University of Science and Technology, Nanjing, China. He is involved in feedback control, intelligent systems, cooperative control systems, and nonlinear systems. He holds six U.S. patents, and has authored numerous journal special issues, journal papers, and 20 books, including Optimal Control, Aircraft Control, Optimal Estimation, and Robot Manipulator Control, which are used as university textbooks worldwide.

Dr. Lewis is a member of the National Academy of Inventors, a fellow of the International Federation of Automatic Control and the Institute of Measurement and Control, U.K., a Texas Board of Professional Engineer, a U.K. Chartered Engineer, and a founding member of the Board of Governors of the Mediterranean Control Association. He was a recipient of the Fulbright Research Award, the NSF Research Initiation Grant, the Terman Award from the American Society for Engineering Education, the Gabor Award from the International Neural Network Society, the Honeywell Field Engineering Medal from the Institute of Measurement and Control, U.K., the Neural Networks Pioneer Award from the IEEE Computational Intelligence Society, the Outstanding Service Award from the Dallas IEEE Section, and the Texas Regents Outstanding Teaching Award in 2013. He was elected as an Engineer of the Year by the Fort Worth IEEE Section. He was listed in the Fort Worth Business Press Top 200 Leaders in Manufacturing.

Jinna Li (M’12) received the M.S. and Ph.D. degrees from Northeastern University, Shenyang, China, 2006 and 2009, respectively. From 2009 to 2011, she held the post-doctoral position with the Laboratory of Industrial Control Networks and Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Beijing, China. From 2014 to 2015, she was a Visiting Scholar granted by the China Scholarship Council with the Energy Research Institute, Nanyang Technological University, Singapore. From 2015 to 2016, she was a Domestic Young Core Visiting Scholar granted by the Ministry of Education of China with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. Since 2017, she has been a Visiting Scholar with The University of Manchester, Manchester, U.K. She is currently an Associate Professor with the Shenyang University of Chemical Technology, Shenyang. Her current research interests include operational optimal control, reinforcement learning, distributed optimization control, and approximate dynamic programming.