KYBERNETIKA — VOLUME *55* (2019), NUMBER 4, PAGES 618–640

HANDLING A KULLBACK–LEIBLER DIVERGENCE RANDOM WALK FOR SCHEDULING EFFECTIVE PATROL STRATEGIES IN STACKELBERG SECURITY GAMES

César U. Solis, Julio B. Clempner and Alexander S. Poznyak

This paper presents a new model for computing optimal randomized security policies in non-cooperative Stackelberg Security Games (SSGs) for multiple players. Our framework rests upon the extraproximal method and its extension to Markov chains, within which we explicitly compute the unique Stackelberg/Nash equilibrium of the game by employing the Lagrange method and introducing the Tikhonov regularization method. We also consider a game-theory realization of the problem that involves defenders and attackers performing a discrete-time random walk over a finite state space. Following the Kullback–Leibler divergence the players' actions are fixed and, then the next-state distribution is computed. The player's goal at each time step is to specify the probability distribution for the next state. We present an explicit construction of a computationally efficient strategy under mild defenders and attackers conditions and demonstrate the performance of the proposed method on a simulated target tracking problem.

Keywords: Stackelberg games, security, patrolling, Markov chains

Classification: 91A10, 91A35, 91A80, 91B06, 91B70, 91B74

1. INTRODUCTION

1.1. Brief review

We focus on a game theory approach well-suited to adversarial reasoning for security resource allocation and scheduling problems referred to as Stackelberg security games [1, 4, 7, 22, 36]. Our approach is based on multiple-players games in which there exist limited security resources which prevent full security coverage all the time. In the game, defenders aim to protect a set of targets that minimizes their expected utility while attackers aim to assail targets that maximizes their expected utility. A central assumption in the literature on Stackelberg security games is that limited security resources must be deployed strategically considering differences in priorities of targets requiring security coverage and the responses of the adversaries to the security position. In the dynamics of the game defenders commit to a probabilistic defense target and the attackers observe the probabilities with which each target is covered. However, attackers cannot

DOI: 10.14736/kyb-2019-4-0618

observe the actual defense realization. Much of the work on Stackelberg security games focuses on potential uncertainty over the types, capabilities, knowledge and priorities of adversaries faced [18, 19, 27].

One important assumption presented in the literature is that in Stackelberg security games it is possible to consider a security game with multiple defenders and attackers at the same time and, all possible combinations of security decisions for all targets [31]. Real applications consider several defenders and attackers among potential targets to defend or attack. These may be implicit as in defending critical infrastructure (branch banking locations, ports, etc.).

Markov decision processes (MDPs) are a well-liked framework for the realization of sequential decision-making in a random dynamic environment for game theory [11, 25]. The dynamics is as follows: at each time step the defenders and attackers observes the state of the game and choose an action. The game then randomly transitions to its next state considering the transition probability established by the current state and the action chosen. In our MDP game realization, it is assumed that the cost/utility functions and the transition probabilities are known in advance, the policies are previously computed applying the extraproximal method for solving the game, and the optimality criterion is forward-looking. In addition, we choose a more control-oriented approach: routing the game along a state trajectory through actions selected according to a state feedback law determined by the Kullback–Leibler (KL) divergence (or the relative entropy) [24] between the actions. We allow the defenders and attackers to select the state transitions directly, so that actions correspond to fixed probability distributions on the underlying state space. Moreover, for presenting a real-world solution to the problem, the control penalizes defenders' deviation from the attackers position. We prove that the realization converges guaranteeing that the sequential decision-making in the proposed random model is correct.

1.2. Related work

Stackelberg security game models become a critical tool that arises in protecting different types of real-world targets. Conitzer and Sandholm [15] described a method to commit to optimal randomized strategies in Stackelberg security games. Paruchuri et al. [23] focused on Bayesian Stackelberg games suggested a mixed-integer linear programming algorithm for computing a Stackelberg equilibrium. Letchford et al. [21] provided theoretical results on the value of being able to commit and the value of being able to correlate, as well as complexity results about computing Stackelberg strategies in stochastic games. Yang et al. [36] based on bounded rationality computed the optimal strategies of a security game. Yin et al. [37, 38] considered noise in the defender's execution of the suggested mixed strategy and/or the observations made by an attacker. A particular case of Stackelberg security games considers the problem of multi-robot patrolling against intrusions around a given area with the existence of an attacker attempting to penetrate into the area [1, 4]. The authors showed that Nash and Stackelberg strategies are the same in the majority of cases only when the follower attacks just one target. They also proposed an extensive form game model that makes the defender's uncertainty about the attacker's ability to observe explicit. These games are security games between a defender (allocates defensive resources), and an attacker (decide on targets to attack).

For multiple defenders and attackers in Markov games Clempner and Poznyak [9] suggested an approach for conforming coalitions in Stackelberg security games where the coalition of the defenders achieves its synergy by computing the Strong $Lp-$Stackelberg/Nash equilibrium [33]. The security model describes a strategic game in which the defenders cooperate, and attackers do not cooperate. Clempner and Poznyak,[7] presented a shortest-path method to represent the Stackelberg security game as a potential game using the Lyapunov theory. Trejo et al. [31] employed the extraproximal method for computing the Stackelberg/Nash equilibria in the case of one defender and multiple attackers. Solis et al. [29] extended the work presented by [31] including multiple leaders and followers and, presenting a proof of convergence. Clempner and Poznyak [9] suggested a SSG that represents a strategic game where the defenders cooperate, and attackers noncooperate. The same authors in [12] improved the technique described in [7] and using the extraproximal method calculated the Lyapunov equilibrium in SSGs. Clempner [5] presented a method for controlling the patrolling activities involving constraints that involve continuous-time SSGs. Guerrero et al. [16, 17] developed a method for the SSGs solution, which used the bargaining Nash approach for computing the cooperative equilibrium point for the defenders, while the attackers played in a non-cooperative approach. Trejo et al. [32, 35] presented an using repeated cooperative Stackelberg security Markov games. The Reinforcement Learning method. combines prior knowledge and temporal-difference methods. The coalition of the defenders is computed employing the Strong L_p-Stackelberg/Nash equilibrium [33, 34]. Albarran and Clempner [2] provided a novel solution for computing the Stackelberg security games for multiple players, considering finite resource allocation in domains with incomplete information. In our model, we consider several defenders and several attackers for non-cooperative Stackelberg security games in which the realization is based on handling a Kullback–Leibler divergence random walk.

1.3. Main results

This paper presents the following contributions.

- Suggests a new technique for computing optimal randomized security policies in non-cooperative Stackelberg security games for multiple defenders and attackers.
- Considers the extraproximal method and its extension to Markov chains [30], within which we explicitly compute the unique Stackelberg/Nash equilibrium of the game by specifying a natural model employing the Lagrange method and introducing Tikhonov's regularization method [13, 14].
- Proposes a method that computes the optimal security policies of the Stackelberg/Nash game exactly and efficiently presenting a "real-world" solution to the problem. We also consider a game-theory realization of the problem that involves defenders and attackers performing a discrete-time random walk over a finite state space.
- Following the Kullback–Leibler divergence [26] the players' actions are fixed and, then the next-state distribution is computed. The player's goal at each time step is to specify the probability distribution for the next state given the current state.

- Proves the realization converge guaranteeing that the sequential decision-making in a random dynamic model is correct.
- Presents an explicit construction of a computationally efficient strategy, under mild defenders and attackers conditions and demonstrate the performance of the proposed method on a simulated target tracking problem.

An application for protecting a marine canal that suggests patrolling strategies to protect ports validates the proposed method.

1.4. Organization of the paper

The remainder of the paper is organized as follows. Section 2 contains preliminaries on MDPs and game theory. Section 3 then describes our proposed Stackelberg security game model presenting the extrapoximal method which employs the Lagrange method and uses the Tikhonov regularization method. Section 4 suggests a model for random walk based on the Kulback-Leibler divergence studying two models for the defenders one using a classical approach and the other models penalize the defenders' deviation from the attackers' location. As well as, we prove that the synchronization of the random walk of defenders and attackers converge in probability to the product of the individual probabilities. Some simulation results are presented in Section 5. We close by summarizing our contributions in Section 6.

2. PRELIMINARIES

2.1. Controllable Markov process in discrete time

A *controllable Markov decision process* is a 5-tuple $MDP = \{S, A, A(s), \Pi, J\}$ where S is a finite set of states, $S \subset \mathbb{N}$, endowed with discrete topology; A is the set of actions, which is a metric space [6, 25]. For each $s \in S$, $A(s) \subset A$ is the non-empty set of admissible actions at state $s \in S$. Without loss of generality we may take $A = \cup_{s \in S} A(s)$; $\mathbb{K} = \{(s, a) | s \in S, a \in A(s)\}$ is the set of admissible state-action pairs, which is a measurable subset of $S \times A$; $\Pi(k) = \left[\pi_{j|ik}\right]$ is a stationary transition controlled matrix, where

$$\pi_{j|ik} \equiv P\left(s(t+1) = s_j | s(t) = s_i, a(t) = a_k\right)$$

representing the probability associated with the transition from state $s_i \in S$ to state s_j under an action $a_k \in A(s_i)$ $(k = 1, \ldots, M)$ at time $t \in \mathbb{N}$. The relations $\pi_{j|ik} \geq 0$ and $\sum_{j=1}^{N} \pi_{j|ik} = 1$ are satisfied for all i, j, k. Finally, $J : S \times \mathbb{K} \to \mathbb{R}^n$ is the cost function.

The system evolves as follows: at each time $t \in \mathbb{N}$ the decision maker knows the previous states and actions and, observes the current state, says $s(t) = s_i \in S$. Using this information, the controller selects an action $a(t) = a_k \in A(s)$. Then two things happen: a cost J_{ijk} is incurred and, the system at time $t + 1$ moves to a new state $s(t + 1) = s_j \in S$ with probability $\pi_{j|ik}$.

We will restrict attention to stationary policies throughout all the paper. A policy d is a (measurable) rule for choosing actions which, at each time $n \in \mathbb{N}$, may depend on the current state and on the record of previous states and actions; see, for instance, [25] for details. The class of all policies is denoted by D and, given the initial state $s \in S$ and

the policy d being used for choosing actions, the distribution of the state-action process $\{(s(t), a(t))\}$ is uniquely determined. Following, we will denote by P and E respectively the probability measure and the expectation operator induced by the policy d. Next, define $\mathbb{F} := \prod_{s\in S} A(s)$ and notice that $\mathbb{F}$ is a compact metric space in the product topology which consists of all functions $f : S \to A$ such that $f(s) \in A(s)$ for each $s \in S$. A policy d is *stationary* iff there exists $f \in \mathbb{F}$ such that the equality $A(t) = f(s(t))$ is always valid under d, i. e. $d_{k|i}(t) = d_{k|i}$. Also, under the action of any stationary policy $d_{k|i}(t) = d_{k|i}$ the state process is a Markov chain with stationary transition mechanism. For each strategy $d_{k|i}$ the associated transition matrix is defined as:

$$\Pi(d) := [\pi_{j|ik}(d)] = \sum_{k=1}^{M} \pi_{j|ik} d_{k|i}$$

such that on a stationary state distribution for all $d_{k|i}$ and $t \geq 0$.

Our results are based on the following Theorems and Lemmas (for the proof of the following Theorem and Lemmas see [**?**]).

Theorem 2.1. For some state $j_0 \in (1, \ldots, N)$ of a homogeneous (stationary) Markov chain with the transition matrix $\mathbf{\Pi}$ and some $t > 0$, $\xi \in (0, 1)$ for all $i \in \mathcal{G}$ let

$$\pi_{ij_0}(t) := P(s(t) = s_{j_0} | s(0) = s_i) \geq \xi. \tag{1}$$

Then for any initial-state distribution $P\{s(0) = s_i\}$ and for any $i, j = 1, \ldots, N$ there exists the limit

$$p_j^* := \lim_{t\to\infty} \pi_{ij}(t)$$

such that for any $t \geq 0$ this limit is reachable with an **exponential rate**, namely,

$$\left|\pi_{ij}(t) - p_j^*\right| \leq (1-\xi)^t = e^{-\alpha t}$$

where $\alpha := |\ln(1-\xi)|$.

Corollary 2.2. Since $\pi_{ij_0}(t) = (\mathbf{\Pi}^n(ij_0))^{\intercal}$ to verify the property (1) it is sufficient to multiply $\mathbf{\Pi}$ by itself t times up to the moment when all elements of at least one row will be positive.

Corollary 2.3. For an optimal policy $d_{k|i}^*$ the corresponding homogeneous Markov chain with the transition matrix $\mathbf{\Pi}^*$ will be ergodic if the multiplication of $\mathbf{\Pi}^*$ by itself n times up to the moment when all the elements of at least one row will be all positive.

Definition 2.4. For a homogeneous finite Markov chain with transition matrix $\mathbf{\Pi} = [\pi_{ij}]_{i,j=1,\ldots,N}$ the parameter $k_{erg}(t_0)$ defined by

$$k_{erg}(t_0) := 1 - \frac{1}{2} \max_{i,j=1,\ldots,N} \sum_{m=1}^{N} |(\tilde{\pi}_{im}(t_0)) - (\tilde{\pi}_{jm}(t_0))| \in [0, 1)$$

is said to be coefficient of ergodicity of this Markov chain at time t_0, where

$$(\tilde{\pi}_{im}(t_0)) = P\{\, s(t_0) = s_m |\, s(1) = s_i\} = (\mathbf{\Pi}^{n_0}(im))$$

is the probability to evolve from the initial state $s_1 = s_i$ to the state $s_{t_0} = s_m$ after t_0 transitions.

Lemma 2.5. The coefficient of ergodicity $k_{erg}(t_0)$ can be estimated from below as

$$k_{erg}(t_0) \geq \min_{i=1,\ldots,N} \max_{j=1,\ldots,N} \tilde{\pi}_{ij}(t_0).$$

Remark 2.6. If all the elements $\tilde{\pi}_{ij}(t_0)$ of the transition matrix $\mathbf{\Pi}^{t_0}$ are positive, then the coefficient of ergodicity $k_{erg}(t_0)$ is also positive. Notice that there exist ergodic Markov chains with elements $\tilde{\pi}_{ij}(t_0)$ equal to zero, but with a positive coefficient of ergodicity $k_{erg}(t_0)$.

Theorem 2.7. If for a finite Markov chain, which is controllable by the fixed local-optimal policy $d^*_{k|i}$, with positive lower bound estimate of the ergodicity coefficient

$$\chi_{erg} := \inf_{t_0} \max_{j=1,\ldots,N} \min_{i=1,\ldots,N} \tilde{\pi}^*_{ij}(t_0) > 0$$

then the following properties hold:

1) there exists a unique stationary distribution

$$\mathbf{p}^* = \lim_{t\to\infty} \mathbf{p}_t;$$

2) the convergence of the current-state distribution to the stationary one is exponential:

$$|\mathbf{p}_t(i) - \mathbf{p}^*(i)| \leq C \exp\{-Dn\}$$

$$C = \frac{1}{1-\chi^{\iota}_{erg}}, \quad D = \frac{1}{t_0^{\iota *}} \ln C,$$

$$t^*_0 = \arg\min_{t_0} \left[\max_{j=1,\ldots,N} \min_{i=1,\ldots,N} \tilde{\pi}^*_{ij}(t_0) \right].$$

Remark 2.8. Theorem 2.1 ensures that $\mathbf{\Pi}^*$ has a unique everywhere positive invariant distribution P^* and, it is equivalent to the existence of some t_0, such that $\pi^*_{ij}(t_0) > 0$.

Remark 2.9. Theorem 2.7 guarantees that the convergence to P^* is exponentially fast (so that $\pi^*_{ij}(t_0)$ is geometrically ergodic).

2.2. Markov games

The dynamic of the game for Markov chains is described as follows. The game consists of a set of $\mathcal{N} = \{1, \ldots, n\}$ players (denoted by $l = \overline{1,n}$) and begins at the initial state $s(0) = s_i$ which (as well as the states further realized by the process) is assumed to be completely measurable. Each of the players l is allowed to randomize, with distribution $d^l_{k|i}(t)$, over the pure action choices $a_k \in A^l(s_i)$, $i = \overline{1,N}$ and $k = \overline{1,M}$. From now on, we will consider only stationary strategies $d^l_{k|i}(t) = d^l_{k|i}$. These choices induce the state distribution dynamics, which in the ergodic case for any stationary strategy $d^l_{k|i}$ the distributions $P^l(s(t+1){=}s_j)$ exponentially quickly converge to their limits satisfying

$$P^l(s_j) = \sum_{i=1}^{N} \left(\sum_{k=1}^{M} \pi^l_{j|ik} d^l_{k|i} \right) P^l(s_i).$$

The cost function of the optimization problem, depend on the states and actions, are given by the values W_{ik}^l, so that the "average cost function" J^l in the stationary regime can be expressed as

$$J^l\left(c^1,..,c^n\right):=\sum_{i=1}^{N}\sum_{k=1}^{M}W_{ik}^l\prod_{l=1}^{n}c_{ik}^l$$

where $c^l:=\left[c_{ik}^l\right]_{i=\overline{1,N};k=\overline{1,M}}$ is a matrix with elements

$$c_{ik}^l=d_{k|i}^lP^l\left(s_i\right) \tag{2}$$

satisfying

$$c^l\in C_{adm}^l=\left\{\ c^l:\sum_{i=1}^{N}\sum_{k=1}^{M}c_{ik}^l=1,\ c_{ik}^l\geq 0,\text{ and }\sum_{k=1}^{M}c_{jk}^l=\sum_{i=1}^{N}\sum_{k=1}^{M}\pi_{j|ik}^lc_{ik}^l\ \right\}$$

where

$$W_{ik}^l=\sum_{i=1}^{N}\sum_{k=1}^{M}\left(\sum_{j=1}^{N}J_{ijk}\prod_{l=1}^{n}\pi_{j|ik}^l\right).$$

Notice that by (2) it follows that

$$P^l\left(s_i\right)=\sum_{k=1}^{M}c_{ik}^l\quad\text{and}\quad d_{k|i}^l=\tfrac{c_{ik}^l}{\sum_{k=1}^{M}c_{ik}^l}. \tag{3}$$

In the ergodic case $\sum_{k=1}^{M}c_{ik}^l>0$ for all $l=\overline{1,n}$.

3. STACKELBERG SECURITY GAME

3.1. Stackelberg game

Following [8, 20, 29, 30] let us consider a set $\mathcal{N}=\{1,\ldots,n\}$ of *defenders* (leaders) indexed by l, $\left(l=\overline{1,n}\right)$, whose randomized strategies are represented by $u^l\in U^l$. The set U is a convex and compact set where

$$u^l:=col\left(c_{ik}^l\right),\ U^l:=C_{adm}^l,\ U:=\bigotimes_{l=1}^{n}U^l$$

such that the operator *col* is the column operator, which transforms a matrix into a column. Let $u=(u^1,\ldots,u^n)^\top\in U$ be the joint strategy of the defenders and $\hat{u}=u^{-l}$ be the strategy of the complementary players adjoint to u^l,

$$u^{-l}:=\left(u^1,\ldots,u^{l-1},u^{l+1},\ldots,u^n\right)^\top\in U^{-l}$$

where $U^{-l}:=\bigotimes_{h=1,h\neq l}^{n}U^h$, and $u=(u^l,u^{\hat{l}})$.

As well as, let us consider a set $\mathcal{R}=\{1,\ldots,r\}$ of *attackers* (followers) indexed by h, $\left(h=\overline{1,r}\right)$ with randomized strategies $v^h\in V^h$. V is a convex an compact set such that

$$v^h:=col\left(c_{ik}^h\right),\ V^h:=C_{adm}^h,V:=\bigotimes_{h=1}^{r}V^h.$$

Let us denote by $v = (v^1, \ldots, v^r) \in V := \bigotimes_{h=1}^{r} V^h$ the joint strategy of the attackers and $v^{\hat{h}} = v^{-h}$ is a strategy of the rest of the players adjoint to v^h, namely,

$$v^{-h} := \left(v^1, \ldots, v^{h-1}, v^{h+1}, \ldots, v^r\right)^\top \in V^{-h}$$

such that $V^{-h} := \bigotimes_{h=1,\ h\neq r}^{r} V^h$ and $v = (v^h, v^{\hat{h}})$ $\left(h = \overline{1,r}\right)$.

In the Stackelberg game the *defenders* first find a strategy $u^* = \left(u^{1*}, \ldots, u^{n*}\right) \in U$ satisfying for any admissible $u^l \in U^l$ and any $l = \overline{1,n}$

$$\Gamma(u) := \sum_{l=1}^{n} \left[\left(\min_{u^l \in U^l} \psi_l\left(u^l, u^{-l}\right) \right) - \psi_l\left(u^l, u^{-l}\right) \right] \tag{4}$$

[8, 31]. Here $\psi_l\left(u^l, u^{-l}\right)$ is the cost-function of the leader l which plays the strategy $u^l \in U^l$ and the rest of the leaders play the strategy $u^{-l} \in U^{-l}$.

If we consider the utopia point

$$\bar{u}^l := \arg\min_{u^l \in U^l} \psi_l\left(u^l, u^{-l}\right) \tag{5}$$

then, we can rewrite Eq. (4) as follows

$$\Gamma(u) := \sum_{l=1}^{n} \left[\psi_l\left(\bar{u}^l, u^{-l}\right) - \psi_l\left(u^l, u^{-l}\right)\right]. \tag{6}$$

The functions $\psi_l\left(u^l, u^{-l}\right)$ $\left(l = \overline{1,n}\right)$ are assumed to be convex in all their arguments.

The function $\Gamma(u)$ satisfies *the* **Nash condition**

$$\max_{u \in U} g(u) = \sum_{l=1}^{n} \left[\psi_l\left(\bar{u}^l, u^{-l}\right) - \psi_l\left(u^l, u^{-l}\right)\right] \leq 0$$

for any $u^l \in U^l$ and all $l = \overline{1,n}$

A strategy $u^* \in U_{adm}$ is said to be a **Nash equilibrium** if

$$u^* \in Arg \min_{u \in U_{adm}} \{\Gamma(u)\}.$$

If $\Gamma(u)$ is strictly convex then $u^* = \arg\min_{u \in U_{adm}} \{\Gamma(u)\}$. Following the dynamics of the game, the attackers observe the defenders behavior and in equilibrium selects the expected strategy (as a response) $v^* = \left(v^{1*}, \ldots, v^{r*}\right) \in V$ satisfying for any admissible $v^h \in V^h$ and any $h = \overline{1,r}$

$$\Phi(v) := \sum_{h=1}^{r} \left[\left(\min_{v^h \in V^h} \varphi_h\left(v^h, v^{-h}\right) \right) - \varphi_h\left(v^h, v^{-h}\right) \right].$$

Here $\varphi_h\left(v^h, v^{-h}\right)$ is the cost-function of the follower m which plays the strategy $v^h \in V^h$ and the rest of the leaders play the strategy $v^{-h} \in V^{-h}$.

If we consider the utopia point

$$\bar{v}^h := \arg\min_{v^r \in V^h} \varphi_h\left(v^h, v^{-h}\right)$$

then, we can rewrite Eq. (4) as follows

$$\Phi(v) := \sum_{h=1}^{r} \left(\varphi_h\left(\bar{v}^h, v^{-h}\right) - \varphi_h\left(v^h, v^{-h}\right)\right).$$

The functions $\varphi_r\left(v^h, v^{-h}\right)$ $\left(h = \overline{1,r}\right)$ are assumed to be convex in all their arguments.

The function $\Phi(v)$ satisfies *the* **Nash condition**

$$\max_{v^h \in V^h} f(v) = \sum_{h=1}^{r} \left(\varphi_h\left(\bar{v}^h, v^{-h}\right) \text{-} \varphi_h\left(v^h, v^{-h}\right)\right) \leq 0$$

for any $v^h \in V^h$ and all $h = \overline{1,r}$.

Defenders and *attackers* together are in a Stackelberg game: the model involves two non-cooperatively Nash games restricted by a Stackelberg game defined as follows.

Definition 3.1. A *game with n defenders and m* attackers said to be a Stackelberg–Nash game if

$$\Gamma(u|v) := \sum_{l=1}^{n} \left(\psi_l\left(\bar{u}^l, u^{-l}|v\right) - \psi_l\left(u^l, u^{-l}|v\right)\right)$$

where u corresponds to a defender, realizing its strategy based on the restriction v of the attackers, such that

$$\max_{u \in U} g(u|v) = \sum_{l=1}^{n} \left[\psi_l\left(\bar{u}^l, u^{-l}|v\right) \text{-} \psi_l\left(u^l, u^{-l}|v\right)\right] \leq 0$$

where u^{-l} is a strategy of the rest of the defenders adjoint to u^l, namely,

$$u^{-l} := \left(u^1, \ldots, u^{l-1}, u^{l+1}, \ldots, u^n\right) \in U^{-l}$$

where $U^{-l} := \bigotimes_{y=1,\ y \neq l}^{n} U^y$ and $\bar{u}^l := \arg\min_{u^l \in U^l} \psi_l\left(u^l, u^{-l}|v\right)$ such that

$$f(v|u) := \sum_{h=1}^{r} \left(\varphi_h\left(\bar{v}^h, v^{-h}|u\right) - \varphi_h\left(v^h, v^{-h}|u\right)\right)$$

given that v^{-h} is a strategy of the rest of the attackers adjoint to v^h, namely,

$$v^{-h} := \left(v^1, \ldots, v^{h-1}, v^{h+1}, \ldots, v^r\right) \in V^{-h}$$

$V^{-h} := \bigotimes_{q=1,\ q \neq h}^{r} V^q$ and $\bar{v}^h := \arg\min_{v^h \in V^r} \varphi_h\left(v^h, v^{-h}|u\right)$.

3.2. Lagrange method and Tikhonov's regularization

Considering that the loss functions for defenders and attackers admit being non-strictly convex, an equilibrium point in the followers game may not be unique. To provide the

uniqueness of an equilibrium let us associate problem (6) with the so-called regularized problem [13, 14], such that for $\delta > 0$ we have:

$$(u_\delta^*, v_\delta^*) \in \arg\min_{u\in U}\max_{v\in V}\left\{F_\delta(u,u^{-l}|v)|g_\delta(v,v^{-h}|u)\leq 0, f_\delta(u,u^{-l}|v)\leq 0\right\}$$
$$F_\delta(u,u^{-l}|v) := \sum_{l=1}^{n}\left[\psi_l(u^l,u^{-l}|v) - \psi_l\left(u^l,u^{-l}|v\right)\right] + \tfrac{\delta}{2}(\|u\|^2+\|v\|^2). \tag{7}$$

Now, the function $F_\delta(u,u^{-l}|v)$ is *strongly convex* if $\delta > 0$. The existence of the solution to problems (6) and (7) follows from Kakutani's fixed point theorem which is valid under accepted smoothness conditions. It is evident that, for $\delta = 0$, the problem (7) converts to problem (6).

The nonlinear programming problem (7) may be resolved by the Lagrange method implementation. To do this, consider the augmented Lagrange function

$$\mathcal{L}_\delta(u,u^{-l},v,v^{-h},\lambda,\theta) = (1+\theta)f_\delta(u,u^{-l}|v) + \lambda g_\delta(v,v^{-h}|u) - \tfrac{\delta}{2}\left(\lambda^2+\theta^2\right). \tag{8}$$

In view of the strict convexity of (8) for $\delta > 0$, there exists a $\lambda_\delta^* \geq 0$ such that the following saddle-point [24] inequalities hold:

$$\mathcal{L}_\delta(u_\delta^*,u_\delta^{-l*},v,v^{-h},\lambda,\theta) \leq \mathcal{L}_\delta(u_\delta^*,u_\delta^{-[*},v_\delta^*,v_\delta^{-h*},\lambda_\delta^*,\theta_\delta^*) \leq \mathcal{L}_\delta(u,u^{-l},v_\delta^*,v_\delta^{-h*},\lambda_\delta^*,\theta_\delta^*). \tag{9}$$

The vector $(u_\delta^*,u_\delta^{-l*},v_\delta^*,v_\delta^{-h*},\lambda_\delta^*,\theta_\delta^*)$ can be interpreted as the δ approximation of the solution of problem (8). Thus, we can rewrite (7) using (8) as follows

$$(u_\delta^*,u_\delta^{-l*},v_\delta^*,v_\delta^{-h*},\lambda_\delta^*,\theta_\delta^*) = \arg\min_{u\in U}\max_{v\in V,\lambda\geq 0,\theta\geq 0}\{\mathcal{L}_\delta(u,u^{-l},v,v^{-h},\lambda,\theta)\}.$$

3.3. The Extraproximal method

In the *proximal format* (see, [3]) the relation (7) can be expressed as

$$\begin{aligned}
\lambda_\delta^* &= \arg\max_{\lambda\geq 0}\left\{-\tfrac{1}{2}\|\lambda-\lambda_\delta^*\|^2 + \gamma\mathcal{L}_\delta(u_\delta^*,u_\delta^{-l*},v_\delta^*,v_\delta^{-h*},\lambda,\theta_\delta^*)\right\}\\
\theta_\delta^* &= \arg\max_{\theta\geq 0}\left\{-\tfrac{1}{2}\|\theta-\theta_\delta^*\|^2 + \gamma\mathcal{L}_\delta(u_\delta^*,u_\delta^{-l*},v_\delta^*,v_\delta^{-h*},\lambda_\delta^*,\theta)\right\}\\
u_\delta^* &= \arg\min_{u\in U}\left\{\tfrac{1}{2}\|u-u_\delta^*\|^2 + \gamma\mathcal{L}_\delta(u,u_\delta^{-l*},v_\delta^*,v_\delta^{-h*},\lambda_\delta^*,\theta_\delta^*)\right\}\\
u_\delta^{-l*} &= \arg\min_{u^{-l}\in U^{-l}}\left\{\tfrac{1}{2}\|u^{-l}-u_\delta^{-l*}\|^2 + \gamma\mathcal{L}_\delta(u_\delta^*,u^{-l},v_\delta^*,v_\delta^{-h*},\lambda_\delta^*,\theta_\delta^*)\right\}\\
v_\delta^* &= \arg\max_{v\in V}\left\{-\tfrac{1}{2}\|v-v_\delta^*\|^2 + \gamma\mathcal{L}_\delta(u_\delta^*,u_\delta^{-l*},v,v_\delta^{-h*},\lambda_\delta^*,\theta_\delta^*)\right\}\\
v_\delta^{-h*} &= \arg\max_{v^{-h}\in V^{-h}}\left\{-\tfrac{1}{2}\|v^{-h}-v_\delta^{-h*}\|^2 + \gamma\mathcal{L}_\delta(u_\delta^*,u_\delta^{-l*},v_\delta^*,v^{-h},\lambda_\delta^*,\theta_\delta^*)\right\}
\end{aligned} \tag{10}$$

where the solutions u_δ^*, u_δ^{-l*}, v_δ^*, v_δ^{-h*} and λ_δ^* depend on the small parameters $\delta, \gamma > 0$. The parameter γ is step decreasing and controls the extent to which the proximal operator maps points towards the minimum of functional.

The *Extraproximal Method* for the conditional optimization problems (7) was suggested in [3] and applied for Markov chains models in [30]. We design the method for the static Stackelberg-Nash game in a general format. The general format iterative version ($t = 0, 1, \ldots$) of the extraproximal method with some fixed admissible initial values ($u_0 \in U, u_0^{-l} \in U^{-l}$, $v_0 \in V$, $v_0^{-h} \in V^{-h}$, $\lambda_0 \geq 0$, and $\theta_0 \geq 0$) is as follows

1. The *first half-step* (prediction):

$$\begin{aligned}
\overline{\lambda}_t &= \arg\min_{\lambda\geq 0}\left\{\tfrac{1}{2}\|\lambda-\lambda_t\|^2-\gamma\mathcal{L}_\delta(u_t,u_t^{-l},v_t,v_t^{-h},\lambda,\overline{\theta}_t)\right\}\\
\bar{\theta}_t &= \arg\min_{\theta\geq 0}\left\{\tfrac{1}{2}\|\lambda-\lambda_t\|^2-\gamma\mathcal{L}_\delta(u_t,u_t^{-l},v_t,v_t^{-h},\bar{\lambda}_t,\theta)\right\}\\
\overline{u}_t &= \arg\min_{u\in U}\left\{\tfrac{1}{2}\|u-u_t\|^2+\gamma\mathcal{L}_\delta(u,u_t^{-l},v_t,v_t^{-h},\bar{\lambda}_t,\overline{\theta}_t)\right\}\\
\overline{u}_t^{-l} &= \arg\min_{u^{-l}\in U^{-l}}\left\{\tfrac{1}{2}\|u^{-l}-u_t^{-l}\|^2+\gamma\mathcal{L}_\delta(u_t,u^{-l},v_t,v_t^{-h},\bar{\lambda}_t,\overline{\theta}_t)\right\}\\
\overline{v}_t &= \arg\min_{v\in V}\left\{\tfrac{1}{2}\|v-v_t\|^2-\gamma\mathcal{L}_\delta(u_t,u_t^{-l},v,v_t^{-h},\bar{\lambda}_t,\overline{\theta}_t)\right\}\\
\overline{v}_t^{-h} &= \arg\min_{v^{-h}\in V^{-h}}\{\tfrac{1}{2}\|v^{-h}-v_t^{-h}\|^2-\gamma\mathcal{L}_\delta(u_t,u_t^{-l},v_t,v^{-h},\bar{\lambda}_t,\overline{\theta}_t)\}.
\end{aligned} \tag{11}$$

2. The *second* (basic) half-step

$$\begin{aligned}
\lambda_{t+1} &= \arg\min_{\lambda\geq 0}\left\{\tfrac{1}{2}\|\lambda-\lambda_t\|^2-\gamma\mathcal{L}_\delta(\bar{u}_t,\bar{u}_t^{-l},\bar{v}_t,\bar{v}_t^{-h},\lambda,\overline{\theta}_t)\right\}\\
\theta_{t+1} &= \arg\min_{\theta\geq 0}\left\{\tfrac{1}{2}\|\lambda-\lambda_t\|^2-\gamma\mathcal{L}_\delta(u_t,u_t^{-l},v_t,\bar{v}_t^{-h},\bar{\lambda}_t,\theta)\right\}\\
u_{t+1} &= \arg\min_{u\in U}\left\{\tfrac{1}{2}\|u-u_t\|^2+\gamma\mathcal{L}_\delta(u_t,\bar{u}_t^{-l},\bar{v}_t,\bar{v}_t^{-l},\bar{\lambda}_t,\overline{\theta}_t)\right\}\\
u_{n+1}^{-l} &= \arg\min_{u^{-l}\in U^{-l}}\left\{\tfrac{1}{2}\|u^{-l}-u_t^{-l}\|^2+\gamma\mathcal{L}_\delta(\bar{u}_t,u^{-l},\bar{v}_t,\bar{v}_t^{-l},\bar{\lambda}_t,\overline{\theta}_t)\right\}\\
v_{n+1} &= \arg\min_{v\in V}\left\{\tfrac{1}{2}\|v-v_t\|^2-\gamma\mathcal{L}_\delta(\bar{u}_t,\bar{u}_t^{-l},v,\bar{v}_t^{-h},\bar{\lambda}_t,\overline{\theta}_t)\right\}\\
v_{n+1}^{-h} &= \arg\min_{v^{-h}\in V^{-h}}\{\tfrac{1}{2}\|v^{-h}-v_t^{-h}\|^2-\gamma\mathcal{L}_\delta(\bar{u}_t,\bar{u}_t^{-l},\bar{v}_t,v^{-h},\bar{\lambda}_t,\overline{\theta}_t)\}.
\end{aligned} \tag{12}$$

4. KULLBACK–LEIBLER RANDOM WALK

For the realization of the game we present a random walk model where the defenders try to catch the attackers while they both travel from state to state of an ergodic Markov

chain. The ergodicity of the Markov chain allowed players to jump arbitrarily between states: such a jump between states corresponds to a short-cut between two places.

During the realization of the round-based model players have no information about the movement decisions made by their opponents and thus do not know their position (state) in the Markov chain. The only interaction between players occurs when the game ends: a defender catches an attacker when the defender and the attacker are both located on the same state of the Markov chain. Therefore the movement decisions of both players do not depend on each other.

The goal of the defenders is to catch the attackers in as few rounds as possible, whereas the attackers aim to maximize the number of rounds until there are caught. In this setting we study defender strategies as well as attackers strategies on the expected number of rounds until the defenders catches the attackers. The strategies of the players are computed solving the Stackelberg game using the extraproximal method given by $d^{l*}_{k|i}$ and $d^{h*}_{k|i}$ respectively.

We introduce a random walk penalized by the Kullback–Leibler divergence [24] (or the relative entropy) between the strategies of the defenders and the attackers. We consider the distance to capture as $Lc(d^l_{k|i}||d^h_{k|i}) \triangleq \sum_{i=1}^N d^l_{k|i} \log \frac{d^l_{k|i}}{d^h_{k|i}}$ and the distance to escape as $Le(d^h_{k|i}||d^l_{k|i}) \triangleq \sum_{i=1}^N d^h_{k|i} \log \frac{d^h_{k|i}}{d^l_{k|i}}$. For determining the penalization of the defender we consider:

$$Le(d^h_{k|i}||d^l_{k|i}) > Lc(d^l_{k|i}||d^h_{k|i}). \tag{13}$$

for a fixed i. The interpretation is that the perception of the defender and the attacker is different.

For the defender we investigate two models: in one model the defender as usual travel from state to state of the ergodic Markov chain, and in the other model the control penalizes the defenders' deviation from the attackers' location.

The controlled state component is a standard Markov chain. The discrete steps are indexed by $n = 0, 1, \ldots$. We assume that the initial state at step 0 is fixed and denoted $s^l(0)$ for every leader and $s^h(0)$ for every follower. At the $t-$th step, the following happen:

Algorithm without penalization:
while(**not** capture condition (see below Eq. (14)))
    for every leader select random a state s^l from P^l_t
    for every follower select random a state s^h from P^h_t
    Set states s^l and s^h, and draw
end
Algorithm with penalization:
while(**not** capture condition see below (Eq. (14)))
    for every leader select random a state s^l from P^l_t
    for every follower select random a state s^h from P^h_t
    if($Le(d^h_{k|i}||d^l_{k|i}) > Lc(d^l_{k|i}||d^h_{k|i})$)
        select random a state s^l such that $s^l \neq s^h$.
    Set states s^l and s^h, and draw
end

Formally, given $(\Omega, \mathcal{F}, P)$ a probability space (Ω is a sample space; $\mathcal{F}$ is a -algebra of measurable subsets (events) of Ω; and P is a probability measure on $\mathcal{F}$ [24], let us introduce the *capture condition* at time t (defenders and attackers are located at the same state) as follows:

$$\sum_{j=1}^{N} \chi(\alpha : s^l(t) = s_j \wedge s^h(t) = s_j) = \sum_{j=1}^{N} \chi(\alpha : s^l(t) = s_j)\chi(\alpha : s^h(t) = s_j),\ \alpha \in \Omega,$$

where $\alpha \in \Omega$ is a trajectory.

Now, the capture event of all the attackers is given by

$$\sum_{l=1}^{n}\sum_{h=1}^{r}\sum_{j=1}^{N} \chi(\alpha : s^l(t) = s_j)\chi(\alpha : s^h(t) = s_j). \tag{14}$$

A fixed Markov transition matrix $\pi^l_{j|ik}$ is given. Then, the state transitions induced by the strategy $d^{l*}_{k|i}$ are governed by the conditional probability law

$$\Pi^{l*}_{ij}(d) = \sum_{k=1}^{M} \pi^l_{j|ik} d^{l*}_{k|i}.$$

Then, considering that

$$P\left\{\alpha : A \in \mathcal{F}\right\} = E\left\{\chi(\alpha : A \in \mathcal{F})\right\},$$

we have that the total probability P_t of converging to a state j at time n for all the defenders and attackers is given by

$$P_t = \sum_{l=1}^{n}\sum_{h=1}^{r}\sum_{j=1}^{N} P_t\left\{\alpha : s^l(t) = s_j\right\} P_t\left\{\alpha : s^h(t) = s_j\right\},$$

where

$$P^l_{j,n}\left\{\alpha : s^l(t) = s_j\right\} = \sum_{i=1}^{N}\sum_{k=1}^{M} \pi^l_{j|ik} d^{l*}_{k^*|i} P^l_{t-1}\left\{\alpha : s^l(t-1) = s_i\right\},$$

and

$$P^h_{j,t}\left\{\alpha : s^h(t) = s_j\right\} = \sum_{i=1}^{N}\sum_{k=1}^{M} \pi^h_{j|ik} d^{h*}_{k^*|i} P^h_{t-1}\left\{\alpha : s^h(t-1) = s_i\right\}\ .$$

Now, defining

$$\Pi^{l*}_{ij} = \sum_{k=1}^{M} \pi^l_{j|ik} d^{l*}_{k|i}, \quad \Pi^{h*}_{ij} = \sum_{k=1}^{M} \pi^h_{ij|k} d^{h*}_{k|i},\ .$$

we have

$$P^l_t\left\{\alpha : s(t)^l = s_j\right\} = \sum_{i=1}^{n} \Pi^{l*}_{ij} P^l_{t-1}\left\{\alpha : s^l(t-1) = s_i\right\},$$

and

$$P_t^h\left\{\alpha : s(t)^h = s_j\right\} = \sum_{i=1}^{N} \Pi_{ij}^{h*} P_{t-1}^h\left\{\alpha : s^h(t-1) = s_i\right\}.$$

Then, the probability P_t satisfies the following relation

$$P_t = \sum_{l=1}^{n}\sum_{h=1}^{r}\sum_{i=1}^{N}\sum_{i=1}^{N}\left[\Pi_{ij}^{l*} P_{t-1}^l\left\{s_i\right\}\right]\left[\Pi_{ij}^{h*} P_{t-1}^h\left\{s_i\right\}\right].$$

The probability of the state-vector P_t converges to a state j at time t by the Weierstrass Theorem. Indeed, let $\{X_t\}$ and $\{Y_t\}$ be two chains and let X and Y two random variables associated $P_t^l\left\{\alpha : s(t)^l = s_j\right\} = P_t^l\{X = s_j\}$ and $P_t^h\left\{\alpha : s(t)^h = s_j\right\} = P_t^h\{Y = s_j\}$, respectively. Because, $P_t^l\left\{\alpha : s_t^l = s_j\right\}$ and $P_t^h\left\{\alpha : s_t^h = s_j\right\}$ converge, then we will suppose that $\{X_t\}$ converges to X and $\{Y_t\}$ converges to Y in distribution when $t \to \infty$. Let $\left\{X_{\omega(t)}Y_{\alpha(t)}\right\}$ be a subsequence of $\{X_tY_t\}$. We need to show that a subsequence converges to XY. Since $\{X_t\}$ converges to X in probability, there exists ζ such that $\left\{X_{\omega(\zeta(t))}\right\}$ converges to X by the Weierstrass Theorem. As well as, $\{Y_t\}$ converges to Y in probability, there exists ξ such that $\left\{Y_{\omega(\zeta(\xi(t)))}\right\}$ converges to Y. Then, we have that $\left\{X_{\omega(\zeta(\xi(t)))}Y_{\omega(\zeta(\xi(t)))}\right\}$ converges to XY. As a result, the probability of the state-vector P_t converges. Then, the theorem is proved.

5. NUMERICAL EXAMPLE

We present an application for protecting a marine canal suggesting patrolling strategies to protect ports. The mission involves ensuring the safety and security of all passenger, cargo, and vessel operations. Given the particular variety of critical infrastructure that an adversary may attack within the port agencies conducts patrols to protect such infrastructure. Whereas attackers have the opportunity to observe patrol patterns, limited security resources imply that agencies patrols cannot be at every location any time. To support agencies in the process of patrolling resources allocation we employ the proposed Stackelberg-Nash game framework fixing two independents agencies as the defenders against two independent thieves (attackers) that conduct surveillance before potentially launching an attack. We consider five different ports as control points and two actions conceptualized as patrol and surveillance [28]. In surveillance the agencies conduct observations to gain information for particular purposes (that in some case can violate the privacy). It reserves the right to respond in problematic situations or risk appears taking place. It main goals are: a) acquire intelligence information (subject, criminal group, etc.), b) intercepting communications, etc. In patrol the agencies monitors a particular area, alert for suspicious behavior or other types of danger. For instance, a naval task force sailing in a strategic shipping lane. The agencies have special obligations to do: a) locate contraband or places of illegal activities, b) prevent a crime from occurring, etc. The output of the example is a schedule of patrols that includes what port to visit for each agency. The schedule is realized by handling a Kullback–Leibler divergence random walk approach where thieves are pursued by agencies and their detention is determined by a capture condition. The transition matrices, as well as, the cost matrices are empirically defined.

For the Agency 1 we have the transition matrices:

$$\pi^1_{j|i1} = \begin{pmatrix} 0.2464 & 0.2329 & 0.1234 & 0.1111 & 0.2862 \\ 0.4706 & 0.1626 & 0.0996 & 0.0711 & 0.1960 \\ 0.0849 & 0.3656 & 0.2388 & 0.2110 & 0.0997 \\ 0.1647 & 0.1514 & 0.2743 & 0.1847 & 0.2249 \\ 0.1740 & 0.1234 & 0.1499 & 0.2691 & 0.2835 \end{pmatrix}$$

$$\pi^1_{j|i2} = \begin{pmatrix} 0.1413 & 0.0950 & 0.2003 & 0.1243 & 0.4390 \\ 0.1874 & 0.1787 & 0.1502 & 0.2942 & 0.1894 \\ 0.2454 & 0.1733 & 0.0733 & 0.2287 & 0.2793 \\ 0.3365 & 0.1384 & 0.1422 & 0.1848 & 0.1982 \\ 0.2761 & 0.3698 & 0.1000 & 0.1390 & 0.1151 \end{pmatrix}$$

and the cost matrices are given by:

$$u^1_{ij1} = \begin{pmatrix} 4 & 82 & 63 & 59 & 94 \\ 14 & 23 & 1 & 68 & 28 \\ 24 & 16 & 12 & 52 & 3 \\ 10 & 17 & 70 & 69 & 56 \\ 16 & 17 & 24 & 31 & 22 \end{pmatrix} u^1_{ij2} = \begin{pmatrix} 94 & 19 & 36 & 26 & 20 \\ 32 & 79 & 9 & 79 & 47 \\ 34 & 11 & 6 & 41 & 14 \\ 78 & 24 & 28 & 9 & 72 \\ 13 & 67 & 13 & 79 & 18 \end{pmatrix}.$$

For the Agency 2 we have the transition matrices:

$$\pi^2_{j|i1} = \begin{pmatrix} 0.2082 & 0.3760 & 0.1896 & 0.1202 & 0.1061 \\ 0.4259 & 0.1001 & 0.1598 & 0.1549 & 0.1593 \\ 0.3533 & 0.1836 & 0.2729 & 0.0747 & 0.1156 \\ 0.0884 & 0.3181 & 0.3800 & 0.0812 & 0.1324 \\ 0.0783 & 0.1472 & 0.1256 & 0.1694 & 0.4796 \end{pmatrix}$$

$$\pi^2_{j|i2} = \begin{pmatrix} 0.2706 & 0.2719 & 0.2210 & 0.1360 & 0.1004 \\ 0.1628 & 0.1912 & 0.2130 & 0.1275 & 0.3054 \\ 0.1486 & 0.1096 & 0.1501 & 0.3140 & 0.2777 \\ 0.1503 & 0.1042 & 0.2647 & 0.2638 & 0.2170 \\ 0.2587 & 0.1187 & 0.3104 & 0.1279 & 0.1843 \end{pmatrix}$$

and the cost matrices are given by:

$$u^2_{ij1} = \begin{pmatrix} 29 & 88 & 12 & 79 & 33 \\ 15 & 24 & 26 & 11 & 97 \\ 17 & 19 & 24 & 4 & 19 \\ 70 & 11 & 29 & 41 & 24 \\ 17 & 97 & 30 & 27 & 25 \end{pmatrix} u^2_{ij2} = \begin{pmatrix} 62 & 13 & 24 & 74 & 68 \\ 37 & 28 & 28 & 6 & 2 \\ 134 & 1 & 44 & 5 & 26 \\ 20 & 12 & 33 & 12 & 21 \\ 80 & 32 & 231 & 16 & 19 \end{pmatrix}.$$

For the Thief 3 we have the transition matrices:

$$\pi^3_{j|i1} = \begin{pmatrix} 0.1414 & 0.2632 & 0.3927 & 0.1097 & 0.0930 \\ 0.2042 & 0.1825 & 0.1275 & 0.3735 & 0.1123 \\ 0.2082 & 0.4375 & 0.1532 & 0.1312 & 0.0700 \\ 0.4116 & 0.2252 & 0.0882 & 0.1388 & 0.1362 \\ 0.1351 & 0.2024 & 0.1243 & 0.3610 & 0.1772 \end{pmatrix}$$

$$\pi^3_{j|i2} = \begin{pmatrix} 0.1333 & 0.3410 & 0.1076 & 0.2790 & 0.1391 \\ 0.1994 & 0.0775 & 0.1187 & 0.4362 & 0.1682 \\ 0.2444 & 0.1112 & 0.3544 & 0.2146 & 0.0754 \\ 0.2045 & 0.2076 & 0.1967 & 0.2465 & 0.1447 \\ 0.1767 & 0.1285 & 0.1280 & 0.3548 & 0.2119 \end{pmatrix}$$

and the cost matrices are given by:

$$u^3_{i,j|1} = \begin{pmatrix} 6 & 1 & 6 & 46 & 18 \\ 69 & 2 & 52 & 44 & 40 \\ 5 & 82 & 8 & 3 & 4 \\ 8 & 3 & 65 & 9 & 81 \\ 5 & 15 & 1 & 14 & 7 \end{pmatrix} u^3_{i,j|2} = \begin{pmatrix} 40 & 30 & 11 & 6 & 3 \\ 53 & 44 & 38 & 3 & 5 \\ 42 & 2 & 20 & 6 & 9 \\ 66 & 9 & 9 & 74 & 42 \\ 3 & 17 & 34 & 27 & 9 \end{pmatrix}.$$

For the Thief 4 we have the transition matrices:

$$\pi^4_{j|i1} = \begin{pmatrix} 0.3083 & 0.1398 & 0.2809 & 0.0975 & 0.1735 \\ 0.0845 & 0.1905 & 0.2199 & 0.3059 & 0.1993 \\ 0.1984 & 0.0980 & 0.2365 & 0.3485 & 0.1186 \\ 0.1871 & 0.1293 & 0.3087 & 0.1339 & 0.2410 \\ 0.3270 & 0.0741 & 0.2641 & 0.1064 & 0.2285 \end{pmatrix}$$

$$\pi^4_{j|i2} = \begin{pmatrix} 0.1266 & 0.4097 & 0.1374 & 0.0969 & 0.2294 \\ 0.4062 & 0.1854 & 0.2306 & 0.1099 & 0.0679 \\ 0.1198 & 0.1316 & 0.2961 & 0.3518 & 0.1008 \\ 0.1163 & 0.1488 & 0.2993 & 0.3431 & 0.0926 \\ 0.1109 & 0.2108 & 0.1331 & 0.3284 & 0.2167 \end{pmatrix}$$

and the cost matrices are given by:

$$u^4_{ij1} = \begin{pmatrix} 61 & 4 & 7 & 74 & 19 \\ 11 & 36 & 13 & 3 & 6 \\ 14 & 6 & 17 & 17 & 39 \\ 18 & 20 & 34 & 32 & 6 \\ 10 & 3 & 3 & 2 & 7 \end{pmatrix} u^4_{ij2} = \begin{pmatrix} 25 & 1 & 7 & 2 & 5 \\ 11 & 59 & 7 & 16 & 6 \\ 46 & 1 & 9 & 14 & 6 \\ 7 & 54 & 7 & 53 & 3 \\ 17 & 1 & 22 & 19 & 6 \end{pmatrix}.$$

Then, fixing $\gamma_0 = 0.024$ and $\delta_0 = 5.0 \times 10^{-3}$ we have that the resulting equilibrium point of the Stackelberg security game is given by:

$$d^{1*}_{k|i} = \begin{pmatrix} 0.5845 & 0.4155 \\ 0.7237 & 0.2763 \\ 0.9664 & 0.0336 \\ 0.0308 & 0.9692 \\ 0.4245 & 0.5755 \end{pmatrix} \quad d^{2*}_{k|i} = \begin{pmatrix} 0.6132 & 0.3868 \\ 0.7607 & 0.2393 \\ 0.9788 & 0.0212 \\ 0.0360 & 0.9640 \\ 0.1135 & 0.8865 \end{pmatrix}$$

$$d^{3*}_{k|i} = \begin{pmatrix} 0.4860 & 0.5140 \\ 0.5553 & 0.4447 \\ 0.6735 & 0.3265 \\ 0.2198 & 0.7802 \\ 0.7252 & 0.2748 \end{pmatrix} \quad d^{4*}_{k|i} = \begin{pmatrix} 0.6791 & 0.3209 \\ 0.3174 & 0.6826 \\ 0.4775 & 0.5225 \\ 0.3947 & 0.6053 \\ 0.7919 & 0.2081 \end{pmatrix}.$$

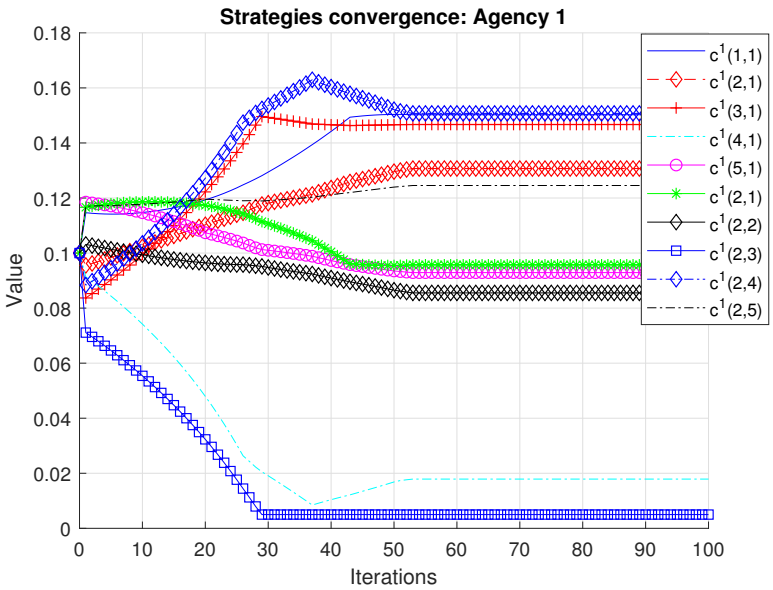

Fig. 1. Convergence for the Agency 1.

Figures 1 and 2 show the convergence of the strategies of the Agency 1 and Agency 2. The Figures 3 and 4 show the convergence of the strategies of the Thief 3 and Thief 4.

The realization of the random walk is a round-based model where the defenders catch the attackers while they both travel from state to state of an ergodic MDP. The realization of the random walk without penalization is shown in Figure 5. This walk can be described as follows. In the course of the random walk without penalization attackers and defenders have no information about the movement decisions made by the other players and then they do not know their position in the MDP. The only interaction between players occurs when the game finishes. In this case Thief 2 is captured at state 5 by agency 1 after two iterations, and Thief 1 is captured at state 1 by agency 1 and agency 2 in cooperation after 14 steps and the realization is over.

On the other hand, the realization of the random walk employing the algorithm with penalization is shown in Figure 6. During the random walk with penalization agencies and thieves have full information about the movement decisions made by the other players. In this case Thief 1 is captured at state 1 after 36 steps and Thief 2 is captured at state 2 after 40 steps and the realization is over. This behavior is in correspondence with the penalization imposed by the selection of the actions of the strategies.

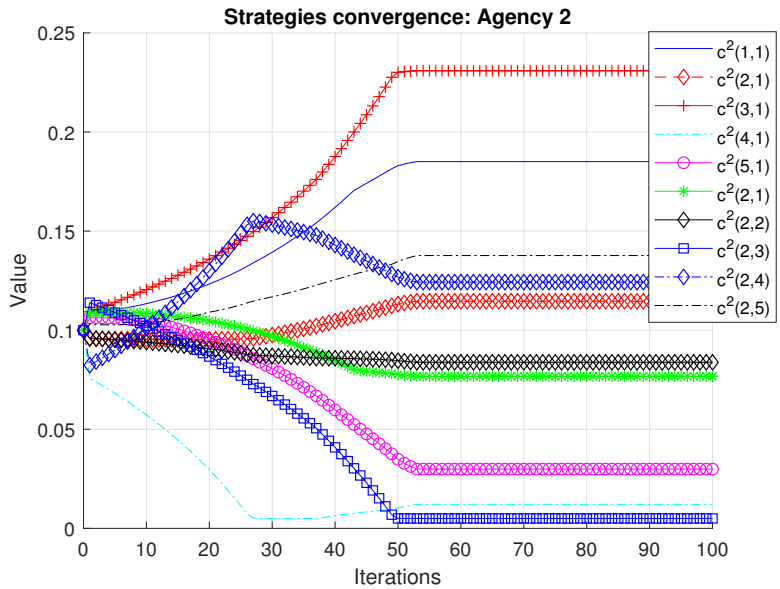

Fig. 2. Convergence for the Agency 2.

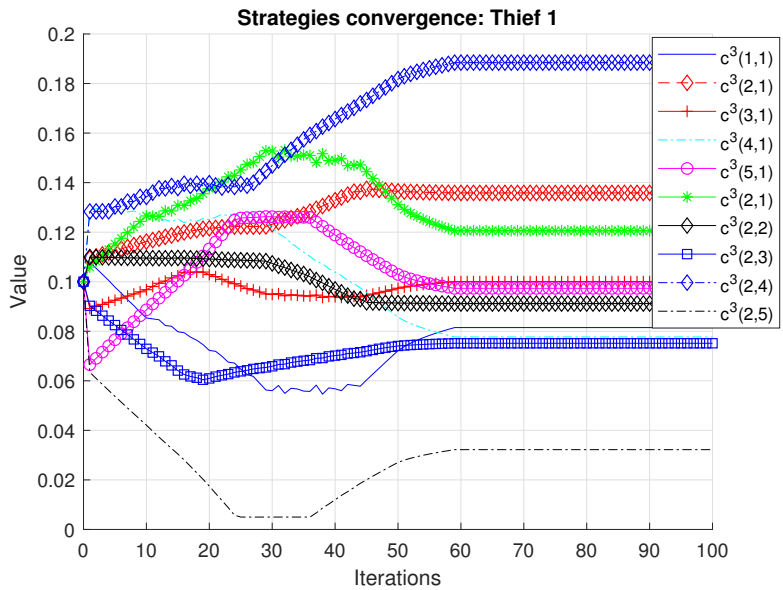

Fig. 3. Convergence for the Thief 3.

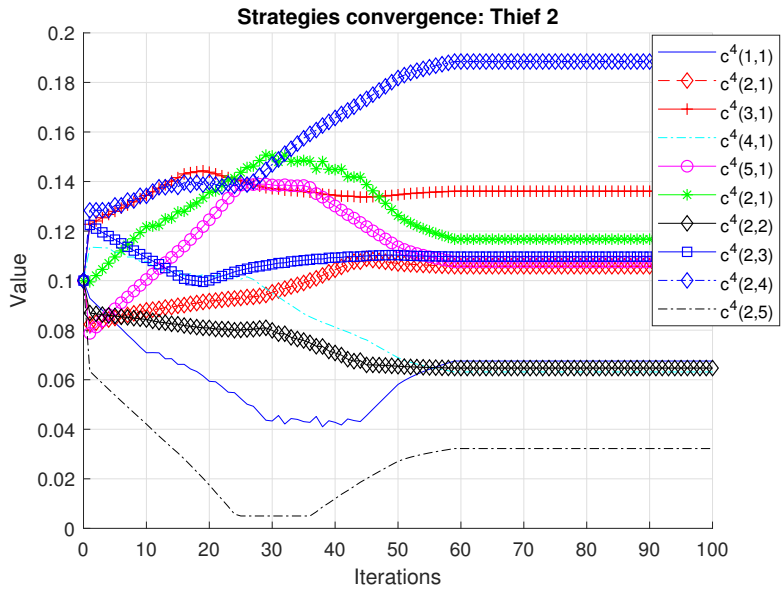

Fig. 4. Convergence for the Thief 4.

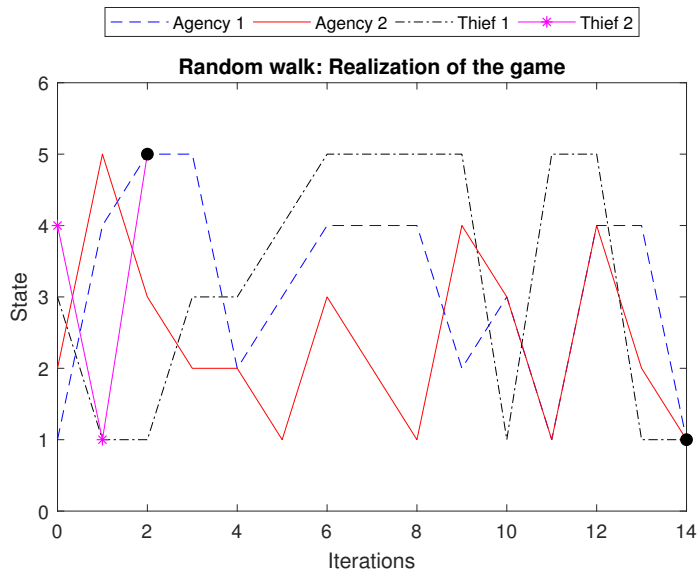

Fig. 5. Random walk realization without penalization.

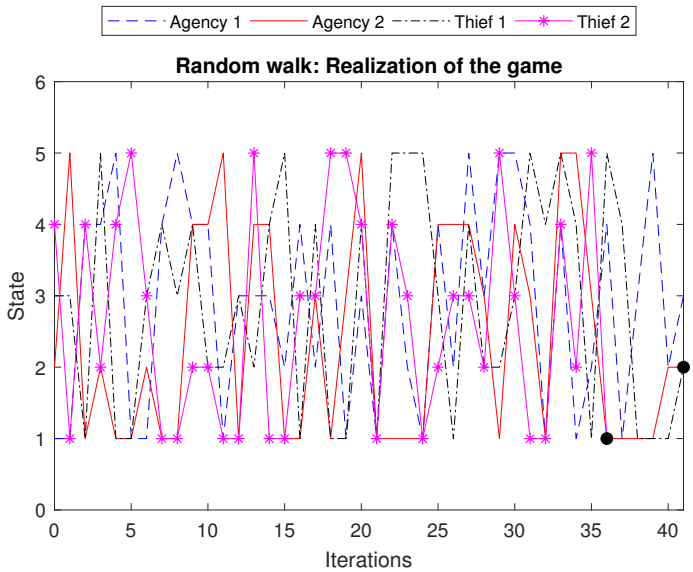

Fig. 6. Random walk realization with penalization.

6. CONCLUSION

The problem studied in this paper combines aspects of both game theory and stochastic control using several ideas and techniques from the theory of MDPs with average cost criterion for solving the game and some new results concerning optimal policies for MDPs with KL for selecting the optimal control law. In particular, the framework presented in this work computes the Stackelberg/Nash equilibrium for multiple players in non-cooperative Stackelberg security games presenting a real-world solution to the problem. For solving the problem, we used the extraproximal method, within which we explicitly compute the unique Stackelberg/Nash equilibrium of the game by specifying a natural model employing the Lagrange method and introducing the Tikhonov regularization method. We introduced a random walk based on the Kulback-Leibler divergence studying two models for the defenders: in one model the defender travel from state to state of the ergodic MDP, and in the other model the control penalizes the defenders' deviation from the attackers' location. We proved that the synchronization of the random walk of defenders and attackers converge in probability to the product of the individual probabilities.

(Received February 18, 2017)

REFERENCES

[1] N. Agmon, G. A. Kaminka, and S. Kraus: Multi-robot adversarial patrolling: facing a full-knowledge opponent. J. Artif. Intell. Res. *42* (2011), 1, 887–916.

[2] S. Albarran and J. B. Clempner: A Stackelberg security Markov game based on partial information for strategic decision making against unexpected attacks. Engrg. Appl. Artif. Intell. *81* (2019), 408–419. DOI:10.1016/j.engappai.2019.03.010

[3] A. S. Antipin: An extraproximal method for solving equilibrium programming problems and games. Comput. Mathematics and Math. Phys. *45* (2005), 11, 1893–1914.

[4] A. Blum, N and; Haghtalab, and A. D. Procaccia: Lazy defenders are almost optimal against diligent attackers. In: Proc. 28th AAAI Conference on Artificial Intelligence, Québec 2014, pp. 573–579.

[5] J. B. Clempner: A continuous-time Markov Stackelberg security game approach for reasoning about real patrol strategies. Int. J. Control *91* (2018), 2494–2510. DOI:10.1080/00207179.2017.1371853

[6] J. B. Clempner and A. S. Poznyak: Simple computing of the customer lifetime value: A fixed local-optimal policy approach. J. Systems Sci. Systems Engrg. *23* (2014), 4, 439–459. DOI:10.1007/s11518-014-5260-y

[7] J. B. Clempner and A. S. Poznyak: Stackelberg security games: Computing the shortest-path equilibrium. Expert Syst. Appl. *42* (2015), 8, 3967–3979. DOI:10.1016/j.eswa.2014.12.034

[8] J. B. Clempner and A. S. Poznyak: Analyzing an optimistic attitude for the leader firm in duopoly models: A strong Stackelberg equilibrium based on a lyapunov game theory approach. Econ. Comput. Econ. Cybern. Stud. Res. *4* (2016), 50, 41–60.

[9] J. B. Clempner and A. S. Poznyak: Conforming coalitions in Stackelberg security games: Setting max cooperative defenders vs. non-cooperative attackers. Appl. Soft Comput. *47* (2016), 1–11. DOI:10.1016/j.asoc.2016.05.037

[10] J. B. Clempner and A. S. Poznyak: Conforming coalitions in Stackelberg security games: Setting max cooperative defenders vs. non-cooperative attackers. Appl. Soft Comput. *47* (2016), 1–11. DOI:10.1016/j.asoc.2016.05.037

[11] J. B. Clempner and A. S. Poznyak: Convergence analysis for pure and stationary strategies in repeated potential games: Nash, lyapunov and correlated equilibria. Expert Systems Appl. *46* (2016), 474–484. DOI:10.1016/j.eswa.2015.11.006

[12] J. B. Clempner and A. S. Poznyak: Using the extraproximal method for computing the shortest-path mixed lyapunov equilibrium in Stackelberg security games. Math. Comput. Simul. *138* (2017), 14–30. DOI:10.1016/j.matcom.2016.12.010

[13] J. B. Clempner and A. S. Poznyak: A Tikhonov regularization parameter approach for solving lagrange constrained optimization problems. Engrg. Optim. *50* (2018), 11, 1996–2012. DOI:10.1080/0305215x.2017.1418866

[14] J. B. Clempner and A. S. Poznyak: A Tikhonov regularized penalty function approach for solving polylinear programming problems. J. Comput. Appl. Math. *328* (2018), 267–286. DOI:10.1016/j.cam.2017.07.032

[15] V. Conitzer and T. Sandholm: Computing the optimal strategy to commit to. In: Seventh ACM Conference on Electronic Commerce, Ann Arbor 2006, pp. 82–90.

[16] D. Guerrero, A. A. Carsteanu, R. Huerta, and J. B. Clempner: An iterative method for solving Stackelberg security games: A Markov games approach. In: 14th International Conference on Electrical Engineering, Computing Science and Automatic Control, Mexico City 2017, pp. 1–6. DOI:10.1109/iceee.2017.8108857

[17] D. Guerrero, A. A. Carsteanu, R. Huerta, and J. B. Clempner: Solving Stackelberg security Markov games employing the bargaining nash approach: Convergence analysis. Computers Security *74* (2018), 240–257. DOI:10.1016/j.cose.2018.01.005

[18] M. Jain, E. Kardes, C. Kiekintveld, F. Ordoñez, and M. Tambe: Security games with arbitrary schedules: A branch and price approach. In: Proc. National Conference on Artificial Intelligence (AAAI), Atlanta 2010. DOI:10.1016/j.cose.2018.01.005

[19] C. Kiekintveld, M. Jain, J. Tsai, J. Pita, F. Ordñez, and M. Tambe: Computing optimal randomized resource allocations for massive security games. In: Proc. Eighth International Conference on Autonomous Agents and Multiagent Systems, volume 1, Budapest 2009, pp. 689–696. DOI:10.1017/cbo9780511973031.008

[20] D. Korzhyk, Z. Yin, C. Kiekintveld, V. Conitzer, and M. Tambe: Stackelberg vs. nash in security games: An extended investigation of interchangeability equivalence, and uniqueness. J. Artif. Intell. Res. *41* (2011), 297–327. DOI:10.1613/jair.3269

[21] J. Letchford, L. MacDermed, V. Conitzer, R. Parr, and C. L. Isbell: Computing optimal strategies to commit to in stochastic games. In: Proc. Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI), Toronto 2012, pp. 1380–1386. DOI:10.1145/2509002.2509011

[22] J. Letchford and Y. Vorobeychik: Optimal interdiction of attack plans. In: Proc. Twelfth International Conference of Autonomous Agents and Multi-agent Systems (AAMAS), Saint Paul 2013, pp. 199–206.

[23] P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordoñez, and S. Kraus: Playing games with security: An efficient exact algorithm for bayesian stackelberg games. In: Proc. Seventh International Conference on Autonomous Agents and Multiagent Systems, Estoril 2008, pp. 895–902.

[24] A. S. Poznyak: Advance Mathematical Tools for Automatic Control Engineers. Vol. 2 Deterministic Techniques. Elsevier, Amsterdam 2008. DOI:10.1016/b978-008044674-5.50015-8

[25] A. S. Poznyak, K. Najim, and E. Gomez-Ramirez: Self-learning Control of Finite Markov Chains. Marcel Dekker, New York 2000.

[26] M. Salgado and J. B. Clempner: Measuring the emotional distance using game theory via reinforcement learning: A kullback-leibler divergence approach. Expert Systems Appl. *97* (2018), 266–275. DOI:10.1016/j.eswa.2017.12.036

[27] E. Shieh, B. An, R. Yang, M. Tambe, C. Baldwin, J. DiRenzo, B. Maule, and G. Meyer: Protect: A deployed game theoretic system to protect the ports of the united states. In: Proc. 11th International Conference on Autonomous Agents and Multiagent Systems, 2012. DOI:10.1609/aimag.v33i4.2401

[28] M. Skerker: Binary Bullets: The Ethics of Cyberwarfare, chapter Moral Concerns with Cyberspionage: Automated Keyword Searches and Data Mining, pp. 251–276. Oxford University Press, NY 2016.

[29] C. Solis, J. B. Clempner, and A. S. Poznyak: Modeling multi-leader-follower non-cooperative Stackelberg games. Cybernetics Systems *47* (2016), 8, 650–673. DOI:10.1080/01969722.2016.1232121

[30] K. K. Trejo, J. B. Clempner, and A. S. Poznyak: Computing the stackelberg/nash equilibria using the extraproximal method: Convergence analysis and implementation details for Markov chains games. Int. J. Appl. Math. Computer Sci. *25* (2015), 2, 337-351. DOI:10.1515/amcs-2015-0026

[31] K. K. Trejo, J. B. Clempner, and A. S. Poznyak: A Stackelberg security game with random strategies based on the extraproximal theoretic approach. Engrg. Appl. Artif. Intell. *37* (2015), 145–153. DOI:10.1016/j.engappai.2014.09.002

[32] K. K. Trejo, J. B. Clempner, and A. S. Poznyak: Adapting strategies to dynamic environments in controllable stackelberg security games. In: IEEE 55th Conference on Decision and Control (CDC), Las Vegas 2016, pp. 5484–5489. DOI:10.1109/cdc.2016.7799111

[33] K. K. Trejo, J. B. Clempner, and A. S. Poznyak: An optimal strong equilibrium solution for cooperative multi-leader-follower Stackelberg Markov chains games. Kybernetika *52* (2016), 2, 258–279. DOI:10.14736/kyb-2016-2-0258

[34] K. K. Trejo, J. B. Clempner, and A. S. Poznyak: Computing the lp-strong nash equilibrium for Markov chains games. Appl. Math. Modell. *41* (2017), 399–418. DOI:10.1016/j.apm.2016.09.001

[35] K. K. Trejo, J. B. Clempner, and A. S. Poznyak: Adapting attackers and defenders preferred strategies: A reinforcement learning approach in stackelberg security games. J. Comput. System Sci. *95* (2018), 35–54. DOI:10.1016/j.jcss.2017.12.004

[36] R. Yang, C. Kiekintveld, F. Ordonez, M. Tambe, and R. John: Improving resource allocation strategy against human adversaries in security games. In: Proc. International Joint Conference on Artificial Intelligence (IJCAI), Barcelona 2011, pp. 458–464.

[37] Z. Yin, M. Jain, M. Tambe, and F. Ordonez: Risk-averse strategies for security games with execution and observational uncertainty. In: Proc. AAAI Conference on Artificial Intelligence (AAAI), San Francisco 2011, pp. 758–763.

[38] Z. Yin and M. Tambe: A unified method for handling discrete and continuous uncertainty in bayesian stackelberg games. In: Proc. Eleventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Valencia 2012, pp. 234–242.

César U. S. Solis, Department of Control Automatics, Center for Research and Advanced Studies, Av. IPN 2508, Col. San Pedro Zacatenco, 07360 Mexico City. Mexico.
e-mail: csolis@ctrl.cinvestav.mx

Julio B. Clempner, Escuela Superior de Fásica y Matemáticas, (School of Physics and Mathematics), Instituto Politécnico Nacional (National Polytechnic Institute), Building 9, Av. Instituto Politécnico Nacional, San Pedro Zacatenco, 07738, Gustavo A. Madero, Mexico City. Mexico.
e-mail: julio@clempner.name

Alexander S. Poznyak, Department of Control Automatics, Center for Research and Advanced Studies Av. IPN 2508, Col. San Pedro Zacatenco, 07360 Mexico City. Mexico.
e-mail: apoznyak@ctrl.cinvestav.mx