IEEE Control Systems Letters paper presented at
2023 American Control Conference (ACC)
San Diego, CA, USA. May 31 - June 2, 2023

Timing-Aware Resilience of Data-driven Off-policy Reinforcement Learning for Discrete-Time Systems

Lijing Zhai[1], Filippos Fotiadis[1], Kyriakos G. Vamvoudakis[1], Jérôme Hugues[2]

***Abstract*— In this paper, we study the impact of clock offsets among different components of cyber-physical systems on data-driven off-policy reinforcement learning (RL) for linear quadratic regulation (LQR). Our results show that under certain conditions the control policies generated by data-driven off-policy RL with clock offsets are stabilizing policies. With clock offsets what directly influences the learning behavior is not only the values of clock offsets but also the dynamics change caused by clock offsets. In particular, larger values of clock offsets do not necessarily lead to non-stabilizing policies. The proposed conclusions are illustrated by numerical simulations.**

I. Introduction

Cyber-physical systems (CPS) are large-scale and highly distributed systems with computer nodes and communication networks connecting various components. Efficient and reliable operation of CPS requires precise timing of all components sharing the same time. Thus, timing is of critical importance to the flawless functionality and resilience of CPS. However, clock asynchronization among components is quite ubiquitous in these distributed systems [1]. Moreover, accurate system models of large-scale CPS are difficult or impossible to obtain, utilizing data generated by systems for control synthesis is a tendency to go. Clock offsets among different components may induce and propagate mismatched signals in the system and thus negatively impact performance [2], [3]. Thus, understanding the impact of asynchronous clocks on data-driven control is an important research topic.

Recently much of the attention has been given to applying data-driven methods for time-delayed systems [4]–[6], and the robustness or resilience of RL under approximation errors [7]–[10], system dynamics with noises [11], [12], corrupted rewards under adversarial attacks [13], [14] and perturbed rewards in noisy environments [15]. The focus of the aforementioned works is mainly on learning algorithms themselves, especially on the robustness of policy iteration (PI) or value iteration (VI) in term of variations of the learning dynamics caused by approximation errors. Despite these significant works, the impact of clock asynchronization on data-driven control is still missing. Though this consideration is similar to time delays, it is noteworthy that clock offsets and time delays are different in essence. Time delays are represented within the system dynamics, usually incorporated in state or control variables. With time delays, the system propagation is changed and thus system trajectories generated by time delayed systems deviate from those without time delays. Instead, clock offsets are reflected only on the data transmitted in the communication channels between network nodes while the system dynamics remain unchanged. In addition, research works on time delays assume that values of time delays are known when deriving stability conditions or optimal control policies [4], [16], [17]. Motivated by this gap, in this work we investigate the impact of clock offsets on RL. Since off-policy RL is proved to be insensitive to bias caused by the addition of probing noise [18], [19], we focus on the impact of clock offsets among network nodes on off-policy RL algorithms without the influence of probing noise. The work of [10] studies the robustness of RL for LQR to errors in the learning process due to inaccurate system dynamics while this work investigates the impact of clock offsets on the performance of data-driven RL. In addition, while [20] studied the effect of clock offsets on data-driven RL, only continuous-time systems were considered.

The contributions of the present paper are threefold. First, we formulate the problem of RL for systems with clock discrepancies among the learning component and other components. Then, we derive a data-driven off-policy RL algorithm under clock offsets. Finally, we analyze the relationship of clock offset and off-policy RL performance. We begin by formulating the problem of data-driven off-policy RL with clock offsets in Section II. Impact of clock offsets on the data-driven off-policy RL is shown in Section III. Section IV presents numerical simulations. The last section concludes and sketches about future work directions. *Notation*: $\mathbb{N}$ ($\mathbb{Z}$) is the natural number (integer) set. $\overline{\lambda}(\cdot)$ ($\underline{\lambda}(\cdot)$) denotes the maximum (minimum) eigenvalue of a matrix. I_n denotes the identity matrix with dimension n. $\otimes$ denotes the Kronecker product. $A \succ 0$ ($A \succeq 0$) denotes a symmetric positive (non-negative) definite matrix. The matrix norm refers to Frobenius norm.

II. Problem Formulation

A. System Setup under Clock Offsets

Consider a discrete-time, linear time-invariant system:

$$x_{k+1} = Ax_k + Bu_k, \tag{1}$$

[1]L. Zhai, F. Fotiadis, and K. G. Vamvoudakis are with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA e-mail: lzhai3@gatech.edu, ffotiadis@gatech.edu, kyriakos@gatech.edu.

[2]J. Hugues is with the Carnegie Mellon University/Software Engineering Institute, Pittsburgh, PA 15213, USA e-mail: jhugues@andrew.cmu.edu.

 This work was supported in part by the Department of Energy under grant No. DE-EE0008453, by ONR Minerva under grant No. N00014 − 18 − 1 − 2160, by NSF under grant Nos. CAREER CPS-1851588 and S&AS 1849198, by the Onassis Foundation-Scholarship ID: F ZQ 064 − 1/2020 − 2021, and by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. DM23-0217.

where $k \in \mathbb{N}$ is the discrete time index, $x_k \in \mathbb{R}^n$ is the state vector, $u_k \in \mathbb{R}^m$ is the control input, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the state matrix and input matrix, respectively. The pair (A, B) is assumed to be stabilizable. The general structure of a CPS that incorporates learning is shown in Figure 1. Each system component is spatially distributed, and all components have their own distinct clocks, which ideally should be synchronized with one another. However, imperfect communication channels and hardware might lead to timing mismatches, which in turn leads to inconsistent data propagation within the system. Such corrupted data can jeopardize the functionality of the learning component, which inherently assumes that all component clocks agree with one another. The true state and control input at time step k are ideally x_k and u_k, respectively. But owing to timing discrepancies between the learning component and other components, the actual state and control input perceived by the learning component at time step k are $x_k^l = x_{k+\delta_x(k)}$ and $u_k^l = u_{k+\delta_u(k)}$, respectively, where $\delta_x : \mathbb{N} \to \mathbb{Z}$ and $\delta_u : \mathbb{N} \to \mathbb{Z}$ are clock offsets for state and control input signals, respectively. We are interested in the impact of offsets between the actual states/control signals and the perceived ones, particularly conditions of achieving tolerable learning behavior under clock offsets for system (1).

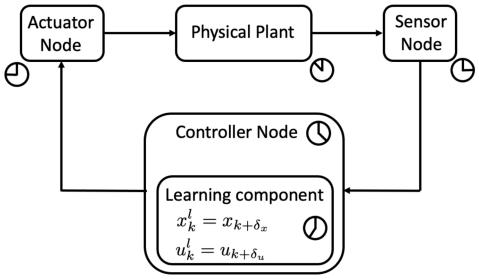

Fig. 1. CPS structure with asynchronous clocks.

B. Preliminaries: Optimal Control and Model-based RL

For a given stabilizing control policy $\mu : \mathbb{R}^n \to \mathbb{R}^m$, define its corresponding performance criterion as $J(x_0;\mu) = \sum_{k=0}^{\infty}\left(x_k^{\mathrm{T}}Qx_k + \mu^{\mathrm{T}}(x_k)R\mu(x_k)\right)$, where $Q \in \mathbb{R}^{n\times n} \succeq 0$ and $R \in \mathbb{R}^{m\times m} \succ 0$, and the pair $(A,\ Q)$ is observable. The value function V^μ corresponding to policy μ denoted by $V^\mu(\cdot) := J(\cdot;\ \mu)$ satisfies the Bellman equation, $\forall k \in \mathbb{N}$:

$$V^\mu(x_k) = x_k^{\mathrm{T}}Qx_k + \mu^{\mathrm{T}}(x_k)R\mu(x_k) + V^\mu(x_{k+1}). \quad (2)$$

The goal is to find a policy $\mu^\star : \mathbb{R}^n \to \mathbb{R}^m$ which minimizes the performance cost and yields the optimal value, i.e., $V^\star(x_k) = \min_\mu J(x_k;\mu)$. Consider the linear state feedback policy and its quadratic value function:

$$\mu(x_k) = -Kx_k,\ V^\mu(x_k) = x_k^{\mathrm{T}}Px_k, \quad (3)$$

with $P \in \mathbb{R}^{n\times n} \succ 0$. Using (3), the Bellman equation (2) can then be simplified to the Lyapunov equation:

$$(A-BK)^{\mathrm{T}}P(A-BK) + Q + K^{\mathrm{T}}RK = P. \quad (4)$$

Now define the Hamiltonian, $\forall x_k,\ \mu,\ k$:

$$H_k = x_k^{\mathrm{T}}Qx_k + \mu^{\mathrm{T}}R\mu + x_{k+1}^{\mathrm{T}}Px_{k+1} - x_k^{\mathrm{T}}Px_k. \quad (5)$$

Then, the optimal control $\mu^\star(x_k) = K^\star x_k$ can be derived by applying the stationarity condition on the Hamiltonian:

$$K^\star = (R + B^{\mathrm{T}}P^\star B)^{-1}B^{\mathrm{T}}P^\star A, \quad (6)$$

where $P^\star$ satisfies the algebraic Riccati equation (ARE):

$$\begin{aligned} 0 =& A^{\mathrm{T}}P^\star A \\ &- P^\star + Q - A^{\mathrm{T}}P^\star B(R + B^{\mathrm{T}}P^\star B)^{-1}B^{\mathrm{T}}P^\star A. \end{aligned} \quad (7)$$

The ARE (7) is a nonlinear function of P and thus is difficult to solve directly. Combine (6) and (7) to get:

$$(A-BK)^{\mathrm{T}}P(A-BK) + Q + (K)^{\mathrm{T}}RK = P. \quad (8)$$

Equation (8) together with equation (6) lays the foundation of model-based PI approach [21] to solving (7) iteratively.

C. Data-driven Off-policy RL without Clock Offsets

Consider off-policy RL [22] to solve (7) in a data-driven manner, which begins by writing the original system (1) as:

$$x_{k+1} = \bar{A}x_k + B(Kx_k + u_k), \quad (9)$$

with $\bar{A} = A - BK$. Here, u_k can be any policy rather than the desired optimal one, hence justifying the "off-policy" label of this algorithm. u_k is generally set as $u_k = \mu_{\mathrm{b}}(x_k) + e_k$, where μ_{b} is a stabilizing behavioral policy, and $e_k \in \mathbb{R}^m$ is an exploration noise added to meet the persistence excitation (PE) condition. This control input is then applied to systems to generate input-state data, which are used to express (8) in a data-based manner. Manipulating (3), (8)-(9) yields the off-policy Bellman equation of PI with iteration number j:

$$\begin{aligned} & x_k^{\mathrm{T}}P^{j+1}x_k - x_{k+1}^{\mathrm{T}}P^{j+1}x_{k+1} \\ =& x_k^{\mathrm{T}}Qx_k + x_k^{\mathrm{T}}(K^j)^{\mathrm{T}}RK^jx_k - (u_k + K^jx_k)^{\mathrm{T}}B^{\mathrm{T}}P^{j+1}\bar{A}x_k \\ & - (u_k + K^jx_k)^{\mathrm{T}}B^{\mathrm{T}}P^{j+1}x_{k+1}. \end{aligned} \quad (10)$$

With (9) and employing the property of Kronecker product, i.e., $x^{\mathrm{T}}Ay = (y^{\mathrm{T}} \otimes x^{\mathrm{T}})\mathrm{vec}(A)$, (10) can be rewritten as:

$$\begin{aligned} 0 =& -(x_k^{\mathrm{T}} \otimes x_k^{\mathrm{T}})\mathrm{vec}(P^{j+1}) + (x_{k+1}^{\mathrm{T}} \otimes x_{k+1}^{\mathrm{T}})\mathrm{vec}(P^{j+1}) \\ & - 2(x_k^{\mathrm{T}} \otimes (u_k + K^jx_k)^{\mathrm{T}})\mathrm{vec}(B^{\mathrm{T}}P^{j+1}A) \\ & - ((u_k - K^jx_k)^{\mathrm{T}} \otimes (u_k + K^jx_k)^{\mathrm{T}})\mathrm{vec}(B^{\mathrm{T}}P^{j+1}B) \\ & + x_k^{\mathrm{T}}Qx_k + x_k^{\mathrm{T}}(K^j)^{\mathrm{T}}RK^jx_k. \end{aligned} \quad (11)$$

Set $W_1^{j+1} = P^{j+1} \in \mathbb{R}^{n\times n}$, $W_2^{j+1} = B^{\mathrm{T}}P^{j+1}A \in \mathbb{R}^{m\times n}$, $W_3^{j+1} = B^{\mathrm{T}}P^{j+1}B \in \mathbb{R}^{m\times m}$, which can be solved by least squares (LS) methods through (11). Thus, P^{j+1} and K^{j+1} can be solved simultaneously without knowing the dynamics. There are n^2+m^2+mn unknown parameters. So the window size for collecting data is selected as $s \geqslant n^2 + m^2 + mn$. Note that W_1^j and W_3^j are symmetric matrices with $n\times(n+1)/2$ and $m\times(m+1)/2$ independent elements, respectively. Hence, only $n\times(n+1)/2$ and $m\times(m+1)/2$ number of elements need to be computed for W_1^j and W_3^j, respectively. Algorithm 1 describes the data-driven off-policy RL.

Algorithm 1 Data-driven Off-policy RL

Select a stabilizing control policy $u_k = \mu_{\rm b}(x_k)$ for data collection. Set iteration number $j = 0$. Select an initial controller gain K^0 and a proper window size s.

1. Solve for W_1^{j+1}, W_2^{j+1}, W_3^{j+1} such that, $\forall k \in \mathbb{N}$, $\psi^j[\mathrm{vec}(W_1^{j+1})^{\rm T}\ \mathrm{vec}(W_2^{j+1})^{\rm T}\ \mathrm{vec}(W_3^{j+1})^{\rm T}]^{\rm T} = \phi^j$, where $\psi^j \in \mathbb{R}^{s\times(n^2+m^2+mn)}$ and $\phi^j \in \mathbb{R}^s$ are given by:

$$\phi^j = \begin{bmatrix} x_k^{\rm T}Qx_k + x_k^{\rm T}(K^j)^{\rm T}RK^jx_k \\ x_{k+1}^{\rm T}Qx_{k+1} + x_{k+1}^{\rm T}(K^j)^{\rm T}RK^jx_{k+1} \\ \vdots \\ x_{k+s-1}^{\rm T}Qx_{k+s-1} + x_{k+s-1}^{\rm T}(K^j)^{\rm T}RK^jx_{k+s-1} \end{bmatrix},$$

$$\psi^j = \begin{bmatrix} H_1^{\rm xx} & H_1^{\rm xu} & H_1^{\rm uu} \\ H_2^{\rm xx} & H_2^{\rm xu} & H_2^{\rm uu} \\ \vdots & \vdots & \vdots \\ H_s^{\rm xx} & H_s^{\rm xu} & H_s^{\rm uu} \end{bmatrix},$$

with, for $i = 1,\ 2,\ \ldots,\ s$, $H_i^{\rm xx} = x_{k+i-1}^{\rm T} \otimes x_{k+i-1}^{\rm T} - x_{k+i}^{\rm T} \otimes x_{k+i}^{\rm T}$, $H_i^{\rm xu} = 2[x_{k+i-1}^{\rm T} \otimes (u_{k+i-1} + K^jx_{k+i-1})^{\rm T}]$, and $H_i^{\rm uu} = (u_{k+i-1} - K^jx_{k+i-1})^{\rm T} \otimes (u_{k+i-1} + K^jx_{k+i-1})^{\rm T}$.

2. Update policy by: $K^{j+1} = (R + W_3^{j+1})^{-1}W_2^{j+1}$.

3. Stop if $\|K^{j+1} - K^j\| \leqslant \xi$ with a positive threshold ξ. Otherwise let $j = j + 1$ and goes to step 1.

The convergence of Algorithm 1 is proved in [22] by showing that the off-policy Bellman equation (10) is equivalent to the model-based Bellman equation (4). So as j goes to ∞, P^j converges to the optimal kernel matrix $P^\star$, i.e., the solution of the ARE (7), and K^j converges to the optimal controller gain $K^\star$ given by (6). The step 1 in Algorithm 1 is implemented by the LS method $\zeta^{j+1} = ((\psi^j)^{\rm T}\psi^j)^{-1}(\psi^j)^{\rm T}\phi^j$ with $\zeta^{j+1} = [\mathrm{vec}(W_1^{j+1})^{\rm T}\ \mathrm{vec}(W_2^{j+1})^{\rm T}\ \mathrm{vec}(W_3^{j+1})^{\rm T}\ ]^{\rm T}$ and the window size $s \geqslant n^2 + m^2 + mn$. PE conditions are required to guarantee ψ^j has full rank [22].

D. Data-driven Off-policy RL with Clock Offsets

Algorithm 1 shows clock offsets influence the matrices ϕ and ψ. Under clock offsets, the state-input data actually utilized by the learning component are $x_k^l = x_{k+\delta_x(k)}$ and $u_k^l = u_{k+\delta_u(k)}$. Then the matrix formulations under clock offsets are given by: $\hat{\phi}^j =$

$$\begin{bmatrix} (x_k^l)^{\rm T}Qx_k^l + (x_k^l)^{\rm T}(K^j)^{\rm T}RK^jx_k^l \\ (x_{k+1}^l)^{\rm T}Qx_{k+1}^l + (x_{k+1}^l)^{\rm T}(K^j)^{\rm T}RK^jx_{k+1}^l \\ \vdots \\ (x_{k+s-1}^l)^{\rm T}Qx_{k+s-1}^l + (x_{k+s-1}^l)^{\rm T}(K^j)^{\rm T}RK^jx_{k+s-1}^l \end{bmatrix}, \tag{12}$$

$$\hat{\psi}^j = \begin{bmatrix} \hat{H}_1^{\rm xx} & \hat{H}_1^{\rm xu} & \hat{H}_1^{\rm uu} \\ \hat{H}_2^{\rm xx} & \hat{H}_2^{\rm xu} & \hat{H}_2^{\rm uu} \\ \vdots & \vdots & \vdots \\ \hat{H}_s^{\rm xx} & \hat{H}_s^{\rm xu} & \hat{H}_s^{\rm uu} \end{bmatrix}, \tag{13}$$

with, for $i = 1,\ 2,\ \ldots,\ s$,

$$\begin{aligned} \hat{H}_i^{\rm xx} =& (x_{k+i-1}^l)^{\rm T} \otimes (x_{k+i-1}^l)^{\rm T} - (x_{k+i}^l)^{\rm T} \otimes (x_{k+i}^l)^{\rm T}, \\ \hat{H}_i^{\rm xu} =& 2[(x_{k+i-1}^l)^{\rm T} \otimes (u_{k+i-1}^l + K^jx_{k+i-1}^l)^{\rm T}], \\ \hat{H}_i^{\rm uu} =& (u_{k+i-1}^l - K^jx_{k+i-1}^l)^{\rm T} \otimes (u_{k+i-1}^l + K^jx_{k+i-1}^l)^{\rm T}. \end{aligned}$$

Accordingly, Algorithm 1 is replaced by Algorithm 2.

Algorithm 2 Data-driven Off-policy RL with Clock Offsets

Select a stabilizing control policy $u_k = \mu_{\rm b}(x_k)$ for data collection. Set iteration number $j = 0$. Select an initial controller gain K^0 and a proper window size s.

1. Solve for $\hat{W}_1^{j+1}$, $\hat{W}_2^{j+1}$, $\hat{W}_3^{j+1}$ such that, $\forall k \in \mathbb{N}$, $\hat{\psi}^j[\mathrm{vec}(\hat{W}_1^{j+1})^{\rm T}\ \mathrm{vec}(\hat{W}_2^{j+1})^{\rm T}\ \mathrm{vec}(\hat{W}_3^{j+1})^{\rm T}]^{\rm T} = \hat{\phi}^j$, with $\hat{\phi}^j \in \mathbb{R}^s$ and $\hat{\psi}^j \in \mathbb{R}^{s\times(n^2+m^2+mn)}$ given by (12) and (13).

2. Update policy by:

$$\hat{K}^{j+1} = (R + \hat{W}_3^{j+1})^{-1}\hat{W}_2^{j+1}. \tag{14}$$

3. Stop if $\|\hat{K}^{j+1} - \hat{K}^j\| \leqslant \xi$ with a positive threshold ξ. Otherwise let $j = j + 1$ and goes to step 1.

Let $\hat{W}_1^{j+1} = \hat{P}^{j+1}$, $\hat{W}_2^{j+1} = B^{\rm T}\hat{P}^{j+1}A$, $\hat{W}_3^{j+1} = B^{\rm T}\hat{P}^{j+1}B$. The LS estimation of ζ with offsets is given by:

$$\hat{\zeta}^{j+1} = ((\hat{\psi}^j)^{\rm T}\hat{\psi}^j)^{-1}(\hat{\psi}^j)^{\rm T}\hat{\phi}^j. \tag{15}$$

III. Main Results

Now we are interested in: with $\hat{\phi}^j$ and $\hat{\psi}^j$ whether Algorithm 1 still generates stabilizing control policies.

Lemma 1. Assume the PE condition is satisfied when collecting data in Algorithm 2. Then the control policies generated by Algorithm 2 are stabilizing policies given that the norm of the learning gap $\|\epsilon\| = \|\zeta^{j+1} - \hat{\zeta}^{j+1}\|$ is sufficiently small.

Proof. Define the $\mathcal{H}$ operator as $\mathcal{H}(K, P, A, B, Q, R, x_k^l) = (x_k^l)^{\rm T}[(A-BK)^{\rm T}P(A-BK)+Q+K^{\rm T}RK]x_k^l$. By model-based Bellman equation (4), $\mathcal{H}(K, P, A, B, Q, R, x_k^l) = (x_k^l)^{\rm T}Px_k^l$. In the $j-$th iteration during the learning process, based on (8) from the model-based RL Algorithm 1, we have $\mathcal{H}(K^j, P^{j+1}, A, B, Q, R, x_k^l) = (x_k^l)^{\rm T}P^{j+1}x_k^l$. It follows from Rayleigh-Ritz inequality for symmetric matrices that $\underline{\lambda}(P^{j+1} - \hat{P}^{j+1})\|x_k^l\|^2 \leqslant (x_k^l)^{\rm T}P^{j+1}x_k^l - (x_k^l)^{\rm T}\hat{P}^{j+1}x_k^l \leqslant \overline{\lambda}(P^{j+1} - \hat{P}^{j+1})\|x_k^l\|^2$. Then we have:

$$|(x_k^l)^{\rm T}P^{j+1}x_k^l - (x_k^l)^{\rm T}\hat{P}^{j+1}x_k^l| \leqslant \varepsilon_1, \tag{16}$$

with $\varepsilon_1 = \max(|\underline{\lambda}(P^{j+1} - \hat{P}^{j+1})|\|x_k^l\|^2,\ |\overline{\lambda}(P^{j+1} - \hat{P}^{j+1})|\|x_k^l\|^2)$. Let $V^{j+1} = (A - BK^j)^{\rm T}P^{j+1}(A - BK^j) + (K^j)^{\rm T}RK^j$ and $\hat{V}^{j+1} = (A - B\hat{K}^j)^{\rm T}\hat{P}^{j+1}(A - B\hat{K}^j) + (\hat{K}^j)^{\rm T}R\hat{K}^j$. Following the same logic, we have $\underline{\lambda}(V^{j+1} - \hat{V}^{j+1})\|x_k^l\|^2 \leqslant \mathcal{H}(K^j, P^{j+1}, A, B, Q, R, x_k^l) - \mathcal{H}(\hat{K}^j, \hat{P}^{j+1}, A, B, Q, R, x_k^l) \leqslant \overline{\lambda}(V^{j+1} - \hat{V}^{j+1})\|x_k^l\|^2$.

$$\begin{aligned} &|\mathcal{H}(K^j, P^{j+1}, A, B, Q, R, x_k^l) \\ &\quad - \mathcal{H}(\hat{K}^j, \hat{P}^{j+1}, A, B, Q, R, x_k^l)| \leqslant \varepsilon_2, \end{aligned} \tag{17}$$

with $\varepsilon_2 \leqslant \max(|\underline{\lambda}(V^{j+1} - \hat{V}^{j+1})|, |\overline{\lambda}(V^{j+1} - \hat{V}^{j+1})|)\|x_k^l\|^2$. According to (5)-(6), the update control policy (14) in each iteration of Algorithm 2 is actually the minimizer of $\mathcal{H}$ operator, i.e., $\hat{K}^{j+1} = \arg\min_{\hat{K}} \mathcal{H}(\hat{K}, \hat{P}^{j+1}, A, B, Q, R, x_k^l)$. Given bounds ε_1 and ε_2 due to clock offsets, $\mathcal{H}(\hat{K}^{j+1}, \hat{P}^{j+1}, A, B, Q, R, x_k^l) < \mathcal{H}(\hat{K}^j, \hat{P}^{j+1}, A, B, Q, R, x_k^l) \leqslant$

$\mathcal{H}(K^j, P^{j+1}, A, B, Q, R, x_k^l) + \varepsilon_2 = (x_k^l)^{\mathrm{T}} P^{j+1} x_k^l + \varepsilon_2 = (x_k^l)^{\mathrm{T}} \hat{P}^{j+1} x_k^l + \varepsilon_1 + \varepsilon_2$. Then $\mathcal{H}(\hat{K}^{j+1}, \hat{P}^{j+1}, A, B, Q, R, x_k^l) = (x_k^l)^{\mathrm{T}}[(A - B\hat{K}^{j+1})^{\mathrm{T}} \hat{P}^{j+1}(A - B\hat{K}^{j+1}) + Q + (\hat{K}^{j+1})^{\mathrm{T}} R \hat{K}^{j+1}] x_k^l < (x_k^l)^{\mathrm{T}} \hat{P}^{j+1} x_k^l + \varepsilon_1 + \varepsilon_2$. With $Q \geq 0$ and $R \succ 0$, for sufficiently small learning gap ϵ and thus sufficiently small ε_1 and ε_2, $(A - B\hat{K}^{j+1})^{\mathrm{T}} \hat{P}^{j+1}(A - B\hat{K}^{j+1}) < \hat{P}^{j+1}$, which implies the largest absolute value of eigenvalues of $A - B\hat{K}^{j+1}$ is smaller than 1. So $\hat{K}^{j+1}$ is stabilizing. ■

Corollary 1. Given the bound ε_1 described by (16) and the bound ε_2 described by (17) due to the learning gap ϵ from clock offsets, assume that $\varepsilon_1 + \varepsilon_2 < (x_k^l)^{\mathrm{T}}[Q + (\hat{K}^{j+1})^{\mathrm{T}} R \hat{K}^{j+1}] x_k^l$ holds at each iteration of the learning process. Then the control policies generated by Algorithm 2 with clock offsets are stabilizing policies.

Proof. It is shown in the proof of *Lemma* 1 that $(x_k^l)^{\mathrm{T}}[(A - B\hat{K}^{j+1})^{\mathrm{T}} \hat{P}^{j+1}(A - B\hat{K}^{j+1}) + Q + (\hat{K}^{j+1})^{\mathrm{T}} R \hat{K}^{j+1}] x_k^l < (x_k^l)^{\mathrm{T}} \hat{P}^{j+1} x_k^l + \varepsilon_1 + \varepsilon_2$. It follows from $\varepsilon_1 + \varepsilon_2 < (x_k^l)^{\mathrm{T}}[Q + (\hat{K}^{j+1})^{\mathrm{T}} R \hat{K}^{j+1}] x_k^l$ that $(A - B\hat{K}^{j+1})^{\mathrm{T}} \hat{P}^{j+1}(A - B\hat{K}^{j+1}) < \hat{P}^{j+1}$. ■

Based on *Lemma* 1, now we derive the direct relationship between clock offsets and the performance of off-policy RL.

Theorem 1. Assume the PE condition is satisfied given the collection of data for Algorithm 2. Then the control policies generated by Algorithm 2 are stabilizing policies given that $\|A^{\delta_x(k)} - I_n\|\mathcal{X} + \|\Sigma_{i=0}^{\delta_x(k)-1} A^i B\|\mathcal{U}$ is sufficient small for $\delta_x(k) > 0$, or $\|A^{\delta_x(k)} - I_n\|\mathcal{X} + \|\Sigma_{i=0}^{-\delta_x(k)-1} A^{i+\delta_x(k)} B\|\mathcal{U}$ is sufficient small for $\delta_x(k) < 0$, where $\mathcal{X}$ is the upper bound of state norm and $\mathcal{U}$ is the upper bound of input norm.

Proof. It follows from $\psi^j \zeta^{j+1} = \phi^j$ and $\hat{\psi}^j \hat{\zeta}^{j+1} = \hat{\phi}^j$ at the j-th iteration of Algorithm 1 and Algorithm 2 that $\hat{\psi}^j \epsilon^{j+1} = \Delta\psi^j \zeta^{j+1} - \Delta\phi^j$ with $\Delta\phi^j = \phi^j - \hat{\phi}^j$, $\Delta\psi^j = \psi^j - \hat{\psi}^j$, $\epsilon^{j+1} = \zeta^{j+1} - \hat{\zeta}^{j+1}$. ϵ is the learning gap which is the difference between the real intermediate variable ζ from Algorithm 1 and the corrupted one $\hat{\zeta}$ from Algorithm 2 under clock offsets. Given that $\hat{\psi}^j$ is invertible under PE condition, if we could prove $\Delta\psi^j \zeta^{j+1}$ and $\Delta\phi^j$ are sufficiently small under certain conditions and so $\|\epsilon^{j+1}\|$ is, then based on *Lemma* 1 Algorithm 2 could generate stabilizing control policies. Consider the difference of state-input data generated with clock offsets and without clock offsets first. Manipulating dynamics (1) recursively to get $x_k = A^k x_0 + \Sigma_{i=0}^{k-1} A^i B u_{k-1-i}$. Accordingly, state data with clock offset $\delta_x(k)$ are given by $x_k^l = x_{k+\delta_x(k)} = A^{k+\delta_x(k)} x_0 + \Sigma_{i=0}^{k+\delta_x(k)-1} A^i B u_{k+\delta_x(k)-1-i}$.

For the case of $\delta_x(k) > 0$, we have $x_k^l - x_k = (A^{\delta_x(k)} - I_n) x_k + \Sigma_{i=0}^{\delta_x(k)-1} A^i B u_{k+\delta_x(k)-1-i}$. The second term on the right-hand side contains a summation of finite terms and thus is bounded. So the norm of difference between state data with clock offsets and those without clock offsets are bounded by:

$$\begin{aligned}
&\|x_k^l - x_k\| \\
&\leqslant \|A^{\delta_x(k)} - I_n\|\|x_k\| + \|\Sigma_{i=0}^{\delta_x(k)-1} A^i B u_{k+\delta_x(k)-1-i}\| \\
&\leqslant \|A^{\delta_x(k)} - I_n\|\mathcal{X} + \|\Sigma_{i=0}^{\delta_x(k)-1} A^i B\|\mathcal{U}, \qquad (18)
\end{aligned}$$

where $\mathcal{X}$ is upper bound of state norm, $\mathcal{U}$ is the upper bound of input norm. Hence, if $\|A^{\delta_x(k)} - I_n\|\mathcal{X} + \|\Sigma_{i=0}^{\delta_x(k)-1} A^i B\|\mathcal{U}$ is sufficiently small, then $\|x_k^l - x_k\|$ is sufficiently small.

Likewise, for the case of $\delta_x(k) < 0$, we have:

$$\begin{aligned}
&\|x_k^l - x_k\| \\
&= \|(A^{\delta_x(k)} - I_n) x_k - \Sigma_{i=0}^{-\delta_x(k)-1} A^{i+\delta_x(k)} B u_{k-1-i}\| \\
&\leqslant \|A^{\delta_x(k)} - I_n\|\|x_k\| + \|\Sigma_{i=0}^{-\delta_x(k)-1} A^{i+\delta_x(k)} B u_{k-1-i}\| \\
&\leqslant \|A^{\delta_x(k)} - I_n\|\mathcal{X} + \|\Sigma_{i=0}^{-\delta_x(k)-1} A^{i+\delta_x(k)} B\|\mathcal{U}. \qquad (19)
\end{aligned}$$

Combine (18) and (19) to get:

$$\begin{aligned}
&\|x_k^l - x_k\| \leqslant \|A^{\delta x(k)} - I_n\|\mathcal{X} \\
&+ \max(\|\Sigma_{i=0}^{\delta_x(k)-1} A^i B\|, \|\Sigma_{i=0}^{-\delta_x(k)-1} A^{i+\delta_x(k)} B\|)\mathcal{U}. \qquad (20)
\end{aligned}$$

For the stabilizing control policy $\mu_{\mathrm{b}}(\cdot)$ during data collection phase, let state corresponding to $u_k^l = u_{k+\delta_u(k)}$ be $\bar{x}_k$. Denote $\bar{x}_k = x_{k+\delta_{\bar{u}}(k)}$, where $\delta_{\bar{u}}(k)$ is a function of $\delta_u(k)$. According to Lipschitz continuity and (20), we have $\|u_k^l - u_k\| = \|\mu_{\mathrm{b}}(\bar{x}_k) - \mu_{\mathrm{b}}(x_k)\| \leqslant L\|x_{k+\delta_{\bar{u}}(k)} - x_k\|$ with L a Lipschitz constant of the function $\mu_{\mathrm{b}}(\cdot)$. Thus, if $\|A^{\delta x(k)} - I_n\|\mathcal{X} + \max(\|\Sigma_{i=0}^{\delta_x(k)-1} A^i B\|, \|\Sigma_{i=0}^{-\delta_x(k)-1} A^{i+\delta_x(k)} B\|)\mathcal{U}$ is sufficiently small, then $\|u_k^l - u_k\|$ is sufficiently small.

For the i-th row and 1st column term of $\Delta\psi^j \zeta^{j+1}$, where $1 \leqslant i \leqslant s$, we have $(H_i^{\mathrm{xx}} - \hat{H}_i^{\mathrm{xx}})\mathrm{vec}(W_1^{j+1}) = x_{k+i-1}^{\mathrm{T}} W_1^{j+1} x_{k+i-1} - x_{k+i}^{\mathrm{T}} W_1^{j+1} x_{k+i} - [(x_{k+i-1}^l)^{\mathrm{T}} W_1^{j+1} x_{k+i-1}^l - (x_{k+i}^l)^{\mathrm{T}} W_1^{j+1} x_{k+i}^l] = (x_{k+i-1} - x_{k+i-1}^l)^{\mathrm{T}} W_1^{j+1} \left(x_{k+i-1} + (x_{k+i-1}^l)\right) + (x_{k+i} - x_{k+i}^l)^{\mathrm{T}} W_1^{j+1} \left(x_{k+i} + (x_{k+i}^l)\right)$. It follows that $\|(H_i^{\mathrm{xx}} - \hat{H}_i^{\mathrm{xx}})\mathrm{vec}(W_1^{j+1})\| \leqslant 2(\|x_{k+i-1} - x_{k+i-1}^l)^{\mathrm{T}}\| + \|(x_{k+i} - x_{k+i}^l)^{\mathrm{T}}\|)\|W_1^{j+1}\|\mathcal{X}$. If $\|x_k^l - x_k\|$ is sufficiently small, so $(H_i^{\mathrm{xx}} - \hat{H}_i^{\mathrm{xx}})\mathrm{vec}(W_1^{j+1})$ is.

For the i-th row and 2nd column term of $\Delta\psi^j \zeta^{j+1}$, where $1 \leqslant i \leqslant s$, $\|(H_i^{\mathrm{xu}} - \hat{H}_i^{\mathrm{xu}})\mathrm{vec}(W_2^{j+1})\| = \|2[u_{k+i-1}^{\mathrm{T}} W_2^{j+1} x_{k+i-1} - (u_{k+i-1}^l)^{\mathrm{T}} W_2^{j+1} x_{k+i-1}^l] + 2[x_{k+i-1}^{\mathrm{T}} (K^j)^{\mathrm{T}} W_2^{j+1} x_{k+i-1} - (x_{k+i-1}^l)^{\mathrm{T}} (K^j)^{\mathrm{T}} W_2^{j+1} x_{k+i-1}^l]\| \leqslant 2\mathcal{U}\|W_2^{j+1}\| \cdot \|x_{k+i-1} - x_{k+i-1}^l\| + 2\mathcal{X}\|(W_2^{j+1})^{\mathrm{T}}\| \cdot \|u_{k+i-1} - u_{k+i-1}^l\| + 2\mathcal{X}\|(K^j)^{\mathrm{T}} W_2^{j+1}\| \cdot \|x_{k+i-1} - x_{k+i-1}^l\| + 2\mathcal{X}\|(W_2^{j+1})^{\mathrm{T}} K^j\| \cdot \|x_{k+i-1} - x_{k+i-1}^l\|$. If $\|x_k^l - x_k\|$ is sufficiently small, then $(H_i^{\mathrm{xu}} - \hat{H}_i^{\mathrm{xu}})\mathrm{vec}(W_2^{j+1})$ is sufficiently small.

For the i-th row and 3rd column term of $\Delta\psi^j \zeta^{j+1}$, where $1 \leqslant i \leqslant s$, $\|(H_i^{\mathrm{uu}} - \hat{H}_i^{\mathrm{uu}})\mathrm{vec}(W_3^{j+1})\| = \|[u_{k+i-1}^{\mathrm{T}} W_3^{j+1} u_{k+i-1} - (u_{k+i-1}^l)^{\mathrm{T}} W_3^{j+1} u_{k+i-1}^l] - [x_{k+i-1}^{\mathrm{T}} (K^j)^{\mathrm{T}} W_3^{j+1} K^j x_{k+i-1} - (x_{k+i-1}^l)^{\mathrm{T}} (K^j)^{\mathrm{T}} W_3^{j+1} K^j x_{k+i-1}^l] - [u_{k+i-1}^{\mathrm{T}} W_3^{j+1} K^j x_{k+i-1} - (u_{k+i-1}^l)^{\mathrm{T}} W_3^{j+1} K^j x_{k+i-1}^l] + [x_{k+i-1}^{\mathrm{T}} (K^j)^{\mathrm{T}} W_3^{j+1} u_{k+i-1} - (x_{k+i-1}^l)^{\mathrm{T}} (K^j)^{\mathrm{T}} W_3^{j+1} u_{k+i-1}^l]\| \leqslant 2\mathcal{U} W_3^{j+1}\| \cdot \|(u_{k+i-1} - u_{k+i-1}^l)\| + 2\mathcal{X}\|(K^j)^{\mathrm{T}} W_3^{j+1} K^j\|\|(x_{k+i-1} - x_{k+i-1}^l)\| + 2\mathcal{U}\|W_3^{j+1} K^j\| \cdot \|(x_{k+i-1} - x_{k+i-1}^l)\| + 2\mathcal{X}\|(K^j)^{\mathrm{T}} W_3^{j+1}\| \cdot \|(u_{k+i-1} - u_{k+i-1}^l)\|$. If $\|x_k^l - x_k\|$ is sufficiently small, $(H_i^{\mathrm{uu}} - \hat{H}_i^{\mathrm{uu}})\mathrm{vec}(W_3^{j+1})$ is sufficiently small.

For the i-th term of $\Delta\phi^j$, where $1 \leqslant i \leqslant s$, $\|\Delta\phi_i^j\| = \|x_{k+i-1}^{\mathrm{T}} Q x_{k+i-1} + x_{k+i-1}^{\mathrm{T}} (K^j)^{\mathrm{T}} R K^j x_{k+i-1}\| - (x_{k+i-1}^l)^{\mathrm{T}} Q x_{k+i-1}^l - (x_{k+i-1}^l)^{\mathrm{T}} (\hat{K}^j)^{\mathrm{T}} R \hat{K}^j x_{k+i-1}^l \leqslant 2\mathcal{X}\|Q + (K^j)^{\mathrm{T}} R K^j\| \|x_{k+i-1} - x_{k+i-1}^l\| + \bar{\lambda}(Q + (K^j)^{\mathrm{T}} R K^j) \|x_{k+i-1} - x_{k+i-1}^l\|^2$. If $\|x_k^l - x_k\|$ is sufficiently small, each term in $\Delta\phi^j$ is sufficiently small. ■

Remark 1. For the sufficient condition of generating stabilizing control policies by *Theorem* 1, the first term $\|A^{\delta_x(k)} - I_n\|\mathcal{X}$ indicates the impact of clock offsets depends on the deviation of dynamics change $A^{\delta_x(k)}$ from identity matrix I_n. The second term $\|\Sigma_{i=0}^{\delta_x(k)-1} A^i B\|\mathcal{U}$ implies the impact depends on the control input during the duration of the non synchronization. If the closed-loop dynamics do not change too fast and clock offsets are small enough, the off-policy RL with clock offsets could still generate stabilizing control policies. The learning gap ϵ reflects the combined effects of system dynamics and clock offsets. □

To analyze the performance loss due to clock offsets, denote the optimal value by $V^\star(x_k) = x_k^{\mathrm{T}} P^\star x_k$ without clock offsets and by $\hat{V}(x_k^l) = (x_k^l)^{\mathrm{T}} \hat{P} x_k^l$ with clock offsets.

Theorem 2. Given a linear state feedback behavior policy K_{b}, and an invertible matrix $(A - BK_{\mathrm{b}})$, under clock offsets $\delta_x(k)$ and $\delta_u(k)$, the performance loss is bounded as $|V^\star(x_k) - \hat{V}(x_k^l)| \leqslant \tilde{\epsilon}\|x_k\|^2$, where $\tilde{\epsilon} = \max(|\underline{\lambda}(P^\star - \hat{P})|, |\overline{\lambda}(P^\star - \hat{P})|) + \max(|\underline{\lambda}(\mathcal{P})|, |\overline{\lambda}(\mathcal{P})|)$ with $\mathcal{P} = \hat{P} - ((A - BK_{\mathrm{b}})^{\delta_x(k)})^{\mathrm{T}} \hat{P} (A - BK_{\mathrm{b}})^{\delta_x(k)}$.

Proof. One can write $|V^\star(x_k) - \hat{V}(x_k^l)| = |x_k^{\mathrm{T}} P^\star x_k - (x_k^l)^{\mathrm{T}} \hat{P} x_k^l| \leqslant |x_k^{\mathrm{T}} P^\star x_k - x_k^{\mathrm{T}} \hat{P} x_k| + |x_k^{\mathrm{T}} \hat{P} x_k - (x_k^l)^{\mathrm{T}} \hat{P} x_k^l|$. Given the behavior policy K_{b}, $x_k^l = (A - BK_{\mathrm{b}})^{\delta_x(k)} x_k$. Denote $\mathcal{P} := \hat{P} - \left((A - BK_{\mathrm{b}})^{\delta_x(k)}\right)^{\mathrm{T}} \hat{P} (A - BK_{\mathrm{b}})^{\delta_x(k)}$. Based on Rayleigh-Ritz inequality for symmetric matrices that $|V^\star(x_k) - \hat{V}(x_k^l)| \leqslant |x_k^{\mathrm{T}} P^\star x_k - x_k^{\mathrm{T}} \hat{P} x_k| + |x_k^{\mathrm{T}} \hat{P} x_k - (x_k^l)^{\mathrm{T}} \hat{P} x_k^l| = |x_k^{\mathrm{T}} P^\star x_k - x_k^{\mathrm{T}} \hat{P} x_k| + |x_k^{\mathrm{T}} \hat{P} x_k - x_k^{\mathrm{T}} \left((A - BK_{\mathrm{b}})^{\delta_x(k)}\right)^{\mathrm{T}} \hat{P} (A - BK_{\mathrm{b}})^{\delta_x(k)} x_k| \leqslant \max(|\underline{\lambda}(P^\star - \hat{P})|, |\overline{\lambda}(P^\star - \hat{P})|)\|x_k\|^2 + \max(|\underline{\lambda}(\mathcal{P})|, |\overline{\lambda}(\mathcal{P})|)\|x_k\|^2$. ■

IV. Simulation Results

Consider the third-order F-16 autopilot aircraft plant [23]:

$$x_{k+1} = \begin{bmatrix} 0.9065 & 0.0816 & -0.0005 \\ 0.0743 & 09012 & -0.0007 \\ 0 & 0 & 0.1327 \end{bmatrix} x_k + \begin{bmatrix} -0.0027 \\ -0.0068 \\ 1 \end{bmatrix} u_k,$$

where the states are $x_k = [\alpha, q, \delta_e]^{\mathrm{T}}$ with α the angle of attack, q the pitch rate, and δ_e the elevator deflection angle, and u is the elevator actuator voltage. Let $Q = I_n$, $R = I_m$, the initial state $x_0 = [10, -10, -3]^{\mathrm{T}}$ and the behavior policy $K_{\mathrm{b}} = [0\ 0.12\ 1]$. To ensure data richness, probing noise $e_k = \sin^2(0.5k) + \sin(k) + \cos(k)$ is added to the system for 50 time steps for data collection. Consider clock offsets between the controller and the learning component. The control input signals received by the learning component are actually $u_{k+\delta_u(k)}$ at k with $\delta_u = \{-1, -2, -3, -4, -5, -6, -7\}$. The data used for Algorithm 2 begin from the time step $k = 8$ with window size $s = 16$. The learning results under clock offsets and state trajectories with learned policies are shown in Figure 2 and Figure 3, respectively. Figure 2 shows that all controller gains K^j and kernel matrices P^j converge in the seven clock offset cases. However, in Figure 3 the learned controller with clock offset $\delta_u = 1$ makes the system unstable and states explode while the learned controller gain with bigger clock offset $\delta_u = 7$ stabilizes the system. As implied from *Theorem* 1, bigger values of clock offsets do not necessarily lead to generating non-stabilizing control policies. Both system dynamics and clock offsets determine the influence. Section III shows the learning gap ϵ from (15) directly determines the influence of clock offsets on learning algorithms. Learning gaps for the seven clock offset cases are presented in Table I. Note that the learning gap for clock offset $\delta_u = 1$ is way much bigger than other cases, which explains the stabilizing differences of learned controller policies among these cases. According to (15), which part of data used for the data-driven off-policy RL Algorithm 2 has an impact on the learning gap ϵ under clock offsets. For example, if the data used for Algorithm 3 begin from $k = 8$ with window size $s = 40$, then for clock offset $\delta_u = 1$, the learned controller is a stabilizing policy as shown in Figure 4, as implied by *Theorem* 1 that both system dynamics and the magnitude of clock offsets together influence the learning behavior.

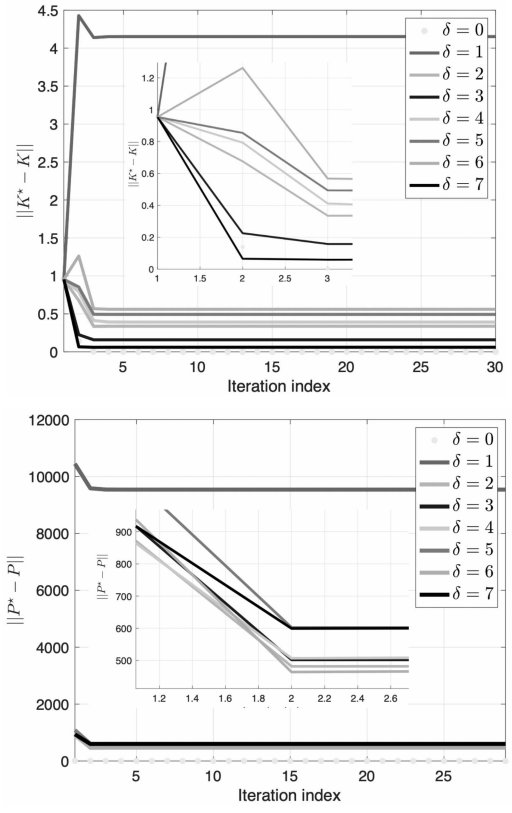

Fig. 2. Learning process of Algorithm 3 under various clock offsets. The inner plots inside display magnified views of the initial iteration stage.

V. Conclusion and Future Work

In this work, we investigate the impact of clock offsets among different components of CPS on the data-driven

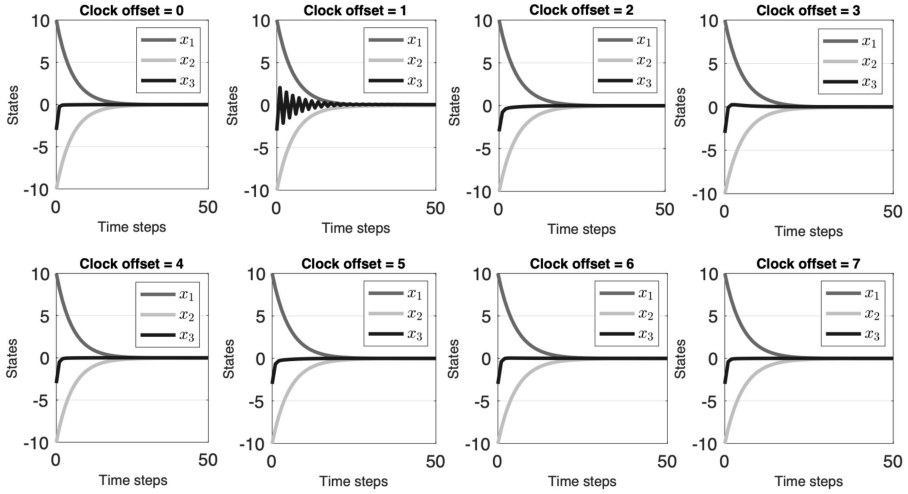

Fig. 3. State trajectories with learned policies under clock offsets on input with window size $s = 16$.

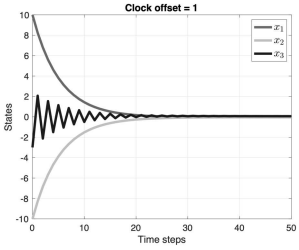

Fig. 4. State trajectory with the learned policy under $\delta_u = 1$ with $s = 40$.

TABLE I

LEARNING GAPS WITH CLOCK OFFSETS OF CONTROL INPUT SIGNALS

δ_u	1	2	3	4	5	6	7
ϵ	8261	417	435	441	522	404	520

off-policy RL algorithm. Certain conditions are derived to determine the policies generated by data-driven off-policy RL with clock offsets are stabilizing policies. With clock offsets, what directly influences the learning behavior is not only the values of clock offsets but also the dynamics change caused by clock offsets. Future work will focus on the influence of clock offsets on other learning algorithms.

REFERENCES

[1] A. Shrivastava, P. Derler, Y.-S. L. Baboud, K. Stanton, M. Khayatian, H. A. Andrade, M. Weiss, J. Eidson, and S. Chandhoke, "Time in cyber-physical systems," in *Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, 2016, pp. 1–10.

[2] R. Singh and V. Gupta, "On LQR control with asynchronous clocks," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE, 2011, pp. 3148–3153.

[3] K. Okano, M. Wakaiki, G. Yang, and J. P. Hespanha, "Stabilization of networked control systems under clock offsets and quantization," *IEEE Transactions on Automatic Control*, vol. 63, no. 6, pp. 1708–1723, 2017.

[4] R. Song, H. Zhang, Y. Luo, and Q. Wei, "Optimal control laws for time-delay systems with saturating actuators based on heuristic dynamic programming," *Neurocomputing*, vol. 73, no. 16-18, pp. 3020–3027, 2010.

[5] T. Fujita and T. Ushio, "Rl-based optimal networked control considering network delay of discrete-time linear systems," in *2015 European Control Conference (ECC)*. IEEE, 2015, pp. 2476–2481.

[6] W. Gao and Z.-P. Jiang, "Adaptive optimal output regulation of time-delay systems via measurement feedback," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 3, pp. 938–945, 2018.

[7] A. Heydari, "Theoretical and numerical analysis of approximate dynamic programming with approximation errors," *Journal of Guidance, Control, and Dynamics*, vol. 39, no. 2, pp. 301–311, 2016.

[8] ——, "Stability analysis of optimal adaptive control using value iteration with approximation errors," *IEEE Transactions on Automatic Control*, vol. 63, no. 9, pp. 3119–3126, 2018.

[9] B. Pang, T. Bian, and Z.-P. Jiang, "Robust policy iteration for continuous-time linear quadratic regulation," *IEEE Transactions on Automatic Control*, vol. 67, no. 1, pp. 504–511, 2021.

[10] B. Pang and Z.-P. Jiang, "Robust reinforcement learning: A case study in linear quadratic regulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9303–9311.

[11] T. Bian, Y. Jiang, and Z.-P. Jiang, "Adaptive dynamic programming for stochastic systems with state and control dependent noise," *IEEE Transactions on Automatic control*, vol. 61, no. 12, pp. 4170–4175, 2016.

[12] F. A. Yaghmaie and F. Gustafsson, "Using reinforcement learning for model-free linear quadratic control with process and measurement noises," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 6510–6517.

[13] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

[14] J. A. Chekan and C. Langbort, "Regret bounds for LQ adaptive control under database attacks (extended version)," *arXiv preprint arXiv:2004.00241*, 2020.

[15] J. Wang, Y. Liu, and B. Li, "Reinforcement learning with perturbed rewards," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6202–6209.

[16] M. Mohammadi, M. M. Arefi, P. Setoodeh, and O. Kaynak, "Optimal tracking control based on reinforcement learning value iteration algorithm for time-delayed nonlinear systems with external disturbances and input constraints," *Information Sciences*, vol. 554, pp. 84–98, 2021.

[17] G. Wang, B. Luo, and S. Xue, "Integral reinforcement learning-based optimal output feedback control for linear continuous-time systems with input delay," *Neurocomputing*, vol. 460, pp. 31–38, 2021.

[18] Y. Yang, Z. Guo, H. Xiong, D.-W. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 12, pp. 3735–3747, 2019.

[19] Y. Wen, H. Zhang, H. Ren, and K. Zhang, "Off-policy based adaptive dynamic programming method for nonzero-sum games on discrete-time system," *Journal of the Franklin Institute*, vol. 357, no. 12, pp. 8059–8081, 2020.

[20] F. Fotiadis, A. Kanellopoulos, K. G. Vamvoudakis, and J. Hugues, "Impact of sensor and actuator clock offsets on reinforcement learning," in *2022 American Control Conference (ACC)*. IEEE, 2022, pp. 2669–2674.

[21] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.

[22] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, "H∞ control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, pp. 144–152, 2017.

[23] Y. Yang, B. Kiumarsi, H. Modares, and C. Xu, "Model-free λ-policy iteration for discrete-time linear quadratic regulation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.