JAISCR, 2013, Vol. 3, No. 3, pp. 175– 187
DOI 10.2478/jaiscr-2014-0012

FIXED FINAL TIME OPTIMAL ADAPTIVE CONTROL OF LINEAR DISCRETE-TIME SYSTEMS IN INPUT-OUTPUT FORM

Qiming Zhao, Hao Xu and S. Jagannathan
Department of Electrical and Computer Engineering, Missouri University of Science and Technology
Rolla, Missouri, USA
qzfyc@mst.edu

Abstract

In this paper, the fixed final time adaptive optimal regulation of discrete-time linear systems with unknown system dynamics is addressed. First, by transforming the linear systems into the input/output form, the adaptive optimal control design depends only on the measured outputs and past inputs instead of state measurements. Next, due to the time-varying nature of finite-horizon, a novel online adaptive estimator is proposed by utilizing an online approximator to relax the requirement on the system dynamics. An additional error term corresponding to the terminal constraint is defined and minimized overtime. No policy/value iteration is performed by the novel parameter update law which is updated once a sampling interval. The proposed control design yields an online and forward-in-time solution which enjoys great practical advantages. Stability of the system is demonstrated by Lyapunov analysis while simulation results verify the effectiveness of the propose approach.

1 Introduction

Optimal regulation of linear systems with quadratic cost function (LQR) has been addressed in the literature by solving the Riccati equation (RE) in a backward in time manner from the terminal weighting matrix S_{N} with known system dynamics [1][2]. To solve the LQR for infinite-horizon case, the algebraic Riccati equation (ARE) is utilized wherein the control gain becomes a constant. In contrast, for the finite-horizon case, the solution to the RE becomes inherently time-varying and the control gain matrix is also varies with time. In the presence of uncertainties in system dynamics, the solution to the RE cannot be found.

Therefore in the recent years, adaptive or neural network (NN)-based scheme to obtain the optimal control has been developed and intensively studied for both linear and nonlinear systems in the case of infinite-horizon [3][4][6][7]. However, fixed final time optimal adaptive control still remains unresolved even for the linear systems when the system dynamics are not known beforehand.

In the past literature, the author in [8] introduced the finite-horizon optimal control problem by solving the so-called generalized Hamilton-Jacobi-Bellman (GHJB) equation. The terminal constaint is incorporated in the control design and the optimal solution is obtained by iteratively solving the GHJB equation backward-in-time from the terminal time t_f via Galerkin method. Later in [9], the authors proposed a NN-based approach to solve the fixed-final time optimal control problem for general affine nonlinear continuous-time systems. The NN, with time-varying weights and state-dependent activations are ultilized to solve the time-varying HJB equation through backward integration from a known terminal NN weights.

Most recently in [10], the authors considered the discrete-time control-constrained finite-horizon optimal regulation problem by using the standard direct heuristic dynamic programming (DHDP)-based offiline training scheme. In constrast with [9], the time-dependency nature of finite-horizon is handled by using a NN with constant weights and time-varying activation functions. The terminal constriant is guranteed to be satisfied by introducing an augmented vector incorporating the terminal value of the co-state $\lambda(N)$.

The aforementioned literature [8][9][10] provided good insights into solving the finite-horizon optimal control problem. However, the solutions developed are either backward-in-time or through iteration-based offline training. Moreover, the algorithm proposed in [8][9][10] requires the knowledge of the system dynamics to update the NN weights which is a major issue.

To relax the requirement on the system dynamics and achieve optimality, approximate dynamic programming (ADP) technique [11][12][13] is developed to solve the optimal control problem by using policy iteration (PI) or value iteration (VI) in a forward-in-time manner. Most PI or VI requires at least partial dynamics of the system [12][13]. In addition, it has been further shown later in [14] that the iteration-based scheme requires significant number of iterations within a sampling interval to guarantee the parameter convergence. Inadequate number of iterations may lead to the instability of the system. In contrast, the authors in [7] developed a novel time-based algorithm to solve the Hamilton-Jacobi-Bellman (HJB) equation for optimal control of a class of general nonlinear affine discrete-time systems without using value and policy iterations. However, the work in [7] mainly addressed the infinite-horizon problem.

On the other hand, the Q-learning methodology [4][15] for discrete-time LQR is used to directly approximate the optimal control gain and such that the full system dynamics are not required. However, both PI/VI-based algorithm and Q-learning require the true system states for the feedback control loop which is another bottleneck. Later in [17][18][19], the authors transformed the system into input-output form such that only the measured data are utilized to find the control input. However, optimality was not considered in [17] and [18], and PI/VI is utilized in [19] without convergence proof.

The effort in [17][18][19] have provided a way to relax the requirement on the system states without designing an observer. However, in contrast to the optimal control gain derived based on standard Q-learning technique [4], the control gain under input-output form cannot be directly obtained from the kernel matrix. In other words, the Kalman gain cannot be expressed directly as the parameter estimation error. In addition, the feedback signal becomes a function of input-output history instead of the system states complicating the closed-loop stability for the time-based ADP scheme when an online-learning algorithm is implemented. In [19], an iterative scheme is developed while in here a time-based ADP technique is derived.

Motivated by the deficiencies mentioned above, in this paper, the ADP techique via reinforcement learning (RL) is used to solve the finite-horizon optimal regulation problem for a linear discrete-time systems with unknown system dynamics in a forward-in-time and online fashion without performing PI or VI. The system is tranformed into an input/output form such that the control policy can be obtained by using only measured input/output data [19]. The Bellman equation is investigated with an online approximator such that the requirement on the system dynamics is relaxed. An additional error term corresponding to the terminal state constraint is defined and minimized overtime so that the terminal constriant can be properly satisfied as the system evolves. Lyapunov stability is utilized to demonstrate the stability of our proposed algorithm.

The main contributions of this paper include the development of the finite-horizon adaptive optimal regulation of unknown linear discrete-time systems by using past inputs and outputs. An online adaptive estimator is introduced to effectively learn the optimal control gain, while the terminal state constraint is incorporated in the novel parameter update law to satisfy the terminal constraint. Convergence proof of the system stability becomes more involved under input-output form with the time-based ADP scheme whereas the boundedness of the regulation and parameter estimation errors is still demonstrated by using Lyapunov analysis. The proposed controller design scheme yields an online and forward-in-time algorithm with no offline training phase. The controller does not perform PI or VI while the cost function and optimal control input are updated once a sampling interval consistent with the standard adaptive control.

In the next section, the optimal regulation problem is formulated based on input/output form such that the system states are not required for the controller design.

2 Problem Formulation

In this paper, consider the following discrete-time linear system given by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k \end{aligned} \tag{1}$$

where $x_k \in \Re^n$, $u_k \in \Re$, $y_k \in \Re^p$ are system states, control inputs and system outputs, respectively. The system matrices $A \in \Re^{n\times n}$, $B \in \Re^n$ and $C \in \Re^{p\times n}$ are assumed to be *unknown*.

Before proceeding, we make the following standard assumption.

Assumption 1: The system (A,B) is controllable and (A,C) is observable.

It should be noted that for the fixed final time, the cost/value function becomes time-varying [1] and denoted as $V(x_k, \mathrm{N}-k)$, which is a function of both system states and time-to-go. The objective of the control design is to determine a feedback control policy that minimizes the cost function in the following form [16]:

$$V(x_k, \mathrm{N}-k) = y_\mathrm{N}^\mathrm{T} P_\mathrm{N} y_\mathrm{N} + \sum_{i=k}^{\mathrm{N}-1} r(y_k, u_k, k) \tag{2}$$

where $r(y_k, u_k, k) = y_k^\mathrm{T} P y_k + u_k^\mathrm{T} R u_k$ is the cost-to-go, with $P = P^\mathrm{T} \geq 0 \in \Re^{p\times p}$ and $R > 0 \in \Re$ being the weighting matrices for the outputs and inputs, and assumed to be positive semi-definite and positive definite, respectively, and P_N is a positive definite matrix conisidered to be the terminal penalty for the outputs at the final stage. Note that for finite-horizon case, the control inputs u_k are essentially time-varying, i.e., $u_k = \mu(x_k, k)$ and hence the cost-to-go becomes time-varying.

The optimal cost, based on Bellman's principle of optimality, is given as

$$\begin{aligned} &V^*(x_k, \mathrm{N}-k) \\ &= \min_{u_k} \left(r(y_k, u_k, k) + V^*(x_{k+1}, \mathrm{N}-k-1)\right) \end{aligned} \tag{3}$$

and the optimal control policy is given by

$$\begin{aligned} &u_k^* = \\ &\arg\min_{u_k} \left(r(y_k, u_k, k) + V^*(x_{k+1}, \mathrm{N}-k-1)\right) \end{aligned} \tag{4}$$

For LQR problem, the cost or value function can be represented in quadratic form in terms of the system states as

$$V(x_k, \mathrm{N}-k) = x_k^\mathrm{T} S_k x_k \tag{5}$$

where$S_k$is the solution sequence to the Riccati equation

$$\begin{aligned} S_k = A^\mathrm{T}[&S_{k+1} - S_{k+1}B(B^\mathrm{T}S_{k+1}B + R)^{-1} \\ &\times B^\mathrm{T}S_{k+1}]A + Q \end{aligned} \tag{6}$$

It is important to note that existence of unique positive RE solution, S_k, is guaranteed if the pair $(A, B, \sqrt{Q})$ is stabilizable and observable [1]. Therefore, by using the stationarity condition [1],the optimal control policy can be finally revealed to be

$$u_k^* = -(B^\mathrm{T}S_{k+1}B + R)^{-1} B^\mathrm{T}S_{k+1}A \cdot x_k \tag{7}$$

From (7), it can be seen that the optimal control can be obtained only when the system state vector x_k is measurable. In practical situation, however, the availability of the entire state vector is normally not possible. Next, the system is transformed into an input/output form [19] so that only the observed input/ouput data can be used to find the optimal control policy.

Consider the current time step k and the time interval $[k-M, k]$, where M is the number of history information and should be selected as $M \geq \nu$, with ν being the observability index. Then the system dynamics at time step k can be represented, by repeated substitution of the system dynamics, as [19]:

$$\begin{aligned} x_k &= A^M x_{k-M} + C_o \bar{u}_{k-1,k-M} \\ \bar{y}_{k-1,k-M} &= O_b x_{k-M} + T_M \bar{u}_{k-1,k-M} \end{aligned} \tag{8}$$

where $\bar{y}_{k-1,k-M} = \begin{bmatrix} y_{k-1} \\ y_{k-2} \\ \vdots \\ y_{k-M+1} \\ y_{k-M} \end{bmatrix} \in \Re^{pM}$,

$$\bar{u}_{k-1,k-M}=\begin{bmatrix}u_{k-1}\\u_{k-2}\\\vdots\\u_{k-M+1}\\u_{k-M}\end{bmatrix}\in\Re^M$$ are the measured input/output data, $C_o=[B\ AB\ \cdots\ A^{M-1}B]$ is the controllability matrix,

$$O_b=[(CA^{M-1})^{\mathrm{T}}\ (CA^{M-2})^{\mathrm{T}}\ \cdots\ (CA)^{\mathrm{T}}\ C^{\mathrm{T}}]^{\mathrm{T}}$$

is the observability matrix, and

$$T_M=\begin{bmatrix}0&CB&CAB&\cdots&CA^{M-2}B\\0&0&CB&\cdots&CA^{M-3}B\\\vdots&\vdots&\ddots&\ddots&\vdots\\0&0&\cdots&0&CB\\0&0&0&\cdots&0\end{bmatrix}$$

is the Toeplitz matrix of the system Markov parameters.

The Bellman equation can be represented in terms of the value function as

$$V^*(x_k,\mathrm{N}-k)=\ r(y_k,u_k,k)+\ V^*(x_{k+1},\mathrm{N}-k-1)\tag{9}$$

Next, define the vector in terms of input/output pair as

$$z_{k-1,k-M}=\begin{bmatrix}\bar{u}_{k-1,k-M}\\\bar{y}_{k-1,k-M}\end{bmatrix}\tag{10}$$

Then, the value function can be further written in terms of $z_{k-1,k-M}$ as

$$\begin{aligned}V(x_k,\mathrm{N}-k)&=\ x_k^{\mathrm{T}}S_kx_k\\&=z_{k-1,k-M}^{\mathrm{T}}\begin{bmatrix}L_u^{\mathrm{T}}\\L_y^{\mathrm{T}}\end{bmatrix}S_k[\ L_u\quad L_y\ ]z_{k-1,k-M}\\&=z_{k-1,k-M}^{\mathrm{T}}\begin{bmatrix}L_u^{\mathrm{T}}S_kL_u&L_u^{\mathrm{T}}S_kL_y\\L_y^{\mathrm{T}}S_kL_u&L_y^{\mathrm{T}}S_kL_y\end{bmatrix}z_{k-1,k-M}\\&\equiv z_{k-1,k-M}^{\mathrm{T}}G_{k-1}z_{k-1,k-M}\end{aligned}\tag{11}$$

where$L_y=L_0$,$L_u=C_o-L_0T_M$,$L_0=A^MO_b^+$, $O_b^+=(O_b^{\mathrm{T}}O_b)^{-1}O_b^{\mathrm{T}}$ is the left inverse of the observability matrix, and

$$G_{k-1}=\begin{bmatrix}L_u^{\mathrm{T}}S_kL_u&L_u^{\mathrm{T}}S_kL_y\\L_y^{\mathrm{T}}S_kL_u&L_y^{\mathrm{T}}S_kL_y\end{bmatrix}\in\Re^{(m+p)M\times(m+p)M}.$$

It can be seen from (11) that the value function is in the quadratic form of $z_{k-1,k-M}$, i.e., the history information of the system inputs and outputs. According to the Bellman equation (9), we have

$$\begin{aligned}&z_{k-1,k-M}^{\mathrm{T}}G_{k-1}z_{k-1,k-M}\\&=r(y_k,u_k,k)+z_{k,k-M+1}^{\mathrm{T}}G_kz_{k,k-M+1}\end{aligned}\tag{12}$$

Partition $z_{k,k-M+1}^{\mathrm{T}}G_kz_{k,k-M+1}$ yields

$$\begin{aligned}&z_{k,k-M+1}^{\mathrm{T}}G_kz_{k,k-M+1}\\&=\begin{bmatrix}\bar{u}_{k,k-M+1}\\\bar{y}_{k,k-M+1}\end{bmatrix}^{\mathrm{T}}G_k\begin{bmatrix}\bar{u}_{k,k-M+1}\\\bar{y}_{k,k-M+1}\end{bmatrix}\\&\equiv\begin{bmatrix}u_k\\\bar{u}_{k-1,k-M}\\\bar{y}_{k,k-M+1}\end{bmatrix}^{\mathrm{T}}G_{\mathrm{IO},k}\begin{bmatrix}u_k\\\bar{u}_{k-1,k-M}\\\bar{y}_{k,k-M+1}\end{bmatrix}\end{aligned}\tag{13}$$

where $G_{\mathrm{IO},k}=\begin{bmatrix}g_k^1&g_k^2&g_k^3\\(g_k^2)^{\mathrm{T}}&*&*\\(g_k^3)^{\mathrm{T}}&*&*\end{bmatrix}$.

Therefore, by taking the partial derivative with respect to u_k, we have

$$Ru_k+g_k^1u_k+g_k^1\bar{u}_{k-1,k-M}+g_k^3\bar{y}_{k,k-M+1}=0\tag{14}$$

The optimal control policy can thus be obtained by [19]

$$u_k^*=-(R+g_k^1)^{-1}(g_k^2\bar{u}_{k-1,k-M}+g_k^3\bar{y}_{k,k-M+1})\tag{15}$$

From equation (15), it is clear that the optimal control policy can be obtained by using only the information of the kernal matrix $G_{\mathrm{IO},k}$ and history information of the system inputs and outputs, thus the system dynamics A, B, C and true system state vector x_k is not required.

The main challenge is to estimate the kernel matrix parameters in order to obtain the control policy. While the estimation of the kernerl matrix parameters is presented in [19] by using the VI or PI, in this paper, they are obtained in a standard adaptive control manner due to wekanesses experienced by the VI or PI scheme.

3 Optimal Regulation Design

In this section, the fixed final time optimal regulation scheme is proposed for the discrete-time linear system with unknown system dynamics. First, an action-dependent function [4][15] is defined and estimated adaptively by using reinforcement learning, which in turn is utilized to design the controller. Next, an additional error term corresponding to the terminal constraint is defined and minimized such that the terminal constraint is properly satisfied. Finally, the stability of the closed-loop system is verified based on the Lyapunov stability theory.

3.1 Model-free Tuning with Online Adaptive Estimator

Based on Section 2, the optimal control policy can be obtained by using (15), where system dynamics are not required and only input/output data are needed. For discrete-time LQR problem, define an action-dependent function $Q(y_k, u_k, \mathrm{N}-k)$ is defined as:

$$Q^*(y_k, u_k, \mathrm{N}-k) =$$

$$= \begin{bmatrix} u_k \\ \bar{u}_{k-1,k-M} \\ \bar{y}_{k,k-M+1} \end{bmatrix}^{\mathrm{T}} G^*_{\mathrm{IO},k} \begin{bmatrix} u_k \\ \bar{u}_{k-1,k-M} \\ \bar{y}_{k,k-M+1} \end{bmatrix} \quad (16)$$

Therefore, the objective is now to approximate the optimal $Q^*(y_k, u_k, \mathrm{N}-k)$, or equivalently, the kernal matrix $G^*_{\mathrm{IO},k}$, which provides the optimal control gain, in an online manner.

Remark 1: In the standard LQR setting when the system states are available, the action-dependent function is referred to as the Q-function and the system dynamics are relaxed by directly utilizing the optimal Kalman gain provided by the Q-function [4][6][11].

Before proceeding, the following standard assumption is introduced.

Assumption 2 (*Linear in the Unknown Parameters*): The slowly time varying kernel matrix, $Q^*(y_k, u_k, \mathrm{N}-k)$, can be expressed as the linear in the unknown parameters (LIP).

By adaptive control theory [22], $Q^*(y_k, u_k, \mathrm{N}-k)$ can be represented in the vector form as

$$\begin{aligned} &Q^*(y_k, u_k, \mathrm{N}-k) \\ &= z^{\mathrm{T}}_{k,k-M+1} G^*_{\mathrm{IO},k} z_{k,k-M+1} = g^{\mathrm{T}}_{\mathrm{IO},k} \bar{z}_{k,k-M+1} \end{aligned} \quad (17)$$

where
$z_{k,k-M+1} = [\bar{u}^{\mathrm{T}}_{k,k-M+1}, \ \bar{y}^{\mathrm{T}}_{k,k-M+1}]^{\mathrm{T}} \in \Re^{(m+p)M=l}$,

$$\begin{aligned} \bar{z}_{k,k-M+1} = [&z^2_{k1,k-M+1}, \cdots, \\ &z_{k1,k-M+1} z_{kl,k-M+1}, z^2_{k2,k-M+1}, \cdots, \\ &z_{kl-1,k-M+1} z_{kl,k-M+1}, z^2_{kl,k-M+1}]^{\mathrm{T}} \end{aligned}$$

is the Kronecker product quadratic polynomial basis vector and $g_{\mathrm{IO},k} = \mathrm{vec}(G_{\mathrm{IO},k})$ with $\mathrm{vec}(\bullet)$ a vector function that acts on a $l \times l$ matrix and gives a $(l+1) \times l/2 = L$ column vector. The output of $\mathrm{vec}(G_{\mathrm{IO},k})$ is constructed by stacking the columns of the squared matrix into a one-column vector with the off-diagonal elements summed as $G^{\mathrm{IO},k}_{mn} + G^{\mathrm{IO},k}_{nm}$.

Based on Assumption 2, define $g_{\mathrm{IO},k}$ as

$$g_{\mathrm{IO},k} = \theta^{\mathrm{T}} \phi(\mathrm{N}-k) \quad (18)$$

where $\theta \in \Re^L$ is target parameter vector and $\phi(\mathrm{N}-k) \in \Re^{L \times L}$ is the time-varying basis function matrix reflecting the time-dependency nature of finite-horizon. From [1], the standard Bellman equation is given in terms of $Q^*(y_k, u_k, \mathrm{N}-k)$ as

$$\begin{aligned} Q^*(y_{k+1}, u_{k+1}, \mathrm{N}-k-1) - Q^*(y_k, u_k, \mathrm{N}-k) \\ + r(y_k, u_k, k) = 0 \end{aligned} \quad (19)$$

However, equation (19) does not hold any longer when the approximated value of $\hat{g}_{\mathrm{IO},k}$ is used. To approximate the time-varying matrix $G_{\mathrm{IO},k}$, or alternatively $g_{\mathrm{IO},k}$, define

$$\hat{g}_{\mathrm{IO},k} = \hat{\theta}^{\mathrm{T}}_k \phi(\mathrm{N}-k) \quad (20)$$

where $\hat{\theta}_k$ is the estimated value of target parameter vectorθ. Therefore, the approximation of $Q^*(y_k, u_k, \mathrm{N}-k)$ can be written as

$$\begin{aligned} \hat{Q}(y_k, u_k, \mathrm{N}-k) &= \hat{g}^{\mathrm{T}}_{\mathrm{IO},k} \bar{z}_{k,k-M+1} \\ = \hat{\theta}^{\mathrm{T}}_k \phi(\mathrm{N}-k) \bar{z}_{k,k-M+1} &= \hat{\theta}^{\mathrm{T}}_k \bar{\Phi}_k \end{aligned} \quad (21)$$

where $\bar{\Phi}_k = \phi(\mathrm{N}-k)\bar{z}_{k,k-M+1} \in \Re^L$ is a time-dependent regression function incorporating the terminal time N while satisfying $\left\|\bar{\Phi}_k\right\| = 0$ for $\bar{z}_{k,k-M+1} = 0$.

Remark 2: In the infinite-horizon case, (18) does not have the time-varying term $\phi(\mathrm{N}-k)$, since the desired value of vector g_{IO} is a constant, or time-invariant [6]. By contrast, in the finite-horizon case, the desired value of g_k is considered to be slowly time-varying. Hence the basis function should be a function of time and can take the form of product of the time-dependent basis function and the system state vector [23].

Using the approximated $\hat{Q}(y_k, u_k, \mathrm{N}-k)$, the estimated Bellman equation can be written as

$$\begin{aligned} e_{k+1} = &\hat{Q}(y_{k+1}, u_{k+1}, \mathrm{N}-k-1) \\ &- \hat{Q}(y_k, u_k, \mathrm{N}-k) + r(y_k, u_k, k) \end{aligned} \quad (22)$$

where e_{k+1} is the estimation error in the Bellman equation *along the system trajectory*. For convenience, using the delayed value from (21) and (22), we have similar to [7] as

$$\begin{aligned} e_k &= r(y_{k-1}, u_{k-1}, k-1) + \hat{\theta}_k^{\mathrm{T}} \bar{\Phi}_k - \hat{\theta}_k^{\mathrm{T}} \bar{\Phi}_{k-1} \\ &= r(y_{k-1}, u_{k-1}, k-1) + \hat{\theta}_k^{\mathrm{T}} \Delta\bar{\Phi}_{k-1} \end{aligned} \tag{23}$$

where $\Delta\bar{\Phi}_{k-1} = \bar{\Phi}_k - \bar{\Phi}_{k-1}$.

The dynamics of the Bellman estimation error can be rewritten by using history information similar to the nonlinear case [7] as

$$e_{k+1} = r(y_k, u_k, k) + \hat{\theta}_{k+1}^{\mathrm{T}} \Delta\bar{\Phi}_k \tag{24}$$

For the fixed final time case, the terminal constraint of the cost function should also be taken in account. Define the approximated value function at the terminal stage as

$$\hat{Q}_k(y_{\mathrm{N}}) = \hat{\theta}_k^{\mathrm{T}} \phi(0) \bar{z}_{\mathrm{N}} \tag{25}$$

In (25), it is important to note that the time-dependent basis function $\phi(\mathrm{N}-k)$ is taken as $\phi(0)$ since from the definition of ϕ, the time index is taken in the reverse order. Next, define the terminal constraint error vector as

$$e_{k,\mathrm{N}} = \hat{g}_{\mathrm{IO},k,\mathrm{N}} - g_{\mathrm{IO,N}} = \hat{\theta}_k^{\mathrm{T}} \phi(0) - g_{\mathrm{IO,N}} \tag{26}$$

with$g_{\mathrm{IO,N}}$ being bounded by $\|g_{\mathrm{IO,N}}\| \leq g_{\mathrm{M}}$.

Remark 3: In fixed final time problems, the error term $e_{k,\mathrm{N}}$, which indicates the difference between the approximated and true value for the terminal constraint, or "target" (in our case, $g_{\mathrm{IO,N}}$), is critical for the controller design. The terminal constraint is satisfied by minimizing $e_{k,\mathrm{N}}$ along the system evolution. Another error term e_k, which can be regarded as temporal difference error (TDE), is always needed for tuning the parameter for both finite-horizon and infinite-horizon case. For infinite-horizon case, see [6] and [7].

The objective of the finite-horizon optimal control design is to achieve optimality as well as satisfying the terminal constraint. Hence define the total error vector as

$$e_{k,\mathrm{total}} = e_k + \|e_{k,\mathrm{N}}\| \tag{27}$$

Next, to reflect the influence of the terminal constraint, the update law for tuning $\hat{\theta}_k$ can be selected as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha \Delta\bar{\Phi}_{k-1} e_k - \alpha \frac{\phi(0) e_{k,\mathrm{N}}^{\mathrm{T}}}{1 + \|\phi(0)\|^2} \tag{28}$$

where $0 < \alpha < 1$ is a design parameter. It also can be seen from (28) that the update law is essentially a gradient descent update [6]. The second and the last terms in the update law essentially use the Bellman and terminal constraint errors in order to tune the parameters.

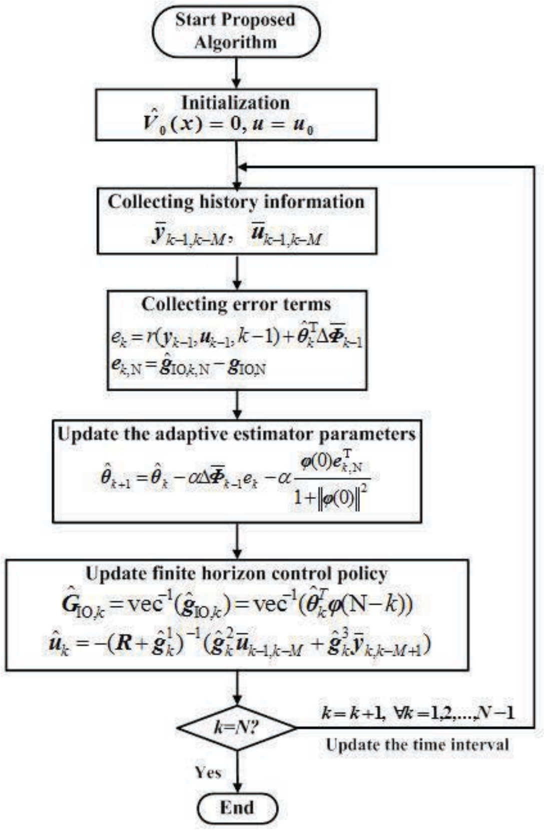

Figure 1. Flowchart of proposed algorithm

Define the parameter estimation error for $\hat{\theta}_k$ as $\tilde{\theta}_k = \theta - \hat{\theta}_k$. Recall from the standard Bellman equation (19) without approximation, we have the true utility vector as $r(y_{k-1}, u_{k-1}, k-1) = -\theta^{\mathrm{T}} \Delta\bar{\Phi}_{k-1}$, which yields the Beman equation error as

$$\begin{aligned} e_k &= r(y_{k-1}, u_{k-1}, k-1) + \hat{\theta}_k^{\mathrm{T}} \Delta\bar{\Phi}_{k-1} \\ &= -\theta^{\mathrm{T}} \Delta\bar{\Phi}_{k-1} + \hat{\theta}_k^{\mathrm{T}} \Delta\bar{\Phi}_{k-1} \\ &= -\tilde{\theta}_k^{\mathrm{T}} \Delta\bar{\Phi}_{k-1} \end{aligned} \tag{29}$$

Moreover, note that $e_{k,\mathrm{N}} = \hat{g}_{\mathrm{IO},k+1,\mathrm{N}} - g_{\mathrm{IO,N}} = \hat{\theta}_k^{\mathrm{T}} \phi(0) - \theta^{\mathrm{T}} \phi(0) = -\tilde{\theta}_k^{\mathrm{T}} \phi(0)$, and similarly $e_{k+1,\mathrm{N}} = -\tilde{\theta}_{k+1}^{\mathrm{T}} \phi(0)$, then the error dynamics for

$\tilde{\theta}_k$ can be finally revealed to be

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - \alpha\Delta\bar{\Phi}_{k-1}\Delta\bar{\Phi}_{k-1}^{\mathrm{T}}\tilde{\theta}_k - \alpha\frac{\phi(0)\phi^{\mathrm{T}}(0)}{1+\|\phi(0)\|^2}\tilde{\theta}_k \tag{30}$$

Remark 5: It is observed from the definition (17) that $Q^*(y_k, u_k, \mathrm{N}-k)$ becomes zero when $\|\bar{z}_{k,k-M+1}\| = 0$. Hence, when the system outputs, which are the inputs to the online approximator, have converged to zero, the approximator is no longer updated. It can be seen as a persistency of excitation (PE) requirement [7] for the inputs to the online approximator wherein the system states must be persistently exiting long enough for the estimator to learn $Q^*(y_k, u_k, \mathrm{N}-k)$. The PE condition requirement can be satisfied by adding exploration noise [21] to the augmented system state vector. In this paper, exploration noise is added to satisfy the PE condition.

Next, the estimation of the optimal feedback control input and the entire scheme is introduced.

3.2 Estimation of the Optimal Feedback Control and Algorithm

The optimal control can be obtained by minimizing the value function [1]. Recall from (15), the optimal control input can be obtained as

$$\hat{u}_k = -(R+\hat{g}_k^1)^{-1}(\hat{g}_k^2\bar{u}_{k-1,k-M} + \hat{g}_k^3\bar{y}_{k,k-M+1}) \tag{31}$$

From (31), the optimal control gain can be calculated based on the information of $\hat{G}_{\mathrm{IO},k}$ matrix, which is obtained by estimating $Q^*(y_k, u_k, \mathrm{N}-k)$. This relaxes the requirement of the system dynamics while the update law (28) relaxes the value and policy iterations. Here $Q^*(y_k, u_k, \mathrm{N}-k)$ and control policy are updated once a sampling interval. To complete this subsection, the flowchart of the proposed algorithm is shown in Fig. 1.

3.3 Stability Analysis

In this subsection, it will be shown that both the estimation error $\tilde{\theta}_k$and the closed-loop system are uniformly ultimately bounded (UUB). Due to the nature of time-dependency, the system becomes essentially non-autonomous in contrast with [10]. First, the boundedness of estimation error $\tilde{\theta}_k$ will be shown in Theorem 1. Before proceeding, the following definition is needed.

Definition [20]: An equilibrium point x_e is said to be uniformly ultimately bounded (UUB) if there exists a compact set $S \subset \Re^n$ so that for all $x_0 \in S$ there exists a bound $\mu > 0$, and a number $N(\mu, x_0)$ such that $\|x_k\| < \mu$ for all $k \geq N$.

Theorem 1: Let the initial conditions for $\hat{g}_{\mathrm{IO},0}$ be bounded in a set. Let $u_0(k)$ be and initial admissible control policy for the linear system (1). Let the update law for tuning $\hat{\theta}_k$ be given by (28). Then, there exists a positive constant α satisfying $0 < \alpha < 1$ such that the system is UUB. Furthermore, when $\mathrm{N} \to \infty$, the parameter estimation error $\tilde{\theta}_k$ will converge to zero asymptotically.

Proof: See Appendix.

Next, we will show the boundedness of the closed-loop system. Before establishing the theorem on system stability, the following lemma is needed which will aid the stabiltiy proof of the overall closed-loop system shown in Theorem 2.

Lemma (*Bounds on the closed-loop dynamics with optimal control signal*): Consider the linear discrete-time system defined in (1), then there exists an optimal control policy u_k^* for (1) such that the closed-loop system dynamics $Ax_k + Bu_k^*$ can be written as

$$\|Ax_k + Bu_k^*\|^2 \leq \rho\,\|x_k\|^2 \tag{32}$$

where $0 < \rho < \frac{1}{2}$ is a constant.

Theorem 2 (*Boundedness of the Closed-loop System*): Let $u_0(k)$ be an initial admissible control policy for the system such that (32) holds with someρ. Let the parameter vector of the online approximator be tuned and estimation control policy be provided by (28) and (31), respectively. Then, with positive constant α satisfying $0 < \alpha < \sqrt{\frac{1}{2}}$, there exists some $\varepsilon > 0$ depending on the initial value $B_{x,0}$ and $B_{\tilde{\theta},0}$ and the terminal stage N, such that for a fixed final time instant N, we have $\|x_k\| \leq \varepsilon(x_k, \mathrm{N})$ and $\|\tilde{\theta}_k\| \leq \varepsilon(\tilde{\theta}_k, \mathrm{N})$. Furhtermore, by geometric theory, when $\mathrm{N} \to \infty$, $\varepsilon(x_k, \mathrm{N})$ and $\varepsilon(\tilde{\theta}_k, \mathrm{N})$ will converge to zero, i.e., the system is asymptotically stable.

Proof: See Appendix.

4 Simulation Results

In this section, the proposed algorithm for finite-horizon optimal regulation problem is tested by an example which does not require the knowledge of the system dynamics and the system states. Consider the following system

$$x_{k+1} = \begin{bmatrix} 1.1 & -0.3 \\ 1 & 0 \end{bmatrix} x_k + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_k \qquad (33)$$
$$y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} x_k$$

It can be easily verify that the system is both controllable and observable, and the observability index is 2. Hence, we choose $M = 2$, and the input/output pair becomes $z_{k-1,k-2} = \begin{bmatrix} \bar{u}_{k-1,k-2} \\ \bar{y}_{k-1,k-2} \end{bmatrix} \in \Re^4$, where $\bar{u}_{k-1,k-2} = \begin{bmatrix} u_{k-1} \\ u_{k-2} \end{bmatrix} \in \Re^2, \bar{y}_{k-1,k-2} = \begin{bmatrix} y_{k-1} \\ y_{k-2} \end{bmatrix} \in \Re^2$, and therefore $\bar{z}_{k-1,k-2} \in \Re^{10}$.

The terminal constraint is given as $G_{\mathrm{IO,N}}$. According to (11),

$$G^*_{\mathrm{IO,N}} = \begin{bmatrix} 5.00 & -0.00 & 5.50 & -1.50 \\ -0.00 & 0.00 & -0.00 & 0.00 \\ 5.50 & -0.00 & 11.05 & -1.65 \\ -1.50 & 0.00 & -1.65 & 0.45 \end{bmatrix} \qquad (34)$$

The weighting matrices in the cost function (2) are selected to be $P = I$, where I denotes the identity matrix. R is selected to be 1 and 5 repectively for comparison purpose. The terminal constraint matrix is selected to be $S_{\mathrm{N}} = 5I$. The designing parameter is selected to be $\alpha = 0.001$.

The time-varying basis function matrix $\phi(\mathrm{N}-k)$ is chosen as a polynomial of time-to-go with saturation. Note that for finite time period, $\phi(\mathrm{N}-k)$ is always bounded. Saturation for $\phi(\mathrm{N}-k)$ is to ensure the magnitude of $\phi(\mathrm{N}-k)$ is within a reasonable range such that the parameter are computable. The initial value for the system states and the admissible control gain is chosen as $x_0 = [-1 \;\; 1]^{\mathrm{T}}$ and $K_0 = [0.6, \;\; -0.6]$. The initial values for $\hat{\theta}_k$ are set to be zeros.

The simulation results are shown as below.

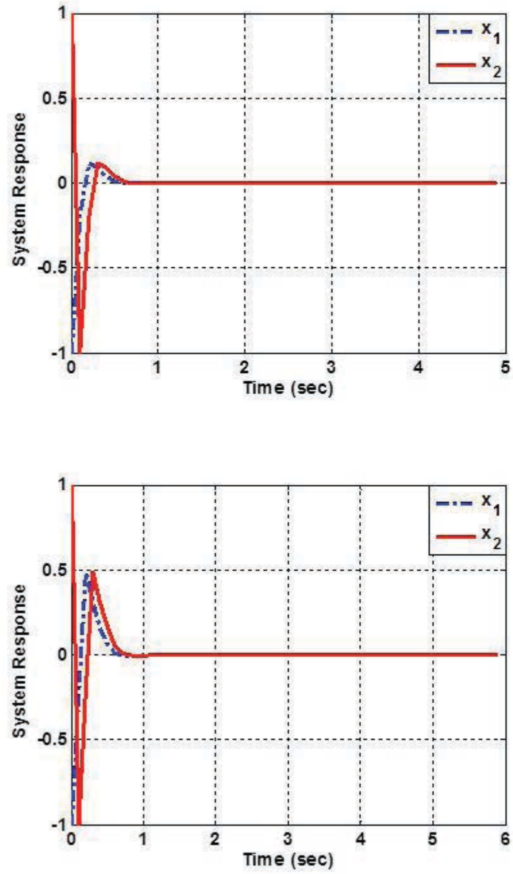

Figure 2. System response with (a): $R = 1$;(b): $R = 5$

First, the response of the system and the control input with our proposed finite-horizon optimal control design scheme are examined in Fig. 2 and 3. It can be seen that both the system states and the control input finally converge close to zero, which verifies the feasibility of our proposed design scheme.

Next, to show the feasibility of the proposed optimal regulation design, the Bellman equation error with the control weighting matrix $R = 1$ is plotted in Fig. 4. From the figure, it clearly shows that the Bellman error eventually converges close to zero. Since Bellman equation (9) only holds when the optimal solution is applied, the convergence of the Bellman equation error verifies that the optimality is indeed achieved.

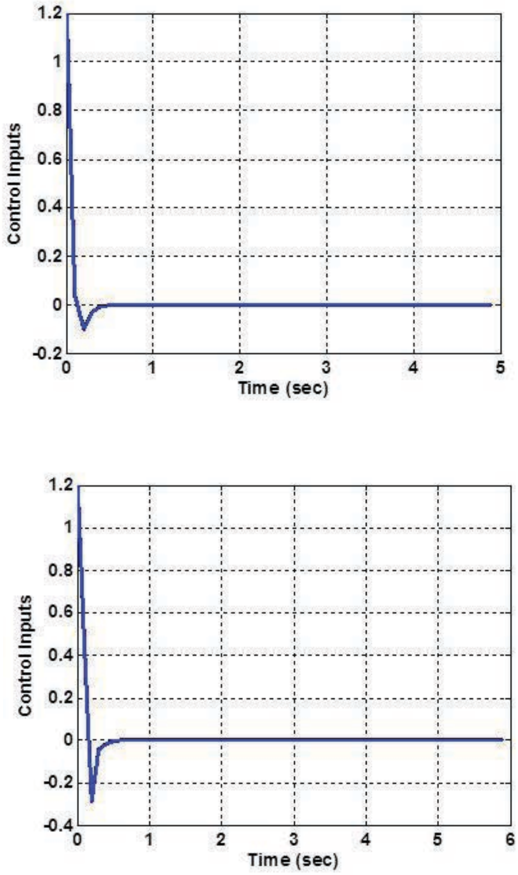

Figure 3. Control inputs with (a): $R=1$;(b): $R=5$

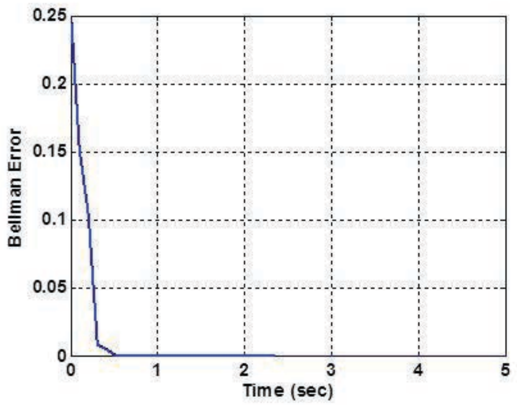

Figure 4. Bellman equation error with $R=1$

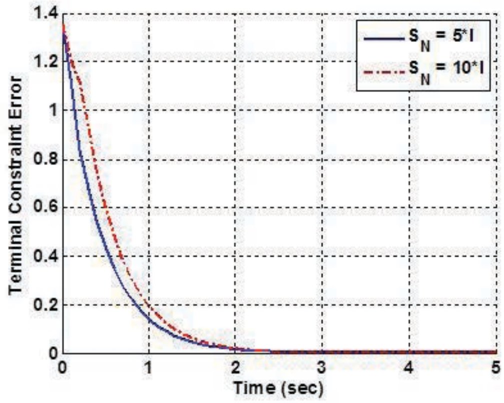

Figure 5. Terminal constraint error with different termnial weighting matrices

Finally, the convergence to the terminal constraint with the proposed finite-horizon optimal control scheme is investigated. To show the effect of terminal constraint, the terminal constraint error with $S_{\mathrm{N}}=5I$ and $S_{\mathrm{N}}=10I$ are plotted in Fig. 5 with the weighting matrices selected to be $P=I$ and $R=1$. It can be seen from the figure that ther terminal constraint errors both converge close to zero very quickly, while the larger S_{N} gives a larger initial error. This illustrates the fact that the terminal constraint is also satisfied with our proposed controller design. In fact, for the case of $S_{\mathrm{N}}=5I$, after 5 seconds of learning, the estimated terminal constraint matrix $\hat{G}_{\mathrm{IO,N}}$ is found to be

$$\hat{G}_{\mathrm{IO,N}}=\begin{bmatrix} 5.00 & 0.00 & 5.50 & -1.50 \\ 0.00 & 0.00 & -0.00 & 0.00 \\ 5.50 & -0.00 & 11.05 & -1.65 \\ -1.50 & 0.00 & -1.65 & 0.45 \end{bmatrix}$$

which exactly converge to the true terminal constraint matrix $G^{*}_{\mathrm{IO,N}}$ in (34).

5 Conclusions

In this paper, the finite-horizon optimal control of linear discrete-time system with unknown system dynamics is addressed by using the ADP technique and through the estimation of a kernel matrix. The input/output form relaxed the need for the true system states whereas the kernel matrix is not same as the standard Q-function. An online adaptive approximator is proposed to learn the kernal matrix $G_{\mathrm{IO},k}$, which in turn provides the optimal control gain and thus relaxes the system dynamics though the Kalman gain cannot be expressed directly as a function of parameter estimation error. Past history of inputs, and current output and its history are utilized to construct the kernel matrix which in turn is used to obtain the control gain matrix provided the update law is carefully selected.

An additional error corresponding to the terminal constraint, besides the Bellman error term, ensures that the terminal constraint is satisfied in the update law. Past history in the novel update law helps in overcoming the policy and/or value iterations. Lyapunov stability appears to be involved than the state feedback case though it was demonstrated. The proposed optimal control design scheme yields an online and forward-in-time scheme which offers many practical benefits.

Appendix

Proof of Theorem 1: Consider the Lyapunov candidate function as

$L(\tilde{\theta}_k) = \tilde{\theta}_k^{\mathrm{T}}\tilde{\theta}_k$ (A.1)

The first difference of $L(\tilde{\theta}_k)$ is given by

$$\begin{aligned}\Delta L(\tilde{\theta}_k) &= \tilde{\theta}_{k+1}^{\mathrm{T}}\tilde{\theta}_{k+1} - \tilde{\theta}_k^{\mathrm{T}}\tilde{\theta}_k \\ &= \left(\tilde{\theta}_k - \alpha\Delta\bar{\Phi}_{k-1}e_k - \alpha\frac{\phi(0)e_{k,\mathrm{N}}^{\mathrm{T}}}{1+\|\phi(0)\|^2}\right)^{\mathrm{T}} \times \\ &\quad \left(\tilde{\theta}_k - \alpha\Delta\bar{\Phi}_{k-1}e_k - \alpha\frac{\phi(0)e_{k,\mathrm{N}}^{\mathrm{T}}}{1+\|\phi(0)\|^2}\right)^{\mathrm{T}} - \tilde{\theta}_k^{\mathrm{T}}\tilde{\theta}_k\end{aligned}$$

(A.3)

Recall from (30) and applying Cauchy-Swartz inequality, we have

$$\begin{aligned}&\Delta L(\tilde{\theta}_k) \le \\ &-2\alpha\tilde{\theta}_k^{\mathrm{T}}\Delta\bar{\Phi}_{k-1}\Delta\bar{\Phi}_{k-1}^{\mathrm{T}}\tilde{\theta}_k - 2\alpha\frac{\tilde{\theta}_k^{\mathrm{T}}\phi(0)\phi^{\mathrm{T}}(0)\tilde{\theta}_k}{1+\|\phi(0)\|^2} \\ &+2\alpha^2\tilde{\theta}_k^{\mathrm{T}}\Delta\bar{\Phi}_{k-1}\Delta\bar{\Phi}_{k-1}^{\mathrm{T}}\Delta\bar{\Phi}_{k-1}\Delta\bar{\Phi}_{k-1}^{\mathrm{T}}\tilde{\theta}_k \\ &+2\alpha^2\frac{\tilde{\theta}_k^{\mathrm{T}}\phi(0)\phi^{\mathrm{T}}(0)\phi(0)\phi^{\mathrm{T}}(0)\tilde{\theta}_k}{1+\|\phi(0)\|^2} \\ &\le -2\alpha(1-\alpha)\tilde{\theta}_k^{\mathrm{T}}\Delta\bar{\Phi}_{k-1}\Delta\bar{\Phi}_{k-1}^{\mathrm{T}}\tilde{\theta}_k \\ &\quad -2\alpha(1-\alpha)\frac{\tilde{\theta}_k^{\mathrm{T}}\phi(0)\phi^{\mathrm{T}}(0)\tilde{\theta}_k}{1+\|\phi(0)\|^2} \\ &\le -2\alpha\left(1-\alpha\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\right)\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\left\|\tilde{\theta}_k\right\|^2 \\ &\quad -2\alpha(1-\alpha)\frac{\|\phi(0)\|^2}{1+\|\phi(0)\|^2}\left\|\tilde{\theta}_k\right\|^2 \le \\ &-2\alpha\left(1-\alpha\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\right)\left\|\tilde{\theta}_k\right\|^2\left\|\Delta\bar{\Phi}_{k-1}\right\|^2 \quad \text{(A.4)}\end{aligned}$$

Therefore, $\Delta L(\tilde{\theta}_k)$ is negative definite while $L(\tilde{\theta}_k)$ is positive definite. By standard Lyapunov stability theory [23], the parameter estimation error $\tilde{\theta}_k$ will converge to zero as $k \to \infty$.

Proof of Theorem 2: Consider the Lyapunov candidate function as

$L = L(\tilde{\theta}_k) + \frac{1}{\Pi}L(x_k)$ (A.5)

where $L(\tilde{\theta}_k)$ is defined in Theorem 1 and $L(x_k) = x_k^{\mathrm{T}}x_k$, with $\Pi = \frac{2B_{\mathrm{M}}^2L_f^2\phi_{\max}^2\alpha L_1}{\phi_{\min}^2}$. Next, we consider each term in (A.5) individually.

The first difference of $L(\tilde{\theta}_k)$ is given by Theorem 1 as

$$\begin{aligned}\Delta L(\tilde{\theta}_k) &= L(\tilde{\theta}_{k+1}) - L(\tilde{\theta}_k) \\ &\le -2\alpha\left(1-\alpha\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\right)\left\|\tilde{\theta}_k\right\|^2\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\end{aligned}$$

(A.6)

Next, considering the term $L(x_k)$ and using Cauchy-Schwartz inequality, the first difference of $L(x_k)$ is given as

$$\begin{aligned}\Delta L(x_k) &= L(x_{k+1}) - L(x_k) \\ &= x_{k+1}^{\mathrm{T}}x_{k+1} - x_k^{\mathrm{T}}x_k \\ &= \|Ax_k + Bu_k - B\tilde{u}_k\|^2 - \|x_k\|^2 \\ &\le 2\|Ax_k + Bu_k\|^2 + 2\|B\tilde{u}_k\|^2 - \|x_k\|^2\end{aligned} \quad \text{(A.7)}$$

where $\tilde{u}_k = u_k^* - \hat{u}_k$ is the difference between the optimal control input and the approximated control signal given by (31). Note that $\|\tilde{u}_k\| = \left\|\tilde{K}_k\begin{bmatrix}\bar{u}_{k-1,k-M} \\ \bar{y}_{k,k-M+1}\end{bmatrix}\right\|$, where $\tilde{K}_k$ is the optimal control gain error and found to be

$$\begin{aligned}\tilde{K}_k = [&(R+\hat{g}_k^1)^{-1}\hat{g}_k^2 - (R+g_k^1)^{-1}g_k^2, \\ &(R+\hat{g}_k^1)^{-1}\hat{g}_k^3 - (R+g_k^1)^{-1}g_k^3]\end{aligned}.$$

Note that

$$\begin{aligned}&(R+\hat{g}_k^1)^{-1}\hat{g}_k^2 - (R+g_k^1)^{-1}g_k^2 \\ &= (R+\hat{g}_k^1)^{-1}\hat{g}_k^2 - (R+g_k^1)^{-1}g_k^2 \\ &\quad + (R+\hat{g}_k^1)^{-1}g_k^2 - (R+\hat{g}_k^1)^{-1}g_k^2 \\ &= \left((R+\hat{g}_k^1)^{-1} - (R+g_k^1)^{-1}\right)g_k^2 - (R+g_k^1)^{-1}\tilde{g}_k^2\end{aligned}$$

(A.8)

Notice that

$(R+\hat{g}_k^1)^{-1} = \frac{1}{R} - \frac{\hat{g}_k^1}{R^2(1+\hat{a})}$, where $\hat{a} = \frac{\hat{g}_k^1}{R}$.

Similarly,

$(R+g_k^1)^{-1} = \frac{1}{R} - \frac{1}{R^2(1+a)}g_k^1$ with $a = \frac{g_k^1}{R}$.

Then we have

$$
\begin{aligned}
&(R+\hat{g}_k^1)^{-1}-(R+g_k^1)^{-1} \\
&=\tfrac{1}{R}-\tfrac{1}{R^2(1+\hat{a})}\hat{g}_k^1-\tfrac{1}{R}+\tfrac{1}{R^2(1+a)}g_k^1 \\
&=\tfrac{1}{R^2(1+a)}g_k^1-\tfrac{1}{R^2(1+\hat{a})}\hat{g}_k^1 \\
&=\tfrac{(1+\hat{a})g_k^1-(1+a)\hat{g}_k^1}{R^2(1+a)(1+\hat{a})} \\
&=\tfrac{\tilde{g}_k^1}{R^2(1+a)(1+\hat{a})}
\end{aligned}
$$

Therefore, (A.8) becomes

$$
\begin{aligned}
&(R+\hat{g}_k^1)^{-1}\hat{g}_k^2-(R+g_k^1)^{-1}g_k^2 \\
&=\tfrac{\tilde{g}_k^1}{R^2(1+a)(1+\hat{a})}g_k^2-(R+g_k^1)^{-1}\tilde{g}_k^2
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&(R+\hat{g}_k^1)^{-1}\hat{g}_k^3-(R+g_k^1)^{-1}g_k^3 \\
&=\tfrac{\tilde{g}_k^1}{R^2(1+a)(1+\hat{a})}g_k^3-(R+g_k^1)^{-1}\tilde{g}_k^3
\end{aligned}
$$

Hence, $\tilde{K}_k$ becomes

$$
\tilde{K}_k=\begin{bmatrix} \tfrac{\tilde{g}_k^1}{R^2(1+a)(1+\hat{a})}g_k^2-(R+g_k^1)^{-1}\tilde{g}_k^2, \\ \tfrac{\tilde{g}_k^1}{R^2(1+a)(1+\hat{a})}g_k^3-(R+g_k^1)^{-1}\tilde{g}_k^3 \end{bmatrix}
$$

Therefore, $\tilde{K}_k$ is a linear function of $(\tilde{g}_k^1,\tilde{g}_k^2,\tilde{g}_k^3)$. In other words, $\tilde{K}_k=f(\tilde{\theta}_k)$ and can be represented as

$$
\begin{aligned}
&\|\tilde{K}_k\|=\|f(g_{\mathrm{IO},k})-f(\hat{g}_{\mathrm{IO},k})\| \\
&\le L_f\|\tilde{g}_{\mathrm{IO},k}\|=L_f\|\tilde{\theta}_k\phi(\mathrm{N}-k)\|\le L_f\phi_{\max}\|\tilde{\theta}_k\|
\end{aligned}
\tag{A.9}
$$

where L_f is the Lipschitz constant.

Next, applying the Lemma and using (A.9) yields

$$
\begin{aligned}
\Delta L(\boldsymbol{x}_k)&\le 2\rho\|\boldsymbol{x}_k\|^2+2\|\boldsymbol{B}\tilde{\boldsymbol{u}}_k\|^2-\|\boldsymbol{x}_k\|^2 \\
&\le -(1-2\rho)\|\boldsymbol{x}_k\|^2+2\|\boldsymbol{B}\tilde{\boldsymbol{u}}_k\|^2 \\
&\le -(1-2\rho)\|\boldsymbol{x}_k\|^2+2B_{\mathrm{M}}^2L_f^2\phi_{\max}^2\|\tilde{\boldsymbol{\theta}}_k\|^2\left\|\begin{bmatrix}\bar{\boldsymbol{u}}_{k-1,k-M}\\ \bar{\boldsymbol{y}}_{k,k-M+1}\end{bmatrix}\right\|^2
\end{aligned}
\tag{A.10}
$$

Combining (A.6) and (A.10), the first difference$\Delta L$is given by

$$
\begin{aligned}
\Delta L&\le -2\alpha\left(1-\alpha\|\Delta\bar{\Phi}_{k-1}\|^2\right)\|\tilde{\theta}_k\|^2\|\Delta\bar{\Phi}_{k-1}\|^2 \\
&-\tfrac{1-2\rho}{\Pi}\|x_k\|^2+\tfrac{2B_{\mathrm{M}}^2L_f^2\phi_{\max}^2}{\Pi}\|\tilde{\theta}_k\|^2\left\|\begin{bmatrix}\bar{u}_{k-1,k-M}\\ \bar{y}_{k,k-M+1}\end{bmatrix}\right\|^2
\end{aligned}
\tag{A.11}
$$

Next, we will find the connection between $\|\tilde{\theta}_k\|^2\|\Delta\bar{\Phi}_{k-1}\|^2$ and $\|\tilde{\theta}_k\|^2\left\|\begin{bmatrix}\bar{u}_{k-1,k-M}\\ \bar{y}_{k,k-M+1}\end{bmatrix}\right\|^2$.

Assume that $\phi_k\bar{z}_k$ satisfies Bi-Lipschitz condition, where $\phi_k=\phi(\mathrm{N}-k)$ for simplicity. Then we have

$$
\frac{1}{L_1}\|\phi_k\bar{z}_k-\phi_{k-1}\bar{z}_{k-1}\|^2\le\|\phi_k\bar{z}_k-\phi_{k-1}\bar{z}_{k-1}\|^2
$$

where L_1 is the Lipschitz constant.

Then, (A.11) can be derived as

$$
\begin{aligned}
\Delta L&\le -2\alpha\left(1-\alpha\|\Delta\bar{\Phi}_{k-1}\|^2\right)\|\tilde{\theta}_k\|^2\|\Delta\bar{\Phi}_{k-1}\|^2 \\
&-\tfrac{1-2\rho}{\Pi}\|x_k\|^2+\tfrac{2B_{\mathrm{M}}^2L_f^2\phi_{\max}^2}{\Pi}\|\tilde{\theta}_k\|^2\left\|\begin{bmatrix}\bar{u}_{k-1,k-M}\\ \bar{y}_{k,k-M+1}\end{bmatrix}\right\|^2 \\
&\le \underbrace{-2\alpha\left(1-\alpha\|\Delta\bar{\Phi}_{k-1}\|^2\right)\tfrac{1}{L_1}\|\tilde{\theta}_k\|^2\times\begin{pmatrix}\|\phi_k\bar{z}_k-\phi_{k-1}\bar{z}_{k-1}\|^2+\\ 2\|\phi_{k-1}\bar{z}_{k-1}-\phi_{k-2}\bar{z}_{k-2}\|^2+\\ \cdots+\|\phi_{k-j}\bar{z}_{k-j}-\phi_{k-j-1}\bar{z}_{k-j-1}\|^2\end{pmatrix}}_{(a)}-\tfrac{1-2\rho}{\Pi}\|x_k\|^2 \\
&+\tfrac{2B_{\mathrm{M}}^2L_f^2\phi_{\max}^2\alpha}{\Pi}\|\tilde{\theta}_k\|^2\times\begin{pmatrix}\|\bar{z}_k\|^2+2\|\bar{z}_{k-1}\|^2+\cdots\\ +2\|\bar{z}_{k-j}\|^2+\|\bar{z}_{k-1-j}\|^2\end{pmatrix}
\end{aligned}
\tag{A.12}
$$

where j is number of columns of ϕ_k.

By the property of norm operator, we further have

$$
\begin{aligned}
&\|\phi_k\bar{z}_k-\phi_{k-1}\bar{z}_{k-1}\|^2= \\
&\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2-2(\phi_k\bar{z}_k)^{\mathrm{T}}(\phi_{k-1}\bar{z}_{k-1})
\end{aligned}
$$

Note that $\phi_k\bar{z}_k\ne\phi_{k-1}\bar{z}_{k-1}$ due to the PE condition. Then there exists $0<\delta_i<1$ for $i=0,2,\cdots,j$ such that

$$
\begin{aligned}
&2(\phi_k\bar{z}_k)^{\mathrm{T}}(\phi_{k-1}\bar{z}_{k-1})=\delta_0\left(\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2\right) \\
&2(\phi_{k-1}\bar{z}_{k-1})^{\mathrm{T}}(\phi_{k-2}\bar{z}_{k-2})=\delta_1\left(\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2\right) \\
&\qquad\vdots \\
&2(\phi_{k-j}\bar{z}_{k-j})^{\mathrm{T}}(\phi_{k-j-1}\bar{z}_{k-j-1})= \\
&\qquad\delta_j\left(\|\phi_{k-j}\bar{z}_{k-j}\|^2+\|\phi_{k-j-1}\bar{z}_{k-j-1}\|^2\right)
\end{aligned}
$$

Therefore, term (a) in (A.12) becomes

$$
\begin{aligned}
&\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2 \\
&+\|\phi_{k-2}\bar{z}_{k-2}\|^2+\cdots+\|\phi_{k-j}\bar{z}_{k-j}\|^2 \\
&+\|\phi_{k-j-1}\bar{z}_{k-j-1}\|^2-2(\phi_k\bar{z}_k)^{\mathrm{T}}(\phi_{k-1}\bar{z}_{k-1}) \\
&-2(\phi_{k-1}\bar{z}_{k-1})^{\mathrm{T}}(\phi_{k-2}\bar{z}_{k-2})-\cdots \\
&-2(\phi_{k-j}\bar{z}_{k-j})^{\mathrm{T}}(\phi_{k-j-1}\bar{z}_{k-j-1})
\end{aligned}
$$

$$
\begin{aligned}
&=\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2 \\
&-\delta_0\left(\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2\right) \\
&+\|\phi_{k-1}\bar{z}_{k-1}\|^2+\|\phi_{k-2}\bar{z}_{k-2}\|^2 \\
&-\delta_1\left(\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2\right)+\cdots
\end{aligned}
$$

$$
\begin{aligned}
&+\left\|\phi_{k-j}\bar{z}_{k-j}\right\|^2+\left\|\phi_{k-j-1}\bar{z}_{k-j-1}\right\|^2\\
&-\delta_j\left(\left\|\phi_{k-j}\bar{z}_{k-j}\right\|^2+\left\|\phi_{k-j-1}\bar{z}_{k-j-1}\right\|^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&=(1-\delta_0)\left(\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2\right)+\\
&(1-\delta_1)\left(\|\phi_k\bar{z}_k\|^2+\|\phi_{k-1}\bar{z}_{k-1}\|^2\right)+\ldots+\\
&(1-\delta_j)\left(\left\|\phi_{k-j}\bar{z}_{k-j}\right\|^2+\left\|\phi_{k-j-1}\bar{z}_{k-j-1}\right\|^2\right)
\end{aligned}
$$

Therefore, the total difference of the Lyapunov function can be represented by

$$
\begin{aligned}
\Delta L \le& -2\alpha\left(1-\alpha\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\right)\tfrac{(1-\delta_{\max})}{L_1}\left\|\tilde{\theta}_k\right\|^2\times\\
&\underbrace{\begin{pmatrix}\|\phi_k\bar{z}_k-\phi_{k-1}\bar{z}_{k-1}\|^2+\\ 2\|\phi_{k-1}\bar{z}_{k-1}-\phi_{k-2}\bar{z}_{k-2}\|^2+\\ \cdots+\left\|\phi_{k-j}\bar{z}_{k-j}-\phi_{k-j-1}\bar{z}_{k-j-1}\right\|^2\end{pmatrix}}_{(a)}-\tfrac{1-2\rho}{\Pi}\|x_k\|^2\\
&+\tfrac{2B_M^2L_f^2\phi_{\max}^2\alpha}{\Pi}\left\|\tilde{\theta}_k\right\|^2\times\begin{pmatrix}\|\bar{z}_k\|^2+2\|\bar{z}_{k-1}\|^2+\cdots\\ +2\left\|\bar{z}_{k-j}\right\|^2+\left\|\bar{z}_{k-1-j}\right\|^2\end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\le& -2\alpha\left(1-\alpha\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\right)\tfrac{(1-\delta_{\max})}{L_1}\phi_{\min}^2\left\|\tilde{\theta}_k\right\|^2\times\\
&\begin{pmatrix}\|\bar{z}_k\|^2+2\|\bar{z}_{k-1}\|^2+\cdots\\ +2\left\|\bar{z}_{k-j}\right\|^2+\left\|\bar{z}_{k-j-1}\right\|^2\end{pmatrix}-\tfrac{1-2\rho}{\Pi}\|x_k\|^2\\
&+\tfrac{2B_M^2L_f^2\phi_{\max}^2\alpha}{\Pi}\left\|\tilde{\theta}_k\right\|^2\begin{pmatrix}\|\bar{z}_k\|^2+2\|\bar{z}_{k-1}\|^2+\cdots\\ +2\left\|\bar{z}_{k-j}\right\|^2+\left\|\bar{z}_{k-1-j}\right\|^2\end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\le& -2\alpha\left(\tfrac{1}{2}-\alpha\left\|\Delta\bar{\Phi}_{k-1}\right\|^2\right)\tfrac{(1-\delta_{\max})}{L_1}\phi_{\min}^2\left\|\tilde{\theta}_k\right\|^2\times\\
&\begin{pmatrix}\|\bar{z}_k\|^2+2\|\bar{z}_{k-1}\|^2+\cdots\\ +2\left\|\bar{z}_{k-j}\right\|^2+\left\|\bar{z}_{k-j-1}\right\|^2\end{pmatrix}-\tfrac{1-2\rho}{\Pi}\|x_k\|^2
\end{aligned}
$$

By geometric sequence theory [25], within finite time, the system states x_k and parameter estimation error $\tilde{\theta}_k$ will be UUB with ultimate bounds depending on the initial conditions $B_{x,0}$ and $B_{\tilde{\theta},0}$ with $\|x_0\|^2\le B_{x,0}$ and $\left\|\tilde{\theta}_0\right\|^2\le B_{\tilde{\theta},0}$ and the terminal time NT_s, i.e.,

$$
\begin{aligned}
&\|x_k\|^2\le\left(\tfrac{1-2\rho}{\Pi}\right)^k B_{x,0}, && \forall k=0,1,\cdots,\mathrm{N}\\
&\left\|\tilde{\theta}_k\right\|^2\le\left(\tfrac{1}{2}-\alpha\Delta\Phi_b^2\right)^k B_{\tilde{\theta},0}, && \forall k=0,1,\cdots,\mathrm{N}
\end{aligned}
$$

where $\Delta\Phi_b^2$ denotes the bound of history system information.

Furthermore, when time goes to infinity, i.e., $\mathrm{N}\to\infty$, the system states x_k and parameter estimation error $\tilde{\theta}_k$ will converge to zero asymptotically.

References

[1] F. L. Lewis and V. L. Syrmos, Optimal Control, 2nd edition. New York: Wiley, 1995.

[2] D. Kirk, Optimal Control Theory: An Introduction, New Jersey, Prentice-Hall, 1970.

[3] Z. Chen and S. Jagannathan, "Generalized Hamilton-Jacobi-Bellman formulation based neural network control of affine nonlinear discrete-time systems", IEEE Trans. Neural Networks, vol. 7, pp. 90–106, 2008.

[4] S. J. Bradtke and B. E. Ydstie, Adaptive linear quadratic control using policy iteration, in Proc. Am Contr. Conf., Baltimore, pp. 3475–3479, 1994.

[5] Z. Qiming, X. Hao and S. Jagannathan, "Finite-horizon optimal control design for uncertain linear discrete-time systems", Proceedings of IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL), Singapore, 2013.

[6] X. Hao, S. Jagannathan and F. L. Lewis, "Stochastic optimal control of unknown networked control systems in the presence of random delays and packet losses," Automatica, vol. 48, pp. 1017–1030, 2012.

[7] T. Dierks and S. Jagannathan, "Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update," IEEE Trans. Neural Networks and Learning Systems, vol. 23, pp. 1118–1129, 2012.

[8] R. Beard, "Improving the closed-loop performance of nonlinear systems," Ph.D. dissertation, Rensselaer Polytechnic Institute, USA, 1995.

[9] T. Cheng, F. L. Lewis, and M. Abu-Khalaf, "A neural network solution for fixed-final-time optimal control of nonlinear systems," Automatica, vol. 43, pp. 482–490, 2007.

[10] A. Heydari and S. N. Balakrishnan, "Finite-horizon Control-Constrained Nonlinear Optimal Control Using Single Network Adaptive Critics," IEEE Trans. Neural Networks and Learning Systems, vol. 24, pp. 145–157, 2013.

[11] P. J. Werbos, "A menu of designs for reinforcement learing over time," J. Neural Network Contr., vol. 3, pp. 835–846, 1983.

[12] J. Si, A. G. Barto, W. B. Powell and D. Wunsch, Handbook of Learning and Approximate Dynamic Programming. New York: Wiley, 2004.

[13] A. Al-Tamimi and F. L. Lewis, ”Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof,” IEEE Trans. Systems, Man and Cybernetics, Part B: Cybernetics, vol. 38, pp. 943–949, 2008.

[14] H. Xu and S. Jagannathan, “Stochastic optimal controller design for uncertain nonlinear networked control system via neuro dynamic programming, IEEE Trans. Neural Netw. And Learning Syst, 24 (2013), pp. 471–484.

[15] C. Watkins, “Learning from delayed rewards,” Ph.D. dissertation, Cambridge University, England, 1989.

[16] W. Aangenent, D. Kostic, B. de Jager, R. van de Molengraft and M. Steinbuch, Data-based optimal control, in Proc. Amer. Control Conf., Portland, OR, 2005, pp. 1460–1465.

[17] R. K. Lim, M. O. Phan, and R. W. Longman, “State-space system identification with identified Hankel matrix,” Dept. Mech. Aerosp. Eng., Princeton Univ., NJ, Tech. Rep. 3045, Sep, 1998.

[18] M. O. Phan, R. K. Lim and R. W. Longman, “Unifying input-output and state-space perspectives of predictive control”, Dept. Mech. Aerosp. Eng., Princeton Univ., NJ, Tech. Rep. 3044, Sep, 1998

[19] F. L. Lewis and K. G. Vamvoudakis, “Reinforcement learning for partial observable dynamic process: adaptive dynamic programming using measured output data”, Trans. On Systems, Man, and Cybernetics – Part B. Vo. 41, pp. 14-25, 2011.

[20] S. Jagannathan, Neural Network Control of Nonlinear Discrete-Time Systems, Boca Raton, FL: CRC Press, 2006.

[21] M. Green and J. B. Moore, “Persistency of excitation in linear systems,” Syst. and Cont. Letter, vol. 7, pp. 351–360, 1986.

[22] K. S. Narendra and A. M. Annaswamy, Stable Adaptive Systems, New Jersey: Prentice-Hall, 1989.

[23] F. L. Lewis, S. Jagannathan, and A. Yesildirek, Neural Network Control of Robot Manipulators and Nonlinear Systems, New York: Taylor & Francis, 1999.

[24] H.K. Khalil, Nonlinear System, 3rd edition, Prentice-Hall, Upper Saddle River, NJ, 2002.

[25] R. W. Brochett, R. S. Millman, and H. J. Sussmann, Differential geometric control theory, Birkhauser, USA, 1983.