On-Off Adversarially Robust Q-Learning

Prachi Pratyusha Sahoo, *Student Member, IEEE*,
and Kyriakos G. Vamvoudakis, *Senior Member, IEEE*

***Abstract*—This letter, presents an "on-off" learning-based scheme to expand the attacker's surface, namely a moving target defense (MTD) framework, while optimally stabilizing an unknown system. We leverage Q-learning to learn optimal strategies with "on-off" actuation to promote unpredictability of the learned behavior against physically plausible attacks. We provide rigorous, theoretical guarantees on the stability of the equilibrium point even when switching. Finally, we develop two adversarial threat models to evaluate the learning agent's ability to generate robust policies based on a distance to uncontrollability.**

***Index Terms*—Adversarial Q-learning, cyber-physical systems.**

I. Introduction

INTERCONNECTED computational subsystems that control physical devices interacting with the operating environment compose a class of platforms called Cyber-Physical Systems (CPS) [1]. Decision making mechanisms, designed to incorporate agility with the help of artificial intelligence (AI) allow self-adaptation, self-healing, and self-optimization of CPS, including military and civilian applications, such as brain-machine interfaces, therapeutic and entertainment robotics, exoskeletons, and prosthetics, power-grids and smart-grids, and self-driving cars [2]–[4]. Decision making mechanisms for CPS perform well for smaller-scale, well-modelled systems. However, in an adversarial environment, and with unknown physical models, these algorithms have robustness, scalability, and optimality problems [5], [6].

Reinforcement learning (RL) schemes generate optimal policies in response to reward signals [7] received from sensors in real-time. Development of data-driven methods allows for on-line approximations, and deployment of optimal decisions that do not have closed-form solutions [8]–[11]. But when learning takes place in an adversarial environment, reinforcement signals may become corrupted and entertain the possibility that the learning agent will eventually learn the wrong policy. The authors of [12] create adversarial examples to modify a deep RL algorithm to lure the learning agent to different equilibrium points. Effectiveness of adversarial attacks on RL are typically measured using the reduction in reward collection [12]. Towards this direction, this letter will propose an alternative measure to track the effectiveness of the attack; the distance of the learning agent's perturbed dynamics from the nearest uncontrollable pair.

Related Work: While severe attacks can disrupt the functionality of the system and cause failure, subtle changes made by an adversary may remain unnoticed if the learning algorithm is not built with adversarial considerations. Many learning tasks, such as intrusion detection, and spam classification [13] become impractical when attackers are stealthy [14]. Another mode of stealthy attacks occur in object recognition tasks in computer vision; the fast gradient value method, described in [15] and [16], generates physically plausible adversarial attacks that destabilize the recognition task by using the gradient of the loss function. In such a case, developing optimal, or sub-optimal, control schemes that can stabilize the system in the presence of adversarial noise [17] is necessary. This letter, incorporates an adversarial noise vector, constructed using the learning agent's utility function, into the system's learning scheme to bolster stability in the face of attacks. Typically, CPS maintain a stationary framework, including: features, network topologies, and communication networks. Such static frameworks provide an attacker with the time to learn and deploy a low cost, destabilizing attack policy. To counter this threat, moving target defense (MTD), a proactive policy, aims to dynamically change the attack surface of CPS, introducing unpredictability [18], [19]. The authors in [20] define key concepts to describe MTD systems and their problems pertaining to selection, adaptation, and timing, to conclude that security increases with higher entropy [21], [22]. We leverage MTD to induce "on-off" actuation and change the attack surface of the CPS.

Contributions: The contribution of this letter is fourfold. First, we develop a data-driven method to create an "on-off" sequence of actuation for the learning system by means of a multi-step intelligent strategy, deployed in a plug-n-play manner. We then leverage the redundancy of the system to induce this "on-off" actuation for ensured security under physically plausible attacks, and provide asymptotic stability and convergence guarantees. We develop two adversarial threat

Manuscript received November 27, 2019; revised February 7, 2020; accepted March 3, 2020. Date of publication March 10, 2020; date of current version May 25, 2020. This work was supported in part by NSF under Grant S&AS-1849264 and Grant CPS-1851588, in part by ONR Minverva under Grant N00014-18-1-2874, and in part by ARO under Grant W911NF-19-1-0270. Recommended by Senior Editor S. Tarbouriech. *(Corresponding author: Prachi Pratyusha Sahoo.)*

Prachi Pratyusha Sahoo is with the Woodruff School for Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30318 USA (e-mail: prachi@gatech.edu).

Kyriakos G. Vamvoudakis is with the Daniel H. Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30318 USA (e-mail: kyriakos@gatech.edu).

Digital Object Identifier 10.1109/LCSYS.2020.2979572

models that will help us produce robust, optimal, and resilient learning schemes. Finally, we characterize the distance to uncontrollability in a model-free manner as a way to assess robustness under physically plausible attacks.

II. Problem Formulation

Consider the continuous-time linear time-invariant system,

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0) = x_0,\ t \geq 0 \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state that is available for feedback, $u(t) \in \mathbb{R}^m$ is the control input, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ are the plant and the input matrices respectively. We want to find a policy u, that minimizes a user-defined energy cost functional, given by

$$J(x(0); u) = \frac{1}{2}\int_0^\infty (x^{\mathrm{T}}Mx + u^{\mathrm{T}}Ru)\mathrm{d}t,\ \forall x(0), \tag{2}$$

where $M \succeq 0$, and $R \succ 0$ are user-defined matrices that penalize the energy of the states and the control inputs during transient behavior. The optimal policy $u^\star$ is the global minimizer, given by the minimization problem $\min_u J(x(0); u)$ with respect to the dynamical system given in equation (1), such that $J(x(0); u^\star) \leq J(x(0); u),\ \forall u,\ x(0)$.

Employing the stationarity condition yields the optimal policy,

$$u^\star(x) = -R^{-1}B^{\mathrm{T}}Px, \forall x, \tag{3}$$

where $P = P^{\mathrm{T}} \succ 0$ is the solution to the Riccati $0 = PA + A^{\mathrm{T}}P - PBR^{-1}B^{\mathrm{T}}P + M$. It is evident that computing the policy (3), requires complete knowledge of the system (1). By augmenting the state vector, x, with the control vector, u, i.e., $U = [x^{\mathrm{T}}\ u^{\mathrm{T}}]^{\mathrm{T}}$, an action dependent utility function can be derived and learned in a model-free manner, using an actor/critic network to yield the optimal policy (see [23] for details). The general framework is summarized in Algorithm 1, with a Q-function compactly written $\forall x,\ u$ as,

$$\begin{aligned} Q(x,u) &= \frac{1}{2}U^{\mathrm{T}}\begin{bmatrix} P + M + PA + A^{\mathrm{T}}P & PB \\ B^{\mathrm{T}}P & R \end{bmatrix}U \\ &:= \frac{1}{2}U^{\mathrm{T}}\begin{bmatrix} Q_{\mathrm{xx}} & Q_{\mathrm{xu}} \\ Q_{\mathrm{ux}} & Q_{\mathrm{uu}} \end{bmatrix}U := \frac{1}{2}U^{\mathrm{T}}\bar{Q}U = W_c^{\mathrm{T}}(U \otimes U), \end{aligned} \tag{4}$$

while the approximators and the estimation errors of the signals will be denoted by $\hat{(\cdot)}$ and $\tilde{(\cdot)}$ respectively and $\otimes$ denotes the Kronecker product.

Theorem 1: Consider the system dynamics given by (1), assume that (A, B) is controllable, $(A, \sqrt{M})$ is detectable, and the learning scheme defined in Algorithm 1. Then the equilibrium point (i.e., the origin) of the closed-loop system for all initial conditions is asymptotically stable given that the tuning gain for the critic α_c is sufficiently larger than the tuning gain for the actor α_a and the following inequality holds, $1 < \alpha_a < \frac{1}{\delta\bar{\lambda}(Q_{\mathrm{uu}}^{-1})}(2\underline{\lambda}(M + Q_{\mathrm{xu}}Q_{\mathrm{uu}}^{-1}Q_{\mathrm{ux}}^{\mathrm{T}}) - \bar{\lambda}(Q_{\mathrm{xu}}Q_{\mathrm{xu}}^{\mathrm{T}}))$, where $\underline{\lambda}(\cdot)$ and $\bar{\lambda}(\cdot)$ denote the minimum and the maximum eigenvalue respectively. Furthermore, the policy $\hat{u}$ given by $\hat{u}(x) = \hat{W}_a^{\mathrm{T}}x$ converges asymptotically to the optimal $u^\star$ given by $u^\star(x) = \arg\min_u Q(x,u) = -Q_{\mathrm{uu}}^{-1}Q_{\mathrm{ux}}x,\ \forall x$.

Proof: The proof follows from [23]. ■

Algorithm 1 Q-Learning

1: Given x_0, $\hat{W}_c(0)$, $\hat{W}_a(0)$.
2: Compute and decompose (by using the half-vectorization operator vech$(\cdot)$)

$$\begin{aligned} \mathrm{vech}(\hat{Q}_{\mathrm{xx}}) &:= \hat{W}_c\left[1{:}\tfrac{n(n+1)}{2}\right] \\ \mathrm{vech}(\hat{Q}_{\mathrm{xu}}) &:= \hat{W}_c\left[\tfrac{n(n+1)}{2} + 1{:}\tfrac{n(n+1)}{2} + nm\right] \\ \mathrm{vech}(\hat{Q}_{\mathrm{uu}}) &:= \hat{W}_c\left[\tfrac{n(n+1)}{2} + nm + 1{:}\tfrac{(n+m)(n+m+1)}{2}\right]. \end{aligned}$$

3: Compute $u(t) = \hat{W}_a^{\mathrm{T}}x(t)$.
4: **if** $t < T_{\mathrm{exp}}$
5: Add probing noise $u(t) \leftarrow u(t) + u_{\mathrm{PE}}(t)$. T_{exp} is the window of time during which high frequency disturbance [23] is added to the control input to employ state space exploration.
6: **end if**
7: Compute e_c, where T is a small window of time to evaluate the error in critic weights, as

$$\begin{aligned} e_c &:= \hat{W}_c^{\mathrm{T}}(U(t) \otimes U(t)) - \hat{W}_c^{\mathrm{T}}(U(t-T) \otimes U(t-T)) \\ &\quad + 0.5\int_{t-T}^{t}(x^{\mathrm{T}}Mx + u^{\mathrm{T}}Ru)\mathrm{d}t,\ T \geq 0. \end{aligned} \tag{5}$$

8: Compute e_a, as

$$e_a := \hat{W}_a^{\mathrm{T}}x + \hat{Q}_{\mathrm{uu}}^{-1}\hat{Q}_{\mathrm{ux}}x,\ \forall x. \tag{6}$$

9: Propagate $\hat{W}_c$ using

$$\dot{\hat{W}}_c = -\alpha_c\frac{\sigma}{(1+\sigma^{\mathrm{T}}\sigma)^2}e_c^{\mathrm{T}}. \tag{7}$$

where $\sigma = U(t) \otimes U(t) - U(t-T) \otimes U(t-T)$.
10: Propagate $\hat{W}_a$ using

$$\dot{\hat{W}}_a = -\alpha_a x e_a^{\mathrm{T}}. \tag{8}$$

11: **if** $e_a \neq 0$ and $e_c \neq 0$
12: Go to step 3.
13: **end if** ▷ $e_a \approx 0$ and $e_c \approx 0$

Problem 1: Note that the Q-learning described in Algorithm 1 eventually learns the optimal strategies, but it assumes a form of model-based controllability, of some potentially physically manipulated model and learning mechanism. To robustify the learning algorithm in the presence of physically plausible adversarial attacks and systematic shifts in the environment that cause performance degradation we need to dynamically diversify the attacking surface by using a proper switching actuation.

III. MTD With On-Off Policies

Selection of an actuator subset that allows our model-free system to be controllable dictates the use of a data-driven and model-free mechanism. Since the controllability Grammian has a data-based nature in the discrete domain we shall find $A_d := e^{\mathrm{A}\Delta t}$, $B_d := A^{-1}(A_d - I_n)B$, where $\Delta t > 0$ and I_n is an identity matrix of order n.

A. Data-Driven Controllability

Without any knowledge of the model, namely the potentially manipulated matrices A and B, setting the policy input, $u_i, \forall i \in \{1, 2, \ldots, m\}$ as unity and propagating the states in a discrete manner for K time steps we obtain $x_{\mathrm{d}}^{[\mathrm{i}]} =$

$\{x^{[i]}(1)\dots x^{[i]}(j)\dots x^{[i]}(K)\}$ with $x^{[i]}(1) = 0$. Upon similar disturbance episodic data collection for all the actuators activated one at a time, restructuring the states corresponding to the jth step yields $\bar{X}(j) = \begin{bmatrix} x^{[1]}(j) & x^{[2]}(j) & \dots x^{[m]}(j), \end{bmatrix}$, $j = 1, 2, \dots$

Defining now W_j, as a measure of a disturbance induced in the state measurements with a different actuator triggering, as $W_j = \bar{X}(j) - \bar{X}(j-1)$, we can write the K-step data-driven controllability Grammian as $W_{\text{dc}} = \sum_{j=1}^{K} W_j W_j^{\text{T}}$, which enables us to approximate the discrete-time model-based controllability Grammian $W_{\text{dc}} = \sum_{m=0}^{\infty} A_d^m B_d B_d^{\text{T}} A_d^{\text{T}m}$ in a model-free manner.

Theorem 2: The system given by (1) is controllable in $K > 0$ steps if and only if, $\underline{\lambda}(W_{\text{dc}}) > 0$.

Proof: It follows the proof [24, Th. 1]. ■

B. Switching Architecture

Consider the allowable actuator sets of the system in (1), as,

$$\mathcal{K}_{\text{a}} = \{u_i \in \eta : \underline{\lambda}(W_{\text{dc}}) > 0\}, \tag{9}$$

wherein η stands for the power set of all the actuators. The actuator sets in $\mathcal{K}_{\text{a}}$, with a cardinality p, maintain controllability of each switching mode. A mode is a categorical dynamic behaviour of the system that changes based on the gain matrix.

Lemma 1: Assume that each of the sets contained in $\mathcal{K}_{\text{a}}$, denoted by S_i $\forall i \in 1, 2, \dots p$ activates one actuating mode from the allowable set. Let k_i, $\forall i \in 1, 2, \dots p$ be the number of the actuators in mode i. Then, the maximum allowable actuation links, elements of the gain matrix K that connect gains with actuators, that can be removed for mode i, while maintaining controllability properties intact, are $s := nm - nk_i$.The matrix K is the feedback gain matrix such that $u = Kx$.

Proof: The proof is a direct consequence of the dimensions of the quantities, the actuator redundancy, and is omitted due to space limitations. ■

The nullification of the rows (termed here as "on-off" strategies) in the policy gain matrix that corresponds to a different policy without destabilizing the system, extends the attacker's surface with MTD as,

$$K[i,:] = \begin{cases} \hat{W}_a[:,i]^{\text{T}}, & \forall i \in S_i, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

wherein $K[i,:]$ stands for the ith row of the gain matrix.

The following definition is now needed.

Definition 1 [25]: The steepest descent directions maximize the decrease of the linear approximation of a function.

To minimize the actuator-related cost, we shall use a tuning law that updates along the steepest gradients of half the squared-norm error with respect to the action approximator weights, given by $g := xe_a^{\text{T}}$. Thus, one can sort the results to obtain the steepest directions of descent and upon combination (with the support of the gain matrix K) write,

$$T_{ij} := \begin{cases} 1, & \text{if } g_{ij} \text{ is one of the } 2s \text{ steepest gradients,} \\ & \text{or } K_{ij} \neq 0, \\ 0, & \text{otherwise} \end{cases}$$

wherein s is defined in Lemma 1.

Remark 1: Appropriate sorting algorithms with average behaviour may be used to identify largest $2s$ elements in g depending on the dimensionality. Note that the gradient descent along the steepest directions maximizes the actuator-related cost reduction.

In other words, the selector matrix T picks the steepest descent directions by following,

$$\dot{\hat{W}}_a = -\alpha_a [x e_a^{\text{T}} \odot T], \tag{11}$$

wherein $\odot$ stands for the Hadamard product. The algorithm shall trigger at most $3s$ actor weight tuning laws.

Note that, arbitrarily switching between these modes using different sets of actuators introduces stability concerns for the learning agent, owing to the hybrid nature of the system dynamics. A dwell time approach will ensure the asymptotic stability of the equilibrium point of the closed-loop learning agent dynamics.

The work of [23] has shown that the value function of the learning agent has the same value with the Q-function, i.e.,

$$V^{\star}(x) = Q^{\star}(x, u^{\star}), \forall x. \tag{12}$$

Hence, the model-free formulation of the continuous-time, algebraic Ricatti equation at steady-state is,

$$P_{\text{data}} = Q_{\text{xx}} - Q_{\text{xu}} Q_{\text{uu}}^{-1} Q_{\text{ux}}. \tag{13}$$

Definition 2 [26]: A switching signal has an average dwell time τ_{dwell} if over any time interval $\begin{bmatrix} t, t_f \end{bmatrix}$, $t_f \geq t$, the number of switches $S(t_f, t)$ is bounded above by,

$$S(t_f, t) \leq S_0 + \frac{t_f - t}{\tau_{\text{dwell}}}, \ t \geq 0 \tag{14}$$

where S_0 serves as the arbitrary chatter bound and τ_{dwell} is the dwell time.

Theorem 3: Consider the system described by (1). The switched system dictated by the switching signal $\sigma(t) = i, i \in \begin{Bmatrix} 1, & \dots & p \end{Bmatrix}$, $t \geq 0$, with an active set of actuators S_i, an actor update law given by (11), and a critic update law given by (7), has an asymptotically stable equilibrium point for every arbitrary switching signal $\sigma(t)$ given that the average dwell time is given by,

$$\hat{\tau}_{\text{dwell}} = \upsilon + \frac{\log(\max_{p,q \in \{1,\dots p\}} \frac{\bar{\lambda}(\hat{Q}_{\text{pxx}} - \hat{Q}_{\text{pxu}} \hat{Q}_{\text{puu}}^{-1} \hat{Q}_{\text{pux}})}{\underline{\lambda}(\hat{Q}_{\text{qxx}} - \hat{Q}_{\text{qxu}} \hat{Q}_{\text{quu}}^{-1} \hat{Q}_{\text{qux}})})}{\min_{i \in \{1,\dots,p\}} \frac{\underline{\lambda}(M_i + \hat{Q}_{\text{ixu}} \hat{Q}_{\text{iuu}}^{-1} \hat{Q}_{\text{iux}})}{\underline{\lambda}(\hat{Q}_{\text{ixx}} - \hat{Q}_{\text{ixu}} \hat{Q}_{\text{iuu}}^{-1} \hat{Q}_{\text{iux}})}}, \tag{15}$$

where $\upsilon \in \mathbb{R}^+$ is a constant of order 1.

Proof: Let K_{S_i} be the "on-off" policy gain matrix corresponding to the S_i set of actuators in the controllability set $\mathcal{K}_a$. The following simplification, eliminates the columns of B based on the rows of K that are nullified by the selection procedure, i.e.,

$$u_i = -K_{S_i} x, \ \forall x,$$
$$B_i = \begin{cases} B[:,i], & \forall i \in S_i \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Between two switches, the system triggers different actuator modes to become active. Thus,

$$\dot{x} = Ax + B_i u_i, \ \forall i \in S_i, \ t \geq 0, \tag{17}$$

with $u_i = -R_i^{-1}B_i^{\mathrm{T}}P_i x$ and an associated Lyapunov equation $V_i = x^{\mathrm{T}}P_i x,\ P_i \succ 0$ bounded as,

$$\underline{\lambda}(P_i)\|x\|_2^2 \leq V_i(x) \leq \bar{\lambda}(P_i)\|x\|_2^2,\ \forall x,\ i \in S_i. \tag{18}$$

Taking the time derivative of V_i, along the trajectories of the system yields $\dot{V}_i = \dot{x}^{\mathrm{T}}P_i x + x^{\mathrm{T}}P_i\dot{x} = x^{\mathrm{T}}(A^{\mathrm{T}}P_i - 2P_iB_iR_i^{-1}B_i^{\mathrm{T}}P_i + P_iA)x$. The Ricatti equation for the ith mode gives $-M - P_iB_iR_i^{-1}B_i^{\mathrm{T}}P_i = A^TP_i + P_iA - 2P_iB_iR_i^{-1}B_i^TP_i$. The derivative of the value function now becomes $\dot{V}_i = -x^{\mathrm{T}}(M + P_iB_iR_i^{-1}B_i^{\mathrm{T}}P_i)x = -x^{\mathrm{T}}(M + Q_{\mathrm{xu}}Q_{\mathrm{uu}}^{-1}Q_{\mathrm{ux}})x := -x^{\mathrm{T}}\bar{H}_i x$ is bounded as $\underline{\lambda}(\bar{H}_i)\|x\|_2^2 \leq x^{\mathrm{T}}\bar{H}_i x \leq \bar{\lambda}(\bar{H}_i)\|x\|_2^2$, assuming we are operating in steady state conditions wherein, our actors and critics have converged. The bounds on the derivative of the value function along the trajectories, given by $\dot{V}_i \leq -\underline{\lambda}(\bar{H}_i)\|x\|_2^2 \leq -\frac{\underline{\lambda}(\bar{H}_i)}{\underline{\lambda}(Q_{\mathrm{ixx}} - Q_{\mathrm{ixu}}Q_{\mathrm{iuu}}^{-1}Q_{\mathrm{iux}})}V_i(x)$ can be found from (18) as,

$$\frac{V_i(x)}{\bar{\lambda}(Q_{\mathrm{ixx}} - Q_{\mathrm{ixu}}Q_{\mathrm{iuu}}^{-1}Q_{\mathrm{iux}})} \leq \|x\|_2^2 \leq \frac{V_i(x)}{\underline{\lambda}(Q_{\mathrm{ixx}} - Q_{\mathrm{ixu}}Q_{\mathrm{iuu}}^{-1}Q_{\mathrm{iux}})},\ \forall x. \tag{19}$$

Considering arbitrary modes, the inequality (19) captures the minimum of all combinations of the smallest eigenvalues of the matrices P_i and $\bar{H}_i$. This can be shown by,

$$\dot{V}_i(x) \leq -\min_{i\in\{1,2,\ldots,p\}} \frac{\underline{\lambda}(\bar{H}_i)}{\underline{\lambda}(Q_{\mathrm{ixx}} - Q_{\mathrm{ixu}}Q_{\mathrm{iuu}}^{-1}Q_{\mathrm{iux}})}V_i(x). \tag{20}$$

From (18), the equations for the modes p and q can be found such that (19) holds for each of the modes, $i \in \{p,\ q\}$. Combining and rearranging, yields $\frac{V_q(x)}{\bar{\lambda}(Q_{\mathrm{qxx}} - Q_{\mathrm{qxu}}Q_{\mathrm{quu}}^{-1}Q_{\mathrm{qux}})} \leq \|x\|_2^2 \leq \frac{V_p(x)}{\underline{\lambda}(Q_{\mathrm{pxx}} - Q_{\mathrm{pxu}}Q_{\mathrm{puu}}^{-1}Q_{\mathrm{pux}})}$ which, can be reduced to $V_q(x) \leq \frac{\bar{\lambda}(P_q)}{\underline{\lambda}(P_p)}V_p(x), \forall x$. Then, the following inequality holds,

$$V_p(x) \leq \max_{p,q\in\{1,\ldots,p\}} \frac{\bar{\lambda}(Q_{\mathrm{pxx}} - Q_{\mathrm{pxu}}Q_{\mathrm{puu}}^{-1}Q_{\mathrm{pux}})}{\underline{\lambda}(Q_{\mathrm{qxx}} - Q_{\mathrm{qxu}}Q_{\mathrm{quu}}^{-1}Q_{\mathrm{qux}})}V_q(x), \forall x,\ p,\ q. \tag{21}$$

Setting now $\nu := \min_{i\in\{1,\ldots p\}} \frac{\underline{\lambda}(\bar{H}_i)}{\underline{\lambda}(Q_{\mathrm{ixx}} - Q_{\mathrm{ixu}}Q_{\mathrm{iuu}}^{-1}Q_{\mathrm{iux}})}$ and $\mu := \max_{p,q\in\{1,\ldots p\}} \frac{\bar{\lambda}(Q_{\mathrm{pxx}} - Q_{\mathrm{pxu}}Q_{\mathrm{puu}}^{-1}Q_{\mathrm{pux}})}{\underline{\lambda}(Q_{\mathrm{qxx}} - Q_{\mathrm{qxu}}Q_{\mathrm{quu}}^{-1}Q_{\mathrm{qux}})}$. Let $T(t_f, 0)$ be the number of switches that take place for $t \in (0, t_f)$ and at specific instances of time t_i, $i \in [0, T(t_f, 0)]$ with $t_i < t_{i+1}$. Let $\sigma(t) = i$ be a piece-wise constant function that activates different modes of actuation in (17). Thus, between any two triggering events, i.e., $t \in [t_i, t_{i+1}]$ the value of $\sigma(t)$ remains constant. Define a function $W(t) = e^{\nu t}V_{\sigma(t)}(x(t)),\ t \geq 0$ and evaluate its derivatives along the trajectories of $\dot{x} = (A - B_{\sigma(t)}Q_{\sigma(t)\mathrm{uu}}^{-1}Q_{\sigma(t)\mathrm{ux}})x(t)$ which is the hybrid system. Between two consecutive switches the derivative of $W(t)$ may be taken as follows $\dot{W}(t) = \nu e^{\nu t}V_{\sigma(t)}(x(t)) + e^{\nu t}\dot{V}_{\sigma(t)}(x(t))$. Using (20) we obtain $\dot{W}(t) \leq \nu e^{\nu t}V_{\sigma(t)}(x(t)) - \nu e^{\nu t}V_{\sigma(t)}(x(t)) \leq 0$ which implies that $W(t)$ is a non-increasing function of time over intervals $t \in [t_i, t_{i+1})$. At switching instances, $t = t_{i+1}$, the inequality $W(t) = e^{\nu t_{i+1}}V_{\sigma(t_{i+1})}(x(t_{i+1})) \leq \mu e^{\nu t_{i+1}}V_{\sigma(t_i)}(x(t_{i+1})) = \mu W(t_{i+1}^-) \leq \mu W(t_i)$ utilizes the relationship shown in (21) to connect $V_{\sigma(t_{i+1})}(x(t_{i+1}))$ with $V_{\sigma(t_i)}(x(t_{i+1}))$ and the non-increasing property of $W(t)$ between two switch instances. Using this relationship and backward stepping from the final switch time $t_{T(t_f,0)} < t_f$ to the initial time t_0 for $T(t_f, 0) - 1$ jumps, we have $W(t_f^-) \leq W(t_{T(t_f,0)}) \leq \mu^{T(t_f,0)}W(0),\ V_{\sigma(t_f^-)}(x(t_f)) \leq \mu^{T(t_f,0)}e^{-\nu t_0}V_{\sigma(t_0^-)}(x(t_0))$. Using the inequality in (14) we have $V_{\sigma(t_f^-)}(x(t_f)) \leq e^{\frac{t_f\log(\mu)-\nu t_0}{\tau_{\mathrm{dwell}}}}V_{\sigma(t_0^-)}(x(t_0))$. To ensure that $V_{\sigma(t_f^-)}(x(t_f)) \to 0$ as $t_f \to 0$, the quantity $\frac{t_f\log(\mu)-\nu t_0}{\tau_{\mathrm{dwell}}}$ must be negative to produce $\tau_{\mathrm{dwell}} > \frac{\log(\mu)}{\nu}$.

This dwell time expression, involving μ and ν which depend on the Q-matrices of every mode can be rewritten using $\hat{Q}_i$-matrices in steady state after convergence; which is a model-free version that can be verified during the learning procedure. An approximate model-free dwell time converges to the ideal dwell time in (15) asymptotically, as

$$\hat{\tau}_{\mathrm{dwell}} > \frac{\log(\max_{p,q\in\{1,2,3,\ldots,p\}} \frac{\bar{\lambda}(\hat{Q}_{\mathrm{pxx}} - \hat{Q}_{\mathrm{pxu}}\hat{Q}_{\mathrm{puu}}^{-1}\hat{Q}_{\mathrm{pux}})}{\underline{\lambda}(\hat{Q}_{\mathrm{qxx}} - \hat{Q}_{\mathrm{qxu}}\hat{Q}_{\mathrm{quu}}^{-1}\hat{Q}_{\mathrm{qux}})})}{\min_{i\in\{1,2,\ldots,p\}} \frac{\underline{\lambda}(M_i + \hat{Q}_{\mathrm{ixu}}\hat{Q}_{\mathrm{iuu}}^{-1}\hat{Q}_{\mathrm{iux}})}{\underline{\lambda}(\hat{Q}_{\mathrm{ixx}} - \hat{Q}_{\mathrm{ixu}}\hat{Q}_{\mathrm{iuu}}^{-1}\hat{Q}_{\mathrm{iux}})}},\ \forall i. \tag{22}$$

Since the above formulation guarantees convergence of the switched system, and $\hat{\tau}_{\mathrm{dwell}}$, in infinite time, we will use a modified version. Let $\upsilon \in \mathbb{R}^+$ be a small enough parameter such that selecting $\hat{\tau}_{\mathrm{dwell}}$ as

$$\hat{\tau}_{\mathrm{dwell}} = \upsilon + \frac{\log(\max_{p,q\in\{1,\ldots p\}} \frac{\bar{\lambda}(\hat{Q}_{\mathrm{pxx}} - \hat{Q}_{\mathrm{pxu}}\hat{Q}_{\mathrm{puu}}^{-1}\hat{Q}_{\mathrm{pux}})}{\underline{\lambda}(\hat{Q}_{\mathrm{qxx}} - \hat{Q}_{\mathrm{qxu}}\hat{Q}_{\mathrm{quu}}^{-1}\hat{Q}_{\mathrm{qux}})})}{\min_{i\in\{1,\ldots,p\}} \frac{\underline{\lambda}(M_i + \hat{Q}_{\mathrm{ixu}}\hat{Q}_{\mathrm{iuu}}^{-1}\hat{Q}_{\mathrm{iux}})}{\underline{\lambda}(\hat{Q}_{\mathrm{ixx}} - \hat{Q}_{\mathrm{ixu}}\hat{Q}_{\mathrm{iuu}}^{-1}\hat{Q}_{\mathrm{iux}})}},\ \forall i, \tag{23}$$

makes the equilibrium point of the the switched system asymptotically stable under arbitrary switching between the active actuators. Finally, the addition of υ ensures finite-time convergence of the model-free dwell time. ■

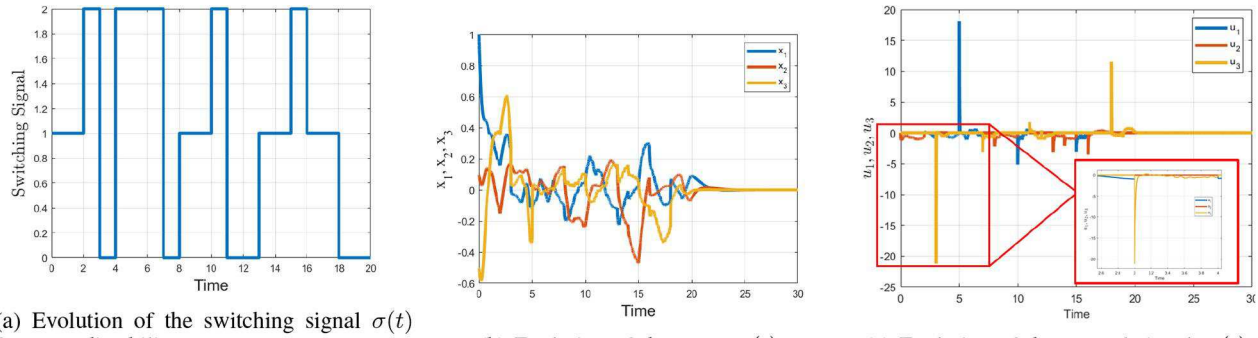
Remark 2: Theorem 3's proof meets all conditions [26, Th. 3.2], which establishes an average dwell time to guarantee stability under arbitrary switching.

IV. A Physically Plausible Adversarial Threat Model

The switching architecture introduced in the learning scheme as described in the previous section, extends the attack surface of the learning agent. However, in the event that the attacker has an information advantage about the physics of the system, one needs to construct bounds within which the proposed framework can still produce policies resilient to such attacks. We shall show that adversarial noise during Q-learning, if within prescribed bounds, allows the learning agent to still produce policy gains that maintain controllability despite these attacks. The gradient of a cumulative state-dependant reward function, defined as $\eta(x)$ with respect to the system's states yields a physically plausible noise vector. Reward functions, such as $\eta(x)$, are dual in nature to value functions; they track cumulative rewards collected between two points in a trajectory while value functions yield the cost-to-go between these points. Comparing $\eta(x)$ and $V(x)$ shall ensure that their gradients are equivalent, and hence by using (12) we can conclude that, $\nabla_x(\eta) := \nabla_x V^\star(x) = \nabla_x Q^\star(x, u)$ with $\nabla_x(\cdot) := \frac{\partial(\cdot)}{\partial x}$. The adversarial noise has the following structure,

$$\begin{aligned} x_{\mathrm{noise}} &= \epsilon\nabla_x Q(x, u) \\ &= \epsilon_x(Px + 2PAx + Mx) + \epsilon_u(PBu), \end{aligned} \tag{24}$$

where $\epsilon_x, \epsilon_u \in \mathbb{R}^+$.



(a) Evolution of the switching signal $\sigma(t)$ for unpredictability.

(b) Evolution of the states $x(t)$.

(c) Evolution of the control signal, $u(t)$.

Fig. 1. Evolution of the switching signal, the states of the system, and the policies.

Lemma 2: Assume that the physically plausible adversarial threat model is given by (24). Then the dynamics, given in (1) become,

$$\dot{\hat{x}} = [A + \epsilon_x \hat{Q}_{\rm xx}]x + [B + \epsilon_u \hat{Q}_{\rm xu}]u, \ \ t \geq 0. \tag{25}$$

Proof: Writing the system dynamics with the added noise (24) yields,

$$\begin{aligned}\dot{\hat{x}} &\triangleq \dot{x} + x_{\rm noise}\\ \dot{\hat{x}} &= [A + \epsilon_x(P + A^{\rm T}P + PA + M)]x + [B + \epsilon_u(PB)]u\\ &= [A + \delta A]x + [B + \delta B]u, \ \ t \geq 0.\end{aligned} \tag{26}$$

In the model-free formulation, by using $P + A^{\rm T}P + PA + M = \hat{Q}_{\rm xx}$ and $PB = \hat{Q}_{\rm xu}$ one gets the required result. ■

Remark 3: Note that, in an adversarial environment wherein the malicious agent has access to the physics of the learning agent's dynamics, we need to estimate, in a model-free manner, the worst-case disturbance that the learning agent's A and B matrices can tolerate.

We shall use two models of physically plausible adversarial perturbation. Those are described as: a random adversarial threat model, and a specific adversarial threat model. The random adversarial threat model is generated by selecting δA in (26) such that $\hat{x}^{\rm A}_{\rm noise} = \hat{\epsilon}_x \hat{Q}_{\rm xx}$ wherein $\hat{\epsilon}_x \in (-\epsilon_x, \epsilon_x)$, and δB in (26) such that $\hat{x}^{\rm B}_{\rm noise} = \hat{\epsilon}_u \hat{Q}_{\rm xu}$ wherein $\hat{\epsilon}_u \in (-\epsilon_u, \epsilon_u)$. Here, the adversarial noise is less predictable, but does not introduce worst-case multipliers, $|\epsilon_x|$ and $|\epsilon_u|$, into the dynamics of the system as process noise. Specific adversarial threat model introduces these worst-case multipliers at all times in the evolution of the system dynamics, which may enable the malicious agent to keep the learning agent uncontrollable at all times by selecting $|\epsilon_x|$ and $|\epsilon_u|$ appropriately. The distance to uncontrollability is a measure of how close a pair (A, B) is to the nearest uncontrollable pair. According to [27], the distance to the nearest uncontrollable pair is given by $d_{\rm uc} = \{\inf(\|\delta A\|_{\rm F}^2 + \|\delta B\|_{\rm F}^2)^{1/2} : \delta A \in \mathbb{R}^{\rm nxn}, \delta B \in \mathbb{R}^{\rm nxm}$ such that $(A + \delta A, B + \delta B)$ is uncontrollable$\}$, where $\|.\|_{\rm F}$ denotes the Frobenius norm of a matrix.

Lemma 3: The distance to uncontrollability $d_{\rm iuc}$ is bounded below as, $\frac{\underline{\lambda}(M_i)}{(\|\hat{Q}_{\rm ixx}\| + \|\hat{Q}_{\rm ixu}\hat{Q}_{\rm iuu}^{-1}\hat{Q}_{\rm iux}\|)(1+\sqrt{1+\frac{\lambda(M)}{\underline{\lambda}(R_i)}})} \leq d_{\rm iuc}, \ \ \forall i.$

Proof: By following [28], one gets $\frac{\underline{\lambda}(M_i)}{d_{\rm iuc}(1+\sqrt{1+\frac{\lambda(M_i)}{\underline{\lambda}(R_i)}})} \leq \|P_i\|$. Using (13), and the triangle inequality, we have $\frac{\underline{\lambda}(M_i)}{(\|\hat{Q}_{\rm ixx}\| + \|\hat{Q}_{\rm ixu}\hat{Q}_{\rm iuu}^{-1}\hat{Q}_{\rm iux}\|)(1+\sqrt{1+\frac{\lambda(M_i)}{\underline{\lambda}(R_i)}})} \leq d_{\rm iuc}, \ \ \forall i.$ ■

Algorithm 2 Adversarial Q-Learning

1: Give x_0, $\hat{W}_c(0)$, $\hat{W}_a(0)$, and the set $\mathcal{K}_a$.
2: **procedure**
3: Compute physically plausible process noise using (25).
4: Propagate t, $x(t)$ using (26).
5: Compute $u(t) = -\hat{K}x(t)$ where $\hat{K} = -\hat{W}_a^{\rm T}$.
6: Select a random set of actuators from $\mathcal{K}_a$.
7: Prune $\hat{K}$ and activate according to (10).
8: **if** $t < T_{\rm exp}$
9: Add probing noise $u(t) \leftarrow u(t) + u_{\rm PE}(t)$
10: **end if**
11: Switch active actuators every $\hat{\tau}_{\rm dwell}$ according to (23).
12: Propagate $\hat{W}_c$ and $\hat{W}_a$ according to (7) and (11) respectively.
13: Estimate e_c and e_a.
14: **if** $e_a \neq 0$ and $e_c \neq 0$
15: Go to step 8.
16: **end if** ▷ $e_a \approx 0$ and $e_c \approx 0$
17: **end procedure**

Remark 4: This lemma provides a lower bound on $d_{\rm iuc}$, $\forall i$, estimating how large of a disturbance each mode of the system can tolerate.

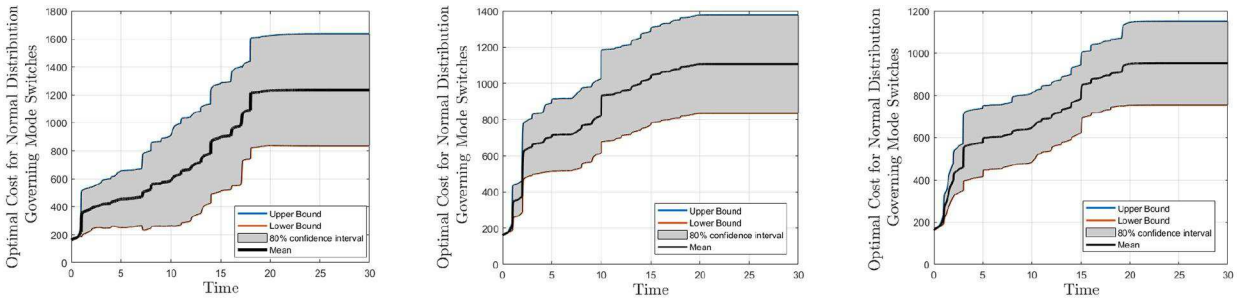
Our proposed framework is described in Algorithm 2.

V. Simulation Results

Consider the model of the F-16 fighter jet [7] given by,

$$\dot{x} = \begin{bmatrix} -1.0189 & -0.9051 & -0.0022 \\ 0.8223 & -1.0774 & -0.1756 \\ 0 & 0 & -1.0000 \end{bmatrix} x + \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} u, \tag{27}$$

and $M = 20I_3$, $R = 0.3I_3$, $\alpha_c = 80$, $\alpha_u = 20$, $T_{\rm exp} = 20$s, $T = 0.01$s The data driven n-step controllability test for this system with $n = 50$ steps, given that all actuators remain activated, suggests that the system is controllable, i.e., $\underline{\lambda}(W_{\rm dc}) = 0.001 > 0$. The set of allowable modes considered here is $\mathcal{K}_a' = \{\{1\}, \{2\}, \{3\}\}$. A uniform distribution governs the switching with dwell time $\hat{\tau}_{\rm dwell} = 1$s to comply with (22). Figure 1a shows the evolution of the switching signal. The convergence of the states under MTD, wherein the switches occur between actuator sets via the "on-off" policy gain matrices is shown in Figure 1b, where one can see that the system remains stable. The policies, capturing the switching behavior are shown in Figure 1c. Given physically plausible adversarial manipulation we want to evaluate the cost degradation. The highest average cost for the specific adversarial noise



(a) Average integral cost for MTD in specific adversarial learning along with error bars to represent an 80% confidence interval.

(b) Average integral cost for MTD in random adversarial learning along with error bars to represent an 80% confidence interval.

(c) Average integral cost for MTD in the absence of threat model along with error bars to represent an 80% confidence interval.

Fig. 2. Evolution of the average integral cost of the associated optimal control problem.

shown in Figure 2a, when compared to the random noise in Figure 2b, arises from using $\epsilon_x = 0.01$ and $\epsilon_u = 0.05$, instead of $\epsilon_x \in (-0.01, 0.01)$ and $\epsilon_u \in (-0.05, 0.05)$. The average cost attained while using MTD without any adversarial noise in Figure 2c is the lowest.

VI. Conclusion and Future Work

This letter devises a framework to diversify the attack surface of Q-learning by exploiting the actuator redundancies of the system while reducing the communication costs by generating "on-off" policy gain matrices that are optimal in nature. Finally, adversarial threat models are devised to "robustify" the learning agent against physically plausible adversarial attacks. Future research efforts will focus on the development of intermittent learning strategies that will further reduce the cost of the communication and actuator resources using operant conditioning reinforcement schedules. Compressed sensing algorithms will be explored to realize these intermittent triggering strategies.

References

[1] R. Baheti and H. Gill, "Cyber-physical systems," *Impact Control Technol.*, vol. 12, no. 1, pp. 161–166, 2011. [Online]. Available: http://ieeecss.org/impact-control-technology-1st-edition

[2] J. Kim, H. Kim, K. Lakshmanan, and R. R. Rajkumar, "Parallel scheduling for cyber-physical systems: Analysis and case study on a self-driving car," in *Proc. ACM/IEEE 4th Int. Conf. Cyber Phys. Syst.*, 2013, pp. 31–40.

[3] I. Lee and O. Sokolsky, "Medical cyber physical systems," in *Proc. IEEE Design Autom. Conf.*, 2010, pp. 743–748.

[4] M. A. Rahman and H. Mohsenian-Rad, "False data injection attacks with incomplete information against smart power grids," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2012, pp. 3153–3158.

[5] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013. [Online]. Available: arXiv:1312.6199.

[6] K. G. Vamvoudakis *et al.*, "Autonomy and machine intelligence in complex systems: A tutorial," in *Proc. IEEE Amer. Control Conf. (ACC)*, 2015, pp. 5062–5079.

[7] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst. Mag.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

[8] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[9] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural Network Design*, Martin Hagan, 2014.

[10] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[12] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," 2017. [Online]. Available: arXiv:1703.06748.

[13] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Proc. Asian Conf. Mach. Learn.*, 2011, pp. 97–112.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014. [Online]. Available: arXiv:1412.6572.

[15] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.

[16] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," *CoRR*, vol. abs/1605.01775, May 2016. [Online]. Available: http://arxiv.org/abs/1605.01775

[17] A. Mandlekar, Y. Zhu, A. Garg, L. Fei-Fei, and S. Savarese, "Adversarially robust policy learning: Active construction of physically-plausible perturbations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 3932–3939.

[18] S. Jajodia, A. K. Ghosh, V. Swarup, C. Wang, and X. S. Wang, *Moving Target Defense: Creating Asymmetric Uncertainty for Cyber Threats*, 1st ed. New York, NY, USA: Springer, 2011.

[19] S. Shetty, X. Yuchi, and M. Song, *Moving Target Defense in Distributed Systems*. New York, NY, USA: Springer, 2016, pp. 1–11.

[20] R. Zhuang, S. A. DeLoach, and X. Ou, "Towards a theory of moving target defense," in *Proc.1st ACM Workshop Moving Target Defense*, New York, NY, USA, 2014, pp. 31–40.

[21] C. E. Shannon, "A note on the concept of entropy," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[22] A. Kanellopoulos and K. G. Vamvoudakis, "A moving target defense control framework for cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 65, no. 3, pp. 1029–1043, Mar. 2020.

[23] K. G. Vamvoudakis, "Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach," *Syst. Control Lett.*, vol. 100, pp. 14–20, Feb. 2017.

[24] H. R. Shaker and S. Lazarova-Molnar, "A new data-driven controllability measure with application in intelligent buildings," *Energy Build.*, vol. 138, pp. 526–529, Mar. 2017.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] D. Liberzon, "Switching in systems and control," in *Systems & Control: Foundations & Applications*. Boston, MA, USA: Birkhäuser, 2003.

[27] P. Gahinet and A. Laub, "Algebraic Riccati equations and the distance to the nearest uncontrollable pair," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 765–786, 1992.

[28] C. He, "On the distance to uncontrollability and the distance to instability and their relation to some condition numbers in control," *Numerische Math.*, vol. 76, no. 4, pp. 463–477, Jun. 1997.