Covert Adversarial Actuators in Finite MDPs

Edoardo David Santi, Gongpu Chen, Deniz Gündüz
Department of Electrical and Electronic Engineering
Imperial College London
London, UK
{eds17, gongpu.chen, d.gunduz}@ic.ac.uk

Asaf Cohen
The School of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel
coasaf@bgu.ac.il

***Abstract*—We consider a Markov decision process (MDP) in which actions prescribed by the controller are executed by a separate actuator, which may behave adversarially. At each time step, the controller selects and transmits an action to the actuator; however, the actuator may deviate from the intended action to degrade the control reward. Given that the controller observes only the sequence of visited states, we investigate whether the actuator can covertly deviate from the controller's policy to minimize its reward without being detected. We establish conditions for covert adversarial behavior over an infinite time horizon and formulate an optimization problem to determine the optimal adversarial policy under these conditions. Additionally, we derive the asymptotic error exponents for detection in two scenarios: (1) a binary hypothesis testing framework, where the actuator either follows the prescribed policy or a known adversarial strategy, and (2) a composite hypothesis testing framework, where the actuator may employ any stationary policy. For the latter case, we also propose an optimization problem to maximize the adversary's performance.**

I. Introduction

There are many scenarios in which an adversary can be stopped, or would suffer negative consequences if identified. Hence, the adversary may wish to limit its adversarial behaviour to stay covert. Conversely, there are scenarios where an agent wishes to complete a task while staying covert so that an adversary cannot interfere or extract information from it. A setting of this type is the problem of covert communication [1]–[3]. It involves designing communication protocols that enable a sender to transmit messages to a receiver without being detected by an adversary monitoring the communication channel. Adversaries in control systems extend these ideas by interfering with the operation of controllers to make their tasks harder. The goal is to influence the state or behaviour of a dynamic system (e.g., a robot, autonomous vehicle, or networked system) while remaining stealthy: avoiding anomalies that might trigger alarms or reveal the presence of external interference. Attacking control systems also has to account for their dynamic and feedback-driven nature, making it a more complex task.

One type of attack involves corrupting the observations of the monitor or the controller of the system to distort its knowledge about the system state, and to derail its actions [4]–[8]. In another class of attacks, the adversary corrupts the control signal [9]–[15]. In [16], the problem of command corruption is studied in multi-armed bandits, where the arms pulled are different from those commanded. In other settings, an agent may wish to stay covert while completing a task. In [17], the authors consider an active sensing scenario, where the goal of the adversary is to estimate a parameter without being detected. In [18], covert best arm detection is studied in a multi-armed bandit setting. In [19], the authors study a reachability problem in an MDP, where the objective is to limit the probability of an adversary observing the states to infer their transition probabilities under the chosen policy.

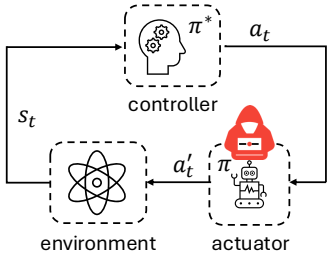

Fig. 1. System model. The controller wants the actuator to follow policy π^*; however, the compromised actuator instead aims to minimize the reward without being detected.

In this paper, we consider a controller interacting with the environment by controlling the actions of an actuator, formulated as an MDP. The adversary either takes over the actuator itself, or infiltrates the communication channel between the controller and the actuator, corrupting the instructions received by the actuator. The adversary aims to modify controller's policy to degrade the system's performance, e.g., to minimize the long-term average reward, while avoiding detection. See Fig. I for an illustration of the system model. We assume that the system dynamics, i.e., the state transition probabilities and reward function, and the current state are known by both the controller and the adversarial actuator, while the average reward is learned only at the end. Hence, the goal of the adversary is to corrupt the actions taken without significantly distorting the statistics of the visited states to avoid detection.

Notation: We denote random variables with upper case letters and realizations with the associated lower case letters. For $n, k \in \mathbb{Z}^+$, we use s_n^{n+k} to represent the sequence $\{s_n, s_{n+1}, \dots, s_{n+k}\}$. s_1^n will be abbreviated as s^n for simplicity and for any symbol, '$_{s^n}$' subscript indicates an empirical estimate using sequence s^n. For any set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of probability distributions over $\mathcal{X}$, while $\mathcal{X}^0$ denotes its interior and $\overline{\mathcal{X}}$ its closure, both in the total

variation metric. $\mathbb{1}\{\cdot\}$ is the indicator function. For two distributions p and q over $\mathcal{X}$ we define the relative entropy as $H(p\|q) = \sum_{i\in\mathcal{X}} p(i)\log\frac{p(i)}{q(i)}$. We adopt $\log = \log_2$.

II. Problem Formulation

A. System model

An MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, T, r, \mu)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, $T : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ is the transition kernel, $r : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is the reward function, and $\mu\in\Delta(\mathcal{S})$ is the initial state distribution. In this work, we focus on MDPs with finite state and action spaces. At every time t, the controller selects an action $a_t\in\mathcal{A}$ based on the current state s_t and transmits a_t to the actuator. Upon the execution of a_t by the actuator, the environment generates a reward $r(s_t, a_t)$ and transitions to the next state s_{t+1} following the distribution $T(\cdot|s_t, a_t)$. We assume that the controller cannot observe the instantaneous reward at each time step; instead, only the average reward is revealed at the very end. A stationary policy for the MDP is a mapping that maps the current state to a distribution over actions. We denote by $\pi(a_t|s_t)$ the probability that policy π selects action a_t in state s_t. The expected average reward of policy π over an infinite time horizon is defined as

$$J(\pi) := \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_\pi\left[\sum_{t=1}^{N} r(s_t, a_t)|s_1\sim\mu\right].$$

Let T_π denote the transition matrix induced by policy π, where each entry is given by $T_\pi(s, s') = \sum_a T(s'|s, a)\pi(a|s)$. Throughout this paper, we assume that the MDP is recurrent, meaning that the Markov chain induced by any stationary policy consists of a single recurrent class. As a result, under any stationary policy π, the system reaches a unique stationary state distribution, denoted by $\bar{\theta}_\pi$. This stationary distribution satisfies $\tau_\pi^\top T_\pi = \tau_\pi^\top$. Moreover, let $\theta_\pi(s, s') = \tau_\pi(s)T_\pi(s, s')$, $\forall s, s'\in\mathcal{S}$, be the stationary distribution over state transitions induced by policy π.

This work investigates a scenario where the controller wants to follow a predefined policy π^*, while the adversarial actuator may not faithfully execute the actions prescribed by π^*. Fig. 1 provides an illustration of the system. For simplicity, we refer to the actuator and the adversary interchangeably as a single entity. We assume that the adversary has full knowledge of the control policy π^* and that both the controller and the adversary have access to the environment state at every time. Consequently, at every time t, the controller prescribes an action a_t using policy π^*; however, the adversary may override this action and instead execute an action a'_t, selected according to an adversarial policy π. To quantify the impact of adversarial behavior, we define the regret induced by the adversarial policy π as:

$$R(\pi) = J(\pi^*) - J(\pi)\ .$$

The objective of the adversary is to find a policy π that maximizes the regret covertly, as being found out might result in negative repercussions. On the other hand, the controller aims to detect with the best possible accuracy whether the actuator is acting faithfully or not.

B. Detection and Covertness

Since the controller has full access to the entire state sequence up to the current time, it can leverage this information to detect any abnormal behavior based on observed state transitions. In this part, we discuss the controller's ability to detect whether the control policy π^* is executed faithfully.

We formulate the controller's detection problem as a hypothesis test, where the null hypothesis $\mathcal{H}^*$ corresponds to the actuator acting faithfully to the controller's policy, i.e. $\pi = \pi^*$, while the alternative hypothesis $\mathcal{H}^{\text{adv}}$ corresponds to adversarial behaviour, i.e. $\pi\neq\pi^*$. For any symbol, let '*' and '$^{\text{adv}}$' superscripts indicate belonging to the two hypotheses, respectively. Let $g_n : \mathcal{S}^n\to\{0,1\}$ denote the controller's decision function at time n, where $g_n(s^n) = 0$ means that $\mathcal{H}^*$ is accepted at time n, and $g_n(s^n) = 1$ means that $\mathcal{H}^{\text{adv}}$ is accepted. Let $P_n^* : \mathcal{S}^n\mapsto[0,1]$ and $P_n^{\text{adv}} : \mathcal{S}^n\mapsto[0,1]$ be the probabilities of sequences of length n under the null and alternative hypothesis, respectively. Define $\mathcal{B}_n\subseteq\mathcal{S}^n$ as the set of sequences for which $g_n(s^n) = 0$, and denote its complement by $\mathcal{B}_n^c$, where $g_n(s^n) = 1$. Then the two probabilities of error are defined as follows:

$$\alpha_n := \Pr\{g_n(s^n) = 1|\mathcal{H}^*\text{ true}\} = P_n^*(\mathcal{B}_n^c)\ ,$$
$$\beta_n := \Pr\{g_n(s^n) = 0|\mathcal{H}^{\text{adv}}\text{ true}\} = P_n^{\text{adv}}(\mathcal{B}_n)\ .$$

Specifically, let $\alpha = \lim_{n\to\infty}\alpha_n$ and $\beta = \lim_{n\to\infty}\beta_n$. We assume the controller performs the optimal detection algorithm, defined as the one minimizing the total probability of error $\alpha+\beta$. An adversarial policy is said to be ϵ-covert if $\alpha+\beta = 1-\epsilon$ under this policy. For any recurrent MDP with finite state and action spaces, for any non-stationary policy there exists a stationary policy that achieves the same long-term average state-action frequencies [20]. We show later that the covertness properties of a policy depend on its long-term state transition frequencies, which are fully determined by its state-action frequencies, which also fully determine the long-term average reward. Hence, we can conclude that for any non-stationary policy, there exists a stationary policy that achieves the same average long-term reward and covertness, and so we can limit ourselves to considering stationary policies.

III. Perfect Covertness over an Infinite Horizon

In this section, we analyze the covertness of the adversarial policy over an infinite time horizon. In this setting, an adversarial policy can either be 0-covert—meaning it will be detected by the controller with probability 1—or 1-covert, indicating perfect covertness. We derive the necessary and sufficient conditions for a policy to be 1-covert, and subsequently formulate the problem of finding the optimal adversarial policy as a linear program.

For any state sequence s^n, we define the following empirical distributions: for any $s, s' \in \mathcal{S}$,

$$\tau_{s^n}(s) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{s_t = s\}, \tag{1}$$

$$\theta_{s^n}(s,s') = \frac{1}{n-1}\sum_{t=1}^{n-1} \mathbb{1}\{s_t = s, s_{t+1} = s'\}, \tag{2}$$

$$T_{s^n}(s,s') = \frac{\theta_{s^n}(s,s')}{\tau_{s^n}(s)}. \tag{3}$$

Suppose that the adversary adopts a stationary policy π^{adv}. Denote by T^* and T^{adv} the transition matrices induced by polices π^* and π^{adv}, respectively. Recall that we assume the MDP is recurrent, hence both T^* and T^{adv} are irreducible.

Theorem 1. *For any stationary policy $\pi^{adv} \neq \pi^*$, π^{adv} is 1-covert if $T^{adv} = T^*$ and 0-covert otherwise.*

Proof. As $n \to \infty$, $T_{s^n} \to T^{\text{adv}}$. Thus, if $T_{s^n} \neq T^*$, it is obvious that $\pi^{\text{adv}} \neq \pi^*$, the controller accepts $\mathcal{H}^{\text{adv}}$ and $\epsilon = 1$. If $T_{s^n} = T^*$, and there exist multiple stationary policies inducing the transition matrix T^*, the controller is uninformed and $\epsilon = 0$. The case where there exists a single stationary policy inducing the transition matrix T^* is degenerate as it implies that $\pi^{\text{adv}} = \pi^*$, in which case the detection is between two identical hypotheses. □

Theorem 1 implies that an adversarial policy π^{adv} is perfectly covert if and only if its transition matrix is identical to that of the original control policy. Denote by Π_1 the set of policies satisfying this condition, i.e., $\Pi_1 := \{\pi^{\text{adv}} : T^{\text{adv}} = T^*\}$. Let T_s be the $|\mathcal{A}|\times|\mathcal{S}|$ matrix with $T_s(a,k) = T(k|s,a)$. Then $\pi \in \Pi_1$ if and only if $T_s^\top(\pi(\cdot|s) - \pi^*(\cdot|s)) = 0$ for all $s \in \mathcal{S}$. Hence Π_1 is convex whenever it is non-empty.

Given Theorem 1, we can set an optimization problem which finds the optimal adversarial policy from the adversary's point of view, with the constraint that $\epsilon = 0$. Define the $(|\mathcal{S}|+1)\times|\mathcal{A}|$ matrix $C = \begin{bmatrix} T_s & \mathbf{1} \end{bmatrix}^\top$, and $\Delta\pi = \pi - \pi^*$. Then, for each $s \in \mathcal{S}$, the best adversarial policy that guarantees $\epsilon = 0$ is given by

$$\begin{aligned} \min_{\Delta\pi(\cdot|s)} \ & \Delta\pi(\cdot|s)^\top r(s,\cdot) \\ s.t.\ & C\Delta\pi(\cdot|s) = \mathbf{0}, \\ & -\pi^*(s,a) \le \Delta\pi(s,a) \le 1 - \pi^*(s,a), \forall a \in \mathcal{A} \end{aligned} \tag{4}$$

We can see that the admissible values of $\Delta\pi$ belong to the null space of C. Thus, if this matrix has the full column rank ($\text{rank}(C) = |\mathcal{A}|$), the null space is trivial, containing only the origin, and the adversary cannot change the policy covertly. Meanwhile, if $\text{rank}(C) < |\mathcal{A}|$, the dimensionality of the space of admissible policies is $|\mathcal{A}| - \text{rank}(C)$.

Remark 1. *Clearly, if we have two actions, a good one, a_g and a bad one a_b such that for all s, s' we have $T(s'|s,a_g) = T(s'|s,a_b)$, but for some s we have $r(s,a_g) > r(s,a_b)$, then the adversary can change the policy and increase the regret, without being noticed by the controller who sees the complete state sequence. So there exist MDPs for which the solution to problem (4) is not only π^* but also a strictly suboptimal policy. This result shows that the adversary can also achieve this in more complex ways, as in each state, it can modify its policy as long as this does not affect the transition probabilities between states. See Appendix A for examples.*

IV. Asymptotic error exponents

In the previous section, we established that as the length of the observed state sequence approaches infinity, the total probability of detection error, ϵ, converges to 1 whenever the adversarial policy induces a transition matrix different from that of the controller's policy. However, in practical scenarios, the time horizon is finite, albeit potentially large. This raises an important question: how rapidly does ϵ increase as a function of time for an adversarial policy π^{adv} when $T^{\text{adv}} \neq T^*$? To address this, we focus on deriving the asymptotic error exponent, which characterizes the rate at which the probability of error decays over time.

We derive the asymptotic results using tools from large deviation theory. Suppose $\{S_t\}$ is the Markov chain induced by a stationary policy π. Then, given a realization s^n of the state sequence, we can construct the sequence of empirical estimates $\{\theta_{s^n}\}$ using (2). Let $P_\pi(\cdot)$ denote the probability distribution of θ_{S^n} under policy π.

Definition 1. *The process $\{\theta_{S^n}\}$ satisfies the large deviation principle (LDP)[1] with rate function I if, for every Borel subset $\Gamma \subseteq \Delta(\mathcal{S}^2)$, we have:*

$$\lim_{n\to\infty} \frac{1}{n} \log P_\pi\{\theta_{S^n} \in \Gamma\} = -\inf_{\nu\in\overline{\Gamma}} I(\nu).$$

where $I : \Delta(\mathcal{S}^2) \to \mathbb{R} \cup \{\infty\}$ is a continuous mapping, and Γ does not have any isolated points, i.e. $\Gamma \subseteq \overline{\Gamma^0}$.

Intuitively, satisfying the LDP means that given a large number of samples, the probability of the sequence of realizations of the process being within a certain set decreases exponentially at a rate depending on the member of the set giving the slowest rate of decrease. Let $\mathcal{M}$ be the set of shift-invariant measures, meaning that $\mathcal{M} = \{\nu \in \Delta(\mathcal{S}^2) : \sum_{j\in\mathcal{S}} \nu_{i,j} = \sum_{j\in\mathcal{S}} \nu_{j,i}\}$. Let $\theta^1, \theta^2 \in \mathcal{M}$ be two shift-invariant distributions over state transitions, τ^1, τ^2 be the corresponding marginal distributions over states, and T^1, T^2 be the corresponding transition matrices. Then, we can define the differential divergence $D_K : \mathcal{M}^2 \mapsto \mathbb{R}_0^+ \cup \{\infty\}$ between state transition distributions:

$$D_K(\theta^1,\theta^2) := \sum_{s\in\mathcal{S}} \tau^1(s) \sum_{s'\in\mathcal{S}} T^1(s,s') \log \frac{T^1(s,s')}{T^2(s,s')}. \tag{5}$$

According to (1)-(3), it is easy to verify that $D_K(\theta^1,\theta^2) = H(\theta^1||\theta^2) - H(\tau^1||\tau^2)$. The following result from the large deviation theory is useful in our analysis:

Theorem 2. *[21, Theorem 3.1.13] Given a Markov process $\{S_t\}$ with irreducible transition matrix T and state transition*

[1] We define the LDP in a stricter sense than in [21].

frequencies θ, the $\Delta(\mathcal{S}^2)$-valued process $\{\theta_{S^n}\}$ satisfies the LDP with the rate function

$$I(\nu) = \begin{cases} D_K(\nu, \theta), & \text{if } \nu \in \mathcal{M} \\ \infty, & \text{if } \nu \notin \mathcal{M} \end{cases}$$

Note that this result is a variation of the classic result given by Sanov's theorem, which is defined for i.i.d. processes instead and has a rate $I(\nu) = D(\nu\|\phi)$, where ϕ is the distribution of the i.i.d. variables.

A. Predefined adversarial policy

We start from a simple setting where the adversary either follows the control policy faithfully or deviates by executing a known adversarial policy π^{adv}, which is known to the controller as well. In this setting, the controller is faced with a binary hypothesis testing problem. The null hypothesis, $\mathcal{H}^*$, represents the scenario where the actuator adheres to the intended control policy, while the adversarial hypothesis, $\mathcal{H}^{\text{adv}}$, corresponds to the scenario where the actuator executes π^{adv}. Both policies are assumed to be stationary and known to both the controller and the adversarial actuator.

Define the normalized log-likelihood ratio of s^n as

$$L(s^n) = \frac{1}{n-1} \log \frac{P_n^*(s^n)}{P_n^{\text{adv}}(s^n)} = D_K(\theta_{s^n}, \theta^{\text{adv}}) - D_K(\theta_{s^n}, \theta^*), \tag{6}$$

where the second equality is shown in Appendix B. We set our detector g_n as the normalized log-likelihood ratio test, which is equivalent to the likelihood ratio test used in binary hypothesis testing. It accepts the null hypothesis ($g_n(s^n) = 0$) when $L(s^n) > \eta$, where η is a threshold, and it accepts the alternative ($g_n(s^n) = 1$) otherwise. The normalized log-likelihood ratio test defines a set $\mathcal{B}_n \subseteq \mathcal{S}^n$, under which the null hypothesis is accepted, and its complement $\mathcal{B}_n^c$, under which the alternative hypothesis is accepted. We can see from (6) that the normalized log-likelihood ratio depends on the trajectory s^n only through the empirical distribution over state transitions θ_{s^n}. Thus, the test equivalently defines a set $\hat{\mathcal{B}}_n \subseteq \Delta(\mathcal{S}^2)$, in which $\mathcal{H}^*$ is accepted, and its complement $\hat{\mathcal{B}}_n^c$, in which $\mathcal{H}^{\text{adv}}$ is accepted, depending on the location of θ_{s^n}.

The Neyman-Pearson lemma states that for a sequence of random variables observed, the optimal detector uses a threshold on the likelihood ratio. The null hypothesis is rejected when the likelihood ratio is lower than a threshold, and it is accepted otherwise. This is optimal in the sense that it is not possible to decrease the probability of type I error (i.e., α_n) without increasing the probability of type II error (i.e., β_n). The threshold can be adjusted higher or lower based on the type of error we prioritize reducing.

1) Error exponents for a fixed threshold: In this part, we analyze the asymptotic decay of α_n and β_n when setting a fixed threshold for all n. For a fixed threshold η, the sets $\hat{\mathcal{B}}_n$ are the same for any $n \geq 2$, i.e., $\hat{\mathcal{B}}_n = \hat{\mathcal{B}}$. Then, we can define the two probabilities of error for all $n \geq 2$ as

$$\alpha_n = \Pr(\theta_{s^n} \in \hat{\mathcal{B}}^c | \mathcal{H}^*), \quad \beta_n = \Pr(\theta_{s^n} \in \hat{\mathcal{B}} | \mathcal{H}^{\text{adv}}).$$

Assuming irreducibility of the hypotheses' transition matrices, which is guaranteed by the recurrent MDP assumption, applying Theorem 2 to these equations, noting that I in this form is continuous and that both $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}}^c$ do not contain isolated points (see Appendix C for the proof), we obtain the following.

Theorem 3. *Under the normalized log-likelihood ratio detector with a fixed threshold $-D_K(\theta^{adv}, \theta^*) \leq \eta \leq D_K(\theta^*, \theta^{adv})$, the asymptotic error exponents are given by*

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = - \inf_{\theta \in \overline{\hat{\mathcal{B}}^c} \cap \mathcal{M}} D_K(\theta, \theta^*),$$
$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = - \inf_{\theta \in \overline{\hat{\mathcal{B}}} \cap \mathcal{M}} D_K(\theta, \theta^{adv}).$$

where the infimum of the empty set is taken to be ∞.

Note that the threshold must take those values as to have a boundary between $\hat{\mathcal{B}}_n$ and $\hat{\mathcal{B}}_n^c$ that is located between the two hypotheses in the space of state transition frequencies. Informally speaking, D_K can be considered a divergence between different distributions over state transitions. In general, the divergences in the above statement grow if the divergence between θ^* and θ^{adv} is greater, leading to higher error exponents due to the processes becoming increasingly different, and thus, easy to distinguish. Additionally, note that when the exponents are 0, the error probabilities do not decay asymptotically. This occurs if $\theta^* = \theta^{\text{adv}}$, meaning that the two Markov chains are indistinguishable, which is the setting described in Section III.

2) Error exponent for a fixed α_n (error type I): In the following, we derive the asymptotic error exponent of β_n (type II error) when instead of fixing the threshold of the detector, we fix an upper bound on the type I error probability and adjust the threshold accordingly. We now provide a version of the Chernoff-Stein lemma, adapted to the Markov chain scenario.

Theorem 4. *Consider the binary hypothesis test when s^n is a Markov chain drawn according to either of the two state transition distributions θ^* and θ^{adv}, respectively, where $D_K(\theta^*, \theta^{adv}) < \infty$. Let $\mathcal{A}_n \subseteq \mathcal{S}^n$ be an acceptance region for the null hypothesis $\mathcal{H}^0$. Let the probabilities of error be*

$$\alpha_n = P_n^*(\mathcal{A}_n^c), \qquad \beta_n = P_n^{adv}(\mathcal{A}_n)$$

and for $0 < \delta < \frac{1}{2}$, define

$$\beta_n^\delta = \min_{\substack{\mathcal{A}_n \subseteq \mathcal{S}^n \\ \alpha_n < \delta}} \beta_n.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n^\delta = -D_K(\theta^*, \theta^{adv}). \tag{7}$$

Proof. The detailed proof is provided in Appendix D. □

Again, it can be seen that the rate of decay of the error probabilities is improved as the divergence between θ^* and θ^{adv} increases, with an error exponent of zero when $\theta^* = \theta^{\text{adv}}$.

B. Unknown adversarial policy

In Section IV-A, we analyzed the performance of a detector with the assumption that the alternative hypothesis $\mathcal{H}^{\text{adv}}$ is known. In this section, we study the setting where the controller has no knowledge of the adversarial policy or its possible distribution, and thus conducts the detection in an 'anomaly detection' fashion: for any state sequence, the detector calculates a statistic of the sequence and compares it with a fixed threshold. If the statistic is above the threshold, the detector does not reject the null hypothesis, otherwise it is rejected. The threshold is chosen depending on the accepted level of Type I error, i.e., the probability of wrongly rejecting the null hypothesis. To decide on the best statistic to use, we define optimality as initially proposed by Hoeffding [21]:

Definition 2. *A test $\mathcal{S}$ is optimal (for a given threshold $\eta > 0$) if, among all tests that satisfy*

$$\limsup_{n\to\infty} \frac{1}{n} \log \alpha_n \leq -\eta,$$

test $\mathcal{S}$ has the maximal exponential rate of error, i.e., uniformly over all possible laws μ_1, $-\limsup_{n\to\infty} \frac{1}{n} \log \beta_n$ is maximal.

Using this definition, we can restrict our focus on tests that rely on the empirical measure of state transitions. This follows from [21, Lemma 3.5.3], whose proof can be readily adapted to accommodate the case of state transition distributions in Markov chains. We use the following theorem to choose the statistic to be used for detection.

Theorem 5. *Let test $\mathcal{F}^*$ consist of the maps*

$$\mathcal{F}^*(s^n) = \begin{cases} 0, & \text{if } D_K(\theta, \theta^*) < \eta, \\ 1, & \text{otherwise.} \end{cases}$$

Then $\mathcal{F}^*$ is an optimal test for η and its error exponents are given by

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = -\eta$$
$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = - \inf_{\{\nu: D_K(\nu,\theta^*)<\eta\}} D_K(\nu, \theta^{adv}).$$

Proof. The proof is provided in Appendix D. □

In this setting, the probability of type I error is a design choice made by the controller and the adversary cannot affect it. The adversary chooses the adversarial policy in order to maximize the regret while also maximizing the probability of type II error. Similarly to the previous settings, the error exponents tend to increase as the D_K divergence between θ^* and θ^{adv} increases. In this case, the relevant divergences are the divergence of θ^* to the decision boundary, which determines the exponent of type I error, and the smallest divergence of θ^{adv} to this boundary, which determines the exponent of type II error. Note that it is possible that both θ^* and θ^{adv} are located on the same side of the decision boundary.

Following Theorem 5, assuming that both entities know the maximum acceptable error exponent corresponding to α_n, the actuator chooses an acceptable level of the error exponent corresponding to β_n, calls it η_β, and solves the following optimization problem:

$$\begin{aligned} &\min_{\pi\in\Pi} \Delta\rho_\pi^\top r \\ &s.t. \inf_{\{\nu: D_K(\nu,\theta^*)<\eta\}} D_K(\nu, \theta^{\text{adv}}) < \eta_\beta \ , \end{aligned} \tag{8}$$

where ρ_π is the $|\mathcal{S}||\mathcal{A}|$-dimensional vector of space-action frequencies induced by policy π, Π is the set of stationary policies, and r is the $|\mathcal{S}||\mathcal{A}|$-dimensional vector of rewards for all space-action combinations. Solution to (4) gives the optimal actuator policy given that the detection occurs after observing an infinitely long sequence of states, whereas (8) provides the optimal actuator policy given that the covertness constraints are given in terms of asymptotic exponents of the two types of error.

C. Discussion

The following connects the results of the two main sections. The actuator aims to maximize the regret $R(\pi) = J(\pi^*) - J(\pi)$. Let π' be the solution of (4). Then, we can write the regret as $R(\pi) = [J(\pi^*) - J(\pi')] + [J(\pi') - J(\pi)]$, where $[J(\pi^*) - J(\pi')]$ represents the regret that can be obtained in a totally covert manner, regardless of the length of the state sequence observed, as this change in policy does not modify the statistics of the induced Markov chain, while $[J(\pi') - J(\pi)]$ is obtained by relaxing the covertness constraint, allowing some information to be 'leaked' to the controller, as this change in policy changes the statistics of the induced Markov chain; and thus, this change must be 0 if detection occurs after observing an infinitely long sequence of states.

Our results show that to achieve a certain degree of covertness, the adversarial cannot use a policy that is 'too far' from π^* in terms of D_K. This means that while the controller may be tempted to simply set π^* as the optimal policy for the system, it is also important to consider the neighbourhood of π^*, as there may be another policy which has lower average rewards, but whose neighbours have better average reward than those of π^*, meaning that the potential covert policies chosen by the actuator could be less damaging to the system.

V. Conclusion

We showed that, given an infinite number of observed states, it is still possible for the adversarial to impact the average reward in certain MDPs without being detected, and the best adversarial policy can be found by solving a linear program. Assuming the number of observed samples is large but finite, the adversary can increase the regret, as the covertness constraint is less tight. We considered error exponents in this setting for adversarial policies known and unknown *a priori* and showed that they take the form of the divergence D_K between the relevant distributions over state transitions. We plan to expand this work by studying the finite time scenario and the setting in which the adversary does not know the dynamics of the MDP but learns in an online fashion by observing the rewards over time.

REFERENCES

[1] B. A. Bash, D. Goeckel, and D. Towsley, "Limits of reliable communication with low probability of detection on awgn channels," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 1921–1930, 2013.

[2] L. Wang, G. W. Wornell, and L. Zheng, "Fundamental limits of communication with low probability of detection," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3493–3503, 2016.

[3] M. R. Bloch, "Covert communication over noisy channels: A resolvability perspective," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2334–2354, 2016.

[4] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2009, pp. 911–918.

[5] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 90–95, Jan. 2011.

[6] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *Preprints of the 1st workshop on Secure Control Systems*, 2010, p. 7.

[7] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 4–13, Mar. 2017, conference Name: IEEE Transactions on Control of Network Systems.

[8] P. Pradhan and P. Venkitasubramaniam, "Stealthy attacks in dynamical systems: Tradeoffs between utility and detectability with application in anonymous systems," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 779–792, 2017.

[9] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *2015 American Control Conference (ACC)*, Jul. 2015, pp. 195–200.

[10] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs," *Automatica*, vol. 82, pp. 251–260, Aug. 2017.

[11] R. Zhang and P. Venkitasubramaniam, "Stealthy control signal attacks in vector lqg systems," in *2016 American Control Conference (ACC)*, Jul. 2016, pp. 1179–1184.

[12] E. Kung, S. Dey, and L. Shi, "The performance and limitations of ϵ-stealthy attacks on higher order systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 941–947, Feb. 2017, conference Name: IEEE Transactions on Automatic Control.

[13] S. Weerakkody, B. Sinopoli, S. Kar, and A. Datta, "Information flow for security in control systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec. 2016, pp. 5065–5072.

[14] B. Amihood and A. Cohen, "Covertly controlling a linear system," in *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 321–326.

[15] B. Amihood and A. Cohen, "Covertly controlling a linear system," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2651–2663, 2024.

[16] M.-C. Chang and M. R. Bloch, "Distributed stochastic bandits with corrupted and defective input commands," in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023, pp. 1318–1323.

[17] M. Tahmasbi and M. R. Bloch, "Active covert sensing," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 840–845.

[18] M.-C. Chang and M. R. Bloch, "Covert best arm identification of stochastic bandits," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 324–329.

[19] M. O. Karabag, M. Ornik, and U. Topcu, "Least inferable policies for markov decision processes," in *2019 American Control Conference (ACC)*, 2019, pp. 1224–1231.

[20] S. Mannor and J. N. Tsitsiklis, "On the empirical state-action frequencies in markov decision processes under general policies," *Mathematics of Operations Research*, vol. 30, no. 3, pp. 545–561, 2005.

[21] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., ser. Stochastic Modelling and Applied Probability, 38. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

APPENDIX

A. Examples of detection at infinity

a) Example 1 - the trivial case: Let $T_s = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \\ 0.8 & 0.2 \end{bmatrix}$ and the policy and the reward function in state s be respectively $\pi(\cdot|s) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $\pi(\cdot|s) = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}$. As the first and the third row of T_s are equal, it is possible to change the policy in this state from deterministically taking the first action to taking the third, without changing the transition matrix of the induced Markov chain. As the reward of the third action is lower and the transition matrix is unchanged, the adversarial covertly affects the total reward of the system.

b) Example 2 - covert adversarial behaviour not possible: Let $T_s = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$ and the policy in state s be $\pi(\cdot|s) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$. Then we get the constraint $\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \Delta\pi(a_0|s) \\ \Delta\pi(a_1|s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$. It is easy to verify that the rank of C is 2 and thus the only solution is the trivial $\Delta\pi(\cdot|s) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

c) Example 3 - covert adversarial behaviour possible: Let $T_s = \begin{bmatrix} 0.8 & 0.5\overline{6} & 0.47\overline{3} \\ 0.1 & 0.21\overline{6} & 0.26\overline{3} \\ 0.1 & 0.21\overline{6} & 0.26\overline{3} \end{bmatrix}$ and the policy in state s be $\pi(\cdot|s) = \begin{bmatrix} 0.\overline{3} \\ 0.\overline{3} \\ 0.\overline{3} \end{bmatrix}$. Then we get the constraint $\begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.5\overline{6} & 0.21\overline{6} & 0.21\overline{6} \\ 0.47\overline{3} & 0.26\overline{3} & 0.26\overline{3} \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \Delta\pi(a_0|s) \\ \Delta\pi(a_1|s) \\ \Delta\pi(a_2|s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$. It easy to show that the matrix C in this case has rank$= 2$ and that the possible solution are in the form $\Delta\pi(\cdot|s) = \begin{bmatrix} 0 \\ \alpha \\ -\alpha \end{bmatrix}$. This gives the adversarial to use any policy $\pi(\cdot|s) = \begin{bmatrix} 0.\overline{3} \\ 0.\overline{3} + \alpha \\ 0.\overline{3} - \alpha \end{bmatrix}$. As we are guaranteeing that the distribution of states in the next step is unaffected, it is sufficient to minimize the immediate expected reward, so given $r(s,\cdot) = \begin{bmatrix} r_0 \\ r_1 \\ r_2 \end{bmatrix}$, the optimal adversarial policy in state s is $\begin{bmatrix} 0.\overline{3} \\ 0.\overline{6} \\ 0 \end{bmatrix}$ if $r_1 < r_2$, $\begin{bmatrix} 0.\overline{3} \\ 0 \\ 0.\overline{6} \end{bmatrix}$ if $r_1 > r_2$ and any policy on the line satisfying the bounds $0 \leq \pi(\cdot|s) \leq 1$ if $r_1 = r_2$.

B. Derivation of the normalized log-likelihood ratio

$$\begin{aligned} L(s_1, s_2, \ldots, s_n) &= \frac{1}{n-1} \log \frac{P_n^*(s_1, s_2, \ldots, s_n)}{P_n^{\text{adv}}(s_1, s_2, \ldots, s_n)} \\ &= \frac{1}{n-1} \sum_{i=1}^{n-1} \log \frac{T^*(s_i, s_{i+1})}{T^{\text{adv}}(s_i, s_{i+1})} \\ &= \frac{1}{n-1} \sum_{a_1, a_2 \in \mathcal{S}^2} (n-1)\theta_{s^n}(a_1, a_2) \log \frac{T^*(a_1, a_2)}{T^{\text{adv}}(a_1, a_2)} \\ &= \sum_{a_1, a_2 \in \mathcal{S}^2} \theta_{s^n}(a_1, a_2) \log \frac{T^*(a_1, a_2)}{T^{\text{adv}}(a_1, a_2)} \frac{T_{s^n}(a_1, a_2)}{T_{s^n}(a_1, a_2)} \\ &= D_K(\theta_{s^n}, \theta^{\text{adv}}) - D_K(\theta_{s^n}, \theta^*) \end{aligned}$$

C. Non-existence of isolated points in log-likelihood ratio decision sets

Let $\theta_{x^n}, \theta_{y^n} \in \Delta(\mathcal{S}^2)$ be two empirical doublet distributions induced by two different sequences of states x^n, y^n. According to Appendix B we have

$$L(x^n) = \sum_{a_1,a_2 \in \mathcal{S}^2} \theta_{x^n}(a_1,a_2) \log \frac{T^*(a_1,a_2)}{T^{\text{adv}}(a_1,a_2)},$$
$$L(y^n) = \sum_{a_1,a_2 \in \mathcal{S}^2} \theta_{y^n}(a_1,a_2) \log \frac{T^*(a_1,a_2)}{T^{\text{adv}}(a_1,a_2)}.$$

Let $\theta_\lambda = \lambda\theta_{x^n} + (1-\lambda)\theta_{y^n}$, where $\lambda \in [0,1]$. If there exists a sequence of states s_λ^n such that $\theta(s_\lambda^n) = \theta_\lambda$, then

$$\begin{aligned} L(s_\lambda^n) &= \sum_{a_1,a_2 \in \mathcal{S}^2} (\lambda\theta_{x^n}(a_1,a_2) + (1-\lambda)\theta_{y^n}(a_1,a_2)) \log \frac{T^{\text{adv}}(a_1,a_2)}{T_2(a_1,a_2)} \\ &= \lambda L(x^n) + (1-\lambda) L(y^n) \end{aligned}$$

Thus, given a constant c, if $L(x^n) > c$ and $L(y^n) > c$, $L(s_\lambda^n) > c$ and conversely, if $L(x^n) < c$ and $L(y^n) < c$, $L(s_\lambda^n) < c$, meaning that the log-likelihood ratio decision sets B and B^c have no isolated points.

D. Proofs of Section IV-A

Lemma 1. *Let s^n be drawn under $\mathcal{H}^*$. Then*

$$\begin{aligned} \frac{1}{n-1} \log \frac{P_n^*(s^n)}{P_n^{adv}(s^n)} &= D_K(\theta_{s^n}, \theta^{adv}) - D_K(\theta_{s^n}, \theta^*) \\ &\to D_K(\theta^*, \theta^{adv}) \end{aligned}$$

in probability, where the first equality is due to (6).

Definition 3. *For a fixed n and $\delta > 0$, a sequence $s^n \in \mathcal{S}^n$ is said to be D_K-typical if and only if*

$$D_K(\theta^*, \theta^{adv}) - \delta \le \frac{1}{n-1} \log \frac{P_n^*(s^n)}{P_n^{adv}(s^n)} \le D_K(\theta^*, \theta^{adv}) + \delta.$$

The set of D_K typical sequences is the D_K-typical set $A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})$.

Lemma 2. *Under the null hypothesis,*

1) *For $s^n = (s_1, s_2, \ldots, s_n) \in A_\delta^{(n)}(P_n^*, P_n^{adv})$,*

$$P_n^*(s^n) 2^{-(n-1)(D_K(\theta^*,\theta^{adv})+\delta)} \le P_n^{adv}(s^n) \le P_n^*(s^n) 2^{-(n-1)(D_K(\theta^*,\theta^{adv})-\delta)}.$$

2) *$P_n^*(A_\delta^{(n)}(P_n^*, P_n^{adv})) > 1-\delta$, for n sufficiently large.*
3) *$P_n^{adv}(A_\delta^{(n)}(P_n^*, P_n^{adv})) < 2^{-(n-1)(D_K(\theta^*,\theta^{adv})-\delta)}$.*
4) *$P_n^{adv}(A_\delta^{(n)}(P_n^*, P_n^{adv})) < (1-\delta)2^{-(n-1)(D_K(\theta^*,\theta^{adv})+\delta)}$, for n sufficiently large.*

Proof. The first property comes from the definition of D_K-typical set. The second property comes from Lemma 1. The third property is given by

$$\begin{aligned} P_n^{\text{adv}}(A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})) &= \sum_{s^n \in A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})} P_n^{\text{adv}}(s^n) \\ &\quad \text{by property 1} \\ &\le \sum_{s^n \in A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})} P_n^*(s^n) 2^{-(n-1)(D_K(\theta^*,\theta^{\text{adv}})-\delta)} \\ &= 2^{-(n-1)(D_K(\theta^*,\theta^{\text{adv}})-\delta)} P_n^*(A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})) \\ &\le 2^{-(n-1)(D_K(\theta^*,\theta^{\text{adv}})-\delta)}. \end{aligned}$$

The last property has a similar proof:

$$
\begin{aligned}
P_n^{\text{adv}}(A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})) &= \sum_{s^n \in A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})} P_n^{\text{adv}}(s^n) \\
&\quad \text{by property 1} \\
&\geq \sum_{s^n \in A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})} P_n^*(s^n) 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})+\delta)} \\
&= 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})+\delta)} P_n^*(A_\delta^{(n)}(P_n^*, P_n^{\text{adv}})) \\
&\quad \text{by property 2} \\
&> (1-\delta) 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})\delta)}.
\end{aligned}
$$

□

Lemma 3. *Let $B_n \subset \mathcal{X}^n$ be any set of state sequences such that $P_n^*(B_n) > 1-\delta$. Let P_n^{adv} be any other distribution such that $D_K(\theta^*, \theta^{adv}) < \infty$. Then*

$$
P_n^{adv}(B_n) > (1-2\delta) 2^{-(n-1)(D_K(\theta^*, \theta^{adv})+\delta)}
$$

Proof. Denote $A_\delta^{(n)}(P_n^{\text{adv}}, P_2)$ by A_n. Combining $P_n^*(B_n) > 1-\delta$ and $P_n^*(A_n) > 1-\delta$ from (2) gives $P_n^*(A_n \cap B_n) > 1-2\delta$. Thus,

$$
\begin{aligned}
P_n^{\text{adv}}(B_n) &\geq P_n^{\text{adv}}(A_n \cap B_n) \\
&= \sum_{s^n \in A_n \cap B_n} P_n^{\text{adv}}(s^n) \\
&\geq \sum_{s^n \in A_n \cap B_n} P_n^*(s^n) 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})+\delta)} \\
&= 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})+\delta)} \sum_{s^n \in A_n \cap B_n} P_n^*(s^n) \\
&= 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})+\delta)} P_n^*(A_n \cap B_n) \\
&\geq 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})+\delta)} (1-2\delta)
\end{aligned}
$$

□

Proof of Theorem 4. Choose a sequence of sets $\mathcal{A}_n = A_\varepsilon^{(n)}(P_n^*, P_n^{\text{adv}})$, where $0 < \varepsilon \leq \delta$. Lemma 2 states that $P_n^*(\mathcal{A}_n^c) < \varepsilon \leq \delta$ for n large enough, meaning that the relative entropy typical set satisfies the bound on the type I error. The same lemma also shows that

$$
\lim_{n\to\infty} \frac{1}{n-1} \log P_n^{\text{adv}}(\mathcal{A}_n) < -D_K(\theta^*, \theta^{\text{adv}}) + \varepsilon,
$$

providing a lower bound for the error exponent for error type II. Now we show that no other sequence of sets can do better than this bound. Consider any sequence of sets $\mathcal{B}_n$ with $P_n^*(\mathcal{B}_n) > 1-\varepsilon \geq 1-\delta$, then Lemma 3 gives $P_n^{\text{adv}}(\mathcal{B}_n) > (1-2\varepsilon) 2^{-(n-1)(D_K(\theta^*, \theta^{\text{adv}})+\varepsilon)}$, and

$$
\lim_{n\to\infty} \frac{1}{n-1} \log P_n^{\text{adv}}(\mathcal{B}_n) > \lim_{n\to\infty} \frac{1}{n-1} \log(1-2\varepsilon) - (D_K(\theta^*, \theta^{\text{adv}}) + \varepsilon) = -D_K(\theta^*, \theta^{\text{adv}}) - \varepsilon.
$$

Combining these two inequalities and letting $\varepsilon \to 0$ proves (7). □

Proof of Theorem 5. By the upper bound of Theorem 2,

$$
\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log \alpha_n &= \lim_{n\to\infty} \frac{1}{n} \log P_n^*(\theta_{s^n} \in \{\nu : D_K(\nu, \theta^*) \geq \eta\}) \\
&= - \inf_{\{\nu : D_K(\nu, \theta^*) \geq \eta\} \cap \mathcal{M}} D_K(\nu, \theta^*) \\
&= -\eta,
\end{aligned}
$$

as long as meaning that constraint (2) is satisfied. We now analyze the probability of error type II under test S^* for the alternative hypothesis for any distribution P^{adv}. By the same upper bound as before we get

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\beta_n &= \lim_{n\to\infty}\frac{1}{n}\log P_n^{\text{adv}}(\theta_{s^n}\in\{\nu : D_K(\nu,\theta^*)<\eta\})\\ &= -\inf_{\{\nu:D_K(\nu,\theta^*)<\eta\}\in\mathcal{M}} D_K(\theta,\theta^\mu)\\ &\triangleq -J(\eta).\end{aligned}\tag{9}$$

where we let the infimum over an empty set be ∞. We then compare these error exponents with those of a test $\mathcal{S}$ determined by the binary function $\mathcal{F}:\Delta(\mathcal{S}^2)\times\mathbb{Z}^+\mapsto\{0,1\}$ which does not follow the same type of threshold. Suppose that for some $\delta>0$ and for some n, there exists a $\nu\in\mathcal{M}$ such that $D_K(\nu,\theta^*)\leq\eta-\delta$ while $\mathcal{F}(\nu,n)=1$, meaning that the test rejects the null hypothesis even though the distribution of η is "closer" to the null hypothesis's distribution in the sense of D_K than the threshold. Then, applying Theorem 2, with $\Gamma=\{\nu\}$ we obtain

$$\lim_{n\to\infty}\frac{1}{n}\log\alpha_n=\lim_{n\to\infty}\frac{1}{n}\log P_n^*\{\theta_{s^n}=\nu\}=-D_K(\nu,\theta^*)$$

However, to satisfy constraint (2) we require that $\eta-\delta\geq\eta$, which contradicts $\delta>0$. We can conclude that then, for every $\delta>0$ and for all n large enough,

$$\mathcal{M}\cap\{\nu : D_K(\nu,\theta^*)\leq\eta-\delta\}\subseteq\mathcal{M}\cap\{\nu:\mathcal{S}(\nu,n)=0\}$$

And so, for every $\delta>0$,

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\beta_n &\geq \liminf_{n\to\infty}\frac{1}{n}\log\beta_n =\\ &= \liminf_{n\to\infty}\frac{1}{n}\log P_n^{\text{adv}}(\theta_{s^n}\in\{\nu:\mathcal{S}(\nu,n)=0\}\cap\mathcal{M})\\ &\geq \liminf_{n\to\infty}\frac{1}{n}\log P_n^{\text{adv}}(\theta_{s^n}\in\{\nu:D_K(\nu,\theta^*)<\eta-\delta\}\cap\mathcal{M})\\ &\geq -\inf_{\{\nu:D_K(\nu,\theta^*)<\eta-\delta\}\cap\mathcal{M}} D_K(\nu,\theta^{\text{adv}})\end{aligned}$$

Thus, we can select the δ providing the tightest bound, which combined with Theorem 2 gives

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\beta_n &\geq -\inf_{\delta>0}\inf_{\{\nu:D_K(\nu,\theta^*)<\eta-\delta\}\cap\mathcal{M}} D_K(\nu,\theta^{\text{adv}})\\ &\geq -\inf_{\delta>0} J(\eta-\delta)\\ &= -J(\eta)\end{aligned}\tag{10}$$

which states that (10), the exponent achieved by test $\mathcal{F}^*$ is the minimum bound among all possible tests. □