International Journal of Dynamics and Control
https://doi.org/10.1007/s40435-021-00836-x

Adaptive optimal control of unknown discrete-time linear systems with guaranteed prescribed degree of stability using reinforcement learning

Seyed Ehsan Razavi[1] · Mohammad Amin Moradi[2] · Saeed Shamaghdari[2] · Mohammad Bagher Menhaj[1]

Received: 23 October 2020 / Revised: 25 June 2021 / Accepted: 29 June 2021

Abstract
This paper proposes a model-free solution for solving the optimal regulation problem for a discrete-time linear time-invariant system that unlike previous methods, presents a guaranteed convergence rate of the state variables as is needed in a group of problems. Initially, the Linear Quadratic Regulation problem (LQR) with a guaranteed convergence rate of the state is formulated for a system with known dynamics and the associated Riccati equation is derived. Solving the Riccati equation and finding the state feedback gain requires full knowledge of the dynamics of the system. To overcome this problem, the Policy Iteration (PI) Reinforcement Learning (RL) algorithm is formulated to solve the LQR problem with a guaranteed convergence rate, and the optimal state feedback gain is derived without having any knowledge about the dynamics of the system and only through the measurement of the states of the system. Eventually, the validity of the results is shown through simulation.

Keywords Model-free optimal control · Reinforcement learning · Policy iteration · Convergence rate · Degree of stability

1 Introduction

Optimal control theory is a mature mathematical field that finds optimal control policies for dynamic systems by the means of optimization of the cost function that is defined by the user. Primarily, there are two general methods for solving optimal control problems, the Pontryagin method, and the dynamic programming method, that respectively state necessary and sufficient conditions for optimality. In the dynamic programming methods, usually, the problem is solved from end to beginning and in the case of discrete-time systems it yields the Bellman equation and in the case of continuous systems, it leads to the HJB equation and eventually, the optimal control signal and the optimal value function is derived from its solution.

The solution of the classic optimal control, are offline, and require full knowledge of the dynamics of the system, moreover, for a nonlinear system solving HJB and Bellman equations and finding the optimal control signal and the optimal value function in the general form, proves to be difficult; And only in certain cases such as linear systems and quadratic cost functions leads to Riccati equations. Given the fact that ARE is nonlinear, it is often hard to solve it through a direct approach, particularly in the case of high dimensional matrices.

To solve this problem, several great algorithms have been proposed, such as the well-known Kleinman algorithm [1]. Based on the Kleinman algorithm, the solution of the Riccati equation can be numerically approximated through iterative solving of the Lyapunov equations. Notwithstanding, the Kleinman algorithm requires full information about the system. The main approach, when full information of the system is not available, is the adaptive optimal control design, such that initially parameters of the system are identified then the Riccati equation is solved. These algorithms,

✉ Saeed Shamaghdari
shamaghdari@iust.ac.ir

Seyed Ehsan Razavi
ser@aut.ac.ir

Mohammad Amin Moradi
m_moradi97@elec.iust.ac.ir

Mohammad Bagher Menhaj
menhaj@aut.ac.ir

[1] School of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran

[2] School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran

Published online: 24 August 2021

however, respond slowly to the changes in the parameters of the system [2].

By the means of inspiration from the learning behavior in biological systems, Reinforcement learning theories and its more recent formulation known as "Adaptive/Approximative Dynamic Programming" or “ADP”, for solving optimal control problems for systems with uncertainty are generating considerable interest in recent years [3–5].

Reinforcement learning originates from computer science and is a branch of machine learning methods, that makes an agent learn in an interactive environment through trial and error and using the feedback of its actions and its experiences. Even though both supervised learning and Reinforcement learning use mapping between input and output, Reinforcement learning, unlike supervised learning, rewards, and punishments are used as signals for positive and negative behaviors [6]. For the recent past years, Reinforcement learning methods have been used in control theory in which the performance of a dynamic system is measured by a scalar function that is indicative of expended costs by the system over time. [7, 15–22].

For the past decades, the RL and the ADP has been one of the powerful methods to design a control protocol based on data. From the perspective of a control engineer, RL and ADP may be seen as a bridge between classic optimal control and adaptive control methods. The ADP method, unlike classic optimal Control methods, yields a real-time approximation of the solution of the HJB differential equation. On the other hand, despite adaptive control in which no cost function is minimized, the ADP algorithms perform optimally from this point of view [6, 8, 9].

Generally, there are two fundamental parts in RL algorithms: evaluation of the policy and improvement of the policy. In the policy evaluation, the cost function (the value function) regarding the current policy is calculated. In the policy improvement stage, the resulting cost function value is evaluated and updates the current policy. It is worth mentioning that the policy has come to be used to refer to the control signal.

The two main classes of RL algorithms used for these two steps are Policy Iteration (PI) and Value Iteration (VI). The VI and PI algorithms perform policy evaluation and policy improvement iteratively until finding an optimal solution. PI Methods are initialized by a stabilizer control policy (control signal) [23, 24]. And solve a series of Bellman equations to find an optimal control policy. Compared to the PI method, the VI does not require a stable control policy. Whereas the majority of RL based control algorithms are PI, some VI algorithms have been employed to find optimal control solutions.

In recent years, the RL algorithms in control have been used in solving problems. Satisfying a criterion of optimality and the lack of requirement of full knowledge of the dynamic of the system, have gained this method a growing interest in solving the control problem of dynamic systems. The optimal regulation problem aims to design an optimal controller to assure that states or outputs of the system go to the origin or near the origin [4, 9, 10]. While in optimal tracking control problems, it is desired that the controller makes states or outputs of the system follow the desired trajectory [11, 12, 25–30]. The purpose of an optimal synchronization problem in multi-agent systems is the design of distributed control protocols based on local information available in agents so that agents achieve team goals [13, 14, 31–36].

There are some studies related to optimal control with guaranteed convergence rate and these studies are based on the system models, such as minimal energy control with guaranteed convergence rate and Linear-quadratic optimal observers with guaranteed convergence rate and so on. Developing off-policy RL algorithms for discrete-time systems is not straightforward because of the appearance of both system matrix A and control matrix B in the policy update equation. In [37], an off-policy RL algorithm is presented to solve the H∞ control of linear discrete-time systems. This paper aims at generalizing the main result of [37] to computational adaptive optimal control with guaranteed convergence rate of the discrete-time linear systems with completely unknown dynamics. With the help of reinforcement learning.

2 Problem description

Consider the discrete-time linear system in Eq. (1):

$$x(k+1) = Ax(k) + Bu(k) \tag{1}$$

where $x(k) \in R^n$ *and* $u(k) \in R^m$ represent the state of the system and the control input, respectively. This paper aims to solve the optimal regulation problem with a guaranteed convergence rate. In other words, the state feedback gain (F) must be found such that in addition to the optimality, it

Table 1 Online PI algorithm to solve model-based LQR problem

Algorithm 1: Online PI algorithm for solving the LQR problem with a guaranteed prescribed degree of stability

1: **procedure**

2: Given the stabilizing feedback gain α and applying it to the system

3: Solve the Bellman Eq. (13) for the value P^{i+1}

$$x^T(k)P^{i+1}x(k) - \frac{1}{\alpha^2}x^T(k+1)P^{i+1}x(k+1) = x^T(k)Qx(k) + u^{iT}(k)Ru^i(k) \tag{13}$$

4: Update the control signal using:

$$u^{i+1}(k) = -F^{i+1}x(k) = -\left(\alpha^2 R + B^T P^{i+1} B\right)^{-1} B^T P^{i+1} Ax(k) \tag{14}$$

5: If $\left|F^{i+1} - F^i\right| < \epsilon$ then end of the algorithm, else i = i + 1 and go to step 1

6: **end procedure**

guarantees that the eigenvalues of the closed-loop system are placed inside the radius α of the unit circle, without having any knowledge about matrices of the dynamics of the system.

$$u(k) = Fx(k) \tag{2}$$

The goal of optimal regulation is to design an optimal control input to stabilize the system in Eq. (1) while minimizing a predefined cost function. Such energy-related cost functional can be defined as:

$$J(x(k)) = \sum_{k=0}^{\infty} \frac{1}{\alpha^{2k}} \left(x^T(k)Qx(k) + u^T(k)Ru(k)\right) \tag{3}$$

where $Q \geq 0$ and R = R^T > 0. The following lemma is used in order to obtain the optimal state feedback gain.

Lemma 1 *Consider the discrete-time linear system with known dynamics in Eq.* (1) *and the cost function in Eq.* (3) $(0 \leq \alpha \leq 1)$. *In this case, the state feedback gain F satisfying formula in Eq.* (4), *in addition to minimizing the cost function in Eq.* (3), *guarantees that the states of the system converge to zero with the rate of* α^k, *in other words,* $\lim_{k\to\infty} \alpha^k x(k) = 0$.

$$F = -\left(\alpha^2 R + B^T PB\right)^{-1} B^T PA \tag{4}$$

where P is derived from solving the following Riccati equation:

$$A^T PA - \alpha^2 P + A^T PB\left(\alpha^2 R + B^T PB\right)^{-1} B^T PA + \alpha^2 Q = 0 \tag{5}$$

Proof Followed by a change of variables $\bar{u(k)} = \frac{u(k)}{\alpha^k}$ and $\bar{x}(k) = \frac{x(k)}{\alpha^k}$ we can rewrite Eqs. (1) and (3) as:

$$\alpha^{k+1}\bar{x}(k+1) = A\alpha^k \bar{x}(k) + B\alpha^k \bar{u}(k) \tag{6}$$

If we divide both sides by α^{k+1}:

$$\bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}\bar{u}(k) \tag{7}$$

where $\bar{A} = \frac{A}{\alpha}, \bar{B} = \frac{B}{\alpha}$.

Using the optimal control theory, the solution of the LQR problem for the system in Eq. (7) and the cost function in Eq. (3) is described as follows:

$$\bar{u}^*(k) = F\bar{x}(k) \tag{8}$$

where:

$$F = -\left(R + \bar{B}^T P\bar{B}\right)^{-1} \bar{B}^T P\bar{A} \tag{9}$$

And P is the solution to the following Riccati Equation:

$$\bar{A}^T P\bar{A} - P + \bar{A}^T P\bar{B}\left(R + \bar{B}^T P\bar{B}\right)^{-1} \bar{B}^T P\bar{A} + Q = 0 \tag{10}$$

We will have: $\lim_{k\to\infty} \bar{x}(k) = 0$. Since $\bar{x}(k) = \alpha^{-k}x(k)$ we have: $\lim_{k\to\infty} \alpha^{-k}x(k) = 0$. Since $0 \leq \alpha \leq 1$, the term α^{-k} is increasing exponentially, on the other hand, we have $\alpha^{-k}x(k) \to 0$ hence $x(k)$ goes to zero with the minimum rate of α^{-k}.

If we replace $\bar{A} = \frac{A}{\alpha}$ and $\bar{B} = \frac{B}{\alpha}$ in Eqs. (9) and (10):

$$F = -\left(\alpha^2 R + B^T PB\right)^{-1} B^T PA \tag{11}$$

$$A^T PA - \alpha^2 P + A^T PB\left(\alpha^2 R + B^T PB\right)^{-1} B^T PA + \alpha^2 Q = 0 \,\square \tag{12}$$

Remark 1 The Eq. (5) is non-linear with respect to P, therefore solving it directly proves to be difficult particularly in the case of high-dimensional matrices. In order to solve this problem, we use the PI algorithm to solve the regulation problem with a guaranteed convergence rate deriving through an iterative process.

3 Online PI Algorithm for Solving the LQR Problem with a Guaranteed Prescribed Degree of Stability

The PI algorithm for solving the model-based regulation problem with guaranteed convergence rate is summarized in Table 1 (i and j indicate the iteration number and time sample respectively.)

Remark 2 In order to implement algorithm 1, the Bellman Eq. (13) can be rewritten as Eq. (15) using $a^T wb = (b^T \otimes a^T)vec(w)$.

$$\left(x_k^T \otimes x_k^T - \frac{1}{\alpha^2} x_{k+1}^T \otimes x_{k+1}^T\right) vec\left(P^{i+1}\right) = x_k^T Q x_k + u_k^{iT} R u_k^i \tag{15}$$

4 Off-policy RL Ö the LQR problem with a guaranteed prescribed degree of stability

In order to derive the off-policy RL algorithm, we rewrite the system in Eq. (1) as follows:

$$x(k+1) = Ax(k) + Bu(k) \pm BF^i x(k) = A_k x(k) + B(F^i x(k) + u(k)) \tag{16}$$

where $A_k = A - BF^i$.

We refer to $u^i(k) = -F^i x(k)$ as a target policy or an estimated policy in Eq. (16), (the policies that are trained and updated), while $u(k)$ is the policy that is applied to the system and upon which the data is generated which we refer to

as the behavior policy. Considering Bellman Eq. (13) and replacing Eq. (1) we have:

$$x_k^T P^{i+1} x_k - \frac{1}{\alpha^2} x_k^T \left(A - BF^i\right)^T P^{i+1} \left(A - BF^i\right) x_k = x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \tag{17}$$

After expanding the left side of Eq. (17) we have:

$$\begin{aligned} & x_k^T P^{i+1} x_k - \frac{1}{\alpha}^2 x_k^T A^T P^{i+1} A x_k + \frac{2}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} A x_k \\ & - \frac{1}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} B F^i x_k = x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \end{aligned} \tag{18}$$

We require the precedent relation as follows:

$$\begin{aligned} & x_k^T P^{i+1} x_k - \frac{1}{\alpha^2} x_k^T A^T P^{i+1} A x_k \\ & = -\frac{2}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} A x_k \\ & + \frac{1}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} B F^i x_k \\ & + x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \end{aligned} \tag{19}$$

$$\begin{aligned} & x_k^T P^{i+1} x_k - \frac{1}{\alpha^2} x_k^T A^T P^{i+1} A x_k = -\frac{1}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} A x_k \\ & - \frac{1}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} (A - BF^i) x_k + x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \end{aligned} \tag{20}$$

We subtract $\frac{1}{\alpha^2}(u_k^T B^T P^{i+1} B u_k + 2 u_k B^T P^{i+1} A x_k)$ from both sides of Eq. (20):

$$\begin{aligned} & x_k^T P^{i+1} x_k - \frac{1}{\alpha^2} x_k^T A^T P^{i+1} A x_k \\ & - \frac{1}{\alpha^2} (u_k^T B^T P^{i+1} B u_k + 2 u_k B^T P^{i+1} A x_k) \\ & = -\frac{1}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} A x_k \\ & - \frac{1}{\alpha^2} x_k^T F^{iT} B^T P^{i+1} (A - BF^i) x_k + x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \\ & - \frac{1}{\alpha^2} (u_k^T B^T P^{i+1} B u_k + 2 u_k B^T P^{i+1} A x_k) \end{aligned} \tag{21}$$

After performing a series of mathematical operations:

$$\begin{aligned} & x_k^T P^{i+1} x_k - \frac{1}{\alpha^2} \left(A x_k + B u_k\right)^T P^{i+1} (A x_k + B u_k) \\ & = -\frac{1}{\alpha^2} \left(u_k + F^i x_k\right)^T B^T P^{i+1} (A - BF^i) x_k \\ & - \frac{1}{\alpha^2} \left(u_k + F^i x_k\right)^T B^T P^{i+1} (A x_k + B u_k) + x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \end{aligned} \tag{22}$$

Replacing $x_{k+1} = A x_k + B u_k$ and $A_k = A - BF^i$ into Eq. (22) we will have:

$$\begin{aligned} & x_k^T P^{i+1} x_k - \frac{1}{\alpha^2} x_{k+1}^T P^{i+1} x_{k+1} \\ & = -\frac{1}{\alpha^2} \left(u_k + F^i x_k\right)^T B^T P^{i+1} A_k x_k \\ & - \frac{1}{\alpha^2} \left(u_k + F^i x_k\right)^T B^T P^{i+1} x_{k+1} + x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \end{aligned} \tag{23}$$

The off-policy algorithm for solving the model-based regulation problem with guaranteed convergence rate is summarized in Table 2 (i and j indicate the iteration number and time sample respectively.)

Remark 3 The off-policy RL Algorithm 2 converges to the optimal control solution given by Eq. (11) where the matrix P satisfies the GARE (12). The convergence proof is similar to Theorem2 in [37].

In order to implement the model-free off-policy method similar to the on-policy method, the Bellman Eq. (24) is convertible to a linear equation. The Bellman Eq. (24) can be rewritten as follows using Eq. (16):

$$\begin{aligned} & \left(x_k^T \otimes x_k^T - \frac{1}{\alpha^2} x_{k+1}^T \otimes x_{k+1}^T\right) vec\left(P^{i+1}\right) \\ & + \frac{2}{\alpha^2} \left(x_k^T \otimes \left(u_k + F^i x_k\right)^T\right) vec\left(B^T P^{i+1} A\right) \\ & - \frac{1}{\alpha^2} \left(\left(F^i x_k - u_k\right) \otimes \left(u_k + F^i x_k\right)^T\right) vec\left(B^T P^{i+1} B\right) \\ & = x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \end{aligned} \tag{25}$$

Through having the data of the system over different time steps, the Eq. (25) can be written as follows:

$$\phi^i \begin{bmatrix} vec\left(P^{i+1}\right) \\ vec\left(B^T P^{i+1} A\right) \\ vec\left(B^T P^{i+1} B\right) \end{bmatrix} = \xi^i \tag{26}$$

where:

Table 2 Off-Policy algorithm to solve model-based LQR problem

Algorithm 2: Off-Policy RL algorithm for solving the LQR problem with a guaranteed prescribed degree of stability

1: **procedure**

2: Given the stabilizing feedback gain α and applying it to the system

3: Solve the Bellman Eq. (24) for the value P^{i+1} and F^{i+1} simultaneously

$$x_k^T P^{i+1} x_k - \frac{1}{\alpha^2} x_{k+1}^T P^{i+1} x_{k+1} = -\frac{1}{\alpha^2}\left(u_k + F^i x_k\right)^T B^T P^{i+1} A_k x_k - \frac{1}{\alpha^2}\left(u_k + F^i x_k\right)^T B^T P^{i+1} x_{k+1} + x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \quad (24)$$

5: If $\left|F^{i+1} - F^i\right| < \epsilon$ then end of the algorithm, else i = i + 1 and go to step 1

6: **end procedure**

$$\phi^i = \begin{bmatrix} \left(x_k^T \otimes x_k^T - \frac{1}{\alpha^2} x_{k+1}^T \otimes x_{k+1}^T\right) & \frac{2}{\alpha^2}\left(x_k^T \otimes \left(u_k + F^i x_k\right)^T\right) & -\frac{1}{\alpha^2}\left(\left(F^i x_k - u_k\right) \otimes \left(u_k + F^i x_k\right)^T\right) \\ \left(x_{k+1}^T \otimes x_{k+1}^T - \frac{1}{\alpha^2} x_{k+2}^T \otimes x_{k+2}^T\right) & \frac{2}{\alpha^2}\left(x_{k+1}^T \otimes \left(u_{k+1} + F^i x_{k+1}\right)^T\right) & -\frac{1}{\alpha^2}\left(\left(F^i x_{k+1} - u_{k+1}\right) \otimes \left(u_{k+1} + F^i x_{k+1}\right)^T\right) \\ \vdots & \vdots & \vdots \\ \left(x_{k+s-1}^T \otimes x_{k+s-1}^T - \frac{1}{\alpha^2} x_{k+s}^T \otimes x_{k+s}^T\right) & \frac{2}{\alpha^2}\left(x_{k+s-1}^T \otimes \left(u_{k+s-1} + F^i x_{k+s-1}\right)^T\right) & -\frac{1}{\alpha^2}\left(\left(F^i x_{k+s-1} - u_{k+s-1}\right) \otimes \left(u_{k+s-1} + F^i x_{k+s-1}\right)^T\right) \end{bmatrix} \quad (27)$$

$$\xi^i = \begin{bmatrix} x_k^T Q x_k + x_k^T F^{iT} R F^i x_k \\ x_{k+1}^T Q x_{k+1} + x_{k+1}^T F^{iT} R F^i x_{k+1} \\ \vdots \\ x_{k+s-1}^T Q x_{k+s-1} + x_{k+s-1}^T F^{iT} R F^i x_{k+s-1} \end{bmatrix} \quad (28)$$

In the precedent relations, matrices ϕ^i and ξ^i are given. Hence, we can solve the Eq. (25) using the sum of squared errors method and come up with a unique solution for $B^T P^{i+1} B$, $B^T P^{i+1} A$, and P^{i+1}.

Assumption 1 For each $i = 0, 1, \ldots$ there is an integer value sufficiently large s such that ϕ^i is full rank.

Remark 4 In order for ϕ^i to be full rank, in addition to sufficiently large s data measurement, appropriate choice of probe noise plays an essential role.

The off-policy algorithm for solving the model-free regulation problem with guaranteed convergence rate is summarized in Table 3 (i and j indicate the iteration number and time sample respectively.)

5 Simulation results

In this section, in order to simulate the proposed method, the discrete-time plant model of an aircraft dynamics [37], is used as Eq. (30) and it is assumed that matrices A and B are not given.

$$x_{k+1} = A x_k + B u_k \quad (30)$$

The following matrices are assumed for the simulation:

$$A = \begin{bmatrix} 0.9064 & 0.0816 & -0.0005 \\ 0.0743 & 0.9012 & -0.0007 \\ 0 & 0 & 0.1326 \end{bmatrix}$$

$$B = \begin{bmatrix} -0.0015 \\ -0.0096 \\ 0.8673 \end{bmatrix}$$

and the simulation time step is 1 ms.

Table 3 Off-policy algorithm to solve model-free LQR problem

Algorithm 3: The off-policy algorithm for solving the model-free regulation problem with guaranteed convergence rate

1: **procedure**

2: Given the stabilizing feedback gain α and applying it to the system

3: Solve the Bellman Eq. (25) for the value $B^T P^{i+1} B$, $B^T P^{i+1} A$, and P^{i+1}

4: Update the control signal using:

$$F^{i+1} = \left(\alpha^2 R + B^T P^{i+1} B\right)^{-1} B^T P^{i+1} A \quad (29)$$

5: If $\left|F^{i+1} - F^i\right| < \epsilon$ then end of the algorithm, else i = i + 1 and go to step 1

6: **end procedure**

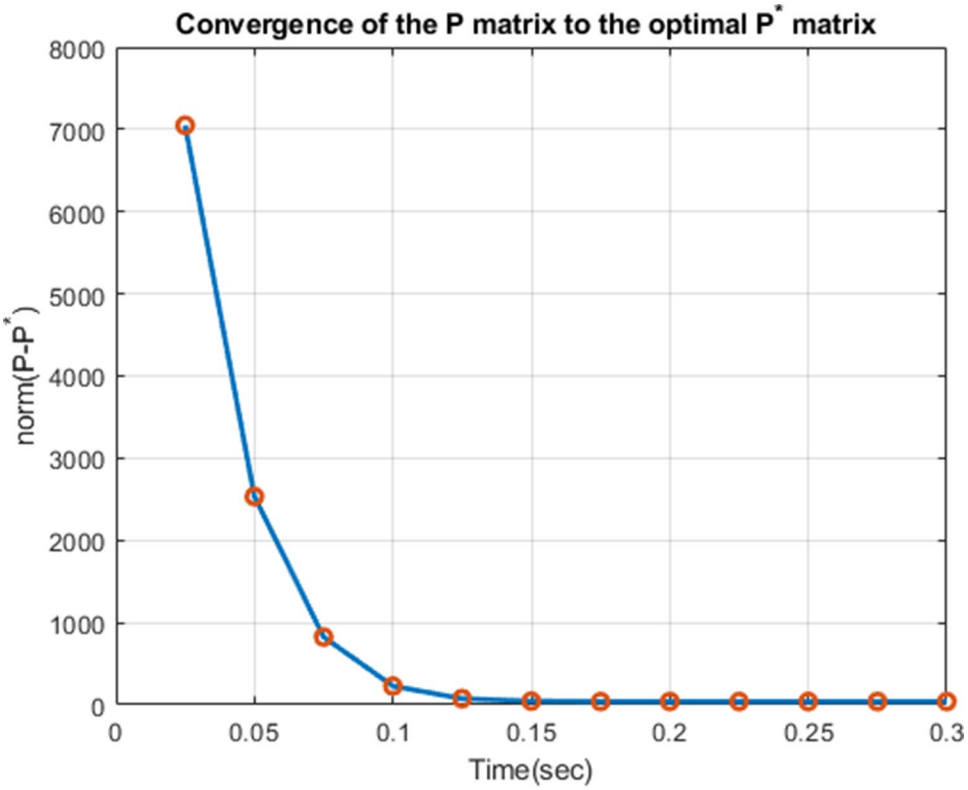

Fig. 1 P matrix convergence to the optimal P^* in Off-Policy RL algorithm

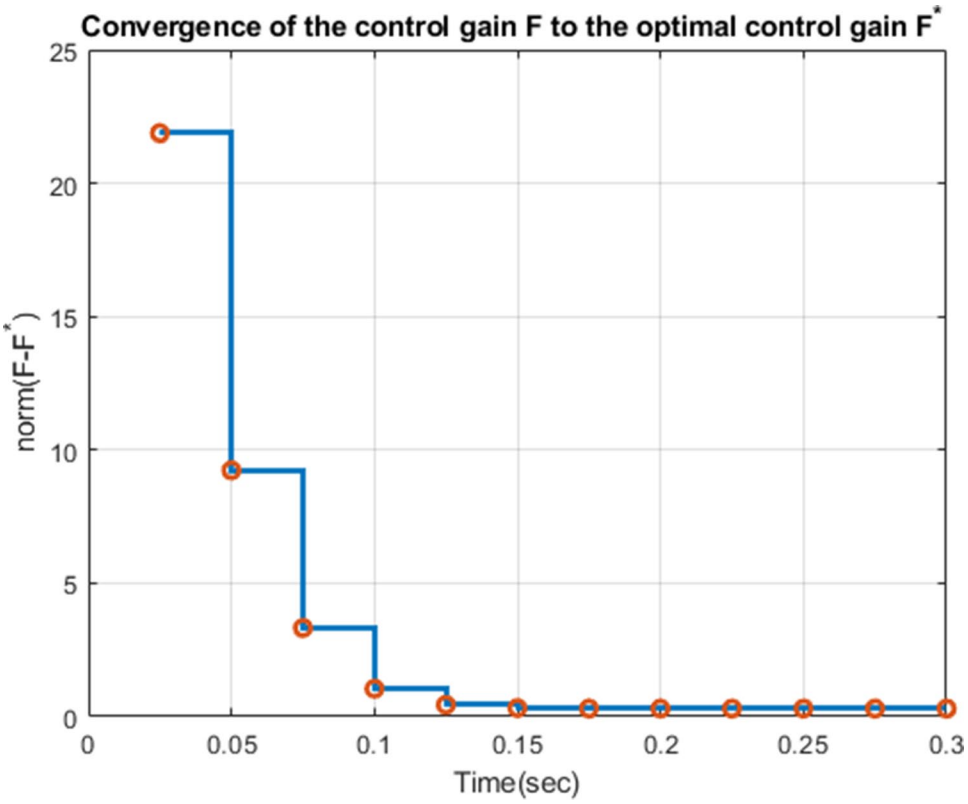

Fig. 2 Control gain convergence to the optimal value in Off-Policy RL algorithm

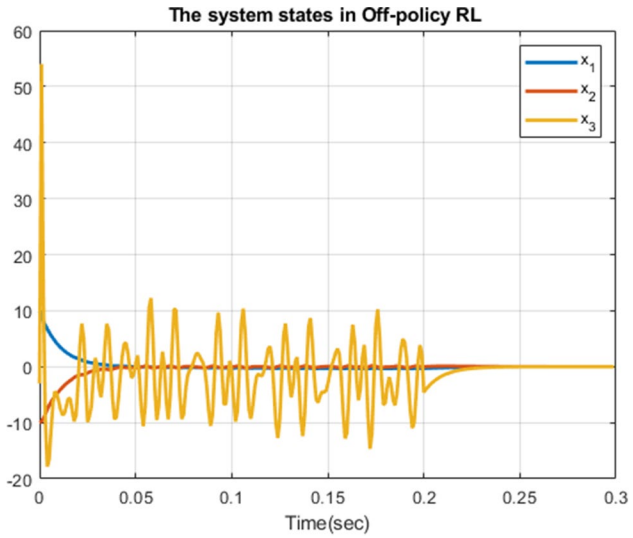

Fig. 3 System's state variables in Off-Policy RL algorithm

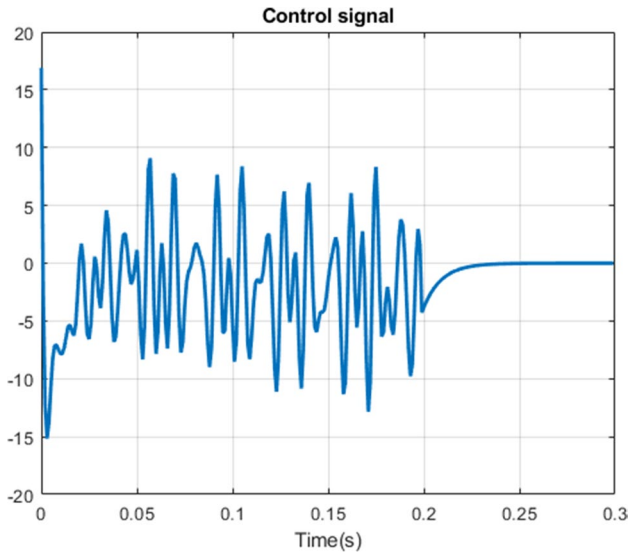

Fig. 4 Control Signal for the regulation problem in Off-Policy RL algorithm

The purpose of the design is to find a state feedback control law for an unknown system in Eq. (30) such that the following cost function is minimized:

$$J(x(k)) = \sum_{k=0}^{\infty} \frac{1}{\alpha^{2k}} (x^T(k)Qx(k) + u^T(k)Ru(k)) \tag{30}$$

where $\alpha = 0.9$, $Q = \boldsymbol{diag}(1, 1, 1)$, and $R = 1$.

For comparison, if the matrices of the system are given, P^* is the solution of the Riccati Eq. (5) and F^* is derived from the Eq. (4) as follows:

$$P^* = \begin{bmatrix} 1.944e3 & 1.989e3 & -2.439 \\ 1.989e3 & 2.016e3 & -2.466 \\ 0 - 2.439 & -2.466 & 1.003 \end{bmatrix}$$

$$F^* = \begin{bmatrix} -12.77 \\ -12.93 \\ 0.016 \end{bmatrix}^T$$

The simulation results using the proposed method in Algorithm 3 are as follows. Initially, the applied control signal to the system is with the probe noise in order to guarantee ϕ^i to be full rank, then after 200 initial step time (equal to 0.2 s) the probe noise will be removed. It is obvious from Figs. 1 and 2, that matrix P and the state feedback gain F derived from Algorithm 3 converge to their optimal value after about 6 iterations (each iteration is assumed to be equal to 25 time steps). Figure 3 shows the states of the system go to zero, as well. The control signal u has been depicted in Fig. 4, where the probe noise can be seen clearly in the beginning.

In order to inspect the convergence speed, the eigenvalues of the closed-loop system have been compared in two cases $\alpha = 1$ and $\alpha = 0.8$.

All of the eigenvalues of the closed-loop system with the optimal feedback gain F derived for $\alpha = 0.8$ are placed inside the unit circle and are closer than 0.96 from the center of the circle, while some of the eigenvalues of the closed-loop system for $\alpha = 1$ are placed outside the circle with the radius of 0.96. Therefore, the convergence rate of the states of the system is higher for $\alpha = 0.8$.

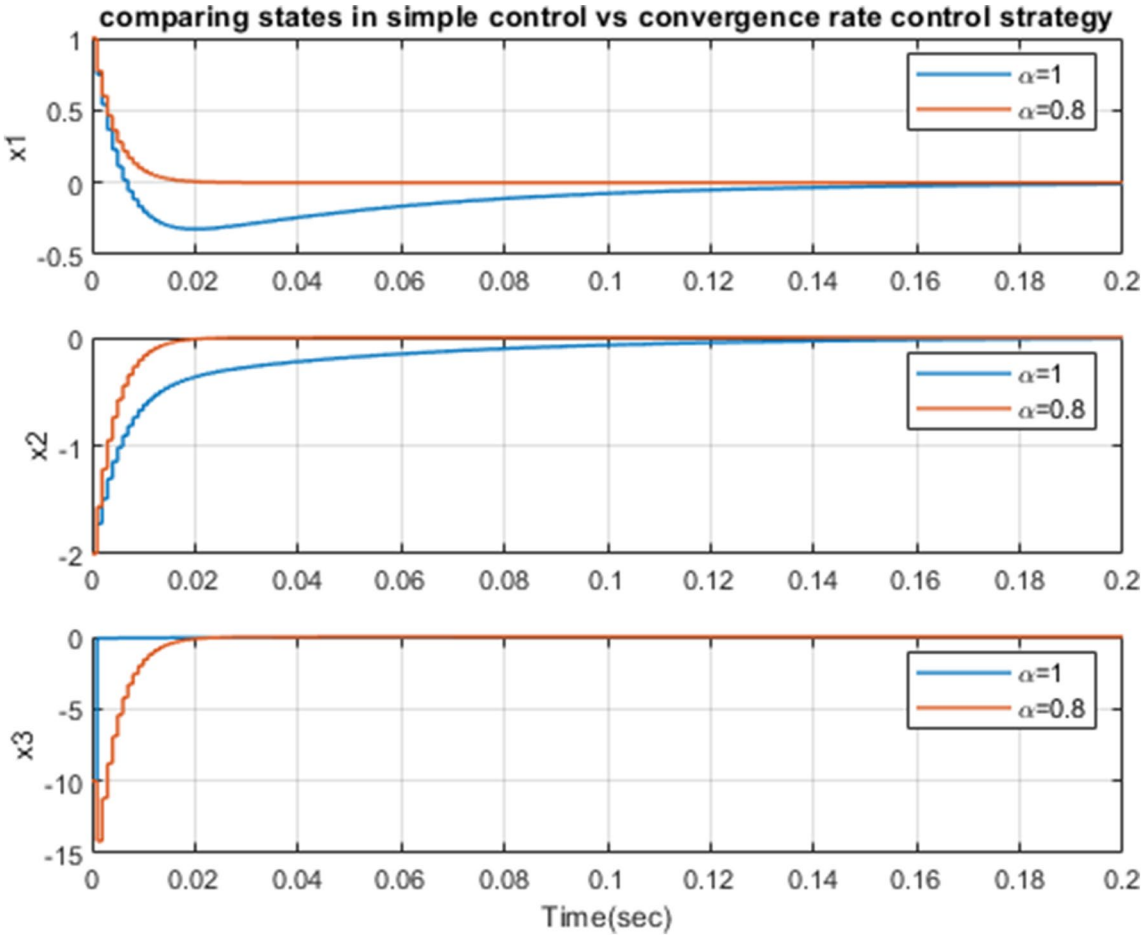

Fig. 5 Comparing states in simple control against convergence rate control strategy

The system states trajectories of the controlled system according to off-policy model-free algorithm presented in Table 3 are depicted in Fig. 5. It can be seen that the degree of stability for a smaller α causes a better convergence rate to the equilibrium point. In optimal control method the settling time is about 0.12 s, in convergence rate control strategy, states converge to zero with settling time less than 0.02 s.

6 Conclusion

In this paper, a model-free solution for solving the optimal regulation problem for discrete-time linear time-invariant system was proposed. Unlike previous methods it also guarantees the convergence rate of the states of the system. In order to achieve this goal, the new cost function and related Riccati equation was derived. The iterative solution of the Riccati Equation and finding the optimal state feedback gain without requiring any knowledge about the dynamics of the system was derived. An example of the optimal problem for an F-16 aircraft with completely unknown dynamics was given. It was seen that the system states trajectories of the controlled system according to off-policy model-free algorithm, converge to the equilibrium point with prescribed convergence rate. The proposed method in this paper can also be used for reaching a consensus algorithm in multi-agent systems with guaranteed convergence rate. Moreover, the next step in order to develop this method is to consider a set of conditions in which all of the states of the system are not available.

Author contributors Seyed Ehsan Razavi, Mohammad Amin Moradi, Saeed Shamaghdari, Mohammad Bagher Menhaj.

Funding There was no funding.

Declarations

Conflicts of interest I declare that there is no conflict of interest in the publication of this article, and that there is no conflict of interest with any other author or institution for the publication of this article.

Ethical Statements I hereby declare that this manuscript is the result of our independent creation under the reviewers' comments. Except for the quoted contents, this manuscript does not contain any research achievements that have been published or written by other individuals or groups. I am the third author of this manuscript. The legal responsibility of this statement shall be borne by me.

Data Availability Not applicable.

Code Availability Not applicable.

References

1. Kleinman D (1968) On an iterative technique for Riccati equation computations. IEEE Trans Autom Control 13:114–115
2. Tao G (2003) Adaptive control design and analysis. Wiley, Hoboken
3. Bhasin S et al (2011) Asymptotic tracking by a reinforcement learning-based adaptive critic controller. J Control Theory Appl 9:400–409
4. Jiang Y et al (2019) Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning. IEEE Trans Cybern
5. Zhang H et al (2017) Data-based adaptive dynamic programming for a class of discrete-time systems with multiple delays. IEEE Trans Syst Man Cybern Syst 1–10
6. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT Press, Cambridge
7. Vamvoudakis KG et al (2017) Game theory-based control system algorithms with real-time reinforcement learning: how to solve multiplayer games online. IEEE Control Syst Mag 37:33–52
8. Kiumarsi B et al (2018) Optimal and autonomous control using reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst 29:2042–2062
9. Lewis FL et al (2012) Optimal control. Wiley, Hoboken
10. Gao W, Jiang Z-P (2016) Adaptive dynamic programming and adaptive optimal output regulation of linear systems. IEEE Trans Autom Control 61:4164–4169
11. Moghadam R, Lewis FL (2019) Output-feedback H∞ quadratic tracking control of linear systems using reinforcement learning. Int J Adapt Control Signal Process 33:300–314
12. Modares H et al (2015) ${H} _ {{\infty}} $ Tracking control of completely unknown continuous-time systems via off-policy reinforcement learning. IEEE Trans Neural Netw Learn Syst 26:2550–2562
13. Zhang H et al (2017) Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method. IEEE Trans Industr Electron 64:4091–4100
14. Modares H et al (2016) Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning. Automatica 71:334–341
15. Barto AG et al (1983) Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybern 834–846.
16. Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3:9–44
17. Werbos PJ (1989) Neural networks for control and system identification. In: Proceedings of the 28th IEEE conference on decision and control. IEEE, pp 260–265
18. Werbos PJ et al (1990) A menu of designs for reinforcement learning over time. In: Neural networks for control, pp 67–95
19. Werbos P (1992) Approximate dynamic programming for realtime control and neural modelling. In: Handbook of intelligent control: neural, fuzzy and adaptive approaches, pp 493–525
20. Bertsekas DP, Tsitsiklis JN (1995) Neuro-dynamic programming: an overview. In: Proceedings of 1995 34th IEEE conference on decision and control. IEEE, pp 560–564
21. Zhang H et al (2009) Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. IEEE Trans Neural Netw 20:1490–1503
22. Lewis FL, Vrabie D (2009) Reinforcement learning and adaptive dynamic programming for feedback control. IEEE Circuits Syst Mag 9:32–50
23. Leake R, Liu R-W (1967) Construction of suboptimal control sequences. SIAM J Control 5:54–63
24. Howard RA (1972) Dynamic programming and markov processes. The MIT Press, Cambridge
25. Kiumarsi-Khomartash B et al (2013) Optimal tracking control for linear discrete-time systems using reinforcement learning, 52nd IEEE Conference on Decision and Control. IEEE, pp 3845–3850
26. Kiumarsi B et al (2014) Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. Automatica 50:1167–1175
27. Kiumarsi B et al (2015) Optimal tracking control of unknown discrete-time linear systems using input-output measured data. IEEE Trans Cybern 45:2770–2779
28. Gao W, Jiang Z-P (2018) Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems. IEEE Trans Neural Netw Learn Syst 29:2614–2624
29. Jiang Y, Jiang Z-P (2012) Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. Automatica 48:2699–2704
30. Modares H et al (2016) Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning. IEEE Trans Cybern 46:2401–2410
31. Zhang H et al (2016) Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method. IEEE Trans Industr Electron 64:4091–4100
32. Zhang H et al (2018) H∞ consensus for linear heterogeneous multiagent systems based on event-triggered output feedback control scheme. IEEE Trans Cybern 1–12
33. Zhang H et al (2018) Data-driven distributed optimal consensus control for unknown multiagent systems with input-delay. IEEE Trans Cybern 1–11.
34. Modares H et al (2018) Optimal synchronization of heterogeneous nonlinear systems with unknown dynamics. IEEE Trans Autom Control 63:117–131
35. Modares H et al (2018) Static output-feedback synchronisation of multi-agent systems: a secure and unified approach. IET Control Theory Appl 12:1095–1106
36. Kiumarsi B, Lewis FL (2017) Output synchronization of heterogeneous discrete-time systems: a model-free optimal approach. Automatica 84:86–94
37. Kiumarsi B et al (2017) H∞ control of linear discrete-time systems: off-policy reinforcement learning. Automatica 78:144–152