2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

# Impact of Sensor and Actuator Clock Offsets on Reinforcement Learning

Filippos Fotiadis[1], *Student Member, IEEE*, Aris Kanellopoulos[1], *Student Member, IEEE*, Kyriakos G. Vamvoudakis[1], *Senior Member, IEEE*, Jérôme Hugues[2], *Member, IEEE*

***Abstract*— In this work, we investigate the effect of sensor-actuator clock offsets on reinforcement learning (RL) enabled cyber-physical systems. In particular, we consider an off-policy RL algorithm that receives data both from the system's sensors and actuators, and uses them to approximate a desired optimal control policy. Nevertheless, owing to timing mismatches, the control-state data obtained from these system components are inconsistent, hence creating the question of how robust RL will be. After an extensive analysis, we show that RL does retain its robustness, in an epsilon-delta sense; given that the sensor-actuator clock offsets are not arbitrarily large, and that the behavioral control input satisfies a Lipschitz continuity condition, RL converges epsilon-close to the desired optimal control policy. Simulations are carried out on a two-link manipulator, which clarify and verify theoretical findings.**

## I. Introduction

Cyber-physical systems (CPS) are large-scale, complex platforms consisting of multiple sensing and actuating elements that are tightly interconnected. From military applications to a variety of civilian ones—such as those related to the healthcare industry [1], autonomous vehicles [2] and the smart grid [3]—CPS are becoming increasingly important to society. Furthermore, due to their operating in human-centric environments, CPS must be safe and secure by design.

Methods for the development of safe-by-design systems has been mostly focused on the quality of the information in the network, i.e., in the mitigation of corrupted signals either due to stochastic faults [4] or due to malicious manipulation by adversarial agents [5]. However, the decentralized nature of a CPS requires the development of methods that take into account timing discrepancies among its components. Issues of timing have been addressed in control systems in order to assess the robustness of their stability properties to such faults [6]; yet, the effects of timing issues on learning mechanisms are rarely considered. Motivated by this fact, the purpose of the present study is to investigate the behavior of a system with RL capabilities under clock offsets. Our main focus is the derivation of guarantees of convergence for the corresponding learning algorithm, given that the CPS suffers from discrepancies in the control and measurement time-stamps.

*Related Work:* RL methods for control systems have been investigated extensively, both from the controls and the learning communities. In particular, a plethora of data-driven RL approaches to optimal control problems have been developed, both model-based [7] and model-free [8]. All of these address the issue of solving a Hamilton-Jacobi equation by leveraging data obtained from the trajectories of the system. One of these algorithms, which will be considered in this work, was proposed in [9], where trajectory data are generated by a pre-specified behavioral policy, different from the target one.

The robustness of RL-enabled CPS, as far as faults and attacks are concerned, has been explored in the literature. For instance, the authors in [10] presented a survey of CPS security issues, as well as the corresponding controls defense techniques that can mitigate them. In [11], the problem of a network of agents communicating via a network under persistent adversarial inputs was investigated using game-theoretic results. Similar results have employed optimization techniques to derive algorithms that take into account injected signals in CPS, such as [12], where an optimal control problem was solved to detect adversarial signals in the system. All of these existing results, however, consider the effect of erroneous information to the CPS itself, rather than to its learning mechanisms. Additionally, while the effect that data manipulation or faults can have on a learning-based system has been studied before (i.e., as in [13]), the effect of timing discrepancies on RL is usually not considered.

The advent of networked and distributed systems has prompted the control community to investigate the topic of clock mismatches. For example, the authors of [14] consider the use of linear feedback controllers for the stabilization of systems with clock offsets between the sensors and the controller. Specifically, sufficient conditions are derived, under which a stabilizing controller exists. In the same line of research, in [15], the dual effect of offsets and quantization is explored. In [16], the authors model the time perceived by the controller as a stochastic process with respect to real, or "calendar," time. Subsequently, they design feedback controllers based on dynamic programming principles. Finally, in [17], the authors demonstrate the loss of optimality of a linear quadratic regulator for a linear control system under clock mismatches. It is worth noting that the aforementioned approaches explored the effects of timing errors to the controlled system itself, rather than to a learning algorithm used to derive the controller; the latter is the purpose of the present work.

[1]F. Fotiadis, A Kanellopoulos, and K. G. Vamvoudakis are with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Email: {ffotiadis, ariskan, kyriakos}@gatech.edu.

[2]J. Hugues is with the Carnegie Mellon University/Software Engineering Institute, Pittsburgh, PA, USA. Email: jjhugues@sei.cmu.edu.

This work was supported in part, by ARO under grant No. W911NF-19 − 1 − 0270, by NSF under grant Nos. CAREER CPS-1851588 and SATC-1801611, 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. DM21-0828

TABLE I

CLOCK MISMATCHES.

| Actual signal | Actuators' Perception | Sensors' Perception |
|---|---|---|
| $u(t)$ | $u(t)$ | $u(t-\delta(t))$ |
| $x(t)$ | $x(t+\delta(t))$ | $x(t)$ |

*Contributions:* The contributions of this paper are threefold. First, we formulate the problem of data-based RL for optimal control, where the system suffers from sensor-actuator clock discrepancies. Subsequently, we derive an off-policy RL algorithm, which depends on inconsistent state-input data to approximate the desired optimal controller. Finally, we prove that, despite the sensor-actuator clock offsets, convergence of the RL algorithm can be guaranteed in an epsilon-delta sense, given some continuity assumptions.

*Notation:* As $I_n$, we will denote the identity matrix of order $n \times n$. For any two matrices $Z_1$ and $Z_2$, $Z_1 \otimes Z_2$ will denote the Kronecker product of $Z_1$ and $Z_2$. In addition, $\text{vec}(Z_1)$ will denote the vectorized form of $Z_1$.

## II. PROBLEM FORMULATION AND PRELIMINARIES

### A. Sensor-Actuator Clock Mismatches

Consider, for all $t \geqslant t_0 \geqslant 0$, the nonlinear system

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t),\ x(t_0) = x_0, \qquad (1)$$

where $x(t) \in \mathbb{R}^n$ denotes the state, $u(t) \in \mathbb{R}^m$ is the control input, and $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are the system's drift and input dynamics functions, respectively. The functions $f,\ g$ are assumed to be unknown, and the origin is assumed to be a fixed point of (1) when $u \equiv 0$.

In this work, we will consider that there is an asynchrony between the clocks of the sensors and the actuators of (1). To be more specific, let us assume that the sensors can measure the state $x(t)$ of (1) at the time instant $t \geqslant t_0$. Then, from the perspective of the actuators, the state $x(t)$ was measured at the time instant $t + \delta(t)$, where $\delta(t) \in \mathbb{R}$ is the clock mismatch; if $\delta(t) = 0$ for all $t \geqslant t_0$, then the sensors' and the actuators' clocks are perfectly synchronized. In summary, the actuators' perception of the measured state is given by:

$$\bar{x}(t) := x(t + \delta(t)),\ \forall t \geqslant t_0.$$

On the other hand, if $u(t) \in \mathbb{R}^m$ is the control input at time $t \geqslant t_0$ from the actuators' perspective, then this particular control vector was implemented in the system at time $t - \delta(t)$ from the sensors' perspective. Specifically, the sensors' perception of the control input at time $t \geqslant t_0$ is given by:

$$\bar{u}(t) = u(t - \delta(t)),\ \forall t \geqslant t_0. \qquad (2)$$

Table I summarizes the information regarding the sensor-actuator clock offsets. In what follows, we will study the robustness of off-policy reinforcement learning algorithms for optimal control, with respect to these offsets.

### B. Optimal Control and Policy Iteration

For a given feedback control policy $\mu : \mathbb{R}^n \to \mathbb{R}^m$, let us define the infinite-horizon performance cost functional:

$$J(x_0, \mu) = \int_{t_0}^{\infty} \Big(Q(x(\tau)) + r\left(\mu(x(\tau))\right)\Big) \mathrm{d}\tau. \qquad (3)$$

**Algorithm 1** Policy Iteration

1: Let $i = 0$, $\Omega \subset \mathbb{R}^n$, $\epsilon > 0$, and pick an admissible control policy $\mu_0 \in \Psi(\Omega)$.
2: **repeat**
3: Solve for $V_i$, $\forall x \in \Omega$ with $V_i(0) = 0$, in

$$\nabla V_i^{\mathrm{T}}(x)(f(x) + g(x)\mu_i(x)) + Q(x) + r(\mu_i(x)) = 0.$$

4: Let the new policy be given by

$$\mu_{i+1}(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V_i(x).$$

5: Set $i = i + 1$.
6: **until** $i \geqslant 2\ \ \&\ \ \sup_{x\in\Omega} |V_{i-1}(x) - V_{i-2}(x)| < \epsilon$.

where $Q : \mathbb{R}^n \to \mathbb{R}$ is a known positive definite function, $r(\star) = \star^{\mathrm{T}} R \star$, $R \in \mathbb{R}^{m\times m}$ is a known, positive definite matrix, and the integration in (3) is over the trajectories of (1) under $u(t) = \mu(x(t))$. The integral (3) is well-defined for any $x_0 \in \Omega \subset \mathbb{R}^n$ if the policy $\mu$ is *admissible* on $\Omega$.

**Definition 1.** A control policy $\mu : \mathbb{R}^n \to \mathbb{R}^m$ will be defined as *admissible* on $\Omega$, and denoted as $\mu \in \Psi(\Omega)$, if it is continuous on $\Omega$ with $\mu(0) = 0$, and, given $x_0 \in \Omega$, $u = \mu(x)$ asymptotically stabilizes (1) to the origin and the cost $J(x_0, \mu)$ is finite. □

Given that it is continuously differentiable, the value function $V := J(\cdot, \mu) : \mathbb{R}^n \to \mathbb{R}$ of an admissible policy $\mu$, with $V(0) = 0$, can be found through the nonlinear equation:

$$\nabla V^{\mathrm{T}}(x)(f(x) + g(x)\mu(x)) + Q(x) + r(\mu(x)) = 0,$$

where the argument of time has been omitted to simplify exposition. In addition, the optimal control $\mu^\star$ is given by

$$\mu^\star(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V^\star(x),$$

where $V^\star := J(\cdot, \mu^\star)$ denotes the optimal value function, which satisfies the Hamilton-Jacobi-Bellman (HJB) equation

$$\nabla V^{\star\mathrm{T}}(x)f(x) - \frac{1}{4}\nabla V^{\star\mathrm{T}}(x)g(x)R^{-1}g^{\mathrm{T}}(x)\nabla V^\star(x) + Q(x) = 0,\ V^\star(0) = 0. \qquad (4)$$

The desired solution of (4) is generally difficult to derive analytically. However, Policy Iteration (PI) [7], [18], described in Algorithm 1, can be used to successively approximate it.

### C. Learning-based PI

Although the PI algorithm provides an alternative to directly solving the HJB equation, it requires knowledge of the system's dynamics in order to be executed. To relax this requirement, the authors in [9] have proposed an off-policy learning-based PI which can approximate the optimal value function $V^\star$, without knowing either $f$ or $g$. However, to carry out this algorithm, state and control input data measured along the system's trajectories are required.

More specifically, given an arbitrary control input $u$, the system dynamics (1) can be expressed as

$$\dot{x}(t) = f(x(t)) + g(x(t))\mu_i(x(t)) + g(x(t))(u(t) - \mu_i(x(t))),$$

**Algorithm 2** Learning-based PI

1: Let $i = 0$, $\Omega \subset \mathbb{R}^n$, $\epsilon > 0$, and pick an admissible control policy $\mu_0 \in \Psi(\Omega)$.
2: **repeat**
3: Solve for $V_i$, $\mu_{i+1}$ over $\Omega$ simultaneously from (7).
4: Set $i = i + 1$.
5: **until** $i \geqslant 2$ & $\sup_{x\in\Omega} |V_{i-1}(x) - V_{i-2}(x)| < \epsilon$.

where $\mu_i$ is the control policy at step $i \in \mathbb{N}$ of Algorithm 1. Taking the time derivative of the corresponding value function $V_i$ along the trajectories of (1), one has

$$\dot{V}_i(x(t)) = \nabla V_i^{\mathrm{T}}(x(t))\left(f(x(t)) + g(x(t))\mu_i(x(t))\right) + \nabla V_i^{\mathrm{T}}(x(t))g(x(t))(u(t) - \mu_i(x(t))). \quad (5)$$

Hence, using the equations in Algorithm 1, (5) yields

$$\dot{V}_i(x(t)) = -Q(x(t)) - \mu_i^{\mathrm{T}}(x(t))R\mu_i(x(t)) - 2\mu_{i+1}^{\mathrm{T}}(x(t))R(u(t) - \mu_i(x(t))). \quad (6)$$

Let $t_k > 0$, $k \in \{0, \ldots, K\} := \mathcal{K}$, be measuring time instants and $T > 0$ be a measuring duration. Then, the integration of (6) over $[t_k,\ t_k + T]$ leads to

$$V_i(x(t_k + T)) - V_i(x(t_k)) = -\int_{t_k}^{t_k+T} \Big(Q(x(\tau)) + r(\mu_i(x(\tau))) + 2\mu_{i+1}^{\mathrm{T}}(x(\tau))R(u(\tau) - \mu_i(x(\tau)))\Big)\mathrm{d}\tau,\ k \in \mathcal{K}. \quad (7)$$

Equation (7) provides a model-free way to express the value function $V_i$, hence leading to the learning-based PI algorithm described in Algorithm 2. Effective methods to implement Algorithm 2 using actor-critic networks have been proposed, with convergence guarantees [9], [19]–[21].

### D. Learning-based PI with Sensor-Actuator Clock Offsets

It is evident that Algorithm 2 assumes perfect synchronization between the sensors' and the actuators' clocks, which motivates us to study its behavior with regards to timing issues. In particular, we will assume that the learning-based Algorithm 2 receives, as input, measured state trajectories $x(t)$ from the sensors over the time intervals $t \in [t_k,\ t_k + T]$, for all $k \in \mathcal{K}$. Additionally, it matches these state trajectories with the control input trajectories received from the controller. However, instead of receiving $u(t)$ for all $t \in [t_k,\ t_k + T]$, the learning component receives $\bar{u}(t) = u(t - \delta(t))$ owing to the clock mismatch between the sensors and the actuators. As a consequence, at each step $i \in \mathbb{N}$ of Algorithm 2, instead of learning the function $V_i$ and the policy $\mu_{i+1}$ satisfying (7), one is forced to learn the function $\bar{V}_i$ and the policy $\bar{\mu}_{i+1}$ that satisfy:

$$\bar{V}_i(x(t_k + T)) - \bar{V}_i(x(t_k)) = -\int_{t_k}^{t_k+T} \Big(Q(x(\tau)) + r(\bar{\mu}_i(x(\tau))) + 2\bar{\mu}_{i+1}^{\mathrm{T}}(x(\tau))R(\bar{u}(\tau) - \bar{\mu}_i(x(\tau)))\Big)\mathrm{d}\tau,\ k \in \mathcal{K}. \quad (8)$$

Notice that $\bar{u}$ from (2) has been used, instead of $u$.

In what follows, we will study whether there exists an upper bound to $\delta(t)$ for which (the approximations of) $\bar{V}_i$ and $\bar{\mu}_{i+1}$ are close to $V_i$ and $\mu_{i+1}$ over $\Omega$. In addition, we will investigate whether $\bar{V}_i$ and $\bar{\mu}_{i+1}$ converge close to the optimal solutions $V^\star,\ \mu^\star$.

*Remark* 1. The preceding discussion implicitly assumes that the learning component of the CPS is synchronized with the sensors' clock, but such an assumption is not restricting due to the time-invariant nature of (8). □

## III. Main Results

### A. Learning Scheme

In practice, in order to approximately solve the set of equations (7) for all $k \in \mathcal{K}$, the infinite dimensionality of $V_i$ and $\mu_{i+1}$ needs to be reduced. To this end, recall that, owing to the Weierstrass approximation theorem [22], it holds that

$$V_i(x) = (w_i^v)^{\mathrm{T}}\phi^v(x) + \epsilon_i^v(x),$$
$$\mu_{i+1}(x) = (w_i^u)^{\mathrm{T}}\phi^u(x) + \epsilon_i^u(x),$$

where $w_i^v \in \mathbb{R}^{N_v}$, $w_i^u \in \mathbb{R}^{N_u \times m}$ are weights, $\phi^v : \mathbb{R}^n \to \mathbb{R}^{N_v}$, $\phi^u : \mathbb{R}^n \to \mathbb{R}^{N_u}$ are basis functions such that $\phi^v(0) = \phi^u(0) = 0$, and $\epsilon_i^v : \mathbb{R}^n \to \mathbb{R}$, $\epsilon_i^u : \mathbb{R}^n \to \mathbb{R}^m$ are the approximation errors. The approximation errors $\epsilon_i^v$, $\epsilon_i^u$ converge to zero, uniformly on $\Omega$, as $N_v, N_u \to \infty$.

Due to the fact that $w_i^v$ and $w_i^u$, $i \in \mathbb{N}$, are not known beforehand, one needs to construct an actor-critic network to approximate $V_i$ and $\mu_{i+1}$, so that:

$$\hat{V}_i(x) = (\hat{w}_i^v)^{\mathrm{T}}\phi^v(x), \quad (9)$$
$$\hat{\mu}_{i+1}(x) = (\hat{w}_i^u)^{\mathrm{T}}\phi^u(x), \quad (10)$$

where $\hat{w}_i^v \in \mathbb{R}^{N_v}$, $\hat{w}_i^u \in \mathbb{R}^{N_u \times m}$ are the critic and the actor weights respectively, and $i \in \mathbb{N}$. Subsequently, the weights $\hat{w}_i^v$, $\hat{w}_i^u$ need to be trained to approximate $V_i$ and $\mu_{i+1}$ through the exploitation of equation (7). However, due to the effect of the clock mismatches, the right-hand side of (7) cannot be constructed, hence one resorts to approximating $\bar{V}_i$ and $\bar{\mu}_{i+1}$ from (8) instead.

To this end, the left-hand side of (8) can be approximated using the critic network as:

$$\hat{V}_i(x(t_k + T)) - \hat{V}_i(x(t_k)) = (\phi^v(x(t_k + T)) - \phi^v(x(t_k)))^{\mathrm{T}}\hat{w}_i^v. \quad (11)$$

In addition, the last term of the right-hand side of (8) can be approximated using the actor network as:

$$\begin{aligned} &2\hat{\mu}_{i+1}^{\mathrm{T}}(x(\tau))R(\bar{u}(\tau) - \hat{\mu}_i(x(\tau))) \\ &\quad = 2(\phi^u(x(\tau)))^{\mathrm{T}}\hat{w}_i^u R(\bar{u}(\tau) - \hat{\mu}_i(x(\tau))) \\ &= 2\Big(\big((\bar{u}(\tau) - \hat{\mu}_i(x(\tau)))^{\mathrm{T}}R\big) \otimes \phi^u(x(\tau))^{\mathrm{T}}\Big)\mathrm{vec}(\hat{w}_i^u). \end{aligned} \quad (12)$$

Then, the error by approximating (8) using (11)-(12) is

$$e_{i,k} := \hat{V}_i(x(t_k + T)) - \hat{V}_i(x(t_k)) + \int_{t_k}^{t_k+T} \Big(Q(x(\tau)) + r(\hat{\mu}_i(x(\tau))) + 2\hat{\mu}_{i+1}^{\mathrm{T}}(x(\tau))R(\bar{u}(\tau) - \hat{\mu}_i(x(\tau)))\Big)\mathrm{d}\tau.$$

One can write $e_{i,k}$ in a compact form, so that

$$e_{i,k} = \Psi_{i,k}\hat{W}_i + \Phi_{i,k}, \quad (13)$$

**Algorithm 3** Learning-based PI with Clock Mismatches

1: Let $i = 0$, $\Omega \subset \mathbb{R}^n$, $\epsilon > 0$, and pick an admissible control policy $\mu_0 \in \Psi(\Omega)$.
2: **repeat**
3: Solve for $\hat{W}_i$ through (14).
4: Set $i = i + 1$.
5: **until** $i \geqslant 2$ & $\left\|\hat{W}_{i-1} - \hat{W}_{i-2}\right\| < \epsilon$.

where $\Psi_{i,k} := [\Psi^v_{i,k}\ \Psi^u_{i,k}]$, $\hat{W}_i := [(\hat{w}^v_i)^{\mathrm{T}}\ \mathrm{vec}(\hat{w}^u_i)^{\mathrm{T}}]^{\mathrm{T}}$, and

$$\Psi^v_{i,k} := \Big(\phi^v(x(t_k+T)) - \phi^v(x(t_k))\Big)^{\mathrm{T}},$$
$$\Psi^u_{i,k} := \int_{t_k}^{t_k+T} 2\big((\bar{u}(\tau) - \hat{\mu}_i(x(\tau)))^{\mathrm{T}} R\big) \otimes \phi^u(x(\tau))^{\mathrm{T}} \mathrm{d}\tau,$$
$$\Phi_{i,k} := \int_{t_k}^{t_k+T} \Big(Q(x(\tau)) + r(\hat{\mu}_i(x(\tau)))\Big)\mathrm{d}\tau.$$

The weight vector $\hat{W}_i$ can then be solved for using least sum of squares, provided that the following assumption holds, which essentially requires the measured state-input data to be sufficiently rich.

**Assumption 1.** There exist constants $\xi > 0$ and $K_0 \in \mathbb{N}$, such that for all $K \geqslant K_0$ it holds that $\frac{1}{K}\sum_{k=0}^{K} \Psi^{\mathrm{T}}_{i,k}\Psi_{i,k} > \xi I_{N_v+mN_u}$. □

Given Assumption 1, the least squares solution to (13) is

$$\hat{W}_i = -\Big(\sum_{k=0}^{K} \Psi^{\mathrm{T}}_{i,k}\Psi_{i,k}\Big)^{-1}\Big(\sum_{k=0}^{K} \Psi^{\mathrm{T}}_{i,k}\Phi_{i,k}\Big). \tag{14}$$

The learning-based PI algorithm with clock mismatches is then given by Algorithm 3.

### B. Convergence Properties

We shall now study the convergence properties of the learning-based PI Algorithm 3 with clock mismatches, as presented previously. Towards this end, let us define, for all $i \in \mathbb{N}$, the function $\tilde{V}_i$ that satisfies the equation:

$$\nabla\tilde{V}_i^{\mathrm{T}}(x)(f(x) + g(x)\hat{\mu}_i(x)) + Q(x) + r(\hat{\mu}_i(x)) = 0,$$

as well as the control law:

$$\tilde{\mu}_{i+1}(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla\tilde{V}_i(x).$$

Notice that $\tilde{V}_i$ is the true value function of $\hat{\mu}_i$. Hence, following similar steps as for $V_i$ in Section II-C, it can be shown, for all $k \in \mathcal{K}$, that:

$$\tilde{V}_i(x(t_k+T)) - \tilde{V}_i(x(t_k)) = -\int_{t_k}^{t_k+T} \Big(Q(x(\tau)) + r(\hat{\mu}_i(x(\tau))) + 2\tilde{\mu}^{\mathrm{T}}_{i+1}(x(\tau))R(u(\tau) - \hat{\mu}_i(x(\tau)))\Big)\mathrm{d}\tau. \tag{15}$$

The following auxiliary lemma shows that there exists an upper bound for the clock mismatch $\delta(t)$, $\forall t \geqslant t_0$, for which $\hat{W}_i$ converges arbitrarily close to the actor-critic weights that approximate $\tilde{V}_i$ and $\tilde{\mu}_{i+1}$ in the PI algorithm. For the results to hold, a continuity assumption will be needed.

**Assumption 2.** The control input $u(t)$ is Lipschitz continuous with respect to time, for all $t \geqslant t_0$. □

**Lemma 1.** *Let Assumptions 1-2 hold, and the compact weight vector $\hat{W}_i$ be trained as in* (14) *for all $i \in \mathbb{N}$. Then, for all $\epsilon > 0$ and $x \in \Omega$, there exist constant integers $N_v^\star,\ N_u^\star > 0$, and an upper clock mismatch bound $\delta^\star > 0$, such that if $N_v \geqslant N_v^\star$, $N_u \geqslant N_u^\star$, and $\|\delta(t)\| \leqslant \delta^\star$ for all $t \geqslant t_0$, it holds that:*

$$\left\|\hat{V}_i(x) - \tilde{V}_i(x)\right\| \leqslant \epsilon, \quad \|\hat{\mu}_{i+1}(x) - \tilde{\mu}_{i+1}(x)\| \leqslant \epsilon.$$

*Proof.* Only a sketch of the proof will be given here. By the Weierstrass approximation theorem, the functions $\tilde{V}_i$ and $\tilde{\mu}_{i+1}$, $i \in \mathbb{N}$, can be uniformly approximated on $\Omega$, so that

$$\begin{aligned}\tilde{V}_i(x) &= (\tilde{w}^v_i)^{\mathrm{T}}\phi^v(x) + \tilde{\epsilon}^v_i(x),\\ \tilde{\mu}_{i+1}(x) &= (\tilde{w}^u_i)^{\mathrm{T}}\phi^u(x) + \tilde{\epsilon}^u_i(x).\end{aligned} \tag{16}$$

The approximation errors $\tilde{\epsilon}^v_i : \mathbb{R}^n \to \mathbb{R}$, $\tilde{\epsilon}^u_i : \mathbb{R}^n \to \mathbb{R}^m$ vanish uniformly on $\Omega$ as $N_v, N_u \to \infty$. Substituting (16) in (15), for $i \in \mathbb{N}$, we derive:

$$0 = \Psi_{i,k}\tilde{W}_i + \Phi_{i,k} + \tilde{E}_{i,k} + M_{i,k},\ k \in \mathcal{K}, \tag{17}$$

where $\tilde{W}_i = [\tilde{w}_i^{v\mathrm{T}}\ \mathrm{vec}(\tilde{w}^u_i)^{\mathrm{T}}]^{\mathrm{T}}$, and

$$\begin{aligned}\tilde{E}_{i,k} &= \tilde{\epsilon}^v_i(x(t_k+T)) - \tilde{\epsilon}^v_i(x(t_k))\\ &\quad + \int_{t_k}^{t_k+T} 2\tilde{\epsilon}^u_i(x(\tau))^{\mathrm{T}}R(u(\tau) - \hat{\mu}_i(x(\tau)))\mathrm{d}\tau,\\ M_{i,k} &= \int_{t_k}^{t_k+T} 2(\phi^u(x(\tau)))^{\mathrm{T}}\tilde{w}^u_i R(u(\tau) - u(\tau - \delta(\tau)))\mathrm{d}\tau.\end{aligned}$$

Since $\hat{W}_i$ is estimated through the least-squares law (14) in order to minimize the sum of squares of the errors in (13), and since Assumption 1 holds, it will hold due to (17) that

$$\sum_{k=0}^{K} e^2_{i,k} \leqslant \sum_{k=0}^{K}(\tilde{E}_{i,k} + M_{i,k})^2,\ i \in \mathbb{N}. \tag{18}$$

Subtracting (17) from (13), we obtain:

$$e_{i,k} + \tilde{E}_{i,k} + M_{i,k} = \Psi_{i,k}(\hat{W}_i - \tilde{W}_i). \tag{19}$$

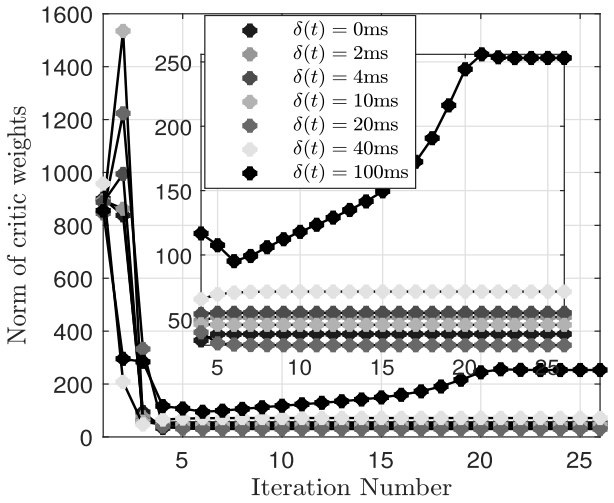
Multiplying (19) by itself and summing over $k$ leads, due to Assumption 1, to

$$\begin{aligned}\sum_{k=0}^{K}(e_{i,k}+\tilde{E}_{i,k}+M_{i,k})^2 &= \sum_{k=0}^{K}(\hat{W}_i-\tilde{W}_i)^{\mathrm{T}}\Psi^{\mathrm{T}}_{i,k}\Psi_{i,k}(\hat{W}_i-\tilde{W}_i)\\ &\geqslant K\xi\left\|(\hat{W}_i - \tilde{W}_i)\right\|^2.\end{aligned} \tag{20}$$

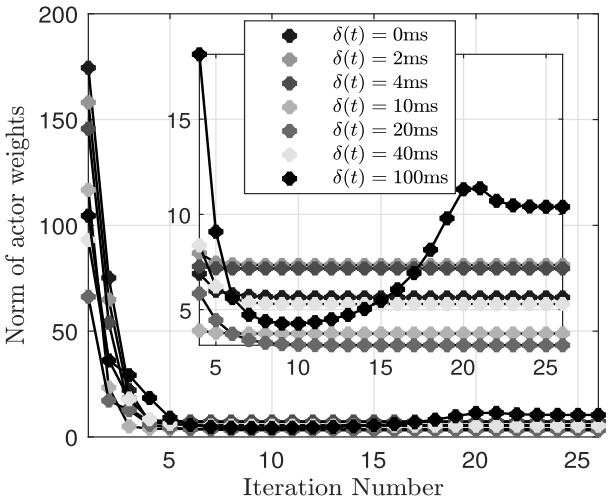
Thus, from (18) and (20):

$$\max_{k\in\mathcal{K}} 4\left(\tilde{E}_{i,k}+M_{i,k}\right)^2 \geqslant \xi\left\|(\hat{W}_i - \tilde{W}_i)\right\|^2. \tag{21}$$

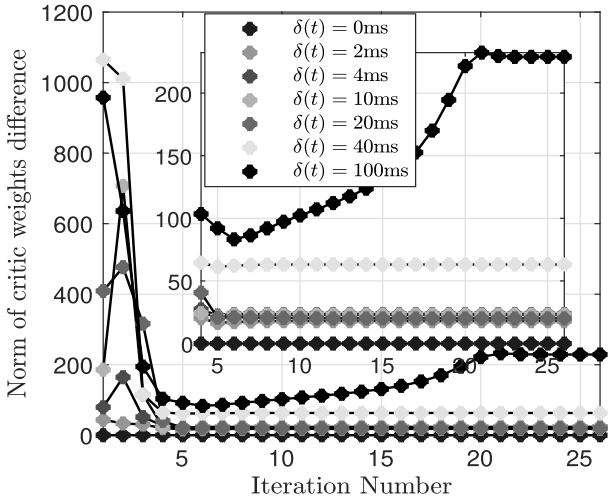
Using Assumptions 1-2 and given that $\|\delta(t)\| \leqslant \bar{\delta}$, $\forall t \geqslant t_0$ with $\bar{\delta} > 0$, one can show that $M_{i,k}$, $\tilde{E}_{i,k}$ and $e_{i,k}$ converge to zero, uniformly on $\Omega$, as $\bar{\delta} \to 0$ and $N_u, N_v \to \infty$. Hence, from (21), $\forall \epsilon_1 > 0$ there exist constants $N_v^m,\ N_u^m > 0$ and an upper clock mismatch bound $\delta^m$, such that if $N_v \geqslant N_v^m$, $N_u \geqslant N_u^m$ and $\|\delta(t)\| \leqslant \delta^m\ \forall t \geqslant t_0$, it holds that



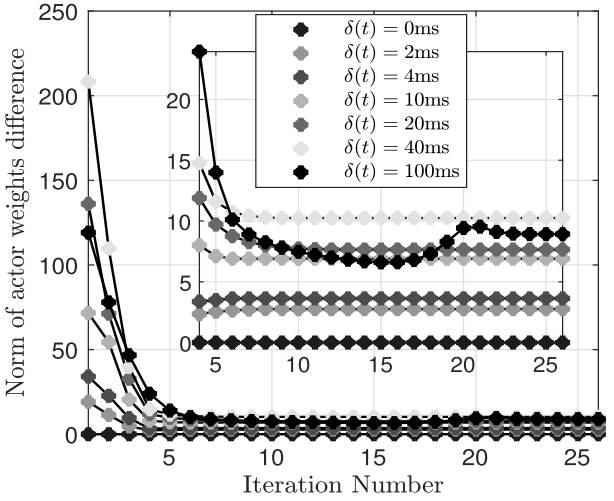

(a) Evolution of the norm of the critic weights at each iteration of the learning-based PI with clock mismatches.




(b) Evolution of the norm of the actor weights at each iteration of the learning-based PI with clock mismatches.




(c) Evolution of the norm of the difference of the critic weights at each scenario with the critic weights in scenario 1, where $\delta(t) = 0,\ \forall t \geqslant 0$.




(d) Evolution of the norm of the difference of the actor weights at each scenario with the critic weights in scenario 1, where $\delta(t) = 0,\ \forall t \geqslant 0$.

Fig. 1. Evolution of the learning-based PI for each clock mismatch scenario.

$\left\|\hat{W}_i - \tilde{W}_i\right\| \leqslant \epsilon_1$. The final result then follows by (9), (16) and the uniform convergence of $\tilde{\epsilon}_i^v$, $\tilde{\epsilon}_i^u$. ■

The upcoming theorem generalizes Lemma 1, and states that there exists an upper bound for the clock mismatches $\delta(t)$, $\forall t \geqslant t_0$, for which the iterative learning law provided by (14) converges arbitrarily close to $V^\star$ and $\mu^\star$.

**Theorem 1.** *Let Assumptions 1-2 hold and $\mu_0 \in \Psi(\Omega)$. Assume that $\hat{W}_i$ is updated according to* (14) *for all $i \in \mathbb{N}$. Then, for all $\epsilon > 0$ and $x \in \Omega$, there exist constant integers $N_v^{\star\star}$, $N_u^{\star\star}$, $i^\star > 0$ and an upper clock mismatch bound $\delta^{\star\star} > 0$, such that if $N_v \geqslant N_v^{\star\star}$, $N_u \geqslant N_u^{\star\star}$, and $\|\delta(t)\| \leqslant \delta^{\star\star}$ for all $t \geqslant t_0$, it holds that:*

$$\left\|\hat{V}_{i^\star}(x) - V^\star(x)\right\| \leqslant \epsilon, \quad \|\hat{\mu}_{i^\star+1}(x) - \mu^\star(x)\| \leqslant \epsilon.$$

*Proof.* The proof is based on the results of Lemma 1, and is omitted due to space limitations. ■

## IV. Simulations

Consider a two-link manipulator [23], with dynamics:

$$M(q)\ddot{q} + V_m(q,\dot{q})\dot{q} + F_d\dot{q} + F_s(\dot{q}) = u, \tag{22}$$

where $q = [q_1\ q_2]^{\mathrm{T}}$ and $\dot{q} = [\dot{q}_1\ \dot{q}_2]^{\mathrm{T}}$ are the angular positions (in rad) and the angular velocities (in rad/s), respectively. As a result, the state vector is given by $x = [x_1\ x_2\ x_3\ x_4]^{\mathrm{T}} = [q_1\ q_2\ \dot{q}_1\ \dot{q}_2]^{\mathrm{T}}$. The matrices $M(q) \in \mathbb{R}^{2\times 2}$ and $V_m(q,\dot{q}) \in \mathbb{R}^{2\times 2}$ are the inertia and the centripetal–Coriolis matrices, while $F_d\dot{q}$ and $F_s(\dot{q})$ model the dynamic and static friction, respectively; all of them are modeled as in [23]. The objective is to approximate the optimal value function $V^\star$ and controller $u^\star$ of (22), where $Q(x) = \|x\|^2$ and $R = I_2$. To this end, the actor-critic network (9)-(10) is employed, with basis functions given by polynomials up to the order of 4.

We will consider 7 different scenarios $i \in \{1,\dots,7\}$, in each of which the clock mismatch $\delta(t) = \delta_i$ is constant. In particular, we choose $\delta_1 = 0$ ms, $\delta_2 = 2$ ms, $\delta_3 = 4$ ms,

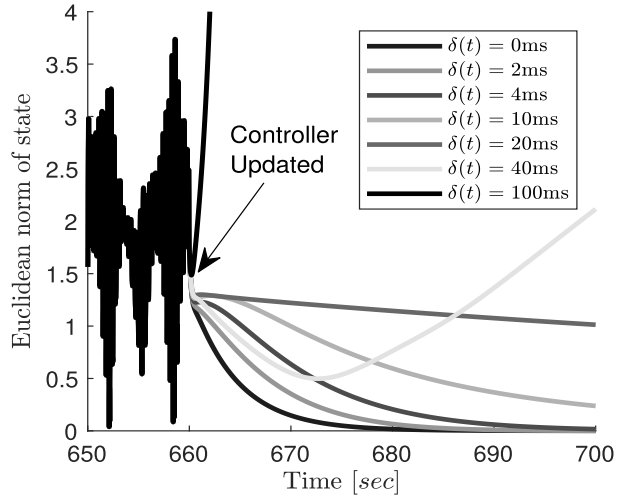



Fig. 2. Evolution of the Euclidean norm of the state over $t \in [650,\ 700]$ seconds, for the 7 different scenarios. Exploration takes place before $t = 660$ $[sec]$, and the controller is subsequently learnt and updated.

$\delta_4 = 10$ ms, $\delta_5 = 20$ ms, $\delta_6 = 40$ ms and $\delta_7 = 100$ ms. As a result, scenario 1 assumes perfect synchronization between the sensors' and the actuators' clocks. In each scenario, the first 660 seconds are used for exploration, in order to gather sufficient state-input data from the system. Subsequently, the learning-based PI with clock mismatches is carried out ($T = 50$ ms), by iteratively solving equations (14). Finally, the exploration noise is terminated at the 660th second, and the controller is changed from the initial one to the one derived by the PI algorithm.

The results are shown in Figures 1-2. It can be seen from Figures 1(a)-1(b) that convergence of the learning-based PI algorithm takes place for all values of the clock mismatch. However, as shown in Figures 1(c)-1(d), the actor-critic weights converge monotonically further away from their nominal values (i.e., their values when the clock mismatch is zero) as $\delta(t)$ is increased. Additionally, for the 100 ms case, the convergence is marginal, with the learning-based PI being close to becoming unstable. Figure 2 shows the evolution of the norm of the state vector for each clock mismatch scenario. While the manipulator does remain stable when $\|\delta(t)\| \leqslant 20$ ms, the trajectories diverge when $\delta(t) = 40$ ms or $\delta(t) = 100$ ms. This is not unexpected; the learning window is equal to $T = 50$ ms, meaning that the control input data over each integration time interval $[t_k,\ t_k + T]$, $k \in \mathcal{K}$, originate from a completely different time interval in these cases.

## V. Conclusion

In this work, we studied the robustness of off-policy actor-critic algorithms with respect to sensor-actuator clock offsets. It was shown that these algorithms remain robust when clock mismatches are present and small. Future work will focus on a more general framework, where different clock mismatches will exist between all distinct sensors and actuators.

## References


[1] Y. Yuehong, Y. Zeng, X. Chen, and Y. Fan, "The internet of things in healthcare: An overview," *Journal of Industrial Information Integration*, vol. 1, pp. 3–13, 2016.

[2] X. Jin, W. M. Haddad, Z.-P. Jiang, A. Kanellopoulos, and K. G. Vamvoudakis, "An adaptive learning and control architecture for mitigating sensor and actuator attacks in connected autonomous vehicle platoons," *International Journal of Adaptive Control and Signal Processing*, vol. 33, no. 12, pp. 1788–1802, 2019.

[3] M. H. Cintuglu, O. A. Mohammed, K. Akkaya, and A. S. Uluagac, "A survey on smart grid cyber-physical system testbeds," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 446–464, 2016.

[4] A. Alan, A. J. Taylor, C. R. He, G. Orosz, and A. D. Ames, "Safe controller synthesis with tunable input-to-state safe control barrier functions," *IEEE Control Systems Letters*, 2021.

[5] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[6] E. Fridman and M. Dambrine, "Control under quantization, saturation and delay: An LMI approach," *Automatica*, vol. 45, no. 10, pp. 2258–2264, 2009.

[7] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[8] B. Luo, D. Liu, T. Huang, and D. Wang, "Model-free optimal tracking control via critic-only Q-learning," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 10, pp. 2134–2144, 2016.

[9] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, 2014.

[10] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of CPS security," *Annual Reviews in Control*, vol. 47, pp. 394–411, 2019.

[11] K. G. Vamvoudakis and J. P. Hespanha, "Cooperative Q-learning for rejection of persistent adversarial inputs in networked linear quadratic systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 4, pp. 1018–1031, 2017.

[12] F. Fotiadis and K. G. Vamvoudakis, "Detection of actuator faults for continuous-time systems with intermittent state feedback," *Systems & Control Letters*, vol. 152, p. 104938, 2021.

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

[14] M. Wakaiki, K. Okano, and J. P. Hespanha, "Stabilization of systems with asynchronous sensors and controllers," *Automatica*, vol. 81, pp. 314–321, 2017.

[15] 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.

[16] A. Lamperski and N. J. Cowan, "Optimal control with noisy time," *IEEE Transactions on Automatic Control*, vol. 61, no. 2, pp. 319–333, 2015.

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

[18] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation," *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[19] Z.-P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations and Trends® in Systems and Control*, vol. 8, no. 3, 2020.

[20] W. Gao and Z.-P. Jiang, "Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2614–2624, 2017.

[21] R. Song, F. L. Lewis, Q. Wei, and H. Zhang, "Off-policy actor-critic structure for optimal control of unknown systems with disturbances," *IEEE transactions on cybernetics*, vol. 46, no. 5, pp. 1041–1050, 2015.

[22] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural networks*, vol. 3, no. 5, pp. 551–560, 1990.

[23] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, 2015.