Sensor Scheduling in Intrusion Detection Games with Uncertain Payoffs

Jayanth Bhargav   JBHARGAV@PURDUE.EDU

Mahsa Ghasemi   MAHSA@PURDUE.EDU

Shreyas Sundaram   SUNDARA2@PURDUE.EDU

Elmore Family School of Electrical and Computer Engineering, Purdue University, IN 47907 USA

Abstract

We study the problem of sensor scheduling for an intrusion detection task. We model this as a 2-player zero-sum game over a graph, where the defender (Player 1) seeks to identify the optimal strategy for scheduling sensor orientations to minimize the probability of missed detection at minimal cost, while the intruder (Player 2) aims to identify the optimal path selection strategy to maximize missed detection probability at minimal cost. The defender's strategy space grows exponentially with the number of sensors, making direct computation of the Nash Equilibrium (NE) strategies computationally expensive. To tackle this, we propose a distributed variant of the Weighted Majority algorithm that exploits the structure of the game's payoff matrix, enabling efficient computation of the NE strategies with provable convergence guarantees. Next, we consider a more challenging scenario where the defender lacks knowledge of the true sensor models and, consequently, the game's payoff matrix. For this setting, we develop online learning algorithms that leverage bandit feedback from sensors to estimate the NE strategies. By building on existing results from perturbation theory and online learning in matrix games, we derive high-probability order-optimal regret bounds for our algorithms. Finally, through simulations, we demonstrate the empirical performance of our proposed algorithms in both known and unknown payoff scenarios.

Keywords: Online Learning, Matrix Games, Sensor Scheduling, Finite-Sample Convergence

1. Introduction

In the modern landscape of security and surveillance, efficient deployment and scheduling of sensors is critical in ensuring comprehensive monitoring and intrusion detection (Guvensan and Yavuz (2011); Murray et al. (2007)). In managing sensing resources, the challenge involves not only their placement, but also dynamically scheduling their orientations to maximize detection efficiency while minimizing resource usage and operational costs (Osais et al.). Traditionally, sensor scheduling has been approached through various optimization techniques, yet these methods often fall short in adversarial settings where an intelligent intruder (non-oblivious) can actively attempt to evade detection. In such scenarios, game theory offers a robust framework for modeling and solving the intricate interactions between the defending system and the intruding adversary (Pirani et al. (2021); An et al. (2017)). Furthermore, the defending system may not have accurate knowledge of sensor performance due to the inherent complexity of modeling certain sensors, such as cameras, or the lack of prior knowledge about detection probabilities. These uncertainties lead to discrepancies between the simulated and real-world game dynamics, requiring the defending system to learn the true game online, and adaptively refine strategies using feedback from sensors to ensure effective detection and maintain performance despite the model uncertainties.

1.1. Related Work

In this section, we review existing works on security games, methods for solving large-scale games, and online learning in games, highlighting how these approaches relate to our contributions.

Security Games: Many works in the literature have modeled security resource allocation tasks using the Stackelberg game framework (Sinha et al. (2018); Korzhyk et al. (2010)). In Paruchuri et al. (2008), the authors model a security patrolling task to protect airports from attackers as a Stackelberg game, where the defender commits to a mixed strategy over an exponentially sized set of pure strategies for allocating patrolling resources (e.g., guards). The attacker observes the defender's mixed-strategy and best-responds with a pure attack strategy. However, we consider the setting where both the players commit to (randomized) mixed-strategies, and operate under concurrent and decentralized decision-making conditions. Since neither the defender nor the intruder can anticipate or respond to the other's exact strategy in real-time, the Stackelberg and Nash equilibria are equivalent in our game setting (Korzhyk et al. (2011)).

Solving Large Games: A line of work focuses on developing efficient techniques for approximately solving games with large strategy spaces (Ganzfried and Sandholm (2015); Lipton et al. (2003); Sandholm (2015)). Li and Wellman (2020) introduce an iterative structure-learning approach to search for approximate solutions of many-player games. In Li et al. (2023), the authors study team-adversary games with combinatorial action spaces. They propose a counterfactual regret minimization framework (also studied in Brown et al. (2019)) for learning individual strategies for each player using regret matching. Along the lines of these works, we propose a distributed variant of the well-known Weighted Majority algorithm (proposed in Freund and Schapire (1997)) which leverages the structure in the game's payoff function, to efficiently and iteratively estimate the Nash equilibrium for a given instance of the game, and establish guarantees for convergence to NE.

Online Learning in Games: In Fiez et al. (2021), the authors study online learning in periodic zero-sum games and establish guarantees for convergence to equilibrium strategies. The authors of Bailey and Piliouras (2019) study online learning in 2×2 zero-sum games and provide convergence guarantees for gradient descent dynamics. However, for a general class of zero-sum games, the dynamics of the equilibrium strategies generated by the learning process can be very intricate and instances may fail to converge (Andrade et al. (2021); Cheung and Piliouras (2019)). Two closest papers to our work on learning in matrix games with bandit feedback are: O'Donoghue et al. (2021) - where the authors propose a UCB style algorithm for no-regret learning in matrix games; and Chen et al. (2024) - where the authors present an adaptation of online mirror descent algorithm with last-iterate convergence rates. The key distinction of our work is the application to the intrusion detection setting, where we exploit the combinatorial structure of the game matrix arising from sensor and path configurations, and leverage its structural properties to establish tighter regret bounds, which guarantee faster convergence to NE strategies.

1.2. Contributions

We begin by modeling the sensor scheduling task as a zero-sum matrix game. First, we propose a distributed variant of the Weighted Majority algorithm that leverages the structure of the game's payoff matrix to efficiently compute Nash Equilibrium (NE) strategies, even in exponentially large strategy spaces, with provable convergence guarantees. Second, recognizing that the defender may lack accurate knowledge of the true sensor models and hence the game's payoff matrix, we develop online learning algorithms that rely on bandit feedback from the sensors to adaptively estimate

the game matrix and refine the defender's sensor scheduling strategies. By building on results from perturbation theory and online learning in matrix games, we derive high-probability order-optimal regret bounds for these algorithms. Finally, we demonstrate the empirical performance of our methods through numerical simulations, for both known and unknown payoff scenarios.

2. Problem Formulation

We model the game environment as a graph $G = (V, E)$, where the set of nodes V denotes the regions of an environment and the set of edges E denotes the possible transitions for the intruder between the regions. Let $\mathcal{S} = \{s_1, \ldots, s_p\}$, with $|\mathcal{S}| = p$, denote the set of sensors deployed on a subset of nodes in the graph. Each sensor has d possible orientations, $\Theta = \{\theta_1, \ldots, \theta_d\}$, with $|\Theta| = d$. Let $c_k^q \in \mathbb{R}_{\geq 0}$ be the cost of orienting the sensor $s_q \in \mathcal{S}$ in the direction $\theta_k \in \Theta$. Each sensor covers a subset of nodes in the graph depending on its orientation, which may not be restricted to its neighboring nodes. Let $I = \{i_1, i_2, \ldots, i_m\}, m = |I|$, denote the set of joint strategies of the defender, i.e., a set of all possible sensor orientations, and let $\{c_1, c_2, \ldots, c_m\}$ be the costs associated with each joint strategy. The cost c_i of a joint strategy $i \in \mathcal{I}$ is the sum of the individual costs associated with each sensor orientation in that strategy, i.e., $c_i = \sum_{q=1}^p c_{k_{q,i}}^q$, where $k_{q,i}$ is the orientation of sensor s_q in the joint strategy i. We note that the number of joint strategies is $m = |\Theta|^{|\mathcal{S}|}$, which is exponential in the number of sensors. We assume that the intruder starts at a source node S and aims to enter a protected facility at a terminal node T in the graph. Let $J = \{j_1, j_2, \ldots, j_n\}, n = |J|$, denote the set of pure strategies of the intruder, i.e., the set of paths in the graph that the intruder can take to reach the node T from node S and let $\{r_1, r_2, \ldots, r_n\}, r_j \in \mathbb{R}_{\geq 0}$, be the associated costs of the respective paths. Without loss of generality, we do not consider cycles, as it will only hurt the intruder, and thus n is finite. Let X and Y denote the set of mixed-strategies for the defender and intruder, respectively. We have the following:

$$X = \left\{ x \in \mathbb{R}^m; \sum_{i=1}^m x_i = 1, x \geq 0 \right\}; Y = \left\{ y \in \mathbb{R}^n; \sum_{i=1}^n y_i = 1, y \geq 0 \right\}. \tag{1}$$

We first consider the case of homogeneous sensors, i.e., all sensors are identical and have the same probability of detection p_{detect}.[1] Specifically, if the intruder traverses through a node which is covered by a sensor, it will be detected by that sensor with a probability p_{detect}, and will go undetected by that sensor with probability $(1 - p_{detect})$. We now make the following assumption.

Assumption 1: The sensors are imperfect with $p_{detect} \in [p_{min}, p_{max}]$, where $0 < p_{min} < p_{max} < 1$. Furthermore, detection events from one sensor are independent of that from other sensors.

The goal of the defender is to minimize the probability of not detecting the intruder. If a certain node is covered by multiple sensors and the intruder has traversed that node, then the overall probability of missed detection at that node is the product of the probability of missed detection of all the sensors covering that node. For a sensor orientation strategy $i \in I$ and a path $j \in J$, let V_{ij} be the number of nodes that are covered by the sensor orientation strategy $i \in I$ and contained in the intruder's path $j \in J$, with repetition. The overall probability of missed detection for a pure strategy pair (i, j) is given by:

$$p_{miss}(i, j) = (1 - p_{detect})^{V_{ij}}. \tag{2}$$

---

1. We later discuss the setting with heterogeneous sensors in Section 4.1.

We consider the log-likelihood representation for the probability of missed detection, given by

$$\log(p_{miss}(i,j)) = V_{ij}\log(1 - p_{detect}). \qquad (3)$$

Let $A, B \in \mathbb{R}^{m \times n}$ denote the payoff matrices for the defender and intruder, respectively. We will refer to the defender as the *x-player* and the intruder as the *y-player*. The x-player plays $x \in X$, the y-player plays $y \in Y$, and they receive payoffs $x^\top A y$ and, $x^\top B y$ respectively. The defender's payoff value for a pure strategy pair (i, j) is given by $A_{ij} = \log(p_{miss}(i,j)) + c_i$. The defender aims to find the mixed strategy x^* that minimizes the expected value of its payoff, given the mixed strategy y of the intruder, i.e., the defender aims to solve the following optimization problem:

$$x^* = \arg\min_{x \in X} x^\top A y. \qquad (4)$$

The intruder's payoff value for a pure strategy pair (i, j) is given by $B_{ij} = -\log(p_{miss}(i,j)) + r_j$. Given the mixed strategy x of the defender, the intruder aims to solve the following problem:

$$y^* = \arg\min_{y \in Y} x^\top B y. \qquad (5)$$

We will denote this bimatrix game by the following tuple: $\mathcal{G} = \{A, B\}$. Note that $\mathcal{G}$ is a non-zero sum game, since $A + B \neq 0$ in general. We will now transform this bimatrix game into an equivalent zero-sum game and establish that these two games have the same Nash equilibrium. A bimatrix game $\mathcal{G} := \{A, B\}$ is zero-sum if $A + B = 0$, and thus can be specified by a single matrix A (since $B = -A$). The solution to a zero-sum game, i.e., the Nash equilibrium strategies (x^*, y^*), can obtained by solving the nested min-max optimization problem below:

$$(x^*, y^*) = \arg\min_{x \in X} \arg\max_{y \in Y} x^\top A y. \qquad (6)$$

Definition 1 (Nash Equilibrium) *A pair of mixed-strategies (x^*, y^*) is a Nash Equilibrium (NE) of the zero-sum game A if $x^{*\top} A y^* \leq x^\top A y^*$ and $x^{*\top} A y^* \geq x^{*\top} A y$, $\forall x \in X, \forall y \in Y$.*

Consider the following equivalent game $\mathcal{G}' := \{A', B'\}$, where matrices A' and B' are such that

$$A'_{ij} = A_{ij} - r_j; \quad B'_{ij} = B_{ij} - c_i. \qquad (7)$$

We have the following result characterizing the equilibrium property of $\mathcal{G}'$. All detailed proofs supporting our theoretical results are deferred to Appendix B.

Proposition 2 *The game $\mathcal{G}' := \{A', B'\}$ is zero-sum and has the same set of Nash Equilibria as $\mathcal{G}$.*

We consider the setting in which the defender has knowledge of c_i's and r_j's for all $i \in I$ and $j \in J$. However, the defender may not know the sensor detection probabilities p_{detect}, consequently the game payoff matrix A, accurately. Proposition 2 implies that an ϵ-NE obtained by a no-regret algorithm applied to the zero-sum game (A', B') is also an ϵ-NE for the original game (A, B). With a slight abuse of notation, we will denote the modified game (which is zero-sum) by $\mathcal{G}$, which can be completely specified by the payoff matrix A, where $A_{ij} = \log(p_{miss}(i,j)) + c_i - r_j$.

Remark 3 *We assume that the entries of A are bounded in $[0, 1]$. If this is not the case, we normalize A using its maximum and minimum values, $A_{min} = \min_{i,j} A_{ij}$ and $A_{max} = \max_{i,j} A_{ij}$, respectively. This normalization preserves the Nash equilibrium strategies, as scaling and shifting the payoffs affect only the game's value, not the equilibrium strategies themselves. Additionally, we note that a ϵ-NE of the normalized game is an ϵ'-NE of the un-normalized game, where $\epsilon' = (A_{max} - A_{min})\epsilon$. Therefore, it is sufficient to restrict all analysis to the normalized zero-sum game, as it fully captures the essential properties and results of the original game.*

3. Distributed Weighted Majority Algorithm

In this section, we propose a fast and scalable algorithm which leverages the structure of the game matrix A for efficiently solving for the Nash equilibrium strategies. For a moderate number of strategies, the min-max optimization problem in (6) can be solved to compute the Nash Equilibrium using a Linear Programming (LP) formulation. However, for the sensor scheduling problem considered in this paper, the defender has an exponentially sized set of scheduling strategies. For example, if $d = |\Theta| = 4$ (i.e., each sensor has 4 possible orientations) and $|\mathcal{S}| = 10$ (i.e., there are 10 sensors), then $|I| = 4^{10} = 1,048,576$. Solving for the exact Nash equilibrium can become computationally intractable in such large-strategy space games. To this end, we present the Distributed Weighted Majority (DWM) algorithm (Algorithm 1), a variant of the *Weighted Majority (WM)* algorithm originally developed by Freund and Schapire (1997). The DWM algorithm leverages the structure of the game to iteratively estimate the Nash equilibrium strategies using a multiplicative weight update rule and enjoys theoretical guarantees for convergence to NE. We present the algorithm for a general case of heterogeneous sensors, where each sensor $s_q \in \mathcal{S}$ has a probability of detection $p_{detect,q}$.

For sensors $s_q \in \mathcal{S}$, we define the set of sub-game payoff matrices $\mathcal{A} = \{A^1, A^2, \ldots, A^p\}$, $A^q \in \mathbb{R}^{d \times n}$, where

$$A_{kj}^q = V_{kj}^q \times \log(1 - p_{detect,q}) + c_k^q, \tag{8}$$

where V_{kj}^q contains all the nodes that are covered by the sensor s_q oriented in the direction θ_k and contained in the intruder path j. At each time step t and for each sensor $s_q \in \mathcal{S}$, we maintain a set of weights $\sigma_t^q \in \mathbb{R}^d$, where $d = |\Theta|$. The joint sensor scheduling strategy $i \in I$ can be denoted as $i = [k_{1,i}, k_{2,i}, \ldots, k_{p,i}]$, where $k_{q,i} \in \{\theta_1, \ldots, \theta_d\}$ is the orientation of the sensor $s_q \in \mathcal{S}$ in the joint strategy $i \in I$. We update the weights for each joint strategy $i \in I$ as follows:

$$w_t(i) = \prod_{q=1}^{p} \sigma_t^q(k_{q,i}), \tag{9}$$

where $\sigma_t^q(k_{q,i})$ is the weight assigned to the orientation θ_k for the sensor s_q in the joint strategy $i \in I$. At time step t, given the intruder's mixed strategy $y_t \in Y$, we define the loss per sensor $(s_q \in \mathcal{S})$ for each individual strategy $\theta_k \in \Theta$ as $\ell_t^q(\theta_k)$, given by

$$\ell_t^q(\theta_k) = \sum_{j \in J} A_{kj}^q y_t(j) - \frac{1}{p} \sum_{j \in J} r_j y_t(j). \tag{10}$$

Algorithm 1 provides mixed strategies for the game $\mathcal{G}$ with the following theoretical guarantees.

Theorem 4 *Let $\{x_1, \ldots, x_T\}$ and $\{y_1, \ldots, y_T\}$ be the mixed strategies returned by Algorithm 1 after T iterations with $\beta = \left(1 + \sqrt{\frac{2|\mathcal{S}|ln|\Theta|}{\tilde{L}}}\right)^{-1}$, where $\tilde{L} \geq \min_{i \in I} \sum_{t=1}^{T} \sum_{q=1}^{p} \ell_t^q(\theta_k)$. We let $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$ and $\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$. Then, $\bar{x}$ and $\bar{y}$ are valid mixed-strategies for the two players, respectively. The pair $(\bar{x}, \bar{y})$ approximates the game's Nash Equilibrium value within ϵ_T, given by*

$$\epsilon_T = \frac{\sqrt{2\tilde{L}|\mathcal{S}|ln|\Theta|}}{T} + \frac{|\mathcal{S}|ln|\Theta|}{T}. \tag{12}$$

Algorithm 1 Distributed Weighted Majority (DWM) Algorithm

Data: Intrusion Detection Game $\mathcal{G}$, Parameter $\beta \in (0,1)$, Number of Iterations T
Result: Mixed-strategies: $\{x_1, \ldots, x_T\}, \{y_1, \ldots, y_T\}$
Initialize: $\sigma_0^q(\theta_k) = 1/|\Theta|, k = 1, \ldots, |\Theta|$; $w_0(i) = 1/|I|, i = 1, \ldots, m$

for $t = 1, \ldots, T$ **do**

    Compute defender's mixed-strategy $x_t(i) = \frac{w_t(i)}{\sum_{i' \in I} w_t(i')}, \forall i \in I$

    Estimate intruder's mixed-strategy $y_t = \max_{y \in Y} x_t^\top A y$

    Compute per sensor loss $\ell_t^q(\cdot)$ as in Equation (10)

    Update weights for each sensor

$$\sigma_{t+1}^q(k_{q,i}) = \sigma_t^q(k_{q,i})\beta^{\ell_t^q(\theta_k)}; \quad k = 1, \ldots, |\Theta| \tag{11}$$

    Update joint strategy weights: $w_{t+1}(i) = \prod_{q=1}^p \sigma_{t+1}^q(k_{q,i}), \forall i \in I$

end

Remark 5 *The DWM algorithm exploits the additive structure of the game's payoff function, achieving a substantial computational advantage by performing only $\mathcal{O}(|\mathcal{S}||\Theta|)$ multiplicative updates in parallel (i.e, $\mathcal{O}(|\Theta|)$ updates for each sensor $s \in \mathcal{S}$ in parallel), to estimate the NE strategy for the defender. In contrast, directly applying the WM algorithm requires $\mathcal{O}(|\Theta|^{|\mathcal{S}|})$ updates, which cannot be parallelized, making DWM far more efficient for large-scale instances.*

4. Online Learning and Adaptation for Unknown Sensor Models

In this section, we extend our study of the sensor scheduling problem to scenarios where the defender lacks knowledge of the true sensor models and, consequently, the game's payoff matrix. We present online learning algorithms with performance guarantees for both homogenous and heterogeneous sensor settings, for refining the defender's strategies based on bandit feedback.

Since the defender does not have accurate knowledge of the game's payoff, it maintains an estimate of the game matrix, which is updated iteratively using feedback from sensor observations. At each time step t, the defender computes a mixed strategy $x_t \in X$ for scheduling sensors, based on its current estimate of the game matrix. The defender and intruder play pure strategies (i_t, j_t), sampled from their mixed-strategy distributions (x_t, y_t) respectively. The defender receives feedback from the sensors monitoring the intruder's path, indicating whether or not the intruder was detected. We will first consider the setting with homogenous sensors, where the game matrix A can be parametrized by a single (unknown) Bernoulli distribution with parameter p_{detect}. We define $\mathcal{F}_t = ((i_1, j_1), \ldots, (i_t, j_t))$ to be the sequence of observations available to the player at round t, and as a shorthand we shall use the notation $\mathbb{E}^t(\cdot) = \mathbb{E}(\cdot \mid \mathcal{F}_t)$. Two aspects of this problem distinguish it from other setups considered in the literature (Cheung and Piliouras (2019); Andrade et al. (2021); O'Donoghue et al. (2021)). Firstly, the defender receives the (pure-strategy) action played by the intruder (j_t) as feedback (as opposed to the mixed strategy y_t), and secondly, the intruder does not have accurate knowledge of the game payoff function/matrix. In particular, our setting is different from the one considered in O'Donoghue et al. (2021), where the players receive a noisy feedback of the payoff, and the noise process is assumed to be 1-sub-Gaussian with a zero mean. At each time

step t, the defender updates its estimate of $\hat{p}^t_{detect}$ of the true parameter p_{detect} and estimates the game matrix $\hat{A}_t$ (see Algorithm 2). Another key distinction from O'Donoghue et al. (2021) is that our problem setting under homogenous sensors does not need complete exploration of the strategy spaces. We introduce a simpler algorithm that eliminates the need for Upper Confidence Bound (UCB) estimates for each entry of the A matrix, as the exploration-driven optimism framework is not required in this scenario. Consequently, we establish tighter regret bounds that do not depend on the strategy spaces, i.e., ($\tilde{\mathcal{O}}(\sqrt{T})$ instead of $\tilde{\mathcal{O}}(\sqrt{|I||J|T})$).

Algorithm 2 Online Learning and Adaptation of Nash Equillbirum Strategies

---

Feedback samples: $\mathcal{H} = \emptyset$
for $t = 1, \ldots, T$ **do**
    Compute sample average $\hat{p}^t_{detect}$ using samples in $\mathcal{H}$
    Clipping estimator: if $(\hat{p}^t_{detect} > p_{max}) : \hat{p}^t_{detect} = p_{max}$; if $(\hat{p}^t_{detect} < p_{min}) : \hat{p}^t_{detect} = p_{min}$
    Solve the estimated game $\hat{A}_t$, where $\hat{A}_t(i, j) = V_{ij} \log(1 - \hat{p}^t_{detect}) + c_i - r_j$:

$$(x^*_t, y^*_t) = \arg\min_{x \in X} \arg\max_{y \in Y} x^\top \hat{A}_t y \tag{13}$$

    Play pure strategy $i_t \sim x^*_t$ and observe intruder's pure strategy $j_t \sim y_t$
    Collect V_{i_t,j_t} samples from sensor feedback into $\mathcal{H}_t$: $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_t$
end

In Algorithm 2, we use a clipped estimator for the Bernoulli parameter $\hat{p}^t_{detect}$, which keeps the estimate within a feasible range $[p_{min}, p_{max}]$; this ensures stability in small sample sizes or noisy data without affecting long-term convergence to the true parameter p_{detect}, since as t grows, the sample mean naturally converges to p_{detect}, making clipping increasingly unnecessary and preserving consistency. Furthermore, we note that, in each iteration of Algorithm 2, the estimated game (as in (13)) can be efficiently solved for NE strategies using Algorithm 1.

We will perform analysis from the perspective of a single player, i.e., the defender, who does not control the actions of the opponent (i.e., the intruder). Let $(x^\dagger, y^\dagger)$ denote the Nash equilibrium strategies of the true game A, and $V^*_A = x^{\dagger\top} A y^\dagger$ be the value of the true game. In order to analyze the performance of the online learning algorithm, we consider the individual regret for the defender, i.e., the difference in expected cumulative payoffs relative to the value of the game $\mathcal{G}$, given by:

$$\mathcal{R}_T = \sum_{t=1}^{T} \left(x^\top_t A y_t - V^*_A \right). \tag{14}$$

We have the following result, characterizing the bound on the cumulative regret defined in (14).

Theorem 6 *Under Assumption 1, for a specified $\alpha \in (0, 1)$, we have the following: With probability at least $1 - \alpha$, Algorithm 2 applied to the intrusion detection game $\mathcal{G}$ has the following guarantees for the cumulative regret (as in (14)):*

1. $\mathcal{R}_T \leq \dfrac{5\sqrt{2}V_{max}}{(1 - p_{max})} \sqrt{T \log \dfrac{2T}{\alpha}}; y_t = y^\dagger;$

2. $\mathcal{R}_T \leq \dfrac{2\sqrt{2}V_{max}}{(1 - p_{max})} \sqrt{T \log \dfrac{2T}{\alpha}}; y_t \neq y^\dagger,$

where $V_{max} = \max\limits_{i,j} V_{ij}$ and p_{max} is the maximum probability of detection of the sensors.

Remark 7 *In Cai et al. (2024), the authors propose an online learning algorithm which is an adaptation of* EXP3-IX*, where they assume the existence of mixed-strategy Nash equilibrium with full support. Particularly, at each iteration t of their algorithm, they restrict the space of NE strategies to $\Omega_t = \{x \in X : x_a \geq \frac{1}{|I|t^2}, \forall a \in I\}$. However, many zero-sum games are not guaranteed to have a mixed-strategy Nash equilibrium with full support, since players typically can have some dominated strategies. In our algorithm, we do not restrict the space of NE strategies for the players. Additionally, our results presented in Theorem 6 provide tighter regret guarantees $\tilde{O}(T^{-\frac{1}{2}})$ compared to $\tilde{O}(T^{-\frac{1}{8}})$ presented in Cai et al. (2024).*

4.1. Intrusion Detection Game with Heterogeneous Sensors

In this section, we consider the setting with heterogeneous sensors, i.e., each sensor has a different probability of detection. In this setting, the game's payoff matrix A can be parameterized by a set of $|\mathcal{S}|$ Bernoulli distributions, with parameters $\{p_{detect,1}, \ldots, p_{detect,p}\}, p = |\mathcal{S}|$, each corresponding to a sensor $s_l \in \mathcal{S}, l = 1, \ldots, p$. The game's payoff matrix in this setting is given by:

$$A_{ij} = \sum_{l=1}^{p} V_{ij,l} \log(1 - p_{detect,l}) + c_i - r_j, \tag{15}$$

where $V_{ij,l}$ is the number of nodes covered by the sensor l in the joint strategy $i \in I$ and contained in the intruder's path $j \in J$, and c_i, r_j are sensor and path costs as specified before, respectively. Define $\max_{ij} V_{ij,l} = V_{max,l}$. Assumption 1 in the heterogeneous sensor setting implies the following: $0 < p_{min,l} \leq p_{detect,l} \leq p_{max,l} < 1$. We note that each entry of A may not depend on all $|\mathcal{S}|$ Bernoulli parameters, as $V_{ij,l}$ can be zero for a subset of sensors in $\mathcal{S}$. This is due to the fact that every intruder path $j \in J$ will intersect nodes covered in the graph by only a subset of the sensors. Thus, unlike the homogenous sensors setting, we will need exploration of strategies of both the players to ensure better estimates of these Bernoulli parameters.

We adapt the Upper Confidence Bound (UCB) algorithm for learning in matrix games proposed in O'Donoghue et al. (2021), by constructing the UCB estimate for the game matrix A based on the confidence bounds of the Bernoulli parameters. We assume that c_i's and r_j's are known.

Lemma 8 *For the game matrix A defined in (15), the following holds with probability at least $1 - \delta$,*

$$A_{ij} \leq \bar{A}_{ij}^t + \sum_{l=1}^{p} \frac{V_{max,l}}{1 - p_{max,l}} \sqrt{\frac{\log(\frac{2}{\delta})}{2(1 \vee n_{ij,l}^t)}} = \tilde{A}_{ij}^t, \tag{16}$$

where $\bar{A}_{ij}^t$ is A_{ij} (as in (15)) evaluated with sample averages $\bar{p}_{detect,l}^t$, $n_{ij,l}^t$ is the number of times the intruder chooses the path j which contains nodes covered by the sensor l in the joint orientation strategy i up to (but not including round t), and we use the notation $(1 \vee \cdot) = \max(1, \cdot)$.

We now have the following result that characterizes the bound on the regret (Equation (14)) when Algorithm 1 in O'Donoghue et al. (2021) is applied to the instance of the intrusion detection game with the UCB estimate $\tilde{A}_{ij}^t$ as defined in Equation (16).

Theorem 9 *Let Assumption 1 hold for all sensors. For $T \geq |\mathcal{S}||\Theta||J| \geq 2$ and $\delta = 1/(2T^2|\mathcal{S}||\Theta||J|)$. When Algorithm 1 in O'Donoghue et al. (2021) is applied to the instance of the intrusion detection game with the UCB estimate $\tilde{A}_{ij}^t$ (as in (16)), the regret (as in (14)) is bounded as:*

$$\mathcal{R}_T \leq 1 + |\mathcal{S}|\sqrt{2|\Theta||J|T\log(4T^2|\mathcal{S}||\Theta||J|)} = \tilde{O}(\sqrt{|\mathcal{S}|^2|\Theta||J|T}). \tag{17}$$

Remark 10 *Applying Algorithm 1 of O'Donoghue et al. (2021) directly would result in a regret bound that depends exponentially on the number of sensors, i.e., $\tilde{O}(\sqrt{|\Theta|^{|\mathcal{S}|}|J|T})$. We leverage the specific structure of the game to achieve significant improvements in the regret bound in Theorem 9, i.e., $\tilde{O}(\sqrt{|\mathcal{S}|^2|\Theta||J|T})$, which will lead to faster convergence to the NE strategy for the defender. Additionally, the parameter δ, which governs the balance between exploration, regret bounds, and high-probability guarantees depend on $|\Theta||\mathcal{S}||J|$ instead of $|\Theta|^{|\mathcal{S}|}|J|$. We present the adapted UCB algorithm (Algorithm 3) and additional discussions in Appendix A.*

5. Experiments

In this section, we evaluate the performance of the proposed algorithms through simulations.

5.1. Empirical Evaluation of Online Learning Algorithm

We consider the intrusion detection game under homogenous sensors, and evaluate the performance of Algorithm 2. We set the true Bernoulli parameter to $p_{detect} = 0.8$ and generate several instances by uniform random sampling of $V_{ij} \in [10, 20], \forall(i,j) \in I \times J$, with $I = \{1, ..., 10\}$ and $J = \{1, ..., 20\}$. We solve for the NE strategies $(x^\dagger, y^\dagger)$ using Algorithm 1 with $\epsilon_T = 10^{-6}$ and estimate the optimal game value $V_A^* = x^{\dagger\top}Ay^\dagger$. We run Algorithm 2 under both settings: (i) $y_t = y^\dagger$; and (ii) y_t being a random point on the probability simplex Y. We plot the cumulative regrets $\mathcal{R}_T$ and $\mathcal{R}_T^\dagger$ by setting $T = 1000$ in Figure 1(a). We observe the sublinear growth of the regret, demonstrating that the learning algorithm is effectively adapting and improving its performance over iterations.
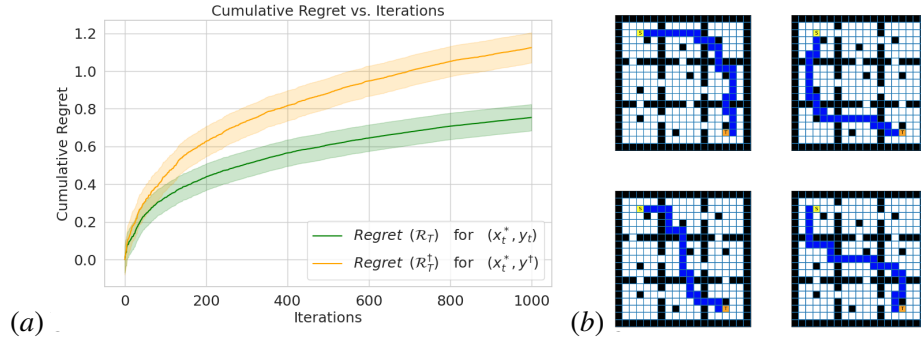

Figure 1: (a) Cumulative regret plots for Algorithm 2; (b) Sample paths (pure strategies) of the intruder in the 9-room grid world

5.2. Empirical Evaluation of Distributed Weighted Majority Algorithm

In this section, we present numerical evaluations to show the computational efficiency of DWM algorithm. We consider the nine-room grid world environment as shown in Figure 2. This is a 19×19 cell environment, which can be represented by a graph. Sensors are installed on the cells (resp. nodes in the graph) marked in red (and numbered), and all cells marked in black are obstacle

cells. We set $|\Theta| = 4$, i.e., each sensor has 4 possible orientations: { East, North, West, South}. We vary the number of sensors $|\mathcal{S}| \in \{2, 3, 4, 5, 6, 7, 8\}$ to generate multiple instances of the game. The defender's strategy space, given by $4^{|\mathcal{S}|}$, scales exponentially with the number of sensors in $\mathcal{S}$. The intruder's strategy space is set to $|J| = 50$, i.e., there are 50 possible paths for the intruder (e.g., see Figure 1(b)) from node S (marked in yellow) to node T (marked in orange) in the grid. All simulations are run on a 2.6 GHz 6-Core Intel Core i7 machine. We set $\epsilon_T = 0.001$, i.e., we wish to compute a ϵ-Nash Equilibrium for the game with $\epsilon = 0.001$. For each instance of the game, we compute the number of iterations T based on $|I|$ and ϵ_T using Equation (12) by setting $\tilde{L} = T$. We solve for the NE

Figure 2: Sample sensor scheduling strategy (pure strategy) of the defender

strategies using three approaches: (i) a standard bi-level Linear Program (LP) (using Gurobi Optimization Solver), (ii) WM algorithm (as in Freund and Schapire (1997)) and (iii) DWM algorithm (Algorithm 1) for T iterations, and record the running times in Table 1. We observe that the WM and DWM algorithms solve the game to within a 0.001 approximation of the game value much faster than the LP. We observe that as the strategy space increases, the DWM algorithm outperforms the WM algorithm, with a significant reduction in running time. Due to limitations on the problem sizes in Gurobi Solver (academic license), we cannot solve LPs for problem instances with strategy spaces more than 2000.
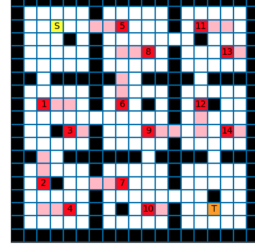
$|\mathcal{S}|$	Strategy Spaces	Linear Program	Weighted Majority	Distributed Weighted Majority
2	16×50	3.22	2.73	2.68
3	64×50	7.90	3.04	3.65
4	256×50	24.88	6.09	4.28
5	1024×50	75.72	12.15	6.13
6	4096×50	-	35.51	7.88
7	$16,384 \times 50$	-	107.65	9.92
8	$65,536 \times 50$	-	365.43	11.83

Table 1: Comparison of running times (in seconds) of Distributed Weighted Majority Algorithm (Algorithm 1) with standard Weighted Majority algorithm and Linear Program

6. Conclusions

In this paper, we address the problem of sensor scheduling for intrusion detection by formulating it as a zero-sum game over a graph. We proposed a distributed algorithm, which leverages the structure of the game's payoff, for efficiently computing equilibrium strategies for instances with exponentially large strategy spaces. We further extend our analysis to scenarios with unknown sensor models, introducing online learning algorithms with high-probability regret guarantees for learning equilibrium strategies under bandit feedback. Our work leverages the structure in the combinatorial action space of the players to achieve significant computational efficiency and scalability. Empirical evaluations demonstrate the practical effectiveness of our approach, showing near-optimal and scalable performance over several instances of the game. In future work, we aim to extend the framework to dynamic game environments, where strategies can adapt to changes in sensor performance, including degradation or reconfiguration, as well as stochastic variations in game payoffs.

Acknowledgments

References

Bo An, Milind Tambe, and Arunesh Sinha. Stackelberg security games (ssg) basics and application overview. *Improving Homeland Security Decisions*, 2:485, 2017.

Gabriel P Andrade, Rafael Frongillo, and Georgios Piliouras. Learning in matrix games can be arbitrarily complex. In *Conference on Learning Theory*, pages 159–185. PMLR, 2021.

James Bailey and Georgios Piliouras. Fast and furious learning in zero-sum games: Vanishing regret with non-vanishing step sizes. *Advances in Neural Information Processing Systems*, 32, 2019.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International conference on machine learning*, pages 793–802. PMLR, 2019.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. Last-iterate convergence of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2409.01447*, 2024.

Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory*, pages 807–834. PMLR, 2019.

Tanner Fiez, Ryann Sim, Stratis Skoulakis, Georgios Piliouras, and Lillian Ratliff. Online learning in periodic zero-sum games. *Advances in Neural Information Processing Systems*, 34:10313–10325, 2021.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Sam Ganzfried and Tuomas Sandholm. Endgame solving in large imperfect-information games. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

M Amac Guvensan and A Gokhan Yavuz. On coverage issues in directional sensor networks: A survey. *Ad Hoc Networks*, 9(7):1238–1255, 2011.

Ravi Kannan and Thorsten Theobald. Games of fixed rank: A hierarchy of bimatrix games. *Economic Theory*, 42:157–173, 2010.

Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. Complexity of computing optimal stackelberg strategies in security resource allocation games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 805–810, 2010.

Dmytro Korzhyk, Zhengyu Yin, Christopher Kiekintveld, Vincent Conitzer, and Milind Tambe. Stackelberg vs. nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *Journal of Artificial Intelligence Research*, 41:297–327, 2011.

Shuxin Li, Youzhi Zhang, Xinrun Wang, Wanqi Xue, and Bo An. Decision making in team-adversary games with combinatorial action space. *CAAI Artificial Intelligence Research*, 2, 2023.

Zun Li and Michael Wellman. Structure learning for approximate solution of many-player games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2119–2127, 2020.

Richard J Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, pages 36–41, 2003.

Richard J Lipton, E Markakis, and Aranyak Mehta. On stability properties of economic solution concepts. *Manuscript*, 2006.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Alan T Murray, Kamyoung Kim, James W Davis, Raghu Machiraju, and Richard Parent. Coverage optimization to support security monitoring. *Computers, Environment and Urban Systems*, 31 (2):133–147, 2007.

Michael Victor Nehme. *Two-person games for stochastic network interdiction: models, methods, and complexities*. The University of Texas at Austin, 2009.

Yahya Osais, Marc St-Hilaire, and F Richard Yu. On sensor placement for directional wireless sensor networks. In *2009 IEEE International Conference on Communications*, pages 1–5. IEEE.

Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback. In *Uncertainty in Artificial Intelligence*, pages 279–289. PMLR, 2021.

Praveen Paruchuri, Jonathan P Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Playing games for security: An efficient exact algorithm for solving bayesian stackelberg games. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 895–902, 2008.

Mohammad Pirani, Ehsan Nekouei, Henrik Sandberg, and Karl Henrik Johansson. A game-theoretic framework for security-aware sensor placement problem in networked control systems. *IEEE Transactions on Automatic Control*, 67(7):3699–3706, 2021.

Tuomas Sandholm. Abstraction for solving large incomplete-information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. Stackelberg security games: Looking beyond a decade of success. IJCAI, 2018.

Appendix A. Upper Confidence Bound Algorithm for Matrix Games

In this section, we present the adaptation of the UCB algorithm in O'Donoghue et al. (2021) for the intrusion detection game studied in this paper. The UCB algorithm and regret analysis in O'Donoghue et al. (2021) is presented with respect to the max-player, however, in our game setting, the defender is the min-player. From Remark 3, we have the following:

$$(x^*, y^*) = \min_{x \in X} \max_{y \in Y} x^\top A y = \max_{x \in X} \min_{y \in Y} x^\top (-A) y \tag{18}$$

$$V_A^* = -V_{-A}^*. \tag{19}$$

Therefore, without loss of generality, we will consider the game matrix to be $-A$ and note that the UCB estimate for analyzing the regret will remain the same, as the following relationship holds for the regret $\mathcal{R}(A, \text{UCB}, T)$ defined in O'Donoghue et al. (2021): $\mathcal{R}(A, \text{UCB}, T) = -\mathcal{R}_T$, where $\mathcal{R}_T$ is the regret measure defined in our paper in (14).

Algorithm 3 UCB Algorithm for Online Learning and Adaptation

Feedback samples: $\mathcal{H} = \emptyset$

for $t = 1, 2, \ldots, T$ **do**

    Compute sample average $\bar{p}_{detect,l}^t$ using samples in $\mathcal{H}$

    Clipping estimator: $\bar{p}_{detect,l}^t = \min(\bar{p}_{detect,l}^t, p_{max,l}); l = 1, \ldots p$

    Compute $\tilde{A}_{ij}^t = \bar{A}_{ij}^t + \sum_{l=1}^{p} \frac{V_{max,l}}{1 - p_{max,l}} \sqrt{\frac{\log(\frac{2}{\delta})}{2(1 \vee n_{ij,l}^t)}}$

    Solve the estimated game $\tilde{A}^t$:

$$(x_t^*, y_t^*) = \arg\max_{x \in X} \arg\min_{y \in Y} x^\top \tilde{A}^t y \tag{20}$$

    Play pure strategy $i_t \sim x_t^*$ and observe intruder's pure strategy $j_t \sim y_t$

    Collect samples from sensor feedback into $\mathcal{H}_t$: $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_t$

end

Appendix B. Theoretical Proofs

B.1. Proposition 2

Proof We have $A_{ij}' = A_{ij} - r_j$ and $B_{ij}' = B_{ij} - c_i$. By substitution, we have $A_{ij}' = \log(p_{miss}(i,j)) + c_i - r_j$ and $B_{ij}' = -\log(p_{miss}(i,j)) + r_j - c_i$. It follows that $A_{ij}' = -B_{ij}'$ and thus $A_{ij}' + B_{ij}' = 0$. Thus, the game $\mathcal{G}' = \{A', B'\}$ is a zero-sum game. The equivalence of Nash Equillbirum strategies for $\mathcal{G}$ and $\mathcal{G}'$ follows directly from the results presented in Section 2.1 in Kannan and Theobald (2010). ∎

B.2. Theorem 4

Proof We establish this result by showing the equivalence between the multiplicative weight updates made by the DWM algorithm (Algorithm 2) to that of the Weighted Majority algorithm in

Freund and Schapire (1997). Consider the per-agent loss function

$$\ell_t^q(\theta_k) = \sum_{j \in J} A_{kj}^q y_t(j) - \frac{1}{p} \sum_{j=1}^n r_j y_t(j). \tag{21}$$

Now, consider the joint strategy weight update rule given by

$$w_{t+1}(i) = \prod_{q=1}^p \sigma_{t+1}^q(k_{q,i}). \tag{22}$$

Substituting for $\sigma_{t+1}^q(k_{q,i})$, we have

$$w_{t+1}(i) = \prod_{q=1}^p \sigma_t^q(k_{q,i}) \beta^{\ell_t^q(i_q)}, \tag{23}$$

$$= \left(\prod_{q=1}^p \sigma_t^q(k_{q,i}) \right) \beta^{\left(\sum_{q=1}^p \ell_t^q(\theta_k) \right)} \tag{24}$$

From definition of A_{kj}^q in (8), we have

$$\sum_{q=1}^p \ell_t^q(\theta_k) = \sum_{q=1}^p \left(\sum_{j \in J} A_{kj}^q y_t(j) - \frac{1}{p} \sum_{j=1}^n r_j y_t(j) \right) \tag{25}$$

$$= \sum_{q=1}^p \sum_{j \in J} \left(V_{kj}^q \times \log(1 - p_{detect,q}) + c_k^q \right) - \sum_{j \in J} r_j y_t(j) \tag{26}$$

$$= \sum_{j \in J} \left(\sum_{q=1}^p V_{kj}^q \times \log \left(1 - p_{detect,q} \right) \right) y_t(j) + \sum_{j \in J} c_i y_t(j) - \sum_{j \in J} r_j y_t(j) \tag{27}$$

$$= \sum_{j \in J} \left(\log(p_{miss}(i,j)) \right) y_t(j) + \sum_{j \in J} c_i y_t(j) - \sum_{j \in J} r_j y_t(j) \tag{28}$$

$$= \sum_{j \in J} \left(\log(p_{miss}(i,j)) + c_i - r_j \right) y_t(j) \tag{29}$$

$$= \sum_{j \in J} A_{ij} y_t(j). \tag{30}$$

From Equations (11) and (9) we have, $w_{t+1}(i) = w_t(i) \beta^{\ell_t(i)}$, where $\ell_t(i) = \sum_{j \in J} A_{ij} y_j$. The weight update rule of the DWM algorithm (Algorithm 1) is equivalent to the multiplicative weight update over the defender's joint strategy space when the Weighted Majority Algorithm (as in Freund and Schapire (1997)) is applied to compute the mixed strategies $x \in X$ for the defender. The guarantees presented in this theorem directly follow from the Theorem 5 for the Hedge(β) algorithm presented in Freund and Schapire (1997). We refer interested readers to Littlestone and Warmuth (1994) and Nehme (2009) for detailed theoretical analysis of the weighted majority algorithm for zero-sum games.

■

B.3. Theorem 6

Proof We begin by establishing a bound on $||\hat{A}_t - A||_\infty$. For each round of interaction t, the players play strategies i_t, j_t and the defender receives V_{i_t,j_t} feedbacks from the sensors, i.e., for each node l traversed in the path j_t and covered by the joint sensor strategy i_t, the sensors get a $0/1$ feedback whether the intruder was detected or not, denoted by the Bernoulli random variable $Z_{ij}^{t,l}$. For round t, the estimate of Bernoulli parameter $\hat{p}_{detect}^t$ is computed as follows:

$$\hat{p}_{detect}^t = \frac{1}{V_{i_t,j_t}} \sum_{l=1}^{V_{i_t,j_t}} Z_{ij}^{t,l}. \tag{31}$$

The sample average for T rounds of interaction is given by

$$\hat{p}_{detect}^T = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{V_{i_t,j_t}} \sum_{l=1}^{V_{i_t,j_t}} Z_{ij}^{t,l} \right). \tag{32}$$

After T rounds, the average number of samples obtained per round is given by

$$\bar{k} = \frac{1}{T} \sum_{t=1}^{T} V_{i_t,j_t}, \tag{33}$$

and the total number of samples if $\bar{k}T$. Let p_{detect} denote the true Bernoulli parameter, i.e., true probability detection of the sensors. Applying Hoeffding's inequality, we have

$$\mathbb{P}\left(|\hat{p}_{detect}^T - p_{detect}| \geq \epsilon_T\right) \leq 2\exp\left(-2\bar{k}T\epsilon_T^2\right). \tag{34}$$

We define $\eta \geq 2\exp\left(-2\bar{k}T\epsilon_T^2\right)$. For a specified ϵ_T, with probability at least $1 - \eta$, for $T > \frac{1}{(2\bar{k}\epsilon_T^2)} \log \frac{2}{\eta}$, we have $|\hat{p}_{detect}^T - p_{detect}| \leq \epsilon_T$.

We will now establish a bound for $||\hat{A}_t - A||_\infty$. By substituting for A_{ij}, we have,

$$|\hat{A}_{t,ij} - A_{ij}| = V_{ij}|\log(1 - \hat{p}_{detect}^t) - \log(1 - p_{detect})|. \tag{35}$$

By applying the Mean Value Theorem to the function $\log(1 - p)$, we have the following bound

$$|\log(1 - \hat{p}_{detect}^t) - \log(1 - p_{detect})| \leq \frac{|\hat{p}_{detect}^t - p_{detect}|}{1 - \max(\hat{p}_{detect}^t, p_{detect})}. \tag{36}$$

Since $V_{ij} \leq V_{max}$ and $1 - \max(\hat{p}_{detect}^t, p_{detect}) \geq 1 - p_{max}$, we have

$$||\hat{A}_t - A||_\infty \leq \frac{V_{max}|\hat{p}_{detect}^t - p_{detect}|}{1 - p_{max}}. \tag{37}$$

Denote $\delta_t = \frac{V_{max}|\hat{p}_{detect}^t - p_{detect}|}{1 - p_{max}}$. We have $\delta_t \leq \frac{V_{max}\epsilon_t}{(1-p_{max})}$.

Since with probability $1 - \eta$, we have $\epsilon_t = \sqrt{\frac{\log(2/\eta)}{2t}}$ (from Hoeffding's bound), we have

$$||\hat{A}_t - A||_\infty \leq \frac{V_{max}}{\sqrt{2t}(1 - p_{max})} \sqrt{\log \frac{2}{\eta}}. \tag{38}$$

We will now analyze the regret $\mathcal{R}_T$ as defined in (14). Consider Case 1, where $y_t = y^\dagger$, i.e., the intruder knows the true game matrix A and plays the NE strategy $y^\dagger$ of the true game. Consider the instantaneous regret $\mathcal{R}_t$ for round t:

$$\mathcal{R}_t = x_t^{*\top} A y^\dagger - V_A^* = x_t^{*\top} A y^\dagger - x^{\dagger\top} A y^\dagger. \tag{39}$$

By adding and subtracting $x_t^* \hat{A}_t y^\dagger$, we have

$$\mathcal{R}_t = \underbrace{x_t^{*\top} A y^\dagger - x_t^{*\top} \hat{A}_t y^\dagger}_{\text{Term-I}} + \underbrace{x_t^{*\top} \hat{A}_t y^\dagger - x^{\dagger\top} A y^\dagger}_{\text{Term-II}}. \tag{40}$$

Term-I:

$$x_t^{*\top} A y^\dagger - x_t^{*\top} \hat{A}_t y^\dagger = x_t^{*\top} \left(A - \hat{A}_t \right) y^\dagger \leq ||\hat{A}_t - A||_\infty = \delta_t. \tag{41}$$

To Term-II, we add and subtract $x^{\dagger\top} \hat{A}_t y^\dagger$, and obtain the following:

$$\text{Term-II} = \underbrace{x_t^{*\top} \hat{A}_t y^\dagger - x^{\dagger\top} \hat{A}_t y^\dagger}_{\text{Term-II-I}} + \underbrace{x^{\dagger\top} \hat{A}_t y^\dagger - x^{\dagger\top} A y^\dagger}_{\text{Term-II-II}} \tag{42}$$

Now consider Term-II-I. We know that (x_t^*, y_t^*) is the NE for $\hat{A}_t$. Since the y-player is maximizing the expected game value, we have $x_t^* \hat{A}_t y_t^* \geq x_t^* \hat{A}_t y, \forall y \in Y$ and thus we have $x_t^* \hat{A}_t y^\dagger \leq x_t^* \hat{A}_t y_t^*$. As a result, we have

$$\text{Term-II-I} \leq x_t^* \hat{A}_t y_t^* - x^{\dagger\top} \hat{A}_t y^\dagger. \tag{43}$$

The right-hand side of the above expression can be viewed as the difference in the game values when the NE strategies of a game A, i.e., $(x^\dagger, y^\dagger)$, are played in a perturbed game $\hat{A}_t$. From Theorem 4 in Lipton et al. (2006), we have the following:

$$\text{Term-II-I} \leq 3||\hat{A}_t - A||_\infty = 3\delta_t. \tag{44}$$

Term-II-II

$$x^{\dagger\top} \hat{A}_t y^\dagger - x^{\dagger\top} A y^\dagger = x^{\dagger\top} (\hat{A}_t - A) y^\dagger \leq ||\hat{A}_t - A||_\infty = \delta_t. \tag{45}$$

Combining Equations (41), (44) and (45), we have the following bound on the instantaneous regret $\mathcal{R}_t$:

$$\mathcal{R}_t \leq \delta_t + 3\delta_t + \delta_t = 5\delta_t. \tag{46}$$

The cumulative regret can be bounded as follows. With probability at least $1 - \alpha$,

$$\mathcal{R}_T = \sum_{t=1}^{T} \mathcal{R}_t \leq \sum_{t=1}^{T} 5\delta_t \tag{47}$$

$$\leq \frac{5V_{max}}{\sqrt{2}(1 - p_{max})} \sqrt{\log \frac{2}{\eta}} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \tag{48}$$

$$\leq \frac{5V_{max}}{\sqrt{2}(1 - p_{max})} \sqrt{\log \frac{2}{\eta}} \int_{t=1}^{T} \frac{1}{\sqrt{t}} \tag{49}$$

$$\leq \frac{5V_{max}}{\sqrt{2}(1 - p_{max})} \sqrt{\log \frac{2}{\eta}} \times 2\sqrt{T} \tag{50}$$

$$\leq \frac{5\sqrt{2}V_{max}}{(1 - p_{max})} \sqrt{T \log \frac{2T}{\alpha}}, \tag{51}$$

where $\alpha = T\eta$. This completes the proof of Part 1.

Now consider the scenario where $y_t \neq y^\dagger$, i.e., when the defender is playing against a suboptimal opponent (intruder), who does not play the NE strategy of the true game. The instantaneous regret is given by

$$\mathcal{R}_t = x_t^{*\top} A y_t - V_A^* = x_t^{*\top} A y_t - x^{\dagger\top} A y^\dagger. \tag{52}$$

By adding and subtracting $x_t^{*\top} \hat{A}_t y_t$, we have

$$\mathcal{R}_t = \underbrace{x_t^{*\top} A y_t - x_t^{*\top} \hat{A}_t y_t}_{\text{Term-I}} + \underbrace{x_t^{*\top} \hat{A}_t y_t - x^{\dagger\top} A y^\dagger}_{\text{Term-II}}. \tag{53}$$

Term-I:

$$x_t^{*\top} A y_t - x_t^{*\top} \hat{A}_t y_t = x^{*\top} \left(A - \hat{A}_t \right) y_t \leq ||\hat{A}_t - A||_\infty = \delta_t. \tag{54}$$

In Term-II, we have $x_t^{*\top} \hat{A}_t y_t \leq x_t^{*\top} \hat{A}_t y_t^*$, since the y-player is maximizing the expected game value, and thus

$$x_t^{*\top} \hat{A}_t y_t - x^{\dagger\top} A y^\dagger \leq x_t^{*\top} \hat{A}_t y_t^* - x^{\dagger\top} A y^\dagger. \tag{55}$$

From Theorem 5 in Lipton et al. (2006), we have that $x_t^{*\top} \hat{A}_t y_t^* - x^{\dagger\top} A y^\dagger < ||\hat{A}_t - A||_\infty$, and thus

$$x_t^{*\top} \hat{A}_t y_t - x^{\dagger\top} A y^\dagger \leq ||\hat{A}_t - A||_\infty = \delta_t. \tag{56}$$

Combining the terms in Equations (54) and (56), we have

$$\mathcal{R}_t \leq \delta_t + \delta_t = 2\delta_t. \tag{57}$$

Using similar arguments as in Case 1, we have with probability at least $1 - \alpha$,

$$\mathcal{R}_T \leq \frac{2\sqrt{2}V_{max}}{(1 - p_{max})} \sqrt{T \log \frac{2T}{\alpha}}. \tag{58}$$

■

B.4. Lemma 8

Proof Consider the following expression for A_{ij}:

$$A_{ij} = \sum_{l=1}^{p} V_{ij,l} \log(1 - p_{detect,l}) + c_i - r_j. \tag{59}$$

We have the following confidence bound estimate for $p_{detect,l}$ which follows from Hoeffding inequality (alternatively a special case of Chernoff's bound): With probability at least $1 - \delta$

$$p_{detect,l} \le \bar{p}_{detect,l}^t + \sqrt{\frac{\log \frac{2}{\delta}}{2(1 \vee n_{ij,l}^t)}} = \tilde{p}_l^t. \tag{60}$$

Using the similar technique as in proof of Theorem 6, we apply the Mean Value Theorem to the function $A_{ij}(p_{detect,l})$ with respect to $p_{detect,l}, l = 1, ..., p$ (as in (59)) to obtain the following UCB estimate for A_{ij}

$$A_{ij} \le \bar{A}_{ij}^t + \sum_{l=1}^{p} \frac{V_{max,l}}{1 - p_{max,l}} \sqrt{\frac{\log(\frac{2}{\delta})}{2(1 \vee n_{ij,l}^t)}} = \tilde{A}_{ij}^t, \tag{61}$$

where $V_{max,l}$ and $p_{max,l}$ are as defined in Section 4.1. ∎

B.5. Theorem 9

Proof We establish this proof by making similar arguments as in Theorem 1 of O'Donoghue et al. (2021). Note that we discuss the equivalence of our regret definition and problem setting with the one presented in O'Donoghue et al. (2021) in Appendix A.

Since the upper confidence matrix (61) overestimates the true matrix, we have $V_{\tilde{A}_t}^* \ge V_A^*$ (see Proof of Theorem 1 in O'Donoghue et al. (2021)). Let E_t be the event that there exists a pair i, j such that $p_{detect,l} > \tilde{p}_l^t$ for $l = 1, ..., p$. By definition, $E_t \in \mathcal{F}_t$. Consider for a moment that E_t does not hold and let x_t^* be the best-response of the x-player at round t. We have the following for the instantaneous regret $\mathcal{R}_t$:

$$\mathcal{R}_t = (V_A^* - x_t^\top A y_t) \le (V_{\tilde{A}_t}^* - x_t^\top A y_t); \tag{62}$$

$$= (x_t^{*\top} \tilde{A}_t y_t - x_t^\top A y_t); \tag{63}$$

$$\le (x_t^\top (\tilde{A}_t - A) y_t); \tag{64}$$

$$= \left[\sum_{l=1}^{p} \frac{V_{max,l}}{1 - p_{max,l}} \sqrt{\frac{\log(\frac{2}{\delta})}{2(1 \vee n_{ij,l}^t)}} \right], \tag{65}$$

where the last step is obtained by substituting the upper confidence bound expression as in (61). The cumulative regret is given by

$$\mathcal{R}_T = \sum_{t=1}^{T} \mathcal{R}_t; \tag{66}$$

$$\le \left[\sum_{t=1}^{T} \sum_{l=1}^{p} \frac{V_{max,l}}{1 - p_{max,l}} \sqrt{\frac{\log(\frac{2}{\delta})}{2(1 \vee n_{ij,l}^t)}} \right] + T\mathbb{P}(\cup_{t=1}^{T} E_t). \tag{67}$$

We bound the second term as follows: $T\mathbb{P}(\cup_{t=1}^{T} E_t) \leq 2T^2|\mathcal{S}||\Theta||J|\delta \leq 1$.
The first term is bounded by

$$\left[\sum_{t=1}^{T}\sum_{l=1}^{p} \frac{V_{max,l}}{1-p_{max,l}}\sqrt{\frac{\log(\frac{2}{\delta})}{2(1 \vee n_{ij,l}^t)}}\right] \leq \sum_{ij}\sum_{l=1}^{p}\sqrt{2n_{ij,l}^T \log\frac{2}{\delta}}; \tag{68}$$

$$\leq |\mathcal{S}|\sqrt{2|\Theta||J|T\log\frac{2}{\delta}}. \tag{69}$$

By substituting for δ, we have with probability at least $1 - \delta$,

$$\mathcal{R}_T \leq 1 + |\mathcal{S}|\sqrt{2|\Theta||J|T\log(4T^2|\mathcal{S}||\Theta||J|)} = \tilde{O}(\sqrt{|\mathcal{S}|^2|\Theta||J|T}). \tag{70}$$

$\blacksquare$