

# Double-Oracle Sampling Method for Stackelberg Equilibrium Approximation in General-Sum Extensive-Form Games

Jan Karwowski, Jacek Mańdziuk

Warsaw University of Technology  
Koszykowa 75, 00-662 Warsaw, Poland  
{jan.karwowski, j.mandziuk}@mini.pw.edu.pl

## Abstract

The paper presents a new method for approximating Strong Stackelberg Equilibrium in general-sum sequential games with imperfect information and perfect recall. The proposed approach is generic as it does not rely on any specific properties of a particular game model. The method is based on iterative interleaving of the two following phases: (1) guided Monte Carlo Tree Search sampling of the Follower’s strategy space and (2) building the Leader’s behavior strategy tree for which the sampled Follower’s strategy is an optimal response. The above solution scheme is evaluated with respect to expected Leader’s utility and time requirements on three sets of interception games with variable characteristics, played on graphs. A comparison with three state-of-the-art MILP/LP-based methods shows that in vast majority of test cases proposed simulation-based approach leads to optimal Leader’s strategies, while excelling the competitive methods in terms of better time scalability and lower memory requirements.

## 1 Introduction

Stackelberg Equilibrium (SE) (Leitmann 1978) defines equilibrium profile for two-player asymmetric games. One player – the Leader – commits to a certain strategy and the other player – the Follower – defines his/her strategy being aware of the Leader’s commitment. The notion of SE, which originated in the field of economy, gained momentum in recent decade thanks to intensive research on Security Games (Sinha et al. 2018) which often use Stackelberg Game (SG) to model interactions between a defender (playing the role of a Leader) and an attacker (being a Follower). We consider the Strong Stackelberg Equilibrium (SSE) (Leitmann 1978) in which (additionally to SE) the Follower breaks ties in favor of the Leader when calculating the optimal response.

Majority of contemporary SG research is focused on developing effective methods for specific games, e.g. (Brazdil, Kucera, and Rehak 2018; Schlenker et al. 2016; Basilico, Gatti, and Amigoni 2012; Wang et al. 2018; Johnson, Fang, and Tambe 2012) and there are just a few works related to finding SE in the case of general SG models.

**Contribution** The main contribution of this paper is a method for approximating SE in a broad and general genre of sequential general-sum imperfect-information games<sup>1</sup>, inspired by a double-oracle approach (Bosansky et al. 2014; Jain et al. 2011).

Despite being rooted in the double-oracle framework, the proposed method presents an entirely different operational principle than those of (Bosansky et al. 2014; Jain et al. 2011) as it relies on iterative Monte Carlo Tree Search (MCTS) (Browne et al. 2012) sampling of the Follower’s strategy alternated with an adjustment of the Leader’s behavior strategy represented in the form of a tree.

Our method is experimentally proven to yield close-to-optimal defender’s strategies while scaling better in time and memory usage than competitive MILP (Mixed Integer Linear Program) based approaches.

To the best of our knowledge there is only one other approach that utilizes MCTS method to solve general-sum extensive-form SGs (Karwowski and Mańdziuk 2015; 2016; 2019a) which, however, adopts a different protocol and relies on iterative adjustment of the Leader’s strategy by means of direct MCTS sampling against gradually changing Follower’s strategy. This method, though, could not be applied to solve games with complex information set (IS) structures, e.g. Search Games (Bosansky and Cermak 2015) considered in this paper.

**Related Work** In the literature, the problem of finding SE is usually considered in the context of some particular game model and therefore majority of proposed approaches are model-specific and cannot be straightforwardly applied to other kinds of SGs. On a general note, existing solution methods usually adapt and tune one of the following well-established techniques: *column and constraint generation* – e.g. (Wang et al. 2018; Jain et al. 2010); *marginal and compact strategies* – exploiting a particular structure of a game and its payoffs, e.g. (Kiekintveld et al. 2009; Schlenker et al. 2016); or *game abstraction* – e.g. (Wang et al. 2018; Basak et al. 2016). Utilization of these techniques requires tailoring a solution method to characteristic game proper-

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>A source code of the presented method is available on our SGs project website: <https://sg.mini.pw.edu.pl/sgHome/src.html>.

ties, what leads to a highly efficient, though game-dedicated algorithm.

An efficient exact approach to generic sequential general-sum SGs was proposed in (Bosansky and Cermak 2015) where the authors considered a sequence-form representation of a sequential game to improve scalability of a corresponding MILP. Another powerful general approach, introduced by Cermak et al. (2016) starts off with finding Stackelberg Extensive Form Correlated Equilibrium of a game using MILP and then restricts it iteratively until the obtained strategy profile corresponds to SE. Yet another general approach to extensive-form games (Cerny, Bosansky, and Kiekintveld 2018) starts from a smaller (restricted) game and gradually expands the game tree to compute the SSE. These three state-of-the-art generic methods are used as reference points in experimental evaluation of the approximate approach proposed in this paper.

Our method (referred to as *O2UCT* – double-oracle UCT sampling) relies on a guided sampling of the Follower’s strategy space and finding a feasible Leader’s strategy using double-oracle method, and does not involve solving Linear Program (LP) of any kind. Application of *O2UCT* leads to scalability performance boost similar to that of using column and constraint generation method in LP/MILP, albeit *with no direct reference to any specific game model properties, in the solution method.*

## 2 Imperfect Information Stackelberg Games

A sequential non-zero-sum game with imperfect information can be defined using an *extensive form*. An Extensive Form (EF) game is an 8-tuple  $G = (\mathcal{N}, \mathcal{S}, \mathcal{Z}, \rho, \mathcal{A}, u, \mathcal{T}, \mathcal{I})$ , where  $\mathcal{N} = \{L, F\}$  is a set of players (the Leader and the Follower in the case of a two player SG).  $\mathcal{S}$  and  $\mathcal{Z}$  are sets of non-terminal and terminal game states, resp.,  $\rho : \mathcal{S} \rightarrow \mathcal{N}$  is a function defining which player acts in a given state.  $\mathcal{A} = \bigcup_{s \in \mathcal{S}} A_s$  is a family of sets  $A_s$ , where  $A_s$  is a set of all actions available to an active player in state  $s$ .  $u : \{\mathcal{Z} \times \mathcal{N}\} \rightarrow [0, 1]$  is a utility function which provides utilities for all players in terminal states.  $\mathcal{T}$  is a set of transition functions  $T_s : A_s \rightarrow \mathcal{S} \cup \mathcal{Z}$  such that for every  $s \in \mathcal{S}$ ,  $T_s(a)$  is a state resulting from playing action  $a$  in state  $s$ .  $\mathcal{I}$  is a family of information sets  $I_k \subseteq \mathcal{S}$  which satisfies the standard definition (Kuhn 1950).

Moreover, all considered games have *perfect recall* property, i.e. an active player is fully aware of his/her past actions and ISs he/she encountered before reaching the current state.

Let’s denote by  $A_{I_k}$  a set of actions available in a given information set  $I_k$  and by  $\mathcal{I}^n$  a family of information sets in which player  $n$  is an active player. A *pure strategy* of player  $n$  (denoted by  $\pi_n$ ) is an assignment of one of the allowed actions per each IS in  $\mathcal{I}^n$ . A *mixed strategy*  $\delta_n$  is a probability distribution over all possible pure strategies  $\pi_n$  of player  $n$ . In EF games a *behavior strategy* is additionally defined, as a function that assigns a probability distribution over all available actions to each IS. In *perfect recall* games mixed and behavior strategies are pairwise equivalent (Kuhn 1950) and therefore in the remainder of the paper we will denote behavior strategies of player  $n$  by  $\delta_n$  and treat them equivalently to mixed strategies.

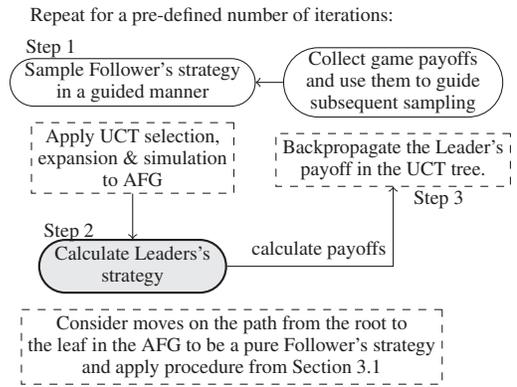


Figure 1: An outline of the *O2UCT* method. Oval frames present the method’s general idea, while rectangular dashed frames summarize particular realization of each step proposed in this work. Implementation of Step 2, which is the most challenging part of *O2UCT*, is described in more detail in Fig. 4.

We will use the notation  $\mathbb{E}U_{\delta_F, \delta_L}^n$  to denote the expected utility of player  $n$  ( $L$  – Leader,  $F$  – Follower) when the Leader and the Follower play strategies  $\delta_L$  and  $\delta_F$ , respectively. An index referring to the Leader’s strategy will be omitted in the contexts in which it does not lead to misunderstandings.

## 3 Double-oracle sampling method (*O2UCT*) for SE approximation

In order to find SE in a perfect recall imperfect-information deterministic multi-act general-sum game the following iterative procedure, depicted in Fig. 1, is applied. In each iteration, in the first step the Follower’s strategy is sampled with a method capable of using the results from previous iterations to guide subsequent sampling. Next, a method for finding the Leader’s strategy, **for which the just-sampled Follower’s strategy is the optimal response**, is applied. In the third step utility values corresponding to obtained strategy profile are collected to adjust the guided sampling procedure (Step 1) in the next iteration.

A distinctive feature of the proposed method is the lack of exhaustive search of the Follower’s strategy space, which is replaced by an iterative guided space sampling procedure.

In order to perform this sampling an Auxiliary Follower’s Game (AFG) is formulated, which is a *one-player* game that yields the Follower’s *restricted pure strategy* (also called *restricted pure realization plan* in EF games) when reaching a terminal state of AFG. AFG is constructed based on the original EF game in the following way:

- The current AFG state is represented in the form of a queue of the Follower’s ISs (from the original ES game). Initially the queue contains ISs observable by the Follower before their first move.
- Each game round consists in taking the first IS from the queue, playing one of the moves available in that IS and

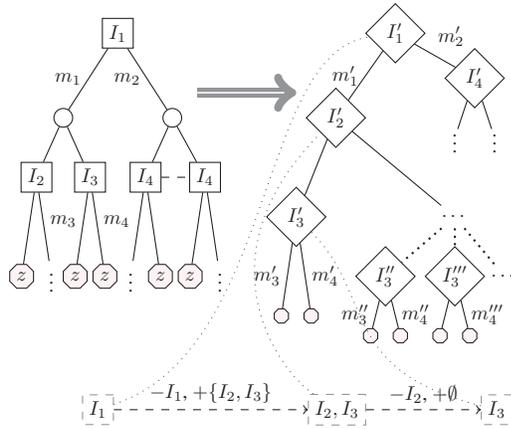


Figure 2: An example of AFG construction. Original two-player EF game (on the left) and the resulting one-player AFG (on the right). Rectangles in the bottom part of the figure present a state of a queue in states  $I'_1$ ,  $I'_2$  and  $I'_3$ , respectively assuming that a path in AFG which is currently considered by the algorithm is the leftmost one. Signs + and - represent *push* and *pop* operations in the queue, resp.

placing in the queue all ISs that may result from the current IS after playing the selected move (i.e. all ISs for which there exists a Leader's strategy directly leading to them).

- The game is played until the queue is empty.

The moves played on the path from the root of AFG to its leaf define a pure Follower's strategy in the original EF game. An example transformation of two-player EF game to one-player AFG is presented in Fig. 2.

In principle, there are no formal requirements for a sampling heuristic method to be used, except for the ability to transfer knowledge related to the sampled space to subsequent iterations. In the experiments an MCTS (Browne et al. 2012) variant called Upper Confidence Bound applied to Trees (UCT) proposed by Kocsis and Szepesvári (Kocsis and Szepesvári 2006) was applied to guide the Follower's strategy space sampling process by means of finding the optimal strategy in AFG, as described in the next section.

In short, each MCTS/UCT iteration (playout) is composed of 4 main phases: *selection*, *expansion*, *simulation*, and *backpropagation* (please consult (Kocsis and Szepesvári 2006) or (Browne et al. 2012) for a detailed description). In our method, selection, expansion and simulation correspond to the first step of an *O2UCT* iteration (guided sampling) and backpropagation phase is implemented in its third step (collection of payoffs). The second step in Fig. 1 refers to reaching the leaf node (a final state of the AFG) and obtaining the Leader's payoff. This payoff is equal to the expected payoff of playing the Leader's strategy and is calculated using a method presented in Section 3.1.

UCT is a powerful and versatile metaheuristic which has proven successful in a wide variety of optimization problems, including General Game Playing (Świechowski, Mańdziuk, and Ong 2016), playing classical board

games (Silver et al. 2017), proactive planning under uncertainty (Wałędzik and Mańdziuk 2018) or combinatorial optimization (Sabharwal, Samulowitz, and Reddy 2012). It is important to note that UCT does not build the entire game tree at once. Instead, the method maintains only a small, relevant part of the tree and expands it gradually, focusing on the most promising paths.

The most challenging part of *O2UCT* is an algorithm (described below) which for the current Follower's strategy  $\pi_f$  finds the respective Leader's strategy  $\delta_l$  for which  $\pi_f$  is an optimal response. The final outcome of *O2UCT* is a pair of strategies  $(\delta_l, \pi_f)$  providing the highest Leader's payoff found across all iterations.

### 3.1 A method of finding the Leader's strategy

The method utilizes a tree structure representation of the Leader's behavior strategy which has the following properties.

- Each node is labeled with the Leader's IS and contains a vector of probabilities of actions available in this IS.
- Root node represents the initial Leader's IS.
- Edges going out of any node are labeled with pairs  $(a, I')$ , where  $a$  is an action and  $I'$  is an IS reachable by playing  $a$  in a given node.

Please note that several ISs may be reachable after playing the same move, depending on an opponent's response. Such a situation is depicted in Fig. 3 presenting an example tree in which playing move  $m6$  may lead to either one of the two ISs ( $s9$  or  $sa$ ) depending on a move played by the opponent. Initially, the tree does not contain all ISs, but only those reachable by an initial strategy profile. Subsequent nodes are added gradually, as explained below.

The algorithm for finding the Leader's strategy is inspired by a double-oracle approach (Bosansky et al. 2014; Jain et al. 2011) and consists of alternating the following two phases: (1) an improvement of the Leader's strategy against a fixed Follower and (2) finding the optimal Follower's response against the current Leader's strategy – based on the Follower's oracle. For a sampled Follower's strategy (Step 1 in Fig. 1) a corresponding Leader's strategy (Step 2 in Fig. 1) must satisfy the following conditions:

- (\*) the optimal Follower's response to that strategy is the same as the sampled Follower's strategy,
- (\*\*) among all Leader's strategies that satisfy the above constraint it is the one that optimizes the Leader's payoff.

Any Leader's strategy satisfying (\*) will be called a *feasible strategy* (a set of *feasible strategies* is also called a *best response region* in (von Stengel and Zamir 2004)).

Let us denote the sampled Follower's strategy by  $\pi_F^r$  ( $r$  stands for the *requested* Follower's strategy). An overview of the method of finding the Leader's strategy that fulfills constraints (\*)–(\*\*) is presented in Fig. 4 and consists of the following steps:

1. Initialize the Leader's strategy.

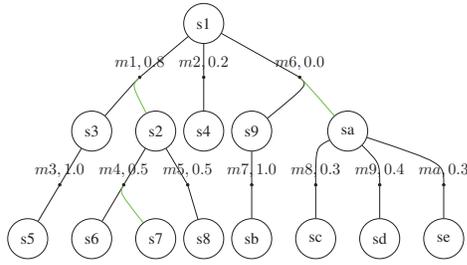


Figure 3: An example of the Leader's behavior strategy tree. Nodes represent the Leader's ISs. Edges are labeled with pairs (move, probability). Playing some moves (e.g.  $m6$ ) may lead to various ISs ( $s9$  or  $sa$ , resp.), depending on the Follower's action.

2. Seek the Follower's strategy yielding better Follower's payoff against the current Leader's strategy using the algorithm described in Section 3.2. If such strategy exists call it  $\pi_F^b$  ( $b$  stands for *better* (in terms of payoff) Follower's strategy).
3. If  $\pi_F^b$  was found, then perform strategy *feasibility pass* (see below) and go to 2, otherwise continue.
4. If stopping condition is not met, perform the Leader's strategy adjustment that increases the Leader's payoff (*positive pass* - see below) and go to 2, otherwise continue.
5. Return the best Leader's strategy among all *feasible strategies* found in step (3).

The Leader's strategy tree is initialized as a single path representing a move sequence maximizing the Leader's payoff against  $\pi_F^r$ . This sequence is found based on a limited number of UCT simulations. All adjustments to the Leader's strategy are performed on a common (continuously evolving) tree-based representation. In Fig. 4 there are two procedures in which these strategy updates occur: *feasibility pass* and *positive pass*. The first one is executed when the current Leader's strategy becomes infeasible, i.e. there exists  $\pi_F^b$  that yields higher Follower's payoff than  $\pi_F^r$ . The latter one is run to improve the Leader's payoff in the case of feasible (Leader's) strategy. In both cases the same procedure, presented in Algorithms 1 and 2, is applied to update the Leader's strategy tree. The only difference lies in a move assessment subroutine which is explained below.

Algorithm 1 starts off from the root node of the Leader's strategy tree and recursively descends to every leaf node of the tree. In each step of recursion one node of the tree, denoted by  $n_c$ , is processed. While in  $n_c$ , the algorithm first recursively calls itself for all child nodes of  $n_c$ . Once adjustment of these nodes (changes in probabilities of available moves) is completed, if there exist moves that are available to play from the IS corresponding to  $n_c$  in the game, but not represented in the tree, a node representing one of them is added to the tree with some probability (equal to 0.3 in the experiments) together with a path expanded from this newly-added node until a leaf node. Next, Algorithm 2 is applied to  $n_c$ .

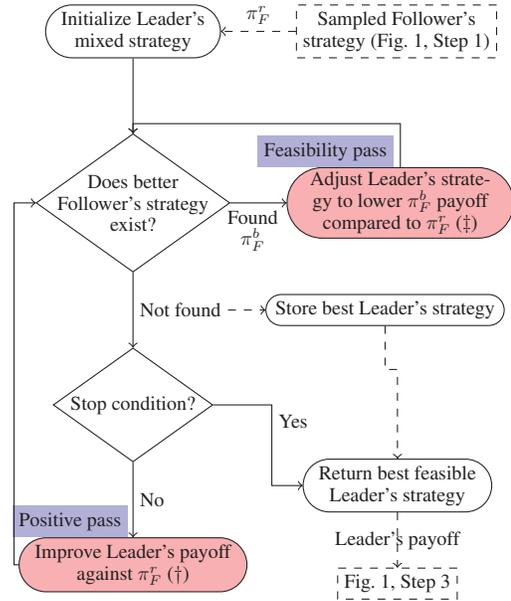


Figure 4: An overview of the method of finding Leader's mixed strategy corresponding to the requested Follower's strategy. Procedures marked in red adjust the current Leader's strategy. Blue labels refer to the names of procedures used in the text. Dashed boxes indicate connection points to the sampling procedure depicted in Fig. 1.

The role of Algorithm 2 is to accumulate direction of strategy changes in two passes: *positive* and *feasibility*. A *momentum* vector is used to store the resultant strategy adjustment stemming from those two passes. The algorithm uses a node *assessment vector* ( $as \in \mathbb{R}^M$ ) to indicate a direction of adjustment of the Leader's strategy and implements this adjustment based on the resultant direction accumulated in all previous iterations in a *momentum vector* ( $mom \in \mathbb{R}^M$ ). First, the momentum vector is updated by adding the assessment vector. A positive value in the assessment vector results in increasing the preference for the respective move, a negative one results in decreasing this preference. Next, a *normalization factor* ( $w \in \mathbb{R}$ ) is increased by adding  $L1$  norm of assessment vector, to confine  $mom/w$  to interval  $[-1, 1]$ . Then the vector of move probabilities is updated with normalized  $mom$  values and normalized to represent a proper probability distribution.

The last element of the method of finding the Leader's strategy is calculation of the assessment vector ( $as$ ) used in Algorithm 2, which is pass-dependent.

- In *positive pass* the goal is to maximize the Leader's payoff. Consequently,  $as$  value for move  $a_i$ :  $as_i = \mathbb{E}U_{\pi_F^r}^L(a_i) - \mathbb{E}U_{\pi_F^r}$  is a difference between the Leader's expected payoff when move  $a_i$  is played in the current state and an expected payoff when moves are played according to the current probabilities (Leader's mixed strategy). The higher the result of playing  $a_i$  compared to the expected result arising from the current probabilities, the greater the  $as_i$  value.

---

**Algorithm 1:** Strategy tree adjustment procedure

---

**Data:**  $n_c$  – a node of a strategy tree currently processed,  $M$  – a set of all moves that are added to the tree in  $n_c$ ,  $I_k$  – an IS corresponding to  $n_c$ ,  $A_{I_k}$  – a set of all moves in  $I_k$ .

- 1 **foreach**  $m \leftarrow M$  **do**
- 2 | recursively adjust successor of  $n_c$  after playing  $m$   
| against  $\pi_F^b$  in *feasibility pass* or  $\pi_F^r$  in *positive pass*
- 3 **end**
- 4 **if** ( $\text{rand}() \leq 0.3$ )  $\wedge$  ( $A_{I_k} \setminus M \neq \emptyset$ ) **then**
- 5 | expand tree with one move  $a' \in A_{I_k} \setminus M$
- 6 **end**
- 7 perform adjustment of  $n_c$  as described in Algorithm 2;

---

---

**Algorithm 2:** Node adjustment with momentum

---

**Data:**  $\text{prob} \in [0, 1]^M$  – a vector of probabilities,  $\text{mom} \in \mathbb{R}^M$  – a momentum vector,  $w \in \mathbb{R}$  – a momentum normalization factor,  $as \in \mathbb{R}^M$  – an assessment vector. In each vector the  $i$ -th position corresponds to the  $i$ -th move.

- 1  $\text{mom} \leftarrow \text{mom} + as$ ;
- 2  $w \leftarrow w + L_1(as)$ ;
- 3  $\text{prob} \leftarrow \max\{\text{prob} + \text{mom}/w, 0\}$  // independent max at each position
- 4  $\text{prob} \leftarrow \text{normalizeOrEqualprob}$  // Normalize vector values so their sum is 1 or, as a fallback, assign equal probability at each position in case all positions equal 0

---

- In *feasibility pass* the goal is to modify the Leader’s strategy in a way that  $\pi_F^r$  will become the corresponding best response strategy. Hence, if the current IS is reachable when playing against both  $\pi_F^r$  and  $\pi_F^b$ , then  $as_i = (\mathbb{E}U_{\pi_F^r}^F(a_i) - \mathbb{E}U_{\pi_F^b}^F(a_i)) - (\mathbb{E}U_{\pi_F^r}^F - \mathbb{E}U_{\pi_F^b}^F)$ , i.e.  $as_i$  is higher for moves that give better result against  $\pi_F^r$  than against  $\pi_F^b$ .  
If the current IS is reachable only when playing against  $\pi_F^b$ , then  $as_i = \mathbb{E}U_{\pi_F^b}^F - \mathbb{E}U_{\pi_F^r}^F(a_i)$ . Note that an order of subtraction operands is reversed compared to a similar equation in the *positive pass*. Here, the weaker the Follower’s payoff when the Leader plays move  $a_i$ , the higher the value of  $as_i$  because the goal is to discourage the Follower from playing strategy  $\pi_F^b$ .

**Stopping condition** The algorithm depicted in Fig. 4 stops when one of the following conditions is reached: a number of executions of step (†) exceeds  $L_{max} = 5000$ , an improvement of the Leader’s payoff in 500 subsequent iterations is less than  $\varepsilon_I = 10^{-5}$ , or a number of subsequent executions of step (‡) without going to step (†) exceeds  $M_{max} = 10000$  (*infeasible strategy*). Values of all steering parameters were selected based on a limited number of preliminary tests.

## 3.2 Follower’s strategy oracle

Implementation of the above-mentioned algorithm requires an ability to find the Follower’s strategy that yields better Follower’s payoff against the current Leader’s strategy.

The most straightforward approach, suitable for any EF game, would be to iterate through all possible Follower’s strategies and choose the one with the highest expected payoff. Such an approach, however, is excessively slow and in practice hinders application of the method to games longer than 4 steps. In order to address this issue our implementation attempts to avoid examination of all Follower’s strategies in the following way:

- Between any two consecutive questions to the oracle about the best response strategy, a collection of  $Q$  pairs ( $n, \pi_F$ ) is maintained, where  $n$  is the use counter and  $\pi_F$  is the Follower’s pure strategy ( $Q = 50$  was used in the experiments).
- When asked about a better Follower’s strategy the algorithm first iterates over Follower’s strategies from the above-mentioned collection. If there exists a strategy that yields better Follower’s payoff than  $\pi_F^r$  by more than  $\varepsilon_O$  (equal to  $10^{-2}$  in the experiments), then this strategy is returned and its use counter is incremented. Otherwise:
  1. If the collection is filled up (contains  $Q$  strategies), a strategy with the smallest use counter is removed.
  2. Examination of all Follower’s pure strategies is performed and the best one is selected and added to the collection with  $n = 1$ .

## 4 Experimental evaluation

Evaluation of *O2UCT* was performed on three game sets: Warehouse Games proposed in (Karwowski and Mańdziuk 2019a), its modified version with more diverse payoffs, and Search Games used in (Bosansky and Cermak 2015).

**Warehouse Games (WHG)** WHG model interactions between the attacker and the defender in a warehouse/office building. A game graph includes the following three types of distinguished vertices: one defender’s starting point, one attacker’s starting point, and several asset locations (targets). The game is sequential and in each turn each of the players can either stay in the current vertex or move to any adjacent vertex. A full description of a game model is presented in (Karwowski and Mańdziuk 2019a). In the experiments 25 WHG instances were used.

**Modified Warehouse Games (WNZ)** WHG setting is relatively close to zero-sum games, the average Pearson’s correlation between the Leader’s and the Follower’s payoffs equals  $-0.82$ . To provide a more challenging setting we used a game generator from (Karwowski and Mańdziuk 2019a) to obtain benchmark games with more diverse payoffs. The following ranges for uniform distributions were applied: attacker’s penalty in targets:  $[-1, 0.2]$  and regular vertices:  $[-1, 0]$  (when caught by the defender), attacker’s reward in targets:  $[-0.2, 1]$  and the corresponding defender’s penalty in targets:  $[-1, 0.2]$  (successful attack). Defender’s reward (for catching the attacker) in non-target and target

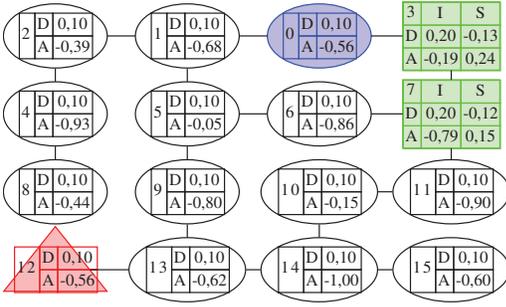


Figure 5: Example WNZ graph. All games are defined on a  $4 \times 4$  grid. Rectangular vertices are targets, a triangle vertex is evader’s starting point, a blue circle vertex is defender’s starting point. Values denote payoffs for the evader and the defender, resp. in the case of evader’s interception in a given vertex. Additional utilities, in case of successful attack, are assigned to targets (the second column).

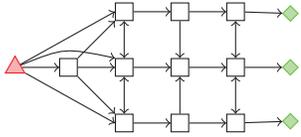


Figure 6: Search Games graph.

vertices was fixed at 0.1 and 0.2, resp. If the game ended due to reaching the round limit, with no interception of the attacker or reaching a target by him/her, a neutral payoff of 0 was assigned to both players. A set of 25 games was generated with the above parameters consisting of games that are less zero-sum like – the average Pearson’s correlation for this set equals  $-0.57$ . An example WNZ graph is depicted in Fig. 5.

**Search Games (SEG)** SEG instances were built on a graph presented in Fig. 6 previously used in (Bosansky and Cermak 2015). Five sets of payoff values in the target vertices for each of the two variants of an attacker’s mobility restrictions (in the first one the attacker can wait in a vertex, in the other one he/she is forced to move in each round) were generated. In each case, game variants with  $T = 4, 5$  and  $6$  were considered, leading to 30 test instances in total.

Similarly to (Bosansky and Cermak 2015), the following distributions of payoffs were used in the tests. In case of catching the attacker the reward for the defender equaled 1 and the penalty for the attacker was equal to  $-1$ . The attacker’s rewards in targets in case of a successful attack were sampled uniformly from  $[1, 2]$  and the respective defender’s penalty was equal to  $-1$ . Otherwise, when the game ended due to reaching the step limit  $T$ , a neutral payoff equal to 0 was assigned to each side.

**Experimental setup** In the following description *BC2015* refers to the method from (Bosansky and Cermak 2015), *C2016* to approach from (Cermak et al. 2016) (a variant AI-MILP was used) and *CBK2018* to the method from (Cerny,

Bosansky, and Kiekintveld 2018) which, following suggestions from the authors, is implemented in two variants: (a):  $\epsilon = 0.3, \delta = 0.4$  and (b):  $\epsilon = 0.0, \delta = 0.4$  (the latter provides better SSE approximations, though requires longer computation time). The following experiments were performed to evaluate efficiency and scalability of *O2UCT*.

- For each of WHG and WNZ instances (defined by game layout and  $T$ ) 15 trails of *O2UCT* were run and for each SEG 5 *O2UCT* tests were run. Multiple trials were required due to stochastic nature of the method.
- For each WHG, WNZ and SEG instance and each of the MILP-based methods (*BC2015*, *C2016*, *CBK2018(a)*, *CBK2018(b)*) a single trial was made (all methods are deterministic). Obtained results were used as a baseline for *O2UCT* assessment.

All experiments were run on Intel Xeon Silver 4116 @ 2.10GHz with 256GB RAM. Experiments involving *O2UCT* were run in parallel, each with 8GB RAM assigned. The remaining tests were run in sequential manner with full memory available to a single process. Each test was run with a time limit of 200 hours and was forcibly terminated if did not finish within the allotted time or exceeded available memory.

#### 4.1 Results

Performance of *O2UCT* is analyzed in two dimensions: an expected Leader’s payoff and time scalability. In both cases the results are presented separately for WHG, WNZ and SEG and grouped by the number of nodes of an extensive-form game  $|\mathcal{S} \cup \mathcal{Z}|$ :

$$bucket = 10^{round(\log_{10} |\mathcal{S} \cup \mathcal{Z}|)}, \quad (1)$$

where *round* rounds a number to the nearest integer. Consequently, games are grouped by the orders of magnitude of game nodes. Such a grouping combines two sources of game complexity: the structure of an underlying game graph and the game length.

**Payoffs** Fig. 7 presents the Leader’s payoffs averaged for all game instances in the respective benchmark sets, calculated at the end of each test against the worst-case Follower (found by examining all possible Follower’s pure strategies). Particular points are plotted if at least 70% of games from a given bucket were solved by a given method within allotted time. Otherwise the respective points are omitted as their comparison would be meaningless. Plots of *BC2015* and *C2016* overlap as both refer to exact methods. Generally speaking, the average Leader’s payoffs calculated by *O2UCT* are very close to optimal results while both variants of *CBK2018* visibly diverge from optimal outcomes, specifically for larger games.

In summary, we believe that the quality of strategies (the Leaders’s payoffs) found by *O2UCT* are very encouraging, as the average results are only slightly worse than the optimal ones.

**Computation times** Computation time analysis is presented in Fig. 8. In the case of reaching computation time

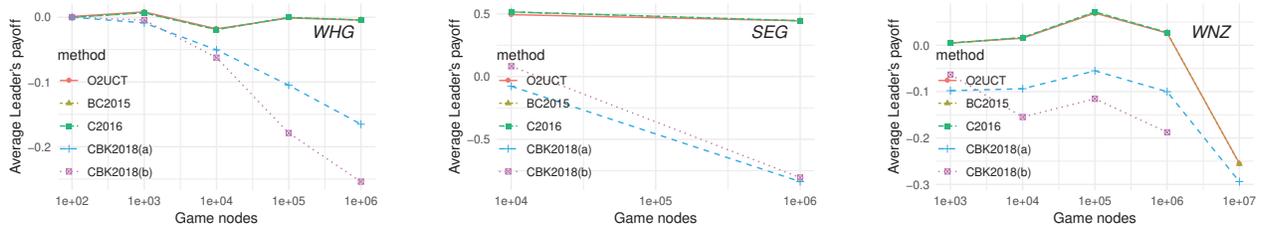


Figure 7: The average Leader's payoffs obtained by the tested methods for three benchmark sets w.r.t. the number of game nodes. Plots are cut at the buckets for which exact methods were still able to solve at least 70% of the respective game instances within 200h time limit. In these borderline cases the average payoffs are calculated for the subsets of solved games only.

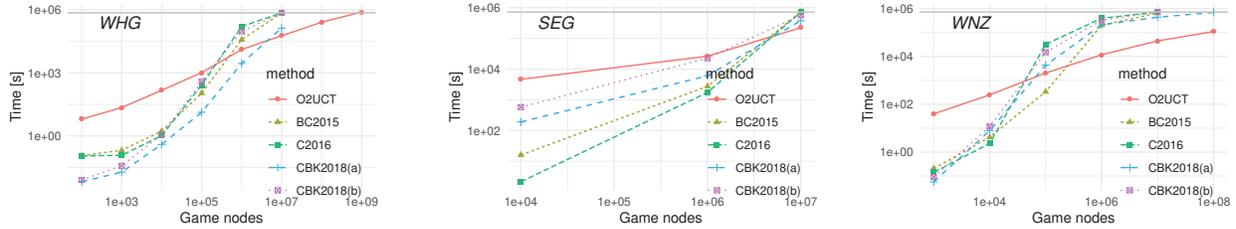


Figure 8: The average computation times for three benchmark sets. *BC2015* and *C2016* were unable to solve some WHG, SEG, WNZ game instances from  $10^7$ ,  $10^7$ ,  $10^6$  buckets, respectively due to hitting the time limit of 200h (marked as a gray line). In such cases time limit value was used in place of the respective computation times.

limit for a given game instance, a value of 200h (time limit) was used in place of the respective result (and averaged with the remaining times in the bucket).

For small WHG and WNZ instances *O2UCT* is the slowest method, however for larger games it starts to outperform MILP approaches. For  $10^7$  and  $10^8$  buckets *O2UCT* managed to solve all WHG/WNZ instances within allotted time and minimally exceeded time limit in games from  $10^9$  bucket. *CBK2018(a)* solved 23/13 WHG/WNZ games from  $10^7$  bucket, resp., and none from  $10^8$  bucket. The remaining methods hit the time limit in almost all  $10^7$  bucket games.

For SEG games *O2UCT* was the slowest for instances with up to  $10^6$  nodes but for  $10^7$  nodes its computation times are already the shortest (all runs were completed within the time limit). Both *CBK2018* variants hit the time limit in at least 40% of the cases and *BC2015* and *C2016* did not solve a single instance.

In summary, it can be observed in Fig. 8 that for small games the results are in favor of MILP approaches, but for larger games MILP methods scale poorer than *O2UCT*. While all considered methods scale exponentially, extrapolation of results to yet bigger games suggests that *O2UCT* is the best-scaling methods amongst the tested ones.

At the same time, it shouldn't be forgotten that *BC2015* and *C2016* are exact methods that yield theoretically guaranteed SSE utilities, while *O2UCT* is only experimentally proven to yield optimal or close-to-optimal strategies.

While exact measurements of memory usage were not performed (it was not possible because of using Java Virtual Machine and its garbage collection facilities) we noted that *O2UCT* was able to compute results for  $10^9$  game nodes

using 8GB of memory while solver based methods started running out of (256GB) memory for games with  $10^7$  nodes.

## 5 Conclusions

This paper presents a novel double-oracle approach for approximating SSE strategy in sequential games with imperfect information and perfect recall. The method does not rely on solving LP/MILP (which is the most common approach) but consists in iterative MCTS/UCT sampling of the Follower's strategy space alternated with adequate modification of the Leader's behavior strategy.

Experimental evaluation shows that proposed approach provides high-quality solutions (optimal in vast majority of the tests) and scales visibly better than state-of-the-art MILP-based methods used for reference. Moreover *O2UCT* requires substantially less memory resources and is therefore capable of solving more complex game instances. Lower memory requirement stems from two factors: application of a double oracle approach which does not require storing in memory all possible strategy profiles simultaneously and dynamic expansion of the Leader's strategy tree – an approach similar in some aspects to the idea of column generation in LP methods. However, what makes *O2UCT* distinct from column generation is the use of a game-independent UCT metaheuristic (instead of a game-specific heuristic) when searching for the most promising moves.

Good time and memory scalability of *O2UCT* enables its application to larger (than in the case of other approaches) game instances. Furthermore, iterative nature of the method allows an easy adjustment of a balance between computation time and quality of results. The outer sampling procedure

employs UCT, which is an anytime algorithm that can be stopped in any moment, though still returning a high quality solution (the best one found so far). An anytime property makes *O2UCT* particularly well suited to problems with strictly allotted time for finding the Leader's strategy.

The introduced method can be applied to any sequential game as it does not depend on any specific game structure or game property.

## Acknowledgments

This work significantly extends our prior extended abstract (Karwowski and Mańdziuk 2019b). The work was supported by the National Science Centre grant number 2017/25/B/ST6/02061.

## References

- Basak, A.; Fang, F.; Nguyen, T. H.; and Kiekintveld, C. 2016. Abstraction methods for solving graph-based security games. In *Autonomous Agents and Multiagent Systems*, 13–33. Springer International Publishing.
- Basilico, N.; Gatti, N.; and Amigoni, F. 2012. Patrolling security games: Definition and algorithms for solving large instances with single patroller and single intruder. *Artificial Intelligence* 184–185:78–123.
- Bosansky, B., and Cermak, J. 2015. Sequence-form algorithm for computing stackelberg equilibria in extensive-form games. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 805–811. AAAI Press.
- Bosansky, B.; Kiekintveld, C.; Lisy, V.; and Pechoucek, M. 2014. An exact double-oracle algorithm for zero-sum extensive-form games with imperfect information. *J. Artif. Intell. Res.* 51:829–866.
- Brazdil, T.; Kucera, A.; and Rehak, V. 2018. Solving patrolling problems in the internet environment. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 121–127. International Joint Conferences on Artificial Intelligence Organization.
- Browne, C.; Powley, E.; Whitehouse, D.; Lucas, S.; Cowling, P.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; and Colton, S. 2012. A Survey of Monte Carlo Tree Search Methods. *Computational Intelligence and AI in Games, IEEE Transactions on* 4(1):1–43.
- Cermak, J.; Bosansky, B.; Durkota, K.; Lisy, V.; and Kiekintveld, C. 2016. Using correlated strategies for computing stackelberg equilibria in extensive-form games. In *30th AAAI Conference on Artificial Intelligence*, 439–445.
- Cerny, J.; Bosansky, B.; and Kiekintveld, C. 2018. Incremental strategy generation for stackelberg equilibria in extensive-form games. In *Proceedings of the 2018 ACM Conference on Economics and Computation, EC '18*, 151–168. ACM.
- Jain, M.; Kardes, E.; Kiekintveld, C.; Ordóñez, F.; and Tambe, M. 2010. Security games with arbitrary schedules: A branch and price approach. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 792–797.
- Jain, M.; Korzhyk, D.; Vaněk, O.; Conitzer, V.; Pěchouček, M.; and Tambe, M. 2011. A double oracle algorithm for zero-sum security games on graphs. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 327–334. International Foundation for Autonomous Agents and Multiagent Systems.
- Johnson, M. P.; Fang, F.; and Tambe, M. 2012. Patrol strategies to maximize pristine forest area. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 295–301.
- Karwowski, J., and Mańdziuk, J. 2015. A new approach to Security Games. In *Artificial Intelligence and Soft Computing*, volume 9120 of *Lecture Notes in Computer Science*. Springer International Publishing, 402–411.
- Karwowski, J., and Mańdziuk, J. 2016. Mixed strategy extraction from UCT tree in security games. In *European Conference on Artificial Intelligence*. IOS Press, 1746–1747.
- Karwowski, J., and Mańdziuk, J. 2019a. A Monte Carlo Tree Search approach to finding efficient patrolling schemes on graphs. *European Journal of Operational Research* 277(1):255–268.
- Karwowski, J., and Mańdziuk, J. 2019b. Stackelberg Equilibrium Approximation in General-Sum Extensive-Form Games with Double-Oracle Sampling Method. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2045–2047.
- Kiekintveld, C.; Jain, M.; Tsai, J.; Pita, J.; Ordóñez, F.; and Tambe, M. 2009. Computing optimal randomized resource allocations for massive security games. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 689–696.
- Kocsis, L., and Szepesvári, C. 2006. Bandit based Monte-Carlo planning. In *Machine Learning: ECML 2006*. Springer, 282–293.
- Kuhn, H. W. 1950. Extensive games. *Proceedings of the National Academy of Sciences* 36(10):570–576.
- Leitmann, G. 1978. On generalized Stackelberg strategies. *Journal of Optimization Theory and Applications* 26(4):637–643.
- Sabharwal, A.; Samulowitz, H.; and Reddy, C. 2012. Guiding combinatorial optimization with UCT. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*. Springer, 356–361.
- Schlenker, A.; Brown, M.; Sinha, A.; Tambe, M.; and Mehta, R. 2016. Get me to my GATE on time: Efficiently solving general-sum bayesian threat screening games. In *ECAI*, 1476–1484.
- Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–359.
- Sinha, A.; Fang, F.; An, B.; Kiekintveld, C.; and Tambe, M. 2018. Stackelberg security games: Looking beyond a decade of success. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5494–5501. International Joint Conferences on Artificial Intelligence Organization.
- von Stengel, B., and Zamir, S. 2004. Leadership with commitment to mixed strategies. Technical report, CDAM.
- Świechowski, M.; Mańdziuk, J.; and Ong, Y.-S. 2016. Specialization of a UCT-based General Game Playing program to single-player games. *Computational Intelligence and AI in Games, IEEE Transactions on* 8(3):218–228.
- Wałędzik, K., and Mańdziuk, J. 2018. Applying hybrid Monte Carlo Tree Search methods to risk-aware project scheduling problem. *Information Sciences* 460–461:450–468.
- Wang, X.; An, B.; Strobel, M.; and Kong, F. 2018. Catching Captain Jack: Efficient time and space dependent patrols to combat oil-siphoning in international waters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 208–215.