Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief Paper

On the effect of clock offsets and quantization on learning-based adversarial games☆

Filippos Fotiadis [a,*], Aris Kanellopoulos [b], Kyriakos G. Vamvoudakis [a], Jerome Hugues [c]

[a] School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[b] Division of Information Science and Engineering, KTH Royal Institute of Technology Stockholm, Sweden
[c] Carnegie Mellon University/Software Engineering Institute, Pittsburgh, PA, USA

ARTICLE INFO

Article history:
Received 2 May 2022
Received in revised form 1 April 2024
Accepted 1 May 2024
Available online xxxx

Keywords:
Clock offsets
Quantization
Learning
Zero-sum games

ABSTRACT

In this work, we consider systems whose components suffer from clock offsets and quantization and study the effect of those on a reinforcement learning (RL) algorithm. Specifically, we consider an off-policy iterative RL algorithm for continuous-time systems, which uses input and state data to approximate the Nash-equilibrium of a zero-sum game. However, the data used by this algorithm are not consistent with one another, in that each of them originates from a slightly different time instant of the past, hence putting the convergence of the algorithm in question. We prove that, given that these timing inconsistencies remain below a certain threshold, the iterative off-policy RL algorithm will still converge epsilon-closely to the desired Nash policy. However, this result is conditional to a certain Lipschitz continuity and differentiability condition on the input-state data collected, which is indispensable in the presence of clock offsets. A similar result is also derived when quantization of the measured state is considered. Finally, unlike prior work, we provide a sufficiently rich data condition for the execution of the iterative RL algorithm, which can be verified a priori across all iteration indices. Simulations are performed, which verify and clarify theoretical findings.

1. Introduction

Cyber–physical systems (CPS) are systems combining multiple actuating and sensing components, which operate and synchronize via communication and computational devices. However, this synchronization is almost always imperfect, leading to discrepancies – *clock offsets* (Okano, Wakaiki, Yang, & Hespanha, 2017) – in the way each component perceives time. The presence of such offsets can be especially threatening to the proper operation of CPS: unless appropriate precautions have been taken, it can potentially cause instability of the system or its learning components.

Reinforcement learning (RL) techniques have been employed to augment CPS with more advanced autonomous capabilities, allowing them to derive optimal decision-making policies via interactions with the environment (Sutton & Barto, 2018). These can also be used in tandem with game-theoretic concepts (Busoniu, Babuska, & De Schutter, 2008), to obtain policies that are not only optimal, but also resilient against adversarial agents (Vamvoudakis & Hespanha, 2017). In this context, Jiang and Jiang (2014) developed an off-policy reinforcement learning algorithm that gathers input and state data from the CPS, and uses them to approximate an optimal stabilization policy without knowledge of the system's dynamics. The authors in Modares, Lewis, and Jiang (2015) extended this algorithm in a game-theoretic sense, to obtain policies that are also resilient to adversarial injections.

In the two aforementioned works, an analysis was carried out to show that the proposed off-policy RL algorithm is convergent, despite the errors generated from approximating the value function using neural networks. However, in the presence of clock offsets and quantization, this analysis is no longer valid or applicable; no matter how small a clock offset may be, it can lead to the injection of uncontrollably large errors in the RL algorithm, owing to the offset's inherent nonlinear nature. The purpose of this study, therefore, is to fill this gap in the literature: by considering the fact that clock offsets and quantization may have affected the data used by the off-policy RL algorithm, we derive conditions – in the form of certain Lipschitz continuity and differentiability assumptions – the satisfaction of which allows for convergence of the algorithm to still be attained.

☆ The material in this paper was partially presented at the 2022 American Control Conference, June 8-10, 2022, Atlanta, Georgia, USA. This paper was recommended for publication in revised form by Associate Editor Raul Ordonez under the direction of Editor Miroslav Krstic.

* Corresponding author.

E-mail addresses: ffotiadis@gatech.edu (F. Fotiadis), arisk@kth.se (A. Kanellopoulos), kyriakos@gatech.edu (K.G. Vamvoudakis), jjhugues@sei.cmu.edu (J. Hugues).

https://doi.org/10.1016/j.automatica.2024.111762
0005-1098/

Related Work: Various RL algorithms can be traced in the literature, often employing deep neural networks for model extraction (Li et al., 2022; Zhao, Wu, Li, Chen, & Zheng, 2022). One famous family of RL methods, known as Adaptive Dynamic Programming (ADP), comprises algorithms that collect input-state data from the trajectories of the system, and use them to solve the underlying Hamilton–Jacobi–Bellman (HJB) equation. This procedure can be done either with knowledge of the system model (Abu-Khalaf & Lewis, 2005) or in a model-free manner (Jiang & Jiang, 2014; Kiumarsi, Lewis, Modares, Karimpour, & Naghibi-Sistani, 2014; Vamvoudakis, 2017), and provably leads to the estimation of the optimal control policy for the system. It can also be extended to solve multi-agent optimal decision-making problems, often formulated as dynamical games among rational agents (Nowé, Vrancx, & Hauwere, 2012). Particularly, by solving an associated Hamilton–Jacobi–Isaacs (HJI) equation, it can be used to obtain policies resilient to adversarial inputs, which form a Nash equilibrium for a user-specified cost (Johnson, Kamalapurkar, Bhasin, & Dixon, 2014; Modares et al., 2015; Vamvoudakis & Hespanha, 2017). In this work, we will be specifically concerned with the off-policy RL algorithm (Jiang & Jiang, 2014; Modares et al., 2015), used to estimate the Nash equilibrium from input-state measurements.

The robustness of the off-policy RL algorithm is an issue of interest in the literature. Specifically, Jiang and Jiang (2014), Modares et al. (2015) analyzed the performance of this algorithm in the presence of errors created due to neural network approximation. It was shown that if the number of neurons employed to approximate the value function and the control policy is sufficiently large, then the RL algorithm still attains convergence. However, robustness to clock offsets and quantization, which behave differently from approximation errors, was not considered. In Gao, Deng, Jiang, and Jiang (2022), Gao, Jiang, Jiang, and Chai (2016), an output-based off-policy ADP approach was designed, which yields policies robust to dynamic uncertainties and Denial-of-Service attacks; however, the effect of clock offsets and quantization at the learning stage was also not considered in these studies, and the studies were restricted to linear systems. Recently, the robustness of off-policy ADP with respect to noise was revisited in Pang, Bian, and Jiang (2021), but was focused on linear systems and assumed a priori that the noise can be forced to become uniformly small. Hence, the highly nonlinear effect of clock offsets and quantization on nonlinear ADP remains unknown, and the convergence of ADP in the presence of those questionable.

Control theorists have investigated the effect of clock offsets and quantization in the context of distributed and decentralized CPS. For example, in Fridman and Dambrine (2009), the authors study the input-to-state stability properties of a linear system in the presence of quantization errors, delays and saturation via the use of Lyapunov–Krasovskii functionals. The problem of timing discrepancies between actuators and sensors is investigated in Wakaiki, Okano, and Hespanha (2017), where the discrepancies are modeled as parametric uncertainties. Expanding on this research, in Okano et al. (2017), the authors consider both clock mismatches and quantization errors. Finally, in Wakaiki, Cetinkaya, and Ishii (2019), an adversarial scenario is analyzed, in which the effect of Denial-of-Service attacks on the system are considered in tandem with output quantization. All these works, however, explore the effects of timing errors and quantization on the controlled system itself; they do not analyze the impact of those on RL algorithms used to compute the corresponding control policy.

Contributions: The discussion above renders clear that the effect of clock offsets and quantization on the off-policy RL algorithm remains an open question. On the one hand, existing approaches that study convergence of this RL under the effect of neural network approximation errors (Jiang & Jiang, 2014; Modares et al., 2015) cannot be extended to account for the effect of clock offsets and quantization. This is because, unlike neural network approximation errors, the errors from clock offsets and quantization do not necessarily vanish uniformly in the limit, unless specific conditions hold. On the other hand, existing works studying clock offsets and quantization focus only on their effect on the system itself, rather than the corresponding RL component (Okano et al., 2017; Wakaiki et al., 2017). The latter is the purpose of this work, which analyzes the effect of both clock offsets and quantization on the off-policy RL algorithm used to solve a zero-sum game. It is specifically proved that, given that the magnitude of the clock offsets and the quantization error remains below a certain threshold, and given certain Lipschitz continuity and differentiability conditions not conventionally required in existing approaches, RL will still approximate the Nash equilibrium ϵ-closely. A preliminary version of this work appeared in Fotiadis, Kanellopoulos, Vamvoudakis, and Hugues (2022), but i) considered only relative sensor–actuator clock offsets; ii) assumed that all sensors and actuators share the same clock; iii) did not study the effect of quantization; and iv) provided only sketches of the proofs of the main results. As an additional contribution with respect to Jiang and Jiang (2014), Modares et al. (2015), we provide a sufficiently rich data condition for the RL algorithm, which can be verified a priori over all iteration indices.

Notation. $\mathbb{R}$ and $\mathbb{N}$ denote the set of real and natural numbers (including zero), respectively, $\mathbb{R}_+$ denotes the set of non-negative real numbers, and $\mathbb{N}_+$ denotes the set of non-zero natural numbers. The operator ∇ denotes the gradient of a function. For any matrices Z_1 and Z_2, $Z_1 \otimes Z_2$ denotes their Kronecker product. For $Z \in \mathbb{R}^{q\times q}$, $\mathrm{vec}(Z) \in \mathbb{R}^{q^2}$ denotes the vectorized form of Z. For $z \in \mathbb{R}^n$, $\|z\|$ denotes the 2-norm of z, $z \otimes_h z = [z_1^2 \ z_1z_2 \ \dots \ z_1z_n \ z_2^2 \ z_2z_3 \ \dots \ z_n^2]^{\mathrm{T}}$ its half-vectorized Kronecker product, and $C_n \in \mathbb{R}^{n^2\times n(n+1)/2}$ the duplication matrix such that $z \otimes z = C_n(z \otimes_h z)$.

2. Problem formulation and preliminaries

2.1. Characterization of clock offsets

Consider, for all $t \geq t_0$, a nonlinear system of the form:

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t) + h(x(t))a(t), \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the system's state with initial condition $x(t_0) = x_0$, $u(t) \in \mathbb{R}^m$ is the control input, $a(t) \in \mathbb{R}^p$ is an adversarial input, and $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$, $h : \mathbb{R}^n \to \mathbb{R}^{n\times p}$. The functions f, g, and h are considered to be locally Lipschitz, though unknown. In addition, it is assumed that $f(0) = 0$, so that the origin is an equilibrium point of the uncontrolled system.

In complex CPS, which are systems of high heterogeneity, the presence of multiple physical and virtual components is almost certain. As a result, it is difficult to design a unique, centralized clock, with which every system component will be perfectly synchronized (Shrivastava et al., 2016). In fact, it is more common for every component to have its own clock and a slightly different sense of time. Consequently, and owing to noise, quantizations and other uncertainties that are commonplace in practice (Okano et al., 2017), mismatches can occur between two or more component clocks.

In the dynamics (1) of the CPS, one can observe several different kinds of CPS components: the state sensors, which measure the state $x(t)$, for all $t \geq t_0$; the actuators, which implement the control policy $u(t)$, for all $t \geq t_0$; and any potential exogenous input sensors, which measure $a(t)$, for all $t \geq t_0$. To capture the general case regarding these components, let us consider that the CPS is equipped with n state sensors, m actuators, and p exogenous input sensors. Then, each of the sensors $i \in \{1, \ldots, n\} := \mathcal{N}_x$ provides us with measurements of the ith element of the state $x(t)$, which we denote as $\bar{x}_i(t)$, $\forall t \geq t_0$. In addition, each actuator $j \in \{1, \ldots, m\} := \mathcal{N}_u$ provides us with values of the jth element of the control input $u(t)$, which we denote as $\bar{u}_j(t)$, $\forall t \geq t_0$. Finally, each exogenous input sensor $l \in \{1, \ldots, p\} := \mathcal{N}_a$ transmits measured values of the exogenous input $a(t)$, which we denote as $\bar{a}_l(t)$, $\forall t \geq t_0$.

Let us now define a "true/reference clock" $c : [t_0,\ \infty) \to [t_0,\ \infty)$, as the clock that is in agreement with the "real" time; that is, it holds that $c(t) = t$ for all $t \geq t_0$. In case that the sensors and the actuators perceive real time accurately, and are perfectly synchronized with one another as well as with the true clock, it will hold that:

$$\begin{aligned}
&\bar{x}_i(t) = x_i(c(t)) = x_i(t),\ \forall t \in [t_0,\ \infty),\ i \in \mathcal{N}_x,\\
&\bar{u}_j(t) = u_j(c(t)) = u_j(t),\ \forall t \in [t_0,\ \infty),\ j \in \mathcal{N}_u,\\
&\bar{a}_l(t) = a_l(c(t)) = a_l(t),\ \forall t \in [t_0,\ \infty),\ l \in \mathcal{N}_a.
\end{aligned}$$

Nevertheless, the clocks of the components of a CPS are rarely synchronized, so it is more accurate to state that:

$$\begin{aligned}
&\bar{x}_i(t) = x_i(c_i^x(t)) = x_i(t + \delta_i^x(t)),\ i \in \mathcal{N}_x,\\
&\bar{u}_j(t) = u_j(c_j^u(t)) = u_j(t + \delta_j^u(t)),\ j \in \mathcal{N}_u,\\
&\bar{a}_l(t) = a_l(c_l^a(t)) = a_l(t + \delta_l^a(t)),\ l \in \mathcal{N}_a,
\end{aligned} \tag{2}$$

$\forall t \in [t_0,\ \infty)$. In (2), $c_i^x(t), c_j^u(t), c_l^a(t) \in [t_0,\ \infty)$ are the clock functions of each state sensor $i \in \mathcal{N}_x$, actuator $j \in \mathcal{N}_u$, and exogenous/adversarial input sensor $l \in \mathcal{N}_a$, respectively. Accordingly, the functions δ_i^x, δ_j^u and δ_l^a are the *clock offsets* of each of these components from the reference clock, defined for all $t \in [t_0,\ \infty)$ as

$$\begin{aligned}
&\delta_i^x(t) = c_i^x(t) - t,\ i \in \mathcal{N}_x,\\
&\delta_j^u(t) = c_j^u(t) - t,\ j \in \mathcal{N}_u,\\
&\delta_l^a(t) = c_l^a(t) - t,\ l \in \mathcal{N}_a.
\end{aligned} \tag{3}$$

When the offsets (3) are nonzero, it is possible for a learning-based algorithm depending on the data (2) to become non-convergent or yield wrong results. Therefore, it is of interest to investigate whether learning retains any kind of robustness towards clock offsets — at least, in an epsilon-delta sense. In this work, we will be specifically concerned about learning-based algorithms used to solve differential-games in a model-free manner.

2.2. Two-player differential game

A zero-sum game can be defined over the dynamics (1), with the players being: a) the CPS operator, who wants to optimally regulate the system; and b) an adversary or an exogenous input, whose goal is the disruption of this regulation. The utility of such a game can be defined as:

$$J(x_0; u, a) = \int_{t_0}^{\infty} (Q(x(\tau)) + u(\tau)^{\mathrm{T}} R u(\tau) - \gamma^2 a(\tau)^{\mathrm{T}} a(\tau)) \mathrm{d}\tau,$$

where $Q : \mathbb{R}^n \to \mathbb{R}$ is positive definite, $R \succ 0$, and $\gamma > 0$ is an attenuation factor. The game is given by:

$$V^\star(x) = \min_u \max_a J(x;\ u,\ a), \tag{4}$$

where $V^\star$ is the optimal value function. We consider that the game (4) has a unique solution, corresponding to a saddle point/Nash equilibrium. In general, for such a solution to exist, it is necessary that $\gamma > \gamma^\star$, where $\gamma^\star$ is the minimum attainable attenuation factor (Vamvoudakis & Lewis, 2012).

Algorithm 1 Policy Iteration

```
Let $i = 0$, $\Omega \subset \mathbb{R}^n$, $\epsilon > 0$. Start with a tuple of policies $\{u_0,\ a_0\}$ that is stabilizing in $\Omega$.
repeat
    Solve for $V_i$, $\forall x \in \Omega$, in
        $\nabla V_i^{\mathrm{T}}(x)(f(x) + g(x)u_i(x) + h(x)a_i(x)) + Q(x) + u_i(x)^{\mathrm{T}} R u_i(x) - \gamma^2 a_i(x)^{\mathrm{T}} a_i(x) = 0,\ V_i(0) = 0.$   (6)
    Let the new policies be given by
        $u_{i+1}(x) = -\frac{1}{2} R^{-1} g^{\mathrm{T}}(x) \nabla V_i(x),$
        $a_{i+1}(x) = \frac{1}{2\gamma^2} h^{\mathrm{T}}(x) \nabla V_i(x).$   (7)
    Set $i = i + 1$.
until $i \geq 2$ & $\sup_{x \in \Omega} |V_{i-1}(x) - V_{i-2}(x)| < \epsilon$.
```

Following Vamvoudakis and Lewis (2012), the saddle-point $\{u^\star,\ a^\star\}$ of (4) satisfies:

$$\begin{aligned}
&u^\star(x) = -\frac{1}{2} R^{-1} g^{\mathrm{T}}(x) \nabla V^\star(x),\\
&a^\star(x) = \frac{1}{2\gamma^2} h^{\mathrm{T}}(x) \nabla V^\star(x),
\end{aligned}$$

$\forall x \in \mathbb{R}^n$, where $V^\star$ solves the Hamilton–Jacobi–Isaacs (HJI) equation:

$$\begin{aligned}
&\nabla V^{\star\mathrm{T}}(x) f(x) - \frac{1}{4} \nabla V^{\star\mathrm{T}}(x) g(x) R^{-1} g^{\mathrm{T}}(x) \nabla V^\star(x) \qquad (5)\\
&+ \frac{1}{4\gamma^2} \nabla V^{\star\mathrm{T}}(x) h(x) h^{\mathrm{T}}(x) \nabla V^\star(x) + Q(x) = 0,\ V^\star(0) = 0.
\end{aligned}$$

Thus, to find the optimal policy $u^\star$, one needs to solve the HJI Eq. (5), but this is too difficult a task to be carried out analytically. Still, Algorithm 1, which describes the Policy Iteration (PI) procedure, can be used to approximate $V^\star$ over a given compact set $\Omega \subset \mathbb{R}^n$ (Wu & Luo, 2012).

2.3. Learning-based PI

Although Algorithm 1 can be used to effectively solve the HJI equation, it has the drawback of being a model-based procedure. To tackle this issue, the authors in Jiang and Jiang (2014), Modares et al. (2015) proposed a learning-based PI algorithm, which can approximate the optimal value function $V^\star$ without knowing f, g or h, and by using measured input-state data.

To demonstrate how learning-based PI works, notice that the system dynamics (1) can be expressed as:

$$\begin{aligned}
\dot{x} = {} & f(x) + g(x)u_i(x) + h(x)a_i(x)\\
& + g(x)(u - u_i(x)) + h(x)(a - a_i(x)),
\end{aligned} \tag{8}$$

where the functions $u_i,\ a_i$ are the same as those derived in the step $i - 1 \in \mathbb{N}$ of Algorithm 1, and the argument of time has been dropped to simplify notation. Using (6), (7) and (8), the time derivative of V_i along (1) is:

$$\begin{aligned}
\dot{V}_i = & -Q(x) - u_i(x)^{\mathrm{T}} R u_i(x) + \gamma^2 a_i(x)^{\mathrm{T}} a_i(x)\\
& - 2u_{i+1}^{\mathrm{T}}(x) R (u - u_i(x)) + 2\gamma^2 a_{i+1}^{\mathrm{T}}(x)(a - a_i(x))).
\end{aligned} \tag{9}$$

Consider now the time instants $t_k,\ t_k'$, with $k \in \{0, 1, \ldots, K\} := \mathcal{K}$, such that $t_k \geq t_0$ and $t_k' > t_k$ for all $k \in \mathcal{K}$. Then, integrating (9) over $[t_k,\ t_k']$ yields:

$$V_i(x(t_k')) - V_i(x(t_k)) = \int_{t_k}^{t_k'} \Big(-Q(x(\tau))$$

Algorithm 2 Learning-based PI

```
Let $i = 0$, $\Omega \subset \mathbb{R}^n$, $\epsilon > 0$. Start with a tuple of policies $\{u_0,\ a_0\}$
that is stabilizing in $\Omega$.
repeat
    Solve for $V_i$, $u_{i+1}$ and $a_{i+1}$ over $\Omega$ from (10).
    Set $i = i + 1$.
until $i \geq 2$ & $\sup_{x\in\Omega} |V_{i-1}(x) - V_{i-2}(x)| < \epsilon$.
```

$$
\begin{aligned}
&- u_i(x(\tau))^{\mathrm{T}} R u_i(x(\tau)) + \gamma^2 a_i(x(\tau))^{\mathrm{T}} a_i(x(\tau)) \\
&\quad - 2u_{i+1}^{\mathrm{T}}(x(\tau)) R(u(\tau) - u_i(x(\tau))) \\
&\qquad + 2\gamma^2 a_{i+1}^{\mathrm{T}}(x(\tau))(a(\tau) - a_i(x(\tau))) \Big) \,\mathrm{d}\tau,\ \ k \in \mathcal{K}.
\end{aligned} \tag{10}
$$

Eq. (10) provides a data-based method to express the value function V_i and the policies $u_{i+1},\ a_{i+1}$. This gives rise to the learning-based PI procedure, which is described in Algorithm 2, and which can be implemented using actor–critic networks (Gao & Jiang, 2017; Jiang, Bian, & Gao, 2020; Jiang & Jiang, 2014; Modares et al., 2015; Song, Lewis, Wei, & Zhang, 2015).

2.4. Learning with clock offsets

It is evident that Algorithm 2 assumes perfect synchronization between all components of the CPS. As a consequence, it is not certain whether it will behave well, given even a relatively small clock offset; Algorithm 2 is, after all, a highly nonlinear process.

Motivated by the preceding, in this work, we will study whether the learning-based Algorithm 2 is robust – in an epsilon-delta sense – with respect to clock offsets in the CPS components. In particular, we will assume that the CPS components provide Algorithm 2 with the data (2), which have been corrupted by timing discrepancies. Consequently, at each $t \geq t_0$, Algorithm 2 receives the following corrupted versions of $x(t)$, $u(t)$ and $a(t)$:

$$
\begin{aligned}
\bar{x}(t) &= [\bar{x}_1(t)\ \bar{x}_2(t)\ \ldots\ \bar{x}_n(t)]^{\mathrm{T}}, \\
\bar{u}(t) &= [\bar{u}_1(t)\ \bar{u}_2(t)\ \ldots\ \bar{u}_m(t)]^{\mathrm{T}}, \\
\bar{a}(t) &= [\bar{a}_1(t)\ \bar{a}_2(t)\ \ldots\ \bar{a}_p(t)]^{\mathrm{T}}.
\end{aligned}
$$

Notice that the clock offsets that have corrupted the measured data are both unknown and time-varying, hence it is not possible to cross them out. As a result, it is not possible to construct Eq. (10) for Algorithm 2 and learn the functions V_i, u_{i+1} and a_{i+1} directly; rather, one is forced to learn the functions $\bar{V}_i$, $\bar{u}_{i+1}$ and $\bar{a}_{i+1}$, which satisfy the clock-offseted version of (10):

$$
\begin{aligned}
\bar{V}_i(\bar{x}(t_k')) - \bar{V}_i(\bar{x}(t_k)) &= \int_{t_k}^{t_k'} \Big(-Q(\bar{x}(\tau)) \\
&- \bar{u}_i(\bar{x}(\tau))^{\mathrm{T}} R \bar{u}_i(\bar{x}(\tau)) + \gamma^2 \bar{a}_i(\bar{x}(\tau))^{\mathrm{T}} \bar{a}_i(\bar{x}(\tau)) \\
&\quad - 2\bar{u}_{i+1}^{\mathrm{T}}(\bar{x}(\tau)) R(\bar{u}(\tau) - \bar{u}_i(\bar{x}(\tau))) \\
&\qquad + 2\gamma^2 \bar{a}_{i+1}^{\mathrm{T}}(\bar{x}(\tau))(\bar{a}(\tau) - \bar{a}_i(\bar{x}(\tau))) \Big) \,\mathrm{d}\tau,\ \ k \in \mathcal{K}.
\end{aligned} \tag{11}
$$

As a result of (11), the following question arises: will the "corrupted" value function $\bar{V}_i$, as well as the "corrupted" policies $\bar{u}_{i+1}$, $\bar{a}_{i+1}$ converge close to the true ones? In the upcoming sections, we will see that the answer to this question is positive, given certain assumptions. In particular, if the clock offsets do not exceed a $\delta(\epsilon)$-threshold, and given certain Lipschitz continuity/differentiability assumptions, the functions $\bar{V}_i$, $\bar{u}_{i+1}$ and $\bar{a}_{i+1}$ enter an ϵ-neighborhood of $V^\star$, $u^\star$ and $a^\star$ over a compact set $\Omega \subset \mathbb{R}^n$, after a finite number of iterations in $i \in \mathbb{N}$.

3. Main results

In this section, we will describe the learning-based PI algorithm with clock offseted and/or quantized data and prove that it can be robust, in an epsilon-delta sense, given certain continuity/differentiability assumptions.

3.1. Learning-based PI with clock offsets

Since Eq. (11) is infinite-dimensional, it is difficult to solve explicitly for V_i and u_{i+1}, a_{i+1}, $i \in \mathbb{N}$. Nevertheless, these functions can be expressed as:

$$
\begin{aligned}
V_i(x) &= (w_i^v)^{\mathrm{T}} \phi^v(x) + \epsilon_i^v(x), \\
u_{i+1}(x) &= (w_i^u)^{\mathrm{T}} \phi^u(x) + \epsilon_i^u(x), \\
a_{i+1}(x) &= (w_i^a)^{\mathrm{T}} \phi^a(x) + \epsilon_i^a(x),
\end{aligned}
$$

where $w_i^v \in \mathbb{R}^{N_v}$, $w_i^u \in \mathbb{R}^{N_u \times m}$, $w_i^a \in \mathbb{R}^{N_a \times p}$ are weight matrices, $\phi^v : \mathbb{R}^n \to \mathbb{R}^{N_v}$, $\phi^u : \mathbb{R}^n \to \mathbb{R}^{N_u}$, $\phi^a : \mathbb{R}^n \to \mathbb{R}^{N_a}$ are basis functions such that $\phi^v(0) = \phi^u(0) = \phi^a(0) = 0$, and $\epsilon_i^v : \mathbb{R}^n \to \mathbb{R}$, $\epsilon_i^u : \mathbb{R}^n \to \mathbb{R}^m$, $\epsilon_i^a : \mathbb{R}^n \to \mathbb{R}^p$ are approximation errors. It is known that the approximation errors ϵ_i^v, ϵ_i^u and ϵ_i^a converge to zero uniformly on any compact set $\Omega \subset \mathbb{R}^n$, as $N_v, N_u, N_a \to \infty$.

Still, the weight matrices w_i^v, w_i^u, w_i^a are not known in advance and have to be identified. For this purpose, an actor–critic network structure is employed, which approximates V_i, u_{i+1} and a_{i+1} according to:

$$
\begin{aligned}
\hat{V}_i(x) &= (\hat{w}_i^v)^{\mathrm{T}} \phi^v(x), \\
\hat{u}_{i+1}(x) &= (\hat{w}_i^u)^{\mathrm{T}} \phi^u(x), \\
\hat{a}_{i+1}(x) &= (\hat{w}_i^a)^{\mathrm{T}} \phi^a(x),
\end{aligned} \tag{12}
$$

where $\hat{w}_i^v \in \mathbb{R}^{N_v}$ are the critic weights, and $\hat{w}_i^u \in \mathbb{R}^{N_u \times m}$, $\hat{w}_i^a \in \mathbb{R}^{N_a \times p}$ are the actor weights. The weights $\hat{w}_i^v$, $\hat{w}_i^u$, $\hat{w}_i^a$ then need to be trained through Eq. (10), so that $\hat{V}_i$, $\hat{u}_{i+1}$, $\hat{a}_{i+1}$ become good approximations of V_i, u_{i+1} and a_{i+1}. However, (10) cannot actually be constructed, owing to the clock offsets (3) that have corrupted the measured data. Hence, one is forced to construct the offseted Eq. (11) instead, and approximate the corrupted functions $\bar{V}_i$, $\bar{u}_{i+1}$, $\bar{a}_{i+1}$ in lieu of the actual ones.

Exploiting now the actor–critic structure (12), notice that the left-hand side of (11) can be approximated as:

$$
\hat{V}_i(\bar{x}(t_k')) - \hat{V}_i(\bar{x}(t_k)) = (\phi^v(\bar{x}(t_k')) - \phi^v(\bar{x}(t_k)))^{\mathrm{T}} \hat{w}_i^v. \tag{13}
$$

Additionally, the right-hand side of (11) can be approximated in terms of the offseted measured data, the actor policies at a previous step $i - 1 \in \mathbb{N}$, and the actor neural networks at step $i \in \mathbb{N}$. Specifically, the two right-most terms in (11) can be approximated as:

$$
\begin{aligned}
&2\hat{u}_{i+1}^{\mathrm{T}}(\bar{x}(\tau)) R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau))) \\
&\quad = 2(\phi^u(\bar{x}(\tau)))^{\mathrm{T}} \hat{w}_i^u R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau))) \\
&\quad = 2\Big(((\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))^{\mathrm{T}} R) \otimes \phi^u(\bar{x}(\tau))^{\mathrm{T}} \Big) \mathrm{vec}(\hat{w}_i^u),
\end{aligned} \tag{14}
$$

$$
\begin{aligned}
&2\gamma^2 \hat{a}_{i+1}^{\mathrm{T}}(\bar{x}(\tau))(\bar{a}(\tau) - \hat{a}_i(\bar{x}(\tau))) \\
&\quad = 2\gamma^2 (\phi^a(\bar{x}(\tau)))^{\mathrm{T}} \hat{w}_i^a (\bar{a}(\tau) - \hat{a}_i(\bar{x}(\tau))) \\
&\quad = 2\gamma^2 \Big(((\bar{a}(\tau) - \hat{a}_i(\bar{x}(\tau)))^{\mathrm{T}}) \otimes \phi^a(\bar{x}(\tau))^{\mathrm{T}} \Big) \mathrm{vec}(\hat{w}_i^a).
\end{aligned} \tag{15}
$$

The residual error by approximating (11) with the actor–critic structure is given, for $k \in \mathcal{K}$ and $i \in \mathbb{N}$, by

$$
\begin{aligned}
e_{i,k} = \hat{V}_i(\bar{x}(t_k')) - \hat{V}_i(\bar{x}(t_k)) &+ \int_{t_k}^{t_k'} \Big(Q(\bar{x}(\tau)) \\
&+ \hat{u}_i(\bar{x}(\tau))^{\mathrm{T}} R \hat{u}_i(\bar{x}(\tau)) + 2\hat{u}_{i+1}^{\mathrm{T}}(\bar{x}(\tau)) R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))
\end{aligned}
$$

Algorithm 3 Learning-based PI with Clock Offseted Data

```
Let $i = 0$, $\Omega \subset \mathbb{R}^n$, $\epsilon > 0$. Start with a tuple of policies $\{\hat{u}_0,\ \hat{a}_0\} = \{u_0,\ a_0\}$ that is stabilizing in $\Omega$.
repeat
    Solve for $\hat{W}_i$ through (17) and set $i = i + 1$.
until $i \geq 2$ & $\left\| \hat{W}_{i-1} - \hat{W}_{i-2} \right\| < \epsilon$.
```

$$- \gamma^2 \hat{a}_i^{\mathrm{T}}(\bar{x}(\tau))\hat{a}_i(\bar{x}(\tau)) - 2\gamma^2 \hat{a}_{i+1}^{\mathrm{T}}(\bar{x}(\tau))(\bar{a}(\tau) - \hat{a}_i(\bar{x}(\tau)))\Big)\mathrm{d}\tau.$$

Using (13)–(15), the residual error can be written in a linear form with respect to the actor–critic weights at step $i \in \mathbb{N}$ for all $k \in \mathcal{K}$, according to the formula:

$$e_{i,k} = \bar{\Psi}_{i,k}\hat{W}_i + \bar{\Phi}_{i,k}, \tag{16}$$

where $\hat{W}_i := [(\hat{w}_i^v)^{\mathrm{T}}\ \mathrm{vec}(\hat{w}_i^u)^{\mathrm{T}}\ \mathrm{vec}(\hat{w}_i^a)^{\mathrm{T}}]^{\mathrm{T}}$, $\bar{\Psi}_{i,k} := [\bar{\Psi}_{i,k}^v\ \bar{\Psi}_{i,k}^u\ \bar{\Psi}_{i,k}^a]$, and:

$$\begin{aligned}
\bar{\Psi}_{i,k}^v &:= \Big(\phi^v(\bar{x}(t_k')) - \phi^v(\bar{x}(t_k))\Big)^{\mathrm{T}},\\
\bar{\Psi}_{i,k}^u &:= \int_{t_k}^{t_k'} 2\Big((\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))^{\mathrm{T}}R\Big) \otimes \phi^u(\bar{x}(\tau))^{\mathrm{T}}\mathrm{d}\tau,\\
\bar{\Psi}_{i,k}^a &:= \int_{t_k}^{t_k'} -2\gamma^2\Big((\bar{a}(\tau) - \hat{a}_i(\bar{x}(\tau)))^{\mathrm{T}}\Big) \otimes \phi^a(\bar{x}(\tau))^{\mathrm{T}}\mathrm{d}\tau,\\
\bar{\Phi}_{i,k} &:= \int_{t_k}^{t_k'} \Big(Q(\bar{x}(\tau)) + \hat{u}_i(\bar{x}(\tau))^{\mathrm{T}}R\hat{u}_i(\bar{x}(\tau))\\
&\qquad - \gamma^2\hat{a}_i(\bar{x}(\tau))^{\mathrm{T}}\hat{a}_i(\bar{x}(\tau))\Big)\mathrm{d}\tau.
\end{aligned}$$

A standard assumption (Jiang & Jiang, 2014; Modares et al., 2015) is now needed, requiring the measured data to be sufficiently rich.

Assumption 1. There exist constants $\eta > 0$ and $K_0 \in \mathbb{N}$, such that if $K \geq K_0$ then $\frac{1}{K}\sum_{k=0}^{K} \bar{\Psi}_{i,k}^{\mathrm{T}}\bar{\Psi}_{i,k} \succ \eta I_{N_v+mN_u+pN_a}$, $\forall i \in \mathbb{N}$. □

While Assumption 1 is standard (Jiang & Jiang, 2014; Modares et al., 2015) and easy to check for $i = 0$, it can be difficult to verify a priori for all $i \in \mathbb{N}$. To deal with this issue, consider the i-independent data matrix $\bar{\Psi} = \begin{bmatrix}\bar{\Psi}_1 & \dots & \bar{\Psi}_K\end{bmatrix}^{\mathrm{T}}$, where for all $k = 0, \dots, K$:

$$\bar{\Psi}_k = \begin{bmatrix} \phi^v(\bar{x}(t_k')) - \phi^v(\bar{x}(t_k)) \\ \int_{t_k}^{t_k'} 2\big(R\bar{u}(\tau)\big) \otimes \phi^u(\bar{x}(\tau))\mathrm{d}\tau \\ \int_{t_k}^{t_k'} -2\phi^u(\bar{x}(\tau)) \otimes_h \phi^u(\bar{x}(\tau))\mathrm{d}\tau \\ \int_{t_k}^{t_k'} -2\gamma^2\big(\bar{a}(\tau)\big) \otimes \phi^a(\bar{x}(\tau))\mathrm{d}\tau \\ \int_{t_k}^{t_k'} 2\phi^a(\bar{x}(\tau)) \otimes_h \phi^a(\bar{x}(\tau))\mathrm{d}\tau \end{bmatrix}$$

The following Proposition (proved in the Appendix) provides an i-independent condition for Assumption 1 to hold, for all $i \in \mathbb{N}_+$.

Proposition 1. *If $\frac{1}{K}\lambda_{min}(\bar{\Psi}^{\mathrm{T}}\bar{\Psi}) \geq \eta$ then Assumption 1 holds for all $i \in \mathbb{N}_+$.* □

Given this condition, the weights of the actor–critic structure can be trained at each step $i \in \mathbb{N}$ of the learning process, according to the least-sum-of-squares law:

$$\hat{W}_i = -\Big(\sum_{k=0}^{K} \bar{\Psi}_{i,k}^{\mathrm{T}}\bar{\Psi}_{i,k}\Big)^{-1}\Big(\sum_{k=0}^{K} \bar{\Psi}_{i,k}^{\mathrm{T}}\bar{\Phi}_{i,k}\Big). \tag{17}$$

The overall procedure is described in Algorithm 3.

3.2. Convergence analysis

In what follows, we will show that the learning-based PI algorithm retains robustness with respect to the clock offsets given certain Lipschitz continuity assumptions. To this end, for all $i \in \mathbb{N}$, consider the function $\tilde{V}_i$ that satisfies the Lyapunov equation:

$$\begin{aligned}
\nabla\tilde{V}_i^{\mathrm{T}}(x)(f(x)\ &+g(x)\hat{u}_i(x) + h(x)\hat{a}_i(x))\\
&+Q(x) + \hat{u}_i(x)^{\mathrm{T}}R\hat{u}_i(x) - \gamma^2\hat{a}_i(x)^{\mathrm{T}}\hat{a}_i(x) = 0.
\end{aligned} \tag{18}$$

Additionally, define the following policy tuple:

$$\begin{aligned}
\tilde{u}_{i+1}(x) &= -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla\tilde{V}_i(x),\\
\tilde{a}_{i+1}(x) &= \frac{1}{2\gamma^2}h^{\mathrm{T}}(x)\nabla\tilde{V}_i(x).
\end{aligned} \tag{19}$$

Following the reasoning of Section 2.3, due to (18)–(19), it can be shown that over the trajectories of (1) we have

$$\begin{aligned}
\tilde{V}_i(x(t_k')) - \tilde{V}_i(x(t_k)) = \int_{t_k}^{t_k'} \Big(&-Q(x(\tau))\\
- \hat{u}_i(x(\tau))^{\mathrm{T}}R\hat{u}_i(x(\tau)) &+ \gamma^2\hat{a}_i(x(\tau))^{\mathrm{T}}\hat{a}_i(x(\tau))\\
- 2\tilde{u}_{i+1}^{\mathrm{T}}(x(\tau))&R(u(\tau) - \hat{u}_i(x(\tau)))\\
+ 2\gamma^2\tilde{a}_{i+1}^{\mathrm{T}}(x(\tau))&(a(\tau) - \hat{a}_i(x(\tau)))\Big)\,\mathrm{d}\tau,\ k \in \mathcal{K}.
\end{aligned} \tag{20}$$

The following auxiliary lemma is the first step towards proving that Algorithm 3 converges close to the desired functions, given that the clock offsets are uniformly bounded below a certain threshold, and provided that the number of basis functions is large enough. However, the following additional assumption is also required.

Assumption 2. The following hold:

- The functions $\tilde{V}_i$ are continuously differentiable on Ω, for all $i \in \mathbb{N}$, with locally Lipschitz gradients.
- The trajectories of the state $x(t)$ are Lipschitz on t and confined in Ω for all $t \geq t_0$.
- The trajectories of the control inputs $u(t)$, $a(t)$ are Lipschitz on t and uniformly bounded. □

Define now the function $\Delta(t)$ to be, for all $t \geq t_0$, the greatest of all clock offsets' magnitudes:

$$\Delta(t) = \max\left\{\max_{i\in\mathcal{N}_x}|\delta_i^x(t)|,\ \max_{j\in\mathcal{N}_u}|\delta_j^u(t)|,\ \max_{l\in\mathcal{N}_a}|\delta_l^a(t)|\right\}.$$

Then, the auxiliary lemma (proved in the Appendix) is stated as follows.

Lemma 1. *Let Assumptions 1–2 hold, and consider the iteration provided by Algorithm 3, for all $i \in \mathbb{N}$. Then, for all $\epsilon > 0$, there exist constants $N_v^\star,\ N_u^\star,\ N_a^\star \in \mathbb{N}_+$, and a strictly positive clock offset upper bound $\Delta^\star > 0$, such that if $N_v \geq N_v^\star$, $N_u \geq N_u^\star$, $N_a \geq N_a^\star$ and $\Delta(t) \leq \Delta^\star$ for all $t \geq t_0$, then it holds that:*

$$\begin{aligned}
&\sup_{x\in\Omega}|\hat{V}_i(x) - \tilde{V}_i(x)| \leq \epsilon,\\
&\sup_{x\in\Omega}\left\|\hat{u}_{i+1}(x) - \tilde{u}_{i+1}(x)\right\| \leq \epsilon,\\
&\sup_{x\in\Omega}\left\|\hat{a}_{i+1}(x) - \tilde{a}_{i+1}(x)\right\| \leq \epsilon. \quad \square
\end{aligned}$$

Remark 1. In the work of Jiang and Jiang (2014), it is proved that sufficiently good approximation of $\hat{V}_i$, $\hat{u}_{i+1}$, $\hat{a}_{i+1}$ can be achieved, in the absence of clock offsets, provided there are sufficiently many basis functions. However, in the presence of clock offsets, this result is no longer valid. Instead, as the proof of Lemma 1 demonstrates, a Lipschitz continuity assumption (Assumption 2) is required, which in turn leads to a different and more challenging theoretical analysis than that of Jiang and Jiang (2014). □

Lemma 1 studied how well the learning-based PI procedure can perform at a specific iteration $i \in \mathbb{N}$. The following theorem (proved in the Appendix) uses Lemma 1 to generalize its results and prove the desired epsilon-delta robustness result that we sought for. In particular, it shows that Algorithm 3 converges to an ϵ-suboptimal tuple of policies given that Assumption 2 holds, that the clock offsets remain below an ϵ-dependent threshold, and that the number of actor–critic nodes is sufficient.

Theorem 1. *Let Assumptions 1–2 hold, and consider the procedure provided by Algorithm 3, for all $i \in \mathbb{N}$. Then, for all $\epsilon > 0$, there exist constants $N_v^{\star\star}$, $N_u^{\star\star}$, $N_a^{\star\star}$, $i^\star \in \mathbb{N}_+$, and an upper clock mismatch bound $\Delta^{\star\star} > 0$, such that if $N_v \geq N_v^{\star\star}$, $N_u \geq N_u^{\star\star}$, $N_a \geq N_a^{\star\star}$, and $\Delta(t) \leq \Delta^{\star\star}$ for all $t \geq t_0$, then it holds that:*

$$\sup_{x \in \Omega} \left\| \hat{V}_{i^\star}(x) - V^\star(x) \right\| \leq \epsilon,$$

$$\sup_{x \in \Omega} \left\| \hat{u}_{i^\star+1}(x) - u^\star(x) \right\| \leq \epsilon,$$

$$\sup_{x \in \Omega} \left\| \hat{a}_{i^\star+1}(x) - a^\star(x) \right\| \leq \epsilon. \quad \square$$

3.3. Learning-based PI with clock offsets and quantization

In this subsection, we consider that the measurements (2) of the state do not only suffer from clock offsets, but also from quantization. Specifically, we employ the logarithmic quantizer considered in Elia and Mitter (2001), Okano et al. (2017): given $\alpha_0 > 0$ and $\rho \in (0,\ 1)$, the quantizer $r : \mathbb{R} \to \mathbb{R}$ is defined as:

$$r(y) = \begin{cases} \frac{2\rho}{\rho+1}\rho^\ell \alpha_0 & \text{if } y \in (\rho^{\ell+1}\alpha_0,\ \rho^\ell \alpha_0], \\ 0 & \text{if } y = 0, \\ -\frac{2\rho}{\rho+1}\rho^\ell \alpha_0 & \text{if } y \in [-\rho^\ell \alpha_0,\ -\rho^{\ell+1}\alpha_0), \end{cases} \tag{21}$$

where $\ell \in \mathbb{Z}$ is the integer for which the inequality $\rho^{\ell+1}\alpha_0 \langle |y| \leq \rho^\ell \alpha_0$ holds. Evidently, the quantization becomes finer as $\rho \to 1$, and coarser as $\rho \to 0$. It is also helpful to define as $r^n : \mathbb{R}^n \to \mathbb{R}^n$, the quantizer which applies the operator (21) entry-wise on vectors in $\mathbb{R}^n$.

The quantization further dilutes the measurements of the state signals; instead of measuring $\bar{x}_i(t)$ as in (2), $\forall i \in \mathcal{N}_x$, one now only has access to the quantized values:

$$\hat{x}_i(t) = r(\bar{x}_i(t)) = r(x_i(c_i^x(t))) = r(x_i(t + \delta_i^x(t))),$$

with the whole quantized vector being $\hat{x}(t) = [\hat{x}_1(t)\ \hat{x}_2(t)\ \ldots\ \hat{x}_n(t)]^{\mathrm{T}}$. Notice that a quantization of u or a will not have any impact on the learning-based PI, as both the measured and the actual values of the control input signals implemented in (1) will be quantized. On the other hand, although the measured value of the state x is quantized, its actual value is continuous owing to the continuous-flow nature of (1). As such, discrepancies will exist between state measurements and reality, which can jeopardize the convergence of the learning-based PI.

To analyze learning-based PI under quantization, the following Lemma (proved in the Appendix) is needed.

Lemma 2. *Let $z \in D \subset \mathbb{R}^n$, where D is compact. Then, the quantization error $\|r^n(z) - z\|$ is bounded on D. In addition, $\lim_{\rho \nearrow 1} \|r^n(z) - z\| = 0$ uniformly on D.* □

Algorithm 4 Learning-based PI with Clock Offseted and Quantized Data

```
1: Let $i = 0$, $\Omega \subset \mathbb{R}^n$, $\epsilon > 0$. Start with a tuple of policies
   $\{\hat{u}_0,\ \hat{a}_0\} = \{u_0,\ a_0\}$ that is stabilizing in $\Omega$.
2: repeat
3:     Solve for $\hat{W}_i$ through (22) and set $i = i + 1$.
4: until $i \geq 2$ & $\left\| \hat{W}_{i-1} - \hat{W}_{i-2} \right\| < \epsilon$.
```

Exploiting Lemma 2, the analysis of the previous subsection can be extended to the case of quantization. Particularly, one can show that learning-based PI always converges to an ϵ neighborhood of the optimal control and value function, given a sufficiently fine quantizer. The actual equations of the learning-based PI with quantization will be identical to the quantization-free case, with the exception that quantized data are utilized instead. Particularly, the weights are trained as:

$$\hat{W}_i = -\Big(\sum_{k=0}^{K} \hat{\Psi}_{i,k}^{\mathrm{T}} \hat{\Psi}_{i,k}\Big)^{-1} \Big(\sum_{k=0}^{K} \hat{\Psi}_{i,k}^{\mathrm{T}} \hat{\Phi}_{i,k}\Big), \tag{22}$$

where $\hat{W}_i := [(\hat{w}_i^v)^{\mathrm{T}}\ \mathrm{vec}(\hat{w}_i^u)^{\mathrm{T}}\ \mathrm{vec}(\hat{w}_i^a)^{\mathrm{T}}]^{\mathrm{T}}$, $\Psi_{i,k} := [\hat{\Psi}_{i,k}^v\ \hat{\Psi}_{i,k}^u\ \hat{\Psi}_{i,k}^a]$, and:

$$\hat{\Psi}_{i,k}^v := \Big(\phi^v(\hat{x}(t_k')) - \phi^v(\hat{x}(t_k))\Big)^{\mathrm{T}},$$

$$\hat{\Psi}_{i,k}^u := \int_{t_k}^{t_k'} 2\Big((\bar{u}(\tau) - \hat{u}_i(\hat{x}(\tau)))^{\mathrm{T}} R\Big) \otimes \phi^u(\hat{x}(\tau))^{\mathrm{T}} \mathrm{d}\tau,$$

$$\hat{\Psi}_{i,k}^a := \int_{t_k}^{t_k'} -2\gamma^2 \Big((\bar{a}(\tau) - \hat{a}_i(\hat{x}(\tau)))^{\mathrm{T}}\Big) \otimes \phi^a(\hat{x}(\tau))^{\mathrm{T}} \mathrm{d}\tau,$$

$$\hat{\Phi}_{i,k} := \int_{t_k}^{t_k'} \Big(Q(\hat{x}(\tau)) + \hat{u}_i(\hat{x}(\tau))^{\mathrm{T}} R \hat{u}_i(\hat{x}(\tau)) - \gamma^2 \hat{a}_i(\hat{x}(\tau))^{\mathrm{T}} \hat{a}_i(\hat{x}(\tau))\Big) \mathrm{d}\tau.$$

The procedure is shown in Algorithm 4. For (22) to be properly defined, we require the following assumption.

Assumption 3. There exist constants $\hat{\eta} > 0$ and $\hat{K}_0 \in \mathbb{N}$, such that if $\hat{K} \geq K_0$ then $\frac{1}{\hat{K}} \sum_{k=0}^{\hat{K}} \hat{\Psi}_{i,k}^{\mathrm{T}} \hat{\Psi}_{i,k} \succ \hat{\eta} I_{N_v + mN_u + pN_a}$, $\forall i \in \mathbb{N}$. □

The convergence properties (proved in the Appendix) of the learning-based PI under both clock offsets and quantization are stated next.

Theorem 2. *Let Assumptions 2–3 hold, and consider the procedure provided by Algorithm 4, for all $i \in \mathbb{N}$. Then, for all $\epsilon > 0$, there exist constants $N_v^{\star\star\star}$, $N_u^{\star\star\star}$, $N_a^{\star\star\star}$, $i^\star \in \mathbb{N}_+$, an upper clock mismatch bound $\Delta^{\star\star\star} > 0$ and a low-bound $\rho^\star \in (0, 1)$ for the quantizer's coarseness, such that if $N_v \geq N_v^{\star\star\star}$, $N_u \geq N_u^{\star\star\star}$, $N_a \geq N_a^{\star\star\star}$, $\rho \in (\rho^\star, 1)$ and $\Delta(t) \leq \Delta^{\star\star\star}$ for all $t \geq t_0$, then it holds that:*

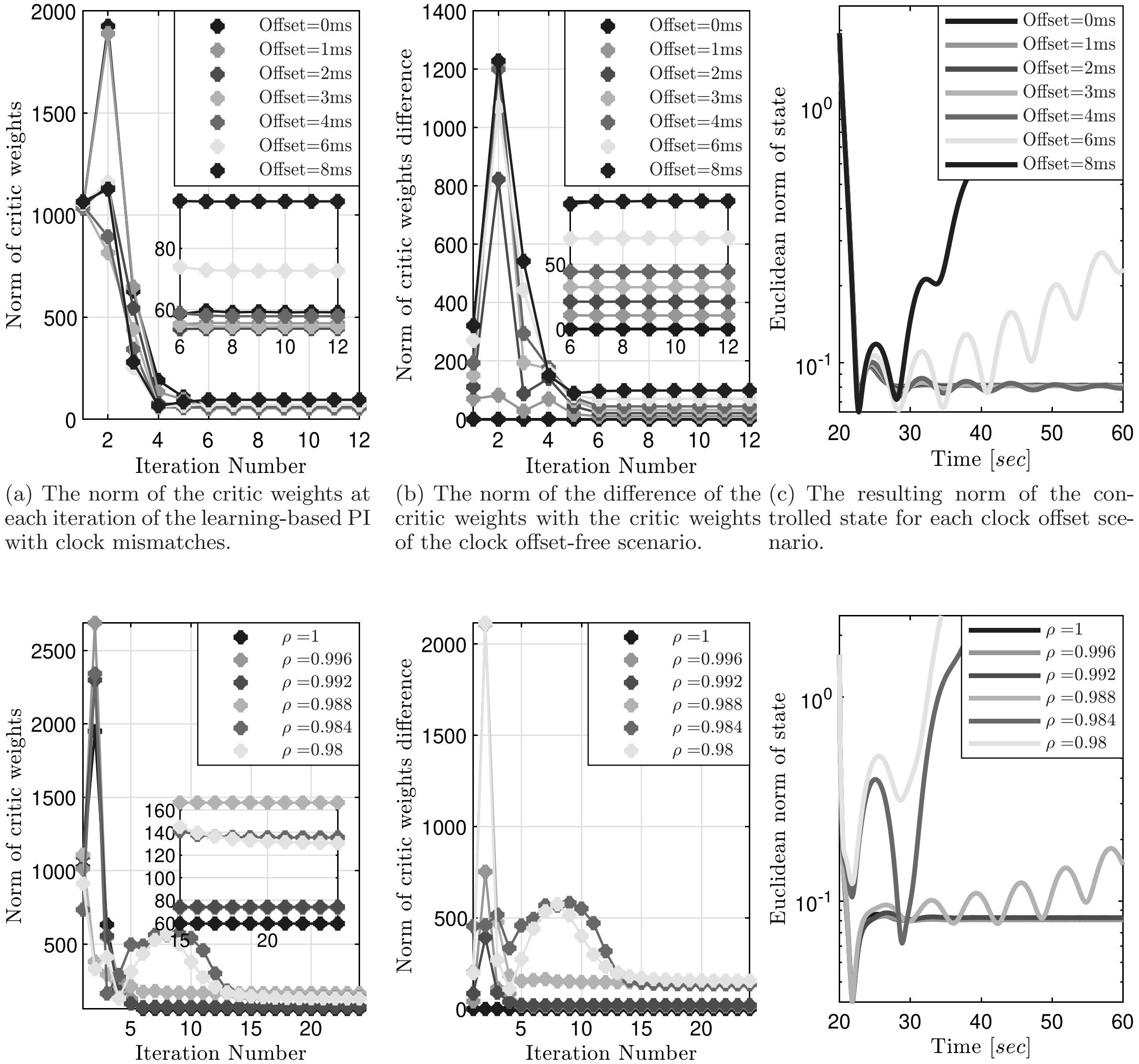
$$\sup_{x \in \Omega} \left\| \hat{V}_{i^\star}(x) - V^\star(x) \right\| \leq \epsilon,$$

$$\sup_{x \in \Omega} \left\| \hat{u}_{i^\star+1}(x) - u^\star(x) \right\| \leq \epsilon,$$

$$\sup_{x \in \Omega} \left\| \hat{a}_{i^\star+1}(x) - a^\star(x) \right\| \leq \epsilon. \quad \square$$

4. Simulation results

We consider a two-link manipulator (Modares et al., 2015), where the state vector is $x = [q^{\mathrm{T}}\ \dot{q}^{\mathrm{T}}]^{\mathrm{T}}$, and $q \in \mathbb{R}^2$ and $\dot{q} \in \mathbb{R}^2$ are the angular positions (in rad) and the angular velocities (in rad/s), respectively. The control input $u \in \mathbb{R}^2$ in this model denotes the torque, while $a \in \mathbb{R}^2$ is the adversarial input (both in Nm). The objective is to approximate the optimal game-based



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

(b) The norm of the difference of the critic weights with the critic weights of the clock offset-free scenario.

(c) The resulting norm of the controlled state for each clock offset scenario.

(d) The norm of the critic weights at each iteration of the learning-based PI with quantization.

(e) The norm of the difference of the critic weights with the critic weights of the quantization-free scenario.

(f) The resulting norm of the controlled state for each quantization scenario.

Fig. 1. Evolution of the learning-based PI in presence of clock offsets (top figures) and quantization (bottom figures).

value function $V^\star$ and controller $u^\star$ of (4), where $Q(x) = 5\|x\|^2$, $R = I_2$ and $\gamma^2 = 20$. To this end, the actor–critic network (12) is employed, with basis functions given by polynomials of order up to 4.

First, we consider a setup where the input data sampled from the system suffers from clock offsets. To showcase the gradual deterioration in performance as the magnitude of these offsets increases, we simulate 7 different cases, in each of which the clock offsets of the input signals $\delta_i^u(t)$, $\delta_i^a(t)$ take the constant values 0 ms, 1 ms, 2 ms, 3 ms, 4 ms, 6 ms and 8 ms, respectively, for all $t \geq t_0$, $i = 1, 2$. In each simulation, the first 20 seconds are used for exploration to gather sufficient state-input data from the system. Subsequently, the learning-based PI with clock mismatches (Algorithm 3) is carried out with $T = 50$ ms, by iteratively solving Eqs. (17). The learnt controller is then implemented in the system for all $t \geq 20$ s, where the adversarial input is $a(t) = 0.5\sin(t)[1\ 1]^{\mathrm{T}}$.

The results are shown in Figs. 1(a)–1(c). It can be seen from Fig. 1(a) that convergence of the learning-based PI algorithm takes place for all values of the clock mismatch. However, as seen from Fig. 1(b) and as expected from Theorem 1, the network weights converge monotonically further away from their nominal values (i.e., their values when the clock mismatch is zero) as the magnitude of the offset is increased. Note here that we cannot directly compare the network weights with their optimal values, because these optimal values are unknown. However, the weights of the offset-free case are known to converge uniformly to the optimal ones (Jiang & Jiang, 2014; Modares et al., 2015), hence Fig. 1(b) gives some information about how close to optimality we are at each offset scenario. Fig. 1(c) shows the evolution of the norm of the state vector for each clock mismatch scenario. In spite of the persistent disturbance, the closed loop remains bounded for offsets ranging up to 4 ms. However, the system's trajectories drift to infinity for offsets equal to 6−8 ms. Finally, Table 1 shows the finite-horizon performance cost over $t \in [20, 60]$, which increases as the offset increases.

Second, we consider a setup where the state data sampled from the system suffers from logarithmic quantization, with $a_0 = 1$. We specifically simulate 6 different cases, in each of which the characteristic value ρ of the quantizer is 1, 0.996, 0.992, 0.988,

Table 1
Performance cost over $t \in [20, 60]$ in each offset scenario.

Offset	0 ms	1 ms	2 ms	3 ms	4 ms	6 ms	8 ms
Cost	16.17	16.18	16.19	16.19	16.2	21.71	1882

0.984 and 0.980, respectively; the value $\rho = 1$ is used with a slight abuse of notation to describe the quantization-free case. Similarly to the previous setup, the first 20 seconds are used for exploration, and subsequently the learning-based PI with quantization (Algorithm 4) is carried out with $T = 50$ ms, by iteratively solving Eqs. (22). The learnt controller is then implemented in the system for all $t \geq 20$ s, where the adversarial input is $a(t) = 0.5\sin(t)[1\ 1]^{\mathrm{T}}$. The results present a behavior similar to the clock offset scenario: while convergence of the learning takes place for all quantizers as shown in Fig. 1(d), the network weights converge further away from their nominal values as the quantizer becomes more coarse (Fig. 1(e)). Moreover, for the more coarse quantizers, the system's state diverges under the learnt control law (Fig. 1(f)).

5. Conclusion

This paper considered the effect of clock offsets and quantization on an RL algorithm used to solve a zero-sum game. Given certain restrictions on the magnitude of these, as well as some Lipschitz continuity and differentiability assumptions, it was shown that the game's solution can still be approximately computed. Possible future directions include studying the effect of clock offsets and quantization on online RL algorithms.

Acknowledgments

This work was supported in part, by ARO, United States under grant No. W911NF-19-1-0270, by NSF, USA under grant Nos. CAREER CPS-1851588 and SATC-1801611, by the Onassis Foundation-Scholarship ID: F ZQ 064-1/2020-2021 (USA). Copyright 2023 Carnegie Mellon University and Georgia Tech. This material is based upon work funded and supported by the Department of Defense, USA 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. DM23-0306

Appendix

Proof of Proposition 1. For all $i \in \mathbb{N}_+$, define the matrix

$$S_i = \begin{bmatrix} I_{N_v} & 0 & 0 \\ 0 & I_{mN_u} & 0 \\ 0 & C_{N_u}^{\mathrm{T}}((\hat{w}_{i-1}^u R) \otimes I_{N_u}) & 0 \\ 0 & 0 & I_{pN_a} \\ 0 & 0 & \gamma^2 C_{N_a}^{\mathrm{T}}((\hat{w}_{i-1}^a) \otimes I_{N_a}) \end{bmatrix},$$

where the 0 entries indicate zero matrices of appropriate dimensions. We notice that $\bar{\Psi}_{i,k} = \bar{\Psi}_k^{\mathrm{T}} S_i$. Therefore, it holds that $\frac{1}{K}\sum_{k=0}^{K} \bar{\Psi}_{i,k}^{\mathrm{T}} \bar{\Psi}_{i,k} = \frac{1}{K}\sum_{k=0}^{K} S_i^{\mathrm{T}} \bar{\Psi}_k \bar{\Psi}_k^{\mathrm{T}} S_i = \frac{1}{K} S_i^{\mathrm{T}} \left(\sum_{k=0}^{K} \bar{\Psi}_k \bar{\Psi}_k^{\mathrm{T}}\right) S_i = \frac{1}{K} S_i^{\mathrm{T}} \bar{\Psi}^{\mathrm{T}} \bar{\Psi} S_i$. This yields

$$\frac{1}{K}\sum_{k=0}^{K} \bar{\Psi}_{i,k}^{\mathrm{T}} \bar{\Psi}_{i,k} \succeq \frac{1}{K} \lambda_{\min}(\bar{\Psi}^{\mathrm{T}} \bar{\Psi}) S_i^{\mathrm{T}} S_i$$
$$\succeq \frac{1}{K} \lambda_{\min}(\bar{\Psi}^{\mathrm{T}} \bar{\Psi}) \lambda_{\min}(S_i^{\mathrm{T}} S_i) I_{N_v + mN_u + pN_a}.$$

But $S_i^{\mathrm{T}} S_i$ satisfies the strict equality $\lambda_{\min}(S_i^{\mathrm{T}} S_i) = 1$, while $\lambda_{\min}(\bar{\Psi}^{\mathrm{T}} \bar{\Psi}) \geq K\eta$ by assumption. Thus, we obtain $\frac{1}{K}\sum_{k=0}^{K} \bar{\Psi}_{i,k}^{\mathrm{T}} \bar{\Psi}_{i,k} \succ \eta I_{N_v + mN_u + pN_a}$, $\forall i \in \mathbb{N}_+$. ■

Proof of Lemma 1. By the Weierstrass approximation theorem, $\tilde{V}_i$, $\tilde{u}_{i+1}$, $\tilde{a}_{i+1}$ can be approximated on Ω as:

$$\begin{aligned} \tilde{V}_i(x) &= (\tilde{w}_i^v)^{\mathrm{T}} \phi^v(x) + \tilde{\epsilon}_i^v(x), \\ \tilde{u}_{i+1}(x) &= (\tilde{w}_i^u)^{\mathrm{T}} \phi^u(x) + \tilde{\epsilon}_i^u(x), \\ \tilde{a}_{i+1}(x) &= (\tilde{w}_i^a)^{\mathrm{T}} \phi^a(x) + \tilde{\epsilon}_i^a(x). \end{aligned} \tag{23}$$

The approximation errors $\tilde{\epsilon}_i^v : \mathbb{R}^n \to \mathbb{R}$, $\tilde{\epsilon}_i^u : \mathbb{R}^n \to \mathbb{R}^m$, $\tilde{\epsilon}_i^a : \mathbb{R}^n \to \mathbb{R}^p$ vanish uniformly on Ω as $N_v, N_u, N_a \to \infty$. Substituting (23) in (20), for $i \in \mathbb{N}$, we derive:

$$0 = \Psi_{i,k} \tilde{W}_i + \Phi_{i,k} + \tilde{E}_{i,k}, \ \ k \in \mathbb{N}, \tag{24}$$

where $\tilde{W}_i = [\tilde{w}_i^{v\mathrm{T}}\ \mathrm{vec}(\tilde{w}_i^u)^{\mathrm{T}}\ \mathrm{vec}(\tilde{w}_i^a)^{\mathrm{T}}]^{\mathrm{T}}$, $\Psi_{i,k} := [\Psi_{i,k}^v\ \Psi_{i,k}^u\ \Psi_{i,k}^a]$, and:

$$\Psi_{i,k}^v = \Big(\phi^v(x(t_k')) - \phi^v(x(t_k))\Big)^{\mathrm{T}},$$

$$\Psi_{i,k}^u = \int_{t_k}^{t_k'} 2\Big((u(\tau) - \hat{u}_i(x(\tau)))^{\mathrm{T}} R\Big) \otimes \phi^u(x(\tau))^{\mathrm{T}} \mathrm{d}\tau,$$

$$\Psi_{i,k}^a = \int_{t_k}^{t_k'} -2\gamma^2\Big((a(\tau) - \hat{a}_i(x(\tau)))^{\mathrm{T}}\Big) \otimes \phi^a(x(\tau))^{\mathrm{T}} \mathrm{d}\tau,$$

$$\Phi_{i,k} = \int_{t_k}^{t_k'} \Big(Q(x(\tau)) + \hat{u}_i(x(\tau))^{\mathrm{T}} R \hat{u}_i(x(\tau)) - \gamma^2 \hat{a}_i(x(\tau))^{\mathrm{T}} \hat{a}_i(x(\tau))\Big) \mathrm{d}\tau,$$

$$\begin{aligned} \tilde{E}_{i,k} = {} & \tilde{\epsilon}_i^v(x(t_k')) - \tilde{\epsilon}_i^v(x(t_k)) \\ & + \int_{t_k}^{t_k'} 2\tilde{\epsilon}_i^u(x(\tau))^{\mathrm{T}} R(u(\tau) - \hat{u}_i(x(\tau))) \mathrm{d}\tau \\ & - \int_{t_k}^{t_k'} 2\gamma^2 \tilde{\epsilon}_i^a(x(\tau))^{\mathrm{T}} (a(\tau) - \hat{a}_i(x(\tau))) \mathrm{d}\tau. \end{aligned}$$

Adding and subtracting identical terms in (24), one has

$$0 = \bar{\Psi}_{i,k} \tilde{W}_i + \bar{\Phi}_{i,k} + \Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k} + \tilde{E}_{i,k}, \tag{25}$$

where $\Delta\Psi_{i,k} = \Psi_{i,k} - \bar{\Psi}_{i,k}$ and $\Delta\Phi_{i,k} = \Phi_{i,k} - \bar{\Phi}_{i,k}$. Note that from (25), we have $\tilde{e}_{i,k} = \bar{\Psi}_{i,k} \tilde{W}_i + \bar{\Phi}_{i,k}$, where $\tilde{e}_{i,k} = -(\Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k} + \tilde{E}_{i,k})$, while (16) specifies $e_{i,k} = \bar{\Psi}_{i,k} \hat{W}_i + \bar{\Phi}_{i,k}$. Since $\hat{W}_i$ is estimated through the least sum of squares law (17) to minimize the sum of squared errors $\sum_{k=0}^{K} e_{i,k}^2$ (valid due to Assumption 1), it must hold that $\sum_{k=0}^{K} e_{i,k}^2 \leq \sum_{k=0}^{K} \tilde{e}_{i,k}^2$. Hence:

$$\sum_{k=0}^{K} e_{i,k}^2 \leq \sum_{k=0}^{K} (\tilde{E}_{i,k} + \Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k})^2, \ \ i \in \mathbb{N}. \tag{26}$$

Subtracting (25) from (16), we obtain:

$$e_{i,k} + \tilde{E}_{i,k} + \Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k} = \bar{\Psi}_{i,k}(\hat{W}_i - \tilde{W}_i). \tag{27}$$

Multiplying (27) by itself and summing over k leads to

$$\begin{aligned} & \frac{1}{K}\sum_{k=0}^{K} (e_{i,k} + \tilde{E}_{i,k} + \Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k})^2 \\ & = \frac{1}{K}\sum_{k=0}^{K} (\hat{W}_i - \tilde{W}_i)^{\mathrm{T}} \bar{\Psi}_{i,k}^{\mathrm{T}} \bar{\Psi}_{i,k} (\hat{W}_i - \tilde{W}_i) \geq \eta \left\| \hat{W}_i - \tilde{W}_i \right\|^2, \end{aligned} \tag{28}$$

where Assumption 1 was used. However, owing to (26):

$$\begin{aligned} & \frac{1}{K}\sum_{k=0}^{K} (e_{i,k} + \tilde{E}_{i,k} + \Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k})^2 \\ & \leq \frac{4}{K}\sum_{k=0}^{K} (\tilde{E}_{i,k} + \Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k})^2 \\ & \leq \max_{1 \leq k \leq K} 4(\tilde{E}_{i,k} + \Delta\Psi_{i,k} \tilde{W}_i + \Delta\Phi_{i,k})^2, \end{aligned}$$

which, combined with (28) yields:

$$\max_{1\leq k\leq K} \frac{4}{\eta}(\tilde{E}_{i,k} + \Delta\Psi_{i,k}\tilde{W}_i + \Delta\Phi_{i,k})^2 \geq \left\|\hat{W}_i - \tilde{W}_i\right\|^2 . \tag{29}$$

Now, notice that given Assumption 2, $\tilde{E}_{i,k} \to 0$ as $N_v, N_u, N_a \to \infty$ (uniformly on any trajectories of x on Ω). Additionally, one has $\Delta\Phi_{i,k} = \tilde{V}_i(x(t'_k)) - \tilde{V}_i(x(t_k)) - \tilde{V}_i(\bar{x}(t'_k)) + \tilde{V}_i(\bar{x}(t_k))$ and $\Delta\Psi_{i,k}\tilde{W}_i = A_{i,k} + B_{i,k}$, where:

$$\begin{aligned}A_{i,k} =&\int_{t_k}^{t'_k}\Big(2(\tilde{u}_{i+1}(x(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(u(\tau) - \hat{u}_i(x(\tau)))\\ &- 2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(\bar{x}(\tau)))^{\mathrm{T}}R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))\Big)\mathrm{d}\tau,\\ B_{i,k} =&\int_{t_k}^{t'_k}\Big(-2\gamma^2(\tilde{a}_{i+1}(x(\tau)) - \tilde{\epsilon}_i^a(x(\tau)))^{\mathrm{T}}(a(\tau) - \hat{a}_i(x(\tau)))\\ &+ 2\gamma^2(\tilde{a}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^a(\bar{x}(\tau)))^{\mathrm{T}}(\bar{a}(\tau) - \hat{a}_i(\bar{x}(\tau)))\Big)\mathrm{d}\tau.\end{aligned}$$

Using Assumption 2, one has:

$$\begin{aligned}|\Delta\Phi_{i,k}| &\leq |\tilde{V}_i(x(t'_k)) - \tilde{V}_i(\bar{x}(t'_k))| + |\tilde{V}_i(x(t_k)) - \tilde{V}_i(\bar{x}(t_k))|\\ &\leq L_{V_i}\left\|x(t'_k) - \bar{x}(t'_k)\right\| + L_{V_i}\left\|x(t_k) - \bar{x}(t_k)\right\|\\ &\leq L_{V_i}\sum_{i=1}^{n}\left(|x_i(t'_k) - \bar{x}_i(t'_k)| + |x_i(t_k) - \bar{x}_i(t_k)|\right)\\ &= L_{V_i}\sum_{i=1}^{n}\Big(|x_i(t'_k) - x_i(t'_k + \delta_i^x(t'_k))|\\ &\qquad + |x_i(t_k) - x_i(t_k + \delta_i^x(t_k))|\Big)\\ &\leq L_{V_i}L_x\sum_{i=1}^{n}(|\delta_i^x(t_k)| + |\delta_i^x(t'_k)|),\end{aligned}$$

where L_{V_i} is the Lipschitz constant of $\tilde{V}_i$ on Ω and L_x is the Lipschitz constant of $x(t)$ on t, which exist owing to Assumption 2. Hence, if $\Delta(t) \leq \Delta_M$ then

$$|\Delta\Phi_{i,k}| \leq 2nL_{V_i}L_x\Delta_M. \tag{30}$$

Inequality (30) implies that $\Delta\Phi_{i,k} \to 0$ (uniformly on any trajectories of x on Ω) as $\Delta_M \searrow 0$ and $N_v, N_u, N_a \to \infty$. In addition:

$$\begin{aligned}|A_{i,k}| \leq&\int_{t_k}^{t'_k}\Big(\Big|2(\tilde{u}_{i+1}(x(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(u(\tau) - \hat{u}_i(x(\tau)))\\ &- 2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(u(\tau) - \hat{u}_i(x(\tau)))\Big|\\ &+ \Big|2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(u(\tau) - \hat{u}_i(x(\tau)))\\ &- 2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau))^{\mathrm{T}}R(\bar{u}(\tau) - \hat{u}_i(x(\tau)))\Big|\\ &+ \Big|2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(\bar{u}(\tau) - \hat{u}_i(x(\tau)))\\ &- 2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))\Big|\\ &+ \Big|2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))\\ &- 2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(\bar{x}(\tau)))^{\mathrm{T}}R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))\Big|\Big)\,\mathrm{d}\tau\\ =&\int_{t_k}^{t'_k}\Big(\Big|2(\tilde{u}_{i+1}(x(\tau)) - \tilde{u}_{i+1}(\bar{x}(\tau)))^{\mathrm{T}}R(u(\tau) - \hat{u}_i(x(\tau)))\Big|\\ &+ \Big|2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(u(\tau) - \bar{u}(\tau))\Big|\\ &+ \Big|2(\tilde{u}_{i+1}(\bar{x}(\tau)) - \tilde{\epsilon}_i^u(x(\tau)))^{\mathrm{T}}R(\hat{u}_i(x(\tau)) - \hat{u}_i(\bar{x}(\tau)))\Big|\\ &+ \Big|2(\tilde{\epsilon}_i^u(x(\tau)) - \tilde{\epsilon}_i^u(\bar{x}(\tau)))^{\mathrm{T}}R(\bar{u}(\tau) - \hat{u}_i(\bar{x}(\tau)))\Big|\Big)\,\mathrm{d}\tau.\end{aligned}$$

After using Assumption 2, one has:

$$\begin{aligned}&\left\|\tilde{u}_{i+1}(x(\tau)) - \tilde{u}_{i+1}(\bar{x}(\tau))\right\| \leq nL_{\tilde{u}_{i+1}}L_x\Delta_M,\\ &\|u(\tau) - \bar{u}(\tau)\| \leq mL_u\Delta_M,\\ &\left\|\hat{u}_i(x(\tau)) - \hat{u}_i(\bar{x}(\tau))\right\| \leq nL_{\hat{u}_i}L_x\Delta_M,\\ &\left\|\tilde{\epsilon}_i^u(x(\tau)) - \tilde{\epsilon}_i^u(\bar{x}(\tau))\right\| \overset{N_u\to\infty}{\to} 0 \quad \text{uniformly}\end{aligned}$$

where $L_{\tilde{u}_{i+1}}, L_{\hat{u}_i}$ are the Lipschitz constants of $\tilde{u}_{i+1}, \hat{u}_i$ on Ω, and L_u is the Lipschitz constant of u on $t \geq t_0$ (Assumption 2). Therefore, since the state is confined in Ω and the control input u is uniformly bounded, it follows that $A_{i,k} \to 0$ (uniformly on any trajectories of x on Ω) as $\Delta_M \searrow 0$ and $N_v, N_u, N_a \to \infty$. The proof that $B_{i,k} \to 0$ is identical, hence it follows that $\Delta\Psi_{i,k} \to 0$ uniformly. Therefore, due to (29), for every ϵ_1 there exist constants $N_v^m,\ N_u^m,\ N_a^m$ and a clock mismatch upper bound Δ^m, such that if $N_v \geq N_v^m$, $N_u \geq N_u^m$, $N_a \geq N_a^m$ and $\Delta(t) \leq \Delta^m$ for all $t \geq t_0$, then $\left\|\hat{W}_i - \tilde{W}_i\right\| \leq \epsilon_1$. The final result follows from this inequality, the uniform convergence the approximation errors to zero and the boundedness of the basis functions on Ω. ■

Proof of Lemma 2. Boundedness of the quantization error on D follows from the boundedness of both z and $r^n(z)$ on D. In proving uniform convergence, let z_i be the ith entry of z, $i \in \mathcal{N}_x$. Then, there exist three cases:

(i) If $z_i = 0$ then $|r(z_i) - z_i| = 0$ by definition.

ii) Let $z_i > 0$. Then $r(z_i) = \frac{2\rho}{\rho+1}\rho^\ell\alpha_0$, where $\ell \in \mathbb{Z}$ is such that $\rho^{\ell+1}\alpha_0 < z_i \leq \rho^\ell\alpha_0$. Consequently:

$$\begin{aligned}&\rho^\ell\alpha_0\left(\frac{2\rho}{\rho+1} - 1\right) \leq r(z_i) - z_i < \rho^\ell\alpha_0\left(\frac{2\rho}{\rho+1} - \rho\right),\\ &\Longrightarrow |r(z_i) - z_i| < \rho^\ell\alpha_0\left(1 - \frac{2\rho}{\rho+1}\right).\end{aligned} \tag{31}$$

Note, $\rho^\ell\alpha_0 = \rho^{-1}\rho^{\ell+1}\alpha_0 < \rho^{-1}z_i \leq \rho^{-1}z_m$, where $z_m = \max_{z\in D}\|z\|_\infty$, and the maximum exists owing to D being compact. Therefore, (31) becomes $|r(z_i) - z_i| < z_m\left(\frac{1}{\rho} - \frac{2}{\rho+1}\right)$. It is now evident that for all $\epsilon > 0$, there exists $\rho^\star \in (0, 1)$, where $\rho^\star = \frac{z_m}{z_m+\epsilon}$, such that if $\rho \in (\rho^\star, 1)$ then $|r(z_i) - z_i| < \epsilon$. Hence, uniform convergence is proved.

iii) The case where $z_i < 0$ is similar to ii). ■

Proof of Theorem 1. Assume that for some $\bar{\Delta} > 0$, $\Delta(t) \leq \bar{\Delta}$, $\forall t \geq t_0$. We will follow an induction.

For $i = 0$, we have $\tilde{V}_0 = V_0$, $\tilde{u}_1 = u_1$ and $\tilde{a}_1 = a_1$, since $\hat{u}_0 = u_0$ and $\hat{a}_0 = a_0$. Hence, it follows from Lemma 1 that $\lim_{N_v,N_u,N_a\to\infty,\bar{\Delta}\searrow 0}\{\hat{V}_0(x), \hat{u}_1(x), \hat{a}_1(x)\} = \{V_0(x), u_1(x), a_1(x)\}$, uniformly on Ω.

Assume now that $\lim_{N_v,N_u,N_a\to\infty,\bar{\Delta}\searrow 0}\{\hat{V}_{i-1}(x), \hat{u}_i(x), \hat{a}_i(x)\} = \{V_{i-1}(x), u_i(x), a_i(x)\}$, uniformly on Ω, for some $i \in \mathbb{N}_+$. Then, by the definitions of V_i, $\tilde{V}_i$ and over the trajectories of $u : \mathbb{R}^n \to \mathbb{R}^m$, $a : \mathbb{R}^n \to \mathbb{R}^p$, we have:

$$\begin{aligned}&\left|V_i(x(t)) - \tilde{V}_i(x(t))\right| =\\ &\Big|\int_t^\infty\Big(u_i(x(\tau))^{\mathrm{T}}Ru_i(x(\tau)) + 2u_{i+1}^{\mathrm{T}}(x(\tau))Rv_i^u(x(\tau))\Big)\mathrm{d}\tau\\ &-\int_t^\infty\Big(\hat{u}_i(x(\tau))^{\mathrm{T}}R\hat{u}_i(x(\tau)) + 2\tilde{u}_{i+1}^{\mathrm{T}}(x(\tau))R\hat{v}_i^u(x(\tau))\Big)\mathrm{d}\tau\\ &-\int_t^\infty\Big(\gamma^2a_i(x(\tau))^{\mathrm{T}}a_i(x(\tau)) + 2\gamma^2a_{i+1}^{\mathrm{T}}(x(\tau))v_i^a(x(\tau))\Big)\mathrm{d}\tau\\ &+\int_t^\infty\Big(\gamma^2\hat{a}_i(x(\tau))^{\mathrm{T}}\hat{a}_i(x(\tau)) + 2\gamma^2\tilde{a}_{i+1}^{\mathrm{T}}(x(\tau))\hat{v}_i^a(x(\tau))\Big)\mathrm{d}\tau\Big|\\ &\leq\left|\int_t^\infty\Big(u_i(x(\tau))^{\mathrm{T}}Ru_i(x(\tau)) - \hat{u}_i(x(\tau))^{\mathrm{T}}R\hat{u}_i(x(\tau))\Big)\mathrm{d}\tau\right|\end{aligned}$$

$$
\begin{aligned}
&+\left|\int_t^{\infty} 2\Big(\big(u_{i+1}(x(\tau)) - \tilde{u}_{i+1}(x(\tau))\big)^{\mathrm{T}} R\hat{v}_i^u(x(\tau))\Big)\mathrm{d}\tau\right| \\
&+\left|\int_t^{\infty} \Big(2u_{i+1}^{\mathrm{T}}(x(\tau))R(\hat{u}_i(x(\tau)) - u_i(x(\tau))\Big)\mathrm{d}\tau\right| \\
&+\left|\int_t^{\infty} \Big(\gamma^2 a_i(x(\tau))^{\mathrm{T}} a_i(x(\tau)) - \gamma^2 \hat{a}_i(x(\tau))^{\mathrm{T}}\hat{a}_i(x(\tau))\Big)\mathrm{d}\tau\right| \\
&+\left|\int_t^{\infty} 2\gamma^2\Big(\big(a_{i+1}(x(\tau)) - \tilde{a}_{i+1}(x(\tau))\big)^{\mathrm{T}}\hat{v}_i^a(x(\tau))\Big)\mathrm{d}\tau\right| \\
&+\left|\int_t^{\infty} \Big(2\gamma^2 a_{i+1}^{\mathrm{T}}(x(\tau))(\hat{a}_i(x(\tau)) - a_i(x(\tau))\Big)\mathrm{d}\tau\right|, \qquad (32)
\end{aligned}
$$

where $v_i^u = u - u_i$, $\hat{v}_i^u = u - \hat{u}_i$, $v_i^a = a - a_i$ and $\hat{v}_i^a = a - \hat{a}_i$. Owing to the assumptions of the induction, as $N_u, N_v, N_a \to \infty$ and $\bar{\Delta} \searrow 0$ we have:

$$
\begin{aligned}
&\int_t^{\infty}\Big(u_i(x(\tau))^{\mathrm{T}}Ru_i(x(\tau)) - \hat{u}_i(x(\tau))^{\mathrm{T}}R\hat{u}_i(x(\tau))\Big)\mathrm{d}\tau \to 0, \\
&\int_t^{\infty} 2u_{i+1}^{\mathrm{T}}(x(\tau))R(\hat{u}_i(x(\tau)) - u_i(x(\tau)))\mathrm{d}\tau \to 0, \\
&\int_t^{\infty}\Big(\gamma^2 a_i(x(\tau))^{\mathrm{T}}a_i(x(\tau)) - \gamma^2\hat{a}_i(x(\tau))^{\mathrm{T}}R\hat{a}_i(x(\tau))\Big)\mathrm{d}\tau \to 0, \\
&\int_t^{\infty} 2\gamma^2 a_{i+1}^{\mathrm{T}}(x(\tau))(\hat{a}_i(x(\tau)) - a_i(x(\tau)))\mathrm{d}\tau \to 0. \qquad (33)
\end{aligned}
$$

In addition, owing to Assumption 1, the definitions of $\tilde{V}_i$, $\tilde{u}_{i+1}$, $\tilde{a}_{i+1}$ and the inductive assumption, as $N_u, N_v, N_a \to \infty$ and $\bar{\Delta} \searrow 0$ it holds on Ω that

$$\tilde{u}_{i+1}(x) \to u_{i+1}(x), \quad \tilde{a}_{i+1}(x) \to a_{i+1}(x). \qquad (34)$$

Hence, from (32), (33) and (34), we conclude that $\lim_{N_u,N_v,N_a\to\infty,\bar{\Delta}\searrow 0}\tilde{V}_i(x) = V_i(x)$ uniformly on Ω. On the other hand, from Lemma 1 we have $\lim_{N_u,N_v,N_a\to\infty,\bar{\Delta}\searrow 0}\{\hat{V}_i(x), \hat{u}_{i+1}(x), \hat{a}_{i+1}(x)\} = \{\tilde{V}_i(x), \tilde{u}_{i+1}(x), \tilde{a}_{i+1}(x)\}$, uniformly on Ω. Combining the two results, we conclude that for all $\epsilon_2 > 0$, there exist constants $N_v^{\star\star}$, $N_u^{\star\star}$, $N_a^{\star\star}$ and a clock mismatch upper bound $\Delta^{\star\star}$, such that if $N_v \geq N_v^{\star\star}$, $N_u \geq N_u^{\star\star}$, $N_a \geq N_a^{\star\star}$, and $\Delta(t) \leq \Delta^{\star\star}$ for all $t \geq t_0$, it holds that:

$$
\begin{aligned}
\left\|\hat{V}_i(x) - V_i(x)\right\| &\leq \left\|\hat{V}_i(x) - \tilde{V}_i(x)\right\| + \left\|\tilde{V}_i(x) - V_i(x)\right\| \\
&\leq \frac{\epsilon_2}{2} + \frac{\epsilon_2}{2} = \epsilon_2, \\
\left\|\hat{u}_{i+1}(x) - u_{i+1}(x)\right\| &\leq \left\|\hat{u}_{i+1}(x) - \tilde{u}_{i+1}(x)\right\| \qquad (35) \\
&\quad + \left\|\tilde{u}_{i+1}(x) - u_{i+1}(x)\right\| \leq \frac{\epsilon_2}{2} + \frac{\epsilon_2}{2} = \epsilon_2, \\
\left\|\hat{a}_{i+1}(x) - a_{i+1}(x)\right\| &\leq \left\|\hat{a}_{i+1}(x) - \tilde{a}_{i+1}(x)\right\| \\
&\quad + \left\|\tilde{a}_{i+1}(x) - a_{i+1}(x)\right\| \leq \frac{\epsilon_2}{2} + \frac{\epsilon_2}{2} = \epsilon_2,
\end{aligned}
$$

which completes the induction. Finally, from Wu and Luo (2012), for all $\epsilon_3 > 0$ there exists $i^\star$, such that if $i \geq i^\star$ and $x \in \Omega$ then:

$$
\begin{gathered}
\left\|V^\star(x) - V_i(x)\right\| \leq \epsilon_3, \\
\left\|u^\star(x) - u_{i+1}(x)\right\| \leq \epsilon_3, \quad \left\|a^\star(x) - a_{i+1}(x)\right\| \leq \epsilon_3.
\end{gathered} \qquad (36)
$$

Hence, the result follows from (35) and (36). ■

Proof of Theorem 2. The proof follows using Lemma 2 and reasoning similar to Lemma 1 and Theorem 1. ■

References

Abu-Khalaf, M., & 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.

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 156–172.

Elia, N., & Mitter, S. K. (2001). Stabilization of linear systems with limited information. *IEEE Transactions on Automatic Control*, *46*(9), 1384–1400.

Fotiadis, F., Kanellopoulos, A., Vamvoudakis, K. G., & Hugues, J. (2022). Impact of sensor and actuator clock offsets on reinforcement learning. In *2022 American control conference* (pp. 2669–2674). IEEE.

Fridman, E., & Dambrine, M. (2009). Control under quantization, saturation and delay: An LMI approach. *Automatica*, *45*(10), 2258–2264.

Gao, W., Deng, C., Jiang, Y., & Jiang, Z.-P. (2022). Resilient reinforcement learning and robust output regulation under denial-of-service attacks. *Automatica*, *142*, Article 110366.

Gao, W., & Jiang, Z.-P. (2017). Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2614–2624.

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

Jiang, Z.-P., Bian, T., & Gao, W. (2020). Learning-based control: A tutorial and some recent results. *Foundations and Trends® in Systems and Control*, *8*(3).

Jiang, Y., & Jiang, Z.-P. (2014). Robust adaptive dynamic programming and feedback stabilization of nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 882–893.

Johnson, M., Kamalapurkar, R., Bhasin, S., & Dixon, W. E. (2014). Approximate *N*-player nonzero-sum game solution for an uncertain continuous nonlinear system. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(8), 1645–1658.

Kiumarsi, B., Lewis, F. L., Modares, H., Karimpour, A., & Naghibi-Sistani, M.-B. (2014). Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Automatica*, *50*(4), 1167–1175.

Li, Z., Su, W., Xu, M., Yu, R., Niyato, D., & Xie, S. (2022). Compact Learning Model for Dynamic Off-Chain Routing in Blockchain-Based IoT. *IEEE Journal on Selected Areas in Communications*, *40*(12), 3615–3630.

Modares, H., Lewis, F. L., & Jiang, Z.-P. (2015). H_∞ Tracking Control of Completely Unknown Continuous-Time Systems via Off-Policy Reinforcement Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(10), 2550–2562.

Nowé, A., Vrancx, P., & Hauwere, Y.-M. D. (2012). Game theory and multi-agent reinforcement learning. In *Reinforcement learning* (pp. 441–470). Springer.

Okano, K., Wakaiki, M., Yang, G., & Hespanha, J. P. (2017). Stabilization of networked control systems under clock offsets and quantization. *IEEE Transactions on Automatic Control*, *63*(6), 1708–1723.

Pang, B., Bian, T., & Jiang, Z.-P. (2021). Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Transactions on Automatic Control*, *67*(1), 504–511.

Shrivastava, A., Derler, P., Baboud, Y.-S. L., Stanton, K., Khayatian, M., Andrade, H. A., et al. (2016). Time in cyber-physical systems. In *Proceedings of the eleventh IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis* (pp. 1–10).

Song, R., Lewis, F. L., Wei, Q., & Zhang, H. (2015). Off-policy actor-critic structure for optimal control of unknown systems with disturbances. *IEEE Transactions on Cybernetics*, *46*(5), 1041–1050.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Vamvoudakis, K. G. (2017). Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach. *Systems & Control Letters*, *100*, 14–20.

Vamvoudakis, K. G., & Hespanha, J. P. (2017). Cooperative Q-learning for rejection of persistent adversarial inputs in networked linear quadratic systems. *IEEE Transactions on Automatic Control*, *63*(4), 1018–1031.

Vamvoudakis, K. G., & Lewis, F. L. (2012). Online solution of nonlinear two-player zero-sum games using synchronous policy iteration. *International Journal of Robust and Nonlinear Control*, *22*(13), 1460–1483.

Wakaiki, M., Cetinkaya, A., & Ishii, H. (2019). Stabilization of networked control systems under DoS attacks and output quantization. *IEEE Transactions on Automatic Control*, *65*(8), 3560–3575.

Wakaiki, M., Okano, K., & Hespanha, J. P. (2017). Stabilization of systems with asynchronous sensors and controllers. *Automatica*, *81*, 314–321.

Wu, H.-N., & Luo, B. (2012). Neural Network Based Online Simultaneous Policy Update Algorithm for Solving the HJI Equation in Nonlinear H_∞ Control. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(12), 1884–1895.

Zhao, H., Wu, J., Li, Z., Chen, W., & Zheng, Z. (2022). Double sparse deep reinforcement learning via multilayer sparse coding and nonconvex regularized pruning. *IEEE Transactions on Cybernetics*, *53*(2), 765–778.

Filippos Fotiadis was born in Thessaloniki, Greece. He received the Diploma (joint B.Sc. and M.Sc. degree) in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece, in 2018. He is currently pursuing the Ph.D. degree in Aerospace Engineering at the Georgia Institute of Technology, where he also received the M.Sc. degree in Aerospace Engineering in 2022, and the M.Sc. degree in Mathematics in 2023. His research interests include optimal and learning-based control, game theory, and their applications to cyber–physical security.

Aris Kanellopoulos received his diploma equivalent to a Master of Science in Mechanical Engineering from the National Technical University of Athens, Greece in 2017. He was awarded a Ph.D. in Aerospace Engineering at the Georgia Institute of Technology where he worked as a Research Engineer in 2021. He is currently a Postdoctoral Researcher at the Royal Institute of Technology, Stockholm, Sweden. His research interests include optimal control, game theory and cyber–physical security.

Kyriakos G. Vamvoudakis was born in Athens, Greece. He received the Diploma (a 5-year degree, equivalent to a Master of Science) in Electronic and Computer Engineering from the Technical University of Crete, Greece in 2006 with highest honors. After moving to the United States of America, he studied at The University of Texas at Arlington with Frank L. Lewis as his advisor, and he received his M.S. and Ph.D. in Electrical Engineering in 2008 and 2011 respectively. From May 2011 to January 2012, he was working as an Adjunct Professor and Faculty Research Associate at the University of Texas at Arlington and at the Automation and Robotics Research Institute. During the period from 2012 to 2016 he was project research scientist at the Center for Control, Dynamical Systems and Computation at the University of California, Santa Barbara. He was an assistant professor at the Kevin T. Crofton Department of Aerospace and Ocean Engineering at Virginia Tech until 2018.

He currently serves as the Dutton-Ducoffe Endowed Professor at The Daniel Guggenheim School of Aerospace Engineering at Georgia Tech. He holds a secondary appointment in the School of Electrical and Computer Engineering. His expertise is in reinforcement learning, control theory, game theory, cyber–physical security, bounded rationality, and safe/assured autonomy.

Dr. Vamvoudakis is the recipient of a 2019 ARO YIP award, a 2018 NSF CAREER award, a 2018 DoD Minerva Research Initiative Award, a 2021 GT Chapter Sigma Xi Young Faculty Award and his work has been recognized with best paper nominations and several international awards including the 2016 International Neural Network Society Young Investigator (INNS) Award, the Best Paper Award for Autonomous/Unmanned Vehicles at the 27th Army Science Conference in 2010, the Best Presentation Award at the World Congress of Computational Intelligence in 2010, and the Best Researcher Award from the Automation and Robotics Research Institute in 2011. He currently is an Associate Editor of: Automatica; IEEE Transactions on Automatic Control; IEEE Transactions on Neural Networks and Learning Systems; IEEE Computational Intelligence Magazine; IEEE Transactions on Systems, Man, and Cybernetics: Systems; IEEE Transactions on Artificial Intelligence; Neurocomputing; Journal of Optimization Theory and Applications; and of Frontiers in Control Engineering-Adaptive, Robust and Fault Tolerant Control. He had also served as a Guest Editor for, IEEE Transactions on Automation Science and Engineering (Special issue on Learning from Imperfect Data for Industrial Automation); IEEE Transactions on Neural Networks and Learning Systems (Special issue on Reinforcement Learning Based Control: Data-Efficient and Resilient Methods); IEEE Transactions on Industrial Informatics (Special issue on Industrial Artificial Intelligence for Smart Manufacturing); and IEEE Transactions on Intelligent Transportation Systems (Special issue on Unmanned Aircraft System Traffic Management).

Jerome Hugues is a Principal Researcher at the Software Engineering Institute on the Assuring Cyber–Physical Systems team. He holds a Habilitation (2017) from INP Toulouse, a Ph.D. (2005) and an engineering degree (2002) both from Telecom ParisTech, and an M.Sc. in Computer Science from UPMC (2002). His research interests focus on the design of software-based real-time and embedded systems and tools to support it. He is a member of the SAE AS-2C committee working on the AADL since 2005. Before joining the CMU/SEI, he was a professor at the Department of Engineering of Complex Systems of ISAE-Supaero in Toulouse France, in charge of teaching curriculum on systems engineering, safety-critical systems, and real-time systems.