Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Online Output-Feedback Optimal Control of Linear Systems Based on Data-Driven Adaptive Learning ⋆

Jun Zhao * **Jing Na** * **Guanbin Gao** * **Shichang Han** *
Qiang Chen ** **Shubo Wang** ***

* *Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming, 650500 China (e-mail:junzhao1993@163.com, najing25@163.com, gbgao@163.com).*
** *College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China. (e-mail: sdnjchq@zjut.edu.cn).*
*** *School of Automation, Qingdao University, Qingdao, 266071, China (e-mail: wangshubo1130@126.com).*

Abstract: This paper proposes a new approach to solve the output-feedback optimal control for linear systems. A modified algebraic Riccati equation (MARE) is constructed by investigating the corresponding relationship with the state-feedback optimal control. To solve the derived MARE, an online data-driven adaptive learning is designed, where the vectorization operation and Kronecker's product are applied to reformulate the output Lyapunov function. Consequently, only the measurable system input and output are used to derive the solution of the MARE. In this case, the output-feedback optimal control solution can be obtained in an online manner without resorting to the unknown system states. Simulation results are provided to demonstrate the efficacy of the suggested method.

Keywords: Adaptive optimal control, output-feedback control, adaptive control, data-driven learning, Kronecker's product

1. INTRODUCTION

The purpose of optimal control (Lewis et al. (2012)) is to find an optimal control law, which can maximize the control efficiency or minimize the control costs. To complete the optimal control design, one needs to address an algebraic Riccati equation (ARE) (for linear systems) or a Hamilton-Jacobi-Bellman (HJB) equation (for nonlinear systems), which are difficult to solve in general. Hence, most existing methods used to solve these optimal control equations are *offline* methods (Allwright (1980)). In order to design the optimal control online, the idea of reinforcement learning (RL) was recently tailored to develop the so-called adaptive dynamic programming (ADP) (Werbos (1992)). The key merit of ADP scheme is to adopt a critic neural network (NN) to estimate the ideal cost function, such that an approximate numerical solution can be obtained for the ARE or HJB equation (Lewis and Vrabie (2009); Modares et al. (2016); Heydari and Balakrishnan (2013)). To relax the assumptions on the fully known system dynamics in the ADP methods, an observer (Zhang et al. (2011)) or identifier (Vamvoudakis and Lewis (2010)) was incorporated into the ADP synthesis, leading to a complex ADP structure. To further reduce the complexity, an identifier-critic structure based ADP algorithm was suggested in Na and Guido (2014); Lv and Ren (2018), where the actor NN is avoided and the control action is calculated based on the critic NN weights directly. However, it is noted that these ADP based optimal control designs again require fully known system states, i.e., they are state-feedback based control methodologies.

In fact, the vast majority of the existing ADP synthesises all require that full system states should be measurable, while the output-feedback optimal control has been rarely considered, which remains as an open questions in the control fields (Syrmos et al. (1997)). Lewis and Vamvoudakis (2011) proposed both policy iteration (PI) and value iteration (VI) algorithms to address output-feedback optimal control for discrete-time systems. On the other hand, by combining the observer and ADP, Zhu et al. (2014) proposed an integral reinforcement learning (IRL) algorithm, where the existence of output-feedback optimal control solution was explored inspired by Gadewadikar et al. (2012). Nevertheless, the cost function used for deriving the IRL depends on the full system states, so that an observer must be used in Zhu et al. (2014) to reconstruct immeasurable system states, leading to a two-step optimal control implementation. Similarly, Modares et al. (2016) adopted a new system state reconstruction method based on the limited measurements of system output over a certain time interval to develop an off-policy method. However, all of these ADP based output-feedback optimal control designs rely on the observer design, and thus can be taken as an *indirect* output-feedback control design method.

⋆ This work was supported by National Natural Science Foundation of China (NSFC) under Grants 61922037 and 61873115. (*Corresponding author: Jing Na*)

Inspired by these discussions, in this paper, an online data-driven learning technique is developed to address the output-feedback optimal control design for linear systems, where only the system input and output data are required. The main idea is to construct a modified algebraic Riccati equation (MARE) for output-feedback optimal control by considering its state-feedback counterpart. Then, to solve this MARE, the vectorization operation and Kronecker's product are applied to reformulate the output Lyapunov function to facilitate the design of adaptive learning using the system input/output only rather than the system states. In this case, the output-feedback optimal control solution can be online calculated based on the solution of MARE. Consequently, the requirements on the immeasurable system states can be removed, and the observer design used in Zhu et al. (2014) is also avoided. Numerical simulation results are also given to verify the efficacy of this proposed method.

2. PROBLEM FORMULATION

2.1 State-feedback Optimal Control

Consider the following continuous-time (CT) linear system

$$\begin{cases} \dot{x} = Ax + Bu \\ y = Cx \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$ is the system states, $u \in \mathbb{R}^m$ is the control action, $y \in \mathbb{R}^p$ is the output, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the system matrix and control matrix, respectively. $C \in \mathbb{R}^{b\times p}$ is the output matrix.

The optimal control problem is to find a control u for system (1) to minimize the following cost function:

$$J(x(t)) = \int_t^\infty r(x(\tau), u(\tau))d\tau \tag{2}$$

with the utility function $r(x(\tau), u(\tau)) = x^T(\tau)Qx(\tau) + u^T(\tau)Ru(\tau)$, where $Q > 0$ and $R > 0$ are the symmetric weight matrices.

Assumption 1. (Lewis et al. (2012)) The pair (A, B) is stabilizable and the pair (A, C) is detectable.

If the full system states x are available or measurable, the state-feedback control can be used to solve the above optimal control problem. We can obtain the state-feedback solution for system (1) with cost function (2) as

$$\begin{aligned} u^* &= -K_x x \\ K_x &= R^{-1}B^T P^* \end{aligned} \tag{3}$$

It is noted that $P^* \in \mathbb{R}^{n\times n}$ is a positive definite matrix, which can be obtained by solving the following ARE:

$$A^T P^* + P^* A + Q - P^* B R^{-1} B^T P^* = 0 \tag{4}$$

Based on the optimality principle (Lewis et al. (2012)) and the optimal control (3), we can obtain the optimal cost function as

$$\begin{aligned} V^*(x(t)) &= \int_t^\infty x^T(\tau)(Q + K_x^T R K_x)x(\tau)d\tau \\ &= x^T(t)P^* x(t) \end{aligned} \tag{5}$$

Remark 1. Recently, the idea of ADP was presented to obtain the optimal control action, where the policy iteration (Jiang and Jiang (2012)), online adaptive learning (Vamvoudakis and Lewis (2010)) and integral reinforcement learning (IRL) (Zhu et al. (2014)) were developed. However, it is noted that the most of existing ADP based optimal control designs were developed based on the full state-feedback control (Na and Guido (2014)) or reconstructed system states via observers (Zhu et al. (2014)). Although for linear systems the incorporation of observer into the control design is trivial, the observer-based output control needs extra computational costs for observer. Hence, the *direct* output-feedback optimal control has not been fully solved yet. Therefore, the main contribution of this paper is to present a *direct* output-feedback optimal control for linear systems without using any observer. This is achieved by exploring the relationship between the state-feedback optimal control and the output-feedback optimal control, and then suggesting an online data-driven learning algorithm to solve the derived optimal control equation.

2.2 Output-feedback Optimal Control

For system (1), the control input can be calculated based on the output measurement associated with an output feedback gain K, that is

$$u^* = -Ky \tag{6}$$

where $K \in \mathbb{R}^{m\times p}$ is the output feedback gain to be calculated.

To derive the feedback gain K in (6), we should rewrite it as $u^* = -Ky = -KCx$, indicating $K_x = KC$. Then, taking (6) into (5), we can further represent the cost function as

$$\begin{aligned} V^*(x(t)) &= \int_t^\infty x^T(\tau)(Q + C^T K^T R K C)x(\tau)d\tau \\ &= x^T(t)P^* x(t) \end{aligned} \tag{7}$$

By calculating the time-derivative of (7), we have

$$\dot{x}^T P^* x + x^T P^* \dot{x} + x^T(Q + C^T K^T R K C)x = 0 \tag{8}$$

Substituting system (1) into (8), we can rewrite it as

$$\begin{aligned} x^T[(A - BKC)^T P^* + P^*(A - BKC) \\ + Q + C^T K^T R K C]x = 0 \end{aligned} \tag{9}$$

Since the equation (9) holds for all $x \in \mathbb{R}^n$, the following modified ARE (MARE) can be given as

$$A_c^T P^* + P^* A_c + Q + C^T K^T R K C = 0 \tag{10}$$

with $A_c = (A - BKC)$.

According to the control actions given in (3) and (6), the following equation can be verified

$$KC = K_x = R^{-1}B^T P^* \tag{11}$$

Since the fact $u^* = -Ky = -KCx$ holds mathematically, by substituting (11) into (10), then the standard ARE (4) can be obtained. Hence, inspired by the analysis given in Zhu et al. (2014), we can prove the following lemma:

Lemma 1. For linear system (1), assume that a control gain K can be found from (10) and fulfills the condition (11), then the control action given in (11) is globally optimal.

Proof. The detailed proof of this claim can be referred to Gadewadikar et al. (2012), which is not given here due to the page limit.

Remark 2. To implement the proposed output-feedback optimal control, the remaining problem is to solve the MARE (10) to obtain an output-feedback control gain. To this end, Gao et al. (2016) proposed both policy iteration and value iteration methods based on a sampled-data method to reconstruct the unmeasurable states. The work of Zhu et al. (2014) proposed an IRL algorithm, where an adaptive observer must be designed to estimate the system states. Different to these results (e.g., Zhu et al. (2014); Gao et al. (2016)), we will introduce a data-driven online learning approach to solve the MARE (10) to further obtain the output-feedback control gain K, where only the system output is needed. This will create an one-step, direct, online output-feedback optimal control algorithm, without using observers or state reconstruction schemes.

3. DATA-DRIVEN ONLINE SOLUTION OF MARE

In this section, a data-driven learning scheme will be introduced to solve the derived MARE (10) by using the output measurement y only so as to implement output-feedback optimal control. The Kronecker's product is first applied on the derived MARE (10) to reformulate the equation in a parameterized form. Then, a novel adaptive law is used to estimate the unknown parameters in the MARE, where the solution P^* can be online calculated.

3.1 Data-driven Reformulation of MARE

To design a data-driven learning method to obtain the solution of (10) via the output y rather than the system states x, we should make further manipulations on (10). To this end, we will carry out the Kronecker's product calculation as Zhao et al. (2019). To avoid using the system states x, both sides of (10) are multiplied by C^TC, then the following equation can be derived

$$\begin{aligned} C^TC(A_c^TP^* + P^*A_c)C^TC = &- C^TC(Q \\ &+ C^TK^TRKC)C^TC \end{aligned} \tag{12}$$

We multiply both sides of (12) by system states x, such that

$$\begin{aligned} x^TC^TC(A_c^TP^* + P^*A_c)C^TCx \\ = -x^TC^TC(Q + C^TK^TRKC)C^TCx \end{aligned} \tag{13}$$

According to the definition of system output $y = Cx$, Eq.(13) can be further written as

$$\begin{aligned} y^TC(A_c^TP^* + P^*A_c)C^Ty \\ = -y^TC(Q + C^TK^TRKC)C^Ty \end{aligned} \tag{14}$$

Remark 3. Comparing (14) with (9), it is clearly shown that we only use the system output y instead of states x. This leads to the exact output-feedback optimal control implementation, which allows to develop new policy iteration methods or adaptive algorithms to solve (15).

Although the system states x are not involved in (14), the matrix A_c in (14) is unknown (since the unknown control gain K is involved in A_c). To avoid using A_c in the following Kronecker and vectorization operations, we need to further decompose A_c. Then, by substituting $A_c = A - BKC$ and $KC = R^{-1}B^TP^*$ into (14), we can rewrite this equation as

$$\begin{aligned} y^TC(A^TP^* + P^*A)C^Ty \\ = -y^TC(Q - P^*BR^{-1}B^TP^*)C^Ty \end{aligned} \tag{15}$$

It is clear now that the unknown matrix A_c is avoided, while the control gain K is replaced by the matrix P^*. Hence, we can online estimate the matrix P^* and then derive the control gain K.

Unlike the existing offline policy iteration or online IRL methods, we will develop a new online adaptive learning scheme to solve (15) to obtain the optimal control solution P^*. For this purpose, the $vec(\cdot)$ operator and the Kronecker's product are applied on both sides of (15) as Zhao et al. (2019). Then, equation (15) can be reformulated as

$$\begin{aligned} &2(C^Ty \otimes AC^Ty)^Tvec(P^*) + (C^Ty \otimes C^Ty)^Tvec(Q) \\ &- (vec(BR^{-1}B^T) \otimes (C^Ty \otimes C^Ty))^Tvec(P^* \otimes P^*) = 0 \end{aligned} \tag{16}$$

It is not difficult to find from (16) that the dimension of $vec(P^* \otimes P^*)$ may be high, which is not preferable in the online learning. For instance, the system control matrix A is with dimension $n \times n$, then the solution of matrix P^* has $n \times n$ parameters, and after applying $vec(\cdot)$ operator and Kronecker's product, the dimension of $vec(P^* \otimes P^*)$ is $n^4 \times 1$, which will increase the computation costs and even may lead to the failure of online learning. To reduce the dimension of online learning parameters to reduce the computational costs, dimension-reduction operations should be employed. For this purpose, we will use the state-feedback control gain K_x to replace the solution P^* of ARE (i.e. $K_x = R^{-1}B^TP^*$). Then, Eq.(15) can be further formulated as

$$y^TC(A^TP^* + P^*A)C^Ty = -y^TC(Q + K_x^TRK_x)C^Ty \tag{17}$$

Then, similar to (16), we use the $vec(\cdot)$ operator and the Kronecker's product on both sides of (17), and have

$$\begin{aligned} &2(C^Ty \otimes AC^Ty)^Tvec(P^*) + (C^Ty \otimes C^Ty)^Tvec(Q) \\ &+ (vec(R) \otimes (C^Ty \otimes C^Ty))^Tvec(K_x \otimes K_x) = 0 \end{aligned} \tag{18}$$

It is shown that (18) represents a linearly parameterized form of the unknown optimal solution P^* and gain K_x after applying $vec(\cdot)$ operator and Kronecker's product. To show this point more clearly, we further reformulate (18) in a compact form as

$$\Theta = -W^T \Xi \tag{19}$$

where $\Theta \in \mathbb{R}$ is the measured output and $\Xi \in \mathbb{R}^{2n^2}$ is the regressor, which can be obtained by the measured system output, and $W \in \mathbb{R}^{2n^2}$ is the unknown parameter vector, which are defined as:

$$\begin{aligned} \Xi(y, A, C) =& [2(C^T y \otimes AC^T y), vec(R) \otimes (C^T y \otimes C^T y)]^T \\ W(P^*) =& [vec(P^*), vec(K_x \otimes K_x)]^T \\ \Theta(y) =& [C^T y \otimes C^T y]^T vec(Q) \end{aligned} \tag{20}$$

After using the vectorization operation, the unknown matrix P^* given in (19) can be considered as a vector. Thus, an adaptive law can be used to estimate the solution P^* online, which will be shown in the next subsection. Moreover, since we use $K_x \otimes K_x$ to replace $P^* \otimes P^*$, the dimension of unknown parameter vector W can be reduced.

3.2 Solving MARE via Online Learning

This section will propose an adaptive law to solve the MARE (10). From the formulation of (20) that the control matrix B is not used, which further relaxes the requirements on the system. Moreover, as shown in (20), the unknown vector W is a function of the unknown matrix P^*, which is the solution of MARE (10) to be solved. Thus, an adaptive law can be used to online estimate W based on (20) so as to obtain the estimate of P^*. To this end, we define an auxiliary regressor matrix $S \in \mathbb{R}^{2n^2 \times 2n^2}$ and vector $\emptyset \in \mathbb{R}^{2n^2}$ as

$$\begin{cases} \dot{S} = -\ell S + \Xi \Xi^T, S(0) = 0 \\ \dot{\emptyset} = -\ell \emptyset + \Xi \Theta, \emptyset(0) = 0 \end{cases} \tag{21}$$

where $\ell > 0$ is a design parameter. Therefore, S and $\emptyset$ can be online obtained via measurable output y.

We define an auxiliary vector $M \in \mathbb{R}^{2n^2}$ according to the derived variables S and $\emptyset$ in (21) as

$$M = S\hat{W} + \emptyset \tag{22}$$

where $\hat{W}$ is the estimate of the unknown vector W.

Therefore, the adaptive law used to online update $\hat{W}$ is designed by

$$\dot{\hat{W}} = -\Gamma M \tag{23}$$

with $\Gamma > 0$ being the adaptive gain set by the designers.

To show the merit of the above learning algorithm (23) over the gradient algorithm, we solve (21) and substitute (19) into (21), then can verify that $\emptyset = -SW$. In this case, we can rewrite (22) as

$$M = S\hat{W} + \emptyset = -S\tilde{W} \tag{24}$$

with $\tilde{W} = W - \hat{W}$ being the learning error. Hence, it is in (23) that the estimates $\hat{W}$ is updated along with the estimation error $\tilde{W}$ extracted by using the measurable system output y. Thus, this adaptive algorithm clearly differs to the gradient descent algorithms used in other ADP literatures (e.g., Abu-Khalaf and Lewis (2005); Vamvoudakis and Lewis (2010) and references therein), which adopt the gradient descent algorithm to minimize the HJB residual error.

Before showing the convergence analysis (faster convergence), we exemplify the positive definiteness of matrix S, which is summarized as:

Lemma 2 (Na and Guido (2014)). The persistent excitation (PE) of vector Ξ in (21) equals to the positive definiteness of S defined in (22).

Lemma 2 implies that the PE condition can be online verified by calculating the minimum eigenvalue of matrix S. The value of Lemma 2 lies in that it provides a feasible technique to online test the PE condition, which remains as an open problem in the field. On the other hand, this PE condition is necessary for retaining the convergence of learning algorithms and has been widely utilized in the ADP literatures (e.g., Abu-Khalaf and Lewis (2005); Vamvoudakis and Lewis (2010)). In practice, a vanishing probing noise can be inserted into the measurements during the transient learning stage to fulfill this condition as proposed in Abu-Khalaf and Lewis (2005); Vamvoudakis and Lewis (2010).

Hence, the main results o this paper can be given as:

Theorem 1. For (19) with adaptive law (23), if the regressor vector Ξ in (21) is PE, then the estimation error $\tilde{W}$ converges to zero exponentially.

Proof. First, Lemma 2 indicates that the matrix S is positive definite under the PE condition of Ξ, i.e., the minimum eigenvalue $\lambda_{min}(S) > \sigma > 0$. Hence, we choose a Lyapunov function $V_1 = \frac{1}{2}(\tilde{W}^T \Gamma^{-1} \tilde{W})$, such that $\dot{V}_1$ can be derived from (23) and (24) as

$$\begin{aligned} \dot{V}_1 &= \tilde{W}^T \Gamma^{-1} \dot{\tilde{W}} = -\tilde{W}^T S \tilde{W} \leq -\sigma \|\tilde{W}\|^2 \\ &\leq -\mu V_1 \end{aligned} \tag{25}$$

where $\mu = 2\sigma/\lambda_{max}(\Gamma^{-1})$ is a positive constant with $\lambda_{max}(\cdot)$ being the maximum eigenvalue. Consequently, one can conclude from the Lyapunov theorem that the estimation error $\tilde{W}$ converges to zero exponentially.

Remark 4. As shown in the above derivation of adaptive law, only the output data y is used for online obtaining the solution of MARE, which is clearly different to available results (Abu-Khalaf and Lewis (2005); Vamvoudakis and Lewis (2010)). In particular, in this section, by introducing the adaptive law (23) driven by the estimation error $\tilde{W}$, faster (exponential) convergence of the estimation error is

obtained, outperforming the gradient based adaptive laws used in the existing ADP designs.

According to the above derived solution of MARE (10), we can extract the estimated matrix $\hat{P}$ of the ideal solution P^*. Hence, we can obtain the actual optimal control action for system (1) as

$$u = -\hat{K}y \tag{26}$$

Based on Theorem 1, the estimation error $\tilde{W} \to 0$ provided that the regressor Ξ is PE. Hence, the derived control in (26) converges to its optimal control (6), i.e., $\|u-u^*\| \to 0$. In this case, based on Lemma 1, we know that for linear system (1) with control action (26) and adaptive law (23), if the initial control $u(0)$ is admissible, then the controlled system is stable.

4. SIMULATION

In this section, we use a practical industrial system (e.g., power systems (Tang et al. (2017); Mu et al. (2017); Zhao et al. (2020))) to verify the validity of the proposed output-feedback optimal control. In this application, the micro-grids include the distributed and renewable energies. However, the frequency deviation may occur due to the imbalance between load consumption and power generation. Therefore, it is very important to ensure the stability of micro-grids. For this purpose, we consider a practical power system consists of a turbine generator, a system load, and an automatic generation control.

The purpose is to find an optimal output-feedback control law $u = -Ky$ such that the closed-loop system is stable, and the predefined cost function is minimized. In order to facilitate simulations, we consider ζ_f, ζ_g and ζ_G as the incremental change of the frequency deviation, the generator output, and the governor value position, respectively. Moreover, we take the control input u to represent the incremental speed change of positive deviation. Then, one defines $x = [\zeta_f;\ \zeta_g;\ \zeta_G] \in \mathbb{R}^3$ as the state vector, where $x_1 = \zeta_f$, $x_2 = \zeta_g$ and $x_3 = \zeta_G$. Hence, we write the state-space model of this power system as

$$\dot{x} = \begin{bmatrix} -\frac{1}{T_G} & 0 & -\frac{1}{F_r T_G} \\ \frac{K_t}{T_t} & -\frac{1}{T_t} & 0 \\ 0 & \frac{K_g}{T_g} & -\frac{1}{T_g} \end{bmatrix} x + \begin{bmatrix} \frac{1}{T_G} \\ 0 \\ 0 \end{bmatrix} u \tag{27}$$
$$y = [1, 0, 0]x$$

where the model parameters can be found in Table 1.

Since the purpose of this simulation is to verify the effectiveness of the proposed method, the parameters can be selected as: the initial conditions $x_0 = [-0.3, 0.5, 1]^T$, and the gains $Q = diag([1,\ 1,\ 1])$ and $R = 1$. Moreover, the parameters used in the adaptive law are set as $\ell = 1$, and $\Gamma = 10$. Since the studied system is in a linear form, the proposed optimal control solution can be obtained by solving the ARE offline. To verify the effectiveness of the developed adaptive learning technology, the offline solution of (4) is first given as

Table 1. PARAMETERS OF POWER SYSTEM

Symbol	Meaning	Values
T_G	Time constant of the governor	5
T_t	Time constant of the turbine model	10
T_g	Time constant of the generator model	10
F_r	Feedback regulation constant	0.5
K_t	Gain constant of the turbine model	1
K_g	Gain constant of the generator model	1

$$P^* = \begin{bmatrix} 2.5817 & 1.4963 & -1.7575 \\ 1.4963 & 7.7916 & 3.2394 \\ -1.7575 & 3.2394 & 11.4121 \end{bmatrix} \tag{28}$$

Fig. 1 indicates the response of the approximated solution of the MARE, i.e., $\hat{P}$, with the adaptive law (23), which shows very fast convergence. To verify the convergence to the exact value, the estimation error between $\hat{P}$ and the ideal value P^* is given in Fig. 2 (For easier reading, Fig. 2 is given in the log scale in the y-axis). We can find that the online updated $\hat{P}$ converges to a very small set around the ideal solution P^*, which illustrates the correctness of the theoretical analysis.

With the estimated matrix $\hat{P}$, the actual output-feedback control gain $\hat{K}$ ($\hat{K} = R^{-1}B^T\hat{P}C^T(CC^T)^{-1}$) can be online calculated as

$$\hat{K} = 0.5163 \tag{29}$$

Fig. 3 indicates the profiles of system states with the derived control action $u = -\hat{K}y$, which shows that the control system is stable. Moreover, the derived control is also shown in Fig. 3, which is bounded and smooth.

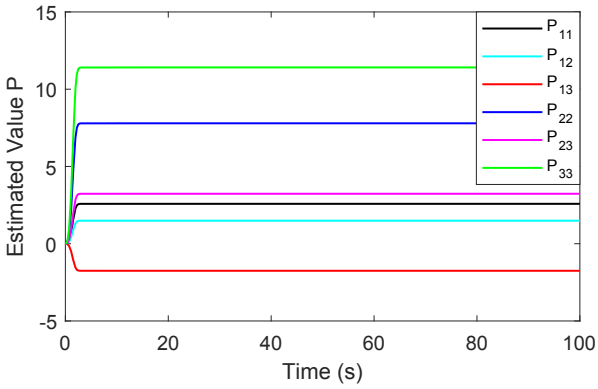

Fig. 1. The profile of the online updated matrix $\hat{P}$.

5. CONCLUSION

This paper is concerned with solving the output-feedback optimal control for linear systems with output measurement only. The main idea is to develop an online data-driven learning method to solve the derived optimal equations. Hence, a MARE is first constructed for the output-feedback optimal control by further investigating the corresponding relationship with the state-feedback optimal control. After using the vectorization operator and Kronecker's product to reformulate this MARE, a new adaptive learning method is developed to obtain the

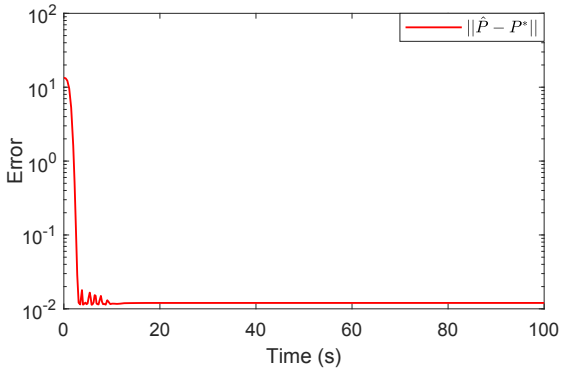

Fig. 2. The norm error between P^* and $\hat{P}$.

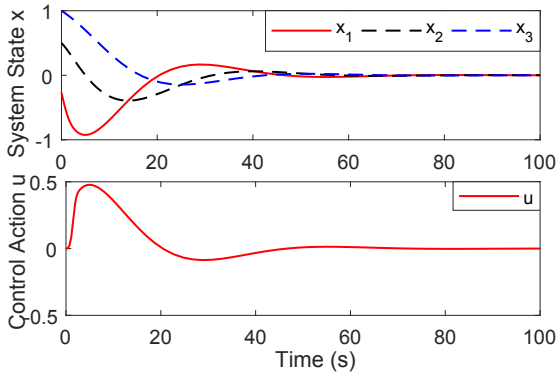

Fig. 3. The profile of controlled system states x and control u.

solution of MARE. In this framework, the observer widely used in the output-feedback control is avoided. Simulation results are provided to illustrate the effectiveness of the suggested algorithm. This idea will be further tailored to output-feedback robust control of uncertain nonlinear systems.

REFERENCES

Abu-Khalaf, M. and Lewis, F.L. (2005). Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach. *Automatica*, 41(5), 779–791.

Allwright, J. (1980). A lower bound for the solution of the algebraic riccati equation of optimal control and a geometric convergence rate for the kleinman algorithm. *IEEE Transactions on Automatic Control*, 25(4), 826–829.

Gadewadikar, J., Abu-Khalaf, M., and Lewis, F.L. (2012). Necessary and sufficient conditions for h-infinity static output-feedback control. *Journal of Guidance Control and Dynamics*, 29(4), 915–920.

Gao, W., Jiang, Y., Jiang, Z.P., and Chai, T. (2016). Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming. *Automatica*, 72, 37–45.

Heydari, A. and Balakrishnan, S.N. (2013). Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics. *IEEE Transactions on Neural Networks and Learning Systems*, 24(1), 145–157.

Jiang, Y. and Jiang, Z.P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 48(10), 2699–2704.

Lewis, F.L. and Vamvoudakis, K.G. (2011). Reinforcement learning for partially observable dynamic processes: adaptive dynamic programming using measured output data. *IEEE Transactions on System Man Cybern B Cybern*, 41(1), 14–25.

Lewis, F.L. and Vrabie, D. (2009). Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 9(3), 32–50.

Lewis, F.L., Vrabie, D., and Syrmos, V.L. (2012). *Optimal control*. John Wiley & Sons.

Lv, Y. and Ren, X. (2018). Approximate nash solutions for multiplayer mixed-zero-sum game with reinforcement learning. *IEEE Transactions on Systems Man and Cybernetics Systems*, PP(99), 1–12.

Modares, H., Lewis, F.L., and Jiang, Z. (2016). Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning. *IEEE Transactions on Cybernetics*, 46(11), 2401–2410.

Mu, C., Tang, Y., and He, H. (2017). Improved sliding mode design for load frequency control of power system integrated an adaptive learning strategy. *IEEE Transactions on Industrial Electronics*, 64(8), 6742–6751.

Na, J. and Guido, H. (2014). Online adaptive approximate optimal tracking control with simplified dual approximation structure for continuous-time unknown nonlinear systems. *IEEE/CAA Journal of Automatica Sinica*, 1(4), 412–422.

Syrmos, V.L., Abdallah, C.T., Dorato, P., and Grigoriadis, K. (1997). Static output feedback-a survey. *Automatica*, 33(2), 125–137.

Tang, Y., He, H., Wen, J., and Liu, J. (2017). Power system stability control for a wind farm based on adaptive dynamic programming. *IEEE Transactions on Smart Grid*, 6(1), 166–177.

Vamvoudakis, K.G. and Lewis, F.L. (2010). Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5), 878–888.

Werbos, P.J. (1992). Approximate dynamic programming for real-time control and neural modeling. *Handbook of Intelligent Control Neural Fuzzy & Adaptive Approaches*.

Zhang, H., Cui, L., Zhang, X., and Luo, Y. (2011). Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Transactions on Neural Networks*, 22(12), 2226–2236.

Zhao, J., Na, J., and Gao, G. (2020). Adaptive dynamic programming based robust control of nonlinear systems with unmatched uncertainties. *Neurocomputing*. doi: https://doi.org/10.1016/j.neucom.2020.02.025.

Zhao, J., Na, J., Gao, G., Xiao, Y., and Song, Z. (2019). Data-driven online adaptive optimal control for linear systems with completely unknown dynamics. In *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*, 557–562.

Zhu, L.M., Modares, H., Gan, O.P., Lewis, F.L., and Yue, B. (2014). Adaptive suboptimal output-feedback control for linear systems using integral reinforcement learning. *IEEE Transactions on Control Systems Technology*, 23(1), 264–273.