AIAA SciTech Forum
8-12 January 2024, Orlando, FL
AIAA SCITECH 2024 Forum

10.2514/6.2024-0125

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

Reinforcement Learning Based Decentralized Weapon-Target Assignment and Guidance

Gleb Merkulov*, Eran Iceland†, Shay Michaeli,‡, Yosef Riechkind§, Oren Gal¶, Ariel Barel‖, Tal Shima**

Multiple-missile attack is one of the simplest ways to saturate and overcome a missile defense system. To increase intercept efficiency against such groups of attackers, it is necessary to allocate the interceptors according to their kinematic limitations. Moreover, such an allocation scheme has to be scalable in order to cope with large scenarios and allow for dynamic reallocation. In this paper we first propose a new formulation of such a Weapon-Target Assignment (WTA) problem and offer a decentralized approach to solve it using Reinforcement Learning (RL) as well as a greedy search algorithm. The engagement is considered from the viewpoint of each pursuer vs. all the targets. Simultaneously, other interceptors engage the target group, and their allocation and success probabilities are available to other team members. To improve the mid-course trajectory shaping, static virtual targets are placed between the pursuers and the incoming adversaries. Each interceptor selects its target dynamically according to a policy that was learned from a large number of scenarios in the computation-efficient simulation environment. The RL input state contains the interceptor reachability coverage of the targets and the probabilities of success of other missiles. The RL reward aggregates the team performance to encourage cooperation on the allocation level. The relevant reachability constraints are obtained analytically by employing kinematic approximations of the interceptor motion. The use of RL ensures real-time scalable and dynamic reallocation for all interceptors. We compare the performance of the proposed RL-based decentralized WTA and guidance scheme against a greedy solution, showing the performance advantage of RL.

I. Introduction

Two problems must be solved in a swarm-attack defense scenario: target allocation and guidance. Target allocation, more generally known as weapon-target allocation (WTA), provides a match between the interceptors and targets, whereas guidance is responsible for the actual intercept trajectory. Clearly, the two processes are intertwined. On the one hand, kinematic constraints dictate trajectory and reachability constraints for the allocation plan; on the other hand, the current allocation decisions limit the scope of future decisions due to kinematics. Therefore, the solution of the multiple interceptor – multiple target WTA and Guidance problem has to exploit the intrinsic relation between the processes to achieve superior performance.

A. Integrated WTA and Guidance

According to the interceptor launch scheme we distinguish between three types of scenarios: (1) single shot, (2) shoot-look-shoot, and (3) shoot-shoot-look. In purely WTA context, the first two problems have been widely studied and the reader can refer to Refs. [1–3]. From the integrated guidance and allocation perspective, the first type was studied by Shalumov and Shima [4] for target-missile-defender (TMD) problem. Several attacker missiles were launched towards the target aircraft, which defended themselves with defender missiles. The authors presented an integrated WTA and Guidance strategy, in which the allocation relied on the miss-distance prediction for the devised cooperative pursuit-evasion guidance law. The miss distance evaluation was based on adjoint mathematics, wherein the system

*Doctoral Student, Faculty of Aerospace Engineering, Technion – Israel Institute of Technology, Haifa, 3200003, Israel; gleb-merkulov@campus.technion.ac.il. (Corresponding author)

†Hebrew University of Jerusalem, Jerusalem, 9190501, Israel; eran.iceland@gmail.com (Equal contribution with corresponding author)

‡AI Researcher; Technion – Israel Institute of Technology, Haifa, 3200003, Israel; shaymichaeli@gmail.com

§Researcher, Faculty of Computer Science; The Open University of Israel, Raanana, 4353701, Israel; ShalomYosefZeev@gmail.com

¶Academic Visitor, Faculty of Computer Science, Technion – Israel Institute of Technology, Haifa, 3200003, Israel; orengal@alumni.technion.ac.il.

‖Academic Visitor, Faculty of Computer Science, Technion – Israel Institute of Technology, Haifa, 3200003, Israel; arielba@technion.ac.il.

**Professor, Faculty of Aerospace Engineering, Technion – Israel Institute of Technology, Haifa, 3200003, Israel; tal.shima@technion.ac.il., Associate Fellow AIAA

Copyright © 2024 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission.

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

linearity is exploited to obtain terminal performance for generic scenarios in a limited number of off-line simulations. The shoot-look-shoot methods can be treated as a two-stage single-shot scenario. After all the engagements of the first wave are terminated we have a look stage to evaluate which targets are still alive, before designating them to new interceptors that are launched in the second shoot stage. The shoot-shoot-look scheme has to be implemented in the case when time constraints do not allow the success evaluation of the first stage before the launch of the second. Namely, the second interceptor wave has to be launched before the outcomes of the first wave engagements are known. The important feature of the scheme is that the number of second-wave interceptors may be smaller than the original number of targets. This concept was mentioned by name in e.g. [5, 6] and studied by Pryluk et al. in [7]. The goal was to find the optimal locations for the backup interceptors at the beginning of the second-stage engagement to minimize the survival probability of the targets. The distinct feature of the work was that the allocation of optimal positions took into account a specific guidance law (proportional navigation (PN)) – the optimal position was selected based on terminal performance prediction for this guidance law. The optimal positions for the backup interceptors were found using brute-force search over the state-space.

To alleviate the computational difficulties of the previous approach, the trailing pursuer guidance was considered as a purely guidance problem by Turetsky et al. in [8] in the optimal control framework. The work dealt with a single pursuer, whose allocation is not known until a certain decision time – delayed-decision guidance (DDG). Assuming engagement probabilities against each of the targets, the before-the-decision guidance law was found to be a weighted sum of minimum-effort one-on-one guidance commands against each of the targets. Weiss et al. extended the results for multiple decision times in [9] such that the strategy could be implemented for several trailing interceptors. The disadvantage of the proposed approaches lies in the necessity to use a complicated non-standard guidance law before the decision and monitoring all the targets by the missile itself. Thus, Merkulov et al. [10] proposed an alternative DDG solution based on the virtual target (VT) approach. Instead of using a complex guidance law, the interceptor was sent towards an optimally selected virtual target using a standard linear impact-angle guidance law [11].

The aforementioned DDG works focused on the minimum-effort trajectory design without taking into account the probabilistic intercept model and the allocation itself. Moreover, the guidance law or the VT computation in the presence of a large number of agents becomes computationally hard. Additionally, taking into account the probabilistic intercept model means that the scenario structure changes due to successful/unsuccessful intercept, making the WTA problem dynamic. Therefore suitable numerical techniques are required to tackle the complexity dynamic WTA problem in real time in a scalable manner. To that end, we propose a decentralized sequential decision scheme and reinforcement learning (RL) methodology as will be discussed later in this paper.

B. RL for WTA

Reinforcement Learning (RL) is a field of machine learning in which an agent learns a policy from experience. RL episode consists of a series of steps starting from an initial state of the environment. In each step, the agent interacts with the environment by a chosen action that changes the state of the environment, and a reward is given for every action. During training the agent learns to choose actions that maximize the cumulative reward. RL has been vastly applied in various applications from games and robotics to WTA. For an overview of RL methods and applications, the reader is referred to [12].

Concerning WTA, Mouton et al. [13] applied Q-learning and MCES techniques for a dynamic WTA problem with 4 weapons and single target and simplified kinematics model. RL solution for a general static WTA problem without integration with guidance was considered by Na et al. [14]. The RL-based strategy provided better assignment results and the trained network could operate in real-time when launching the interceptors. A dynamic WTA problem was considered by Shokoohi et al. [15]. A pursuit-evasion problem with deep integration between assignment and kinematics was considered by Bertram and Wei [16]. There, two teams of aircraft were acting against each other and the RL controller optimized the expected reward associated with the Bellman equation and issued high-frequency discrete control commands and low-frequency allocations. The FastMDP algorithm allowed to process with a trained network large scenarios within a given sample time. Asgharnia et al. [17] considered a variation of TMD scenario in game formulation. The hierarchical fuzzy RL policy decided on control and allocation commands. The resulting policy showed adequate intercept results in small scenarios.

Therefore, we conclude that RL-based policy has a high potential to be applied in WTA problems with dynamics coupling. The distinct feature of the present work is the delayed allocation of the backup interceptors, which has not been addressed previously to the best of the authors' knowledge.

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

C. Contribution and Structure

Thus, in this paper, we propose a scalable RL-based integrated Guidance and WTA methodology for the solution of the dynamic WTA problem in a shoot-shoot-look scenario. The scalability and computational efficiency are achieved by (1) sequential and decentralized decision-making and (2) VT allocation from the interceptor reachable set. The latter limits the search space of the VT as opposed to [7]. Considering the maneuver limitation of the interceptor also enforces the interplay between the Guidance and WTA processes. The proposed RL algorithm selects the next virtual target for the interceptor or the actual target in the final phase. Due to the generality of the reachability analysis, the concrete guidance laws are not of utmost importance. For the sake of simplicity, we use trajectory-shaping guidance to the virtual target and augmented proportional navigation to the actual target. Due to the ability to plan the decision steps ahead, the RL algorithm superiority is demonstrated over the proposed greedy allocation scheme.

This paper is organized as follows. Section II provides the mathematical description of the scenario. In Section III we state the problem objectives and describe the solution approach. Section IV.A describes a benchmark greedy algorithm for dynamic WTA. Section IV.B presents the RL methodology. We compare and discuss the simulation results in Section V. Section VI concludes the findings of this stage of the research.

II. Scenario Description and Modeling

Consider a planar N-vs-M scenario as shown in the Fig. 1. where the evaders are denoted as E_j, $j = 1 \dots M$ and the pursuers (interceptors) are denoted as P_i, $i = 1 \dots N$. We assume that at the initial time, all M pursuers are in the air and no intercepts have happened yet. The pursuit group was launched in two stages according to a shoot-shoot-look scheme [7]. Thus, initially, M out of N interceptors were launched first and allocated one-to-one against all M evaders. The remaining $N - M$ pursuers are trailing behind and will select the evaders to engage based on the results of the intercepts of the first stage. Before that selection is made, the second-stage pursuers are guided towards intermittent virtual targets as shown in Fig. 2. These virtual targets will be selected in a way to enhance the overall intercept performance. The decision-making is decentralized and sequential, i.e. at certain decision times, each pursuer individually chooses a virtual target or evader to engage. The overall goal of the pursuer team is to maximize the expected number of eliminated targets.

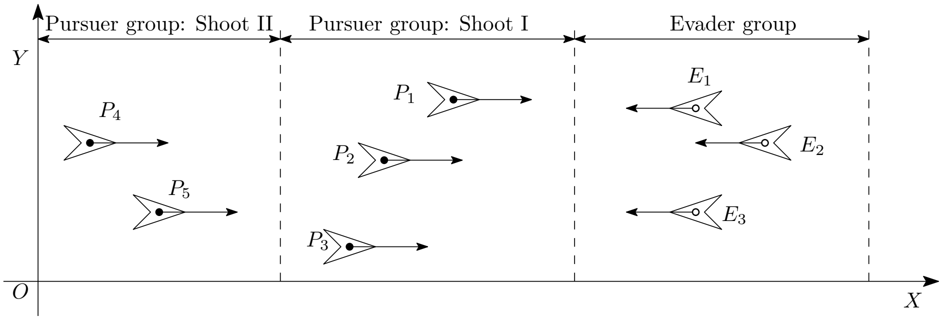


Fig. 1 Shoot-shoot-look engagement

In Fig. 2 pursuers $P_{1\dots3}$ are allocated against $E_{1\dots3}$, respectively. Until the first intercept occurs at $t_{f_{11}}$, the pursuers P_4 and P_5 are guided to virtual targets. By selecting the virtual targets appropriately, the pursuer P_4 can assist in intercepting E_1 and E_2 depending on the engagement outcome between P_1 and E_1. The cooperation between P_4 and P_5 stems from the fact that due to limited maneuverability, P_5 may not be able to assist with the intercept of E_1, or P_4 – with E_3, accordingly.

Remark 1. *Note that unlike Ref. [7], we do not assume that all the engagements of the first stage are over before the second wave is allocated to the targets.*

A. Scenario Modeling

We shall refer to the missiles and targets that are currently alive in the scenario as active. Denote $\mathcal{P}(t)$ to be the set of all active interceptors P_i, $i = 1, 2, \dots$ that are present in the scenario at time t. Analogously, we denote $\mathcal{E}(t)$ to be the set of all active evaders E_j, $j = 1, 2, \dots$ at time t. Naturally, at $t = 0$, $\mathcal{P}_0$ and $\mathcal{E}_0$ have N and M elements, respectively. Additionally, we separate the pursuer set into two: $\mathcal{P}_E(t)$ is the subset of active pursuers allocated to the actual targets at

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

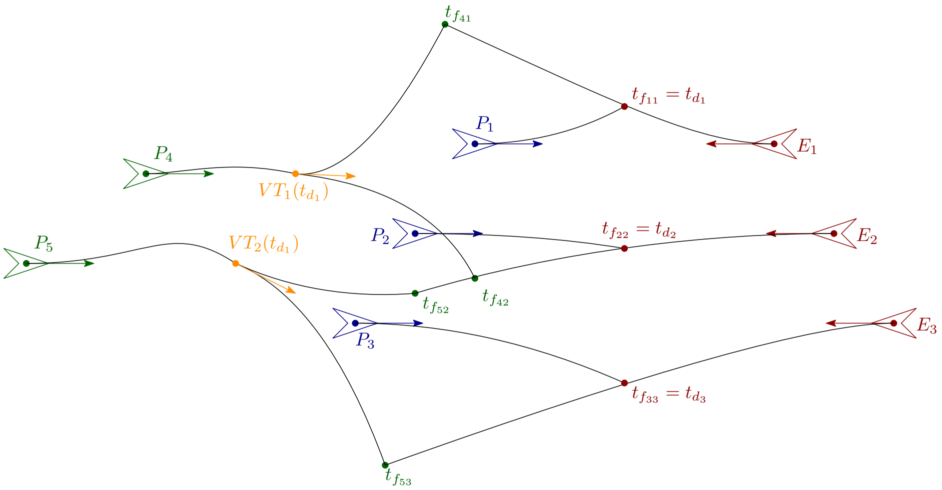


Fig. 2 Guidance to intermittent virtual targets.

time t, whereas the remainder $\mathcal{P}_{VT}(t) = \mathcal{P}(t) \setminus \mathcal{P}_E(t)$ is allocated to intermittent virtual targets. In the sequel, unless explicitly stated otherwise, the short-hand notations $\mathcal{P}$ and $\mathcal{E}$ denote $\mathcal{P}(t)$ and $\mathcal{E}(t)$, respectively. The index function $I(\mathcal{S})$ returns indices of active pursuers or evaders from the input set $\mathcal{S} = \mathcal{P}, \mathcal{E}, \ldots$. The number-of-elements function $n(\mathcal{S})$ returns the number of elements in the set $\mathcal{S} = \mathcal{P}, \mathcal{E}, \ldots$.

For instance, in 5-vs-3 scenario in Fig 1 at $t = 0$ the aforementioned sets are defined as $\mathcal{P}(0) = \{P_i \mid i = 1 \ldots 5\}$, $\mathcal{E}(0) = \{E_j \mid j = 1 \ldots 3\}$, $\mathcal{P}_E(0) = \{P_i \mid i = 1 \ldots 3\}$, $\mathcal{P}_{VT}(0) = \{P_4, P_5\}$. The index function acting on the set $\mathcal{P}_{VT}(0)$ returns the indices of the pursuers in the set, i.e. $I(\mathcal{P}_{VT}(0)) = \{4, 5\}$ and so on.

B. Pursuer and Evader Kinematics

Assume the pursuers and evaders are modeled as point-mass vehicles with constant speed and lateral acceleration steering. Such motion is described by a nonlinear unicycle model

$$\begin{aligned} \dot{x}_{P_i} &= V_{P_i} \cos\theta_{P_i} & & & \dot{x}_{E_j} &= -V_{E_j} \cos\theta_{E_j} & & \\ \dot{y}_{P_i} &= V_{P_i} \sin\theta_{P_i} & & i = 1 \ldots N & \dot{y}_{E_j} &= V_{E_j} \sin\theta_{E_j} & & j = 1 \ldots M \\ \dot{\theta}_{P_i} &= u_{P_i} / V_{P_i} & & & \dot{\theta}_{E_j} &= u_{E_j} / V_{E_j} & & \end{aligned} \tag{1}$$

where $\mathrm{Pos}(k, t) = (x_k(t),\ y_k(t))$ is the position, θ_k is heading, u_k is the lateral acceleration for all $k \in \mathcal{P}, \mathcal{E}$. The engagement geometry between a pursuer and an evader is schematically shown in Fig. 3.

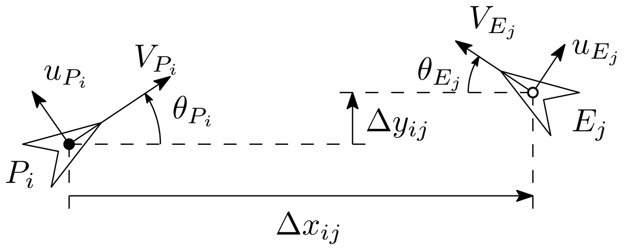


Fig. 3 Engagement geometry

To better convey the essence of the proposed method for N-on-M scenario and simplify the subsequent material presentation, we additionally assume that the motion of the players is in the vicinity of a common reference line Ox, i.e.

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

$\theta \simeq 0$. Thus, the kinematic equations can be written as

$$
\begin{aligned}
\dot{x}_{P_i} &= V_{P_i} & & \dot{x}_{E_j} &= -V_{E_j} & \\
\dot{y}_{P_i} &= V_{P_i}\theta_{P_i} & i = 1\ldots N \qquad & \dot{y}_{E_j} &= V_{E_j}\theta_{E_j} & \qquad j = 1\ldots M \\
\dot{\theta}_{P_i} &= u_{P_i}/V_{P_i} & & \dot{\theta}_{E_j} &= u_{E_j}/V_{E_j} &
\end{aligned}
\tag{2}
$$

The state vector of a player is denoted as $\mathbf{x}_k^T = [x_k \;\; y_k \;\; \theta_k]^T$. The relative positions between pursuers and evaders are denoted as $\Delta x_{ij} = x_{E_j} - x_{P_i}$, $\Delta y_{ij} = y_{E_j} - y_{P_i}$.

Since the equations for x do not depend on the inputs, the x-positions of the players are linearly related to time

$$x_{P_i} = x_{P_i}(0) + V_{P_i} t, \qquad i = 1\ldots N \tag{3}$$

$$x_{E_j} = x_{E_j}(0) - V_{E_j} t, \qquad j = 1\ldots M \tag{4}$$

Thus, all the engagement times of pursuers vs evaders can be determined via the initial x-positions of the pursuers and evaders at the beginning of the engagement

$$t_{f_{ij}} = \frac{\Delta x_{ij}(0)}{V_{E_j} + V_{P_i}}, \quad i \in 1\ldots N,\; j = 1\ldots M \tag{5}$$

Then the overall engagement duration is $T_f = \max_{ij} t_{f_{ij}}$.

In a general form, we assume that the evader control functions u_{E_j}, $j = 1\ldots M$ are known to the pursuer group. In a special case of constant maneuver, the lateral evader motion can be solved analytically as

$$y_{E_j}(t) = y_{E_j}(0) + V_{E_J}\theta_{E_j}(0)\,t + \frac{1}{2}u_{T_j}\,t^2 \tag{6}$$

$$\theta_{E_j}(t) = \theta_{E_j}(0) + \frac{u_{E_j}}{V_{E_j}}\,t \tag{7}$$

Pursuer lateral accelerations belong to a class of piece-wise continuous functions limited by $|u_{P_i}| \leq U_{P_i}$, $i = 1\ldots N$.

III. Problem Statement and Solution Approach

Until the results of the first-wave engagements are known, we propose guiding the pursuers to virtual targets represented by the position and angle. We assume that the guidance law for satisfying impact-angle constraint is available, e.g. trajectory-shaping guidance described in Appendix A. When an engagement with a first-wave pursuer terminates, the backup pursuers are dynamically re-allocated. If the intercept is successful, the backup pursuers are reassigned to new virtual targets. If the first-wave pursuer misses, then one of the backup interceptors is allocated against the surviving evader, and the rest are assigned new virtual targets. Under these conditions, we aim to propose an algorithm for virtual target assignment such that the overall number of surviving evaders is minimal. Next, we formally state this problem.

A. Evader Assignment and Event Timeline

The allocation happens at predefined decision times. The allocation is formalized by the means of the allocation matrix $\mathbf{A}(t) = [A_{ij}(t)]$. If pursuer P_i is allocated against the evader E_j at time t, then $A_{ij}(t) = 1$, otherwise, $A_{ij}(t) = 0$. The engagement with $A_{ij}(t) = 1$ is called active. Since each pursuer can be assigned only one evader at once, the allocation matrix satisfies

$$\sum_{j \in I(\mathcal{E}(t))} A_{ij}(t) = 1, \quad \forall i \in I\left(\mathcal{P}_E(t)\right) \tag{8}$$

Allocation decisions are made sequentially by all second-wave pursuers at decision times t_{d_k}, $k = 0, 1, \ldots$. The first decision round occurs at the beginning of the scenario at $t = 0$. Subsequent decision rounds are made at the times when the scenario structure is changed, i.e. when a pursuer P_i achieves the point of closest approach with the evader E_j at the time $t_{f_{ij}}$, and the intercept outcome is determined. Thus, we define the decision time set as an increasingly ordered sequence of distinct active engagement times together with the initial time. We may formalize the decision time set as

$$\mathcal{T}_d = \left\langle t_{d_k}, t_{d_{k+1}}, \ldots \right\rangle \;:\; \forall t_{d_k} \in \{A_{ij}(t_{f_{ij}})\, t_{f_{ij}},\; i = 1\ldots N,\; j = 1\ldots M\},\; t_{d_k} < t_{d_{k+1}},\; t_{d_k} < T_f,\; k = 0, 1, \ldots \tag{9}$$

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

If $A_{ij} = 0$, then the product $A_{ij} t_{f_{ij}}$ yields 0; hence, $t_{d_0} = 0$ is naturally included in the set. If $A_{ij} = 1$, then $A_{ij} t_{f_{ij}} = t_{f_{ij}}$ – these final times are increasingly ordered to serve as decision times. The requirement $t_{d_k} < T_f$ means that at the last final time, no decisions are made since all pursuers are spent, and the scenario is terminated.

B. Reachable Sets and Virtual Targets

The reachable set $\mathcal{R}(P_i, t_1, t_2)$, $t_1, t_2 \in [0, T_F]$ of the pursuer P_i at time t_2 is defined as a set of all states $\mathbf{x}_{P_i}(t_2)$ that can be achieved using a piece-wise continuous bounded control $|u_{P_i}| \leq U_{P_i}$ from the state $\mathbf{x}_{P_i}(t_1)$ under the dynamic constraint (2). The variable x is calculated as in (3). The variables y, θ are governed by a double integrator sub-model; therefore, the attainable y, θ belong to the interior of the following boundary (based on Theorem 2 and Example 2 in [18])

$$y_{P_i}(t_2) = y_{P_i}(t_1) + V_{P_i}\theta_{P_i}(t_1)(t_2 - t_1) + \frac{U_{P_i}}{2c}\left((t_2 - t_1)^2 - 2(t_2 - \tau)^2\right) \tag{10}$$

$$\theta_{P_i}(t_2) = \theta_{P_i}(t_1) - \frac{U_{P_i}}{V_{P_i} c}\left((t_2 - \tau) - (\tau - t_1)\right) \tag{11}$$

where $c = \pm 1$ and τ is a parameter that varies from t_1 to t_2; i.e. the boundary of the reachable set is obtained by sweeping τ from t_1 to t_2 for $c = 1$ and $c = -1$. The reachable set examples are shown in Fig. 4.

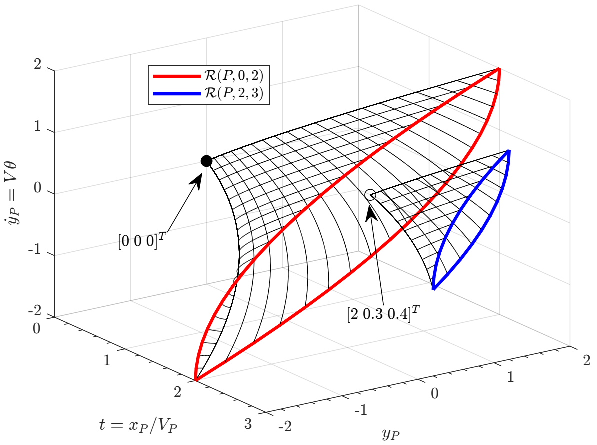


Fig. 4 Consecutive reachable sets of the pursuer

Within the reachable set $\mathcal{R}(P_i, t_1, t_2)$, we define a discrete subset $\mathcal{V}(P_i, t_1, t_2)$ of L virtual targets $VT_l(P_i, t_1, t_2) = \left[x_{P_i}(t_2)\ \ y_{P_i}(t_2)\ \ \theta_{P_i}(t_2)\right]^T$, $l = 1 \ldots L$. Note that similarly to [10], the virtual targets are points (positions and headings) of the pursuer trajectories. Since they are chosen from the pursuer reachable set, we have the guarantee that they can be intercepted using appropriate guidance law that can control impact angle.

The evader E_j can be potentially intercepted by the pursuer P_i if the extrapolated position of the evader is in the reachable set of the pursuer, i.e. $\text{Pos}(E_j, t_{f_{ij}}) \in \mathcal{R}(P_i, t_{d_k}, t_{f_{ij}})$

According to Fig. 5, for each virtual target, we can evaluate which evaders can be intercepted from it as follows. Assume at decision time t_{d_k}, the l-th virtual target $VT_l(P_i, t_{d_k}, t_{d_{k+1}})$ is chosen, which effectively specifies the pursuer state at the next decision time $t_{d_{k+1}}$. Since the evader motion is known, we calculate the evader states $\mathbf{x}_{E_j}$ at times $t_{f_{ij}}$, which are the engagement times against the considered i-th evader using (6,7). Then constructing reachability sets for the pursuer P_i to engagement times $t_{f_{ij}}$, we can determine, whether the evader can be intercepted from the chosen virtual target or not, i.e. a pursuer P_i can intercept the evader E_j if $\text{Pos}(E_j, t_{f_{ij}}) \in \mathcal{R}(P_i, t_{d_{k+1}}, t_{f_{ij}})$. If the evader E_j can be intercepted by the pursuer P_i we say that P_i "covers" E_j.

C. Intercept Model

Assume at time t_{d_k} pursuer P_i is allocated against an evader E_j, i.e. $A_{ij}(t_{d_k}) = 1$. Then the intercept occurs with the given probability p_i if the evader is within the reachable set of the pursuer, otherwise, the intercept cannot happen.

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

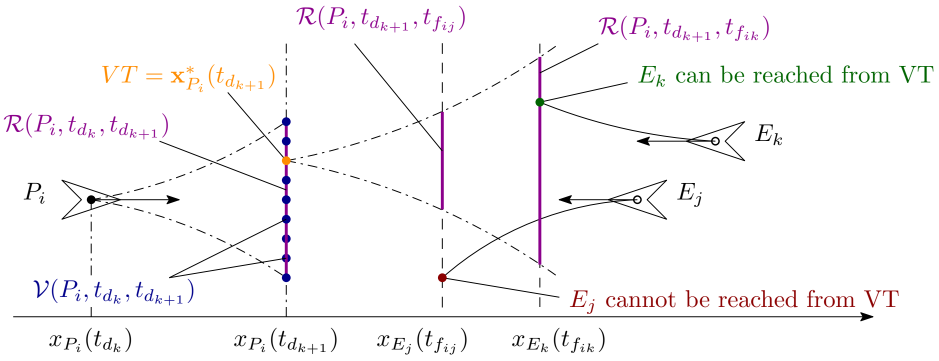

$\mathcal{R}(P_i, t_{d_{k+1}}, t_{f_{ij}})$
$\mathcal{R}(P_i, t_{d_{k+1}}, t_{f_{ik}})$
$VT = \mathbf{x}^*_{P_i}(t_{d_{k+1}})$
E_k can be reached from VT
$\mathcal{R}(P_i, t_{d_k}, t_{d_{k+1}})$
E_k
P_i
E_j
$\mathcal{V}(P_i, t_{d_k}, t_{d_{k+1}})$
E_j cannot be reached from VT
$x_{P_i}(t_{d_k})$
$x_{P_i}(t_{d_{k+1}})$
$x_{E_j}(t_{f_{ij}})$
$x_{E_k}(t_{f_{ik}})$

Fig. 5 Virtual target coverage

Namely,

$$\Pr\{P_i \text{ intercepts } E_j\} = \begin{cases} p_i, & \mathrm{Pos}(E_j, t_{f_{ij}}) \in \mathcal{R}(P_i, t_{d_k}, t_{f_{ij}}) \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

This is illustrated in Fig. 6.

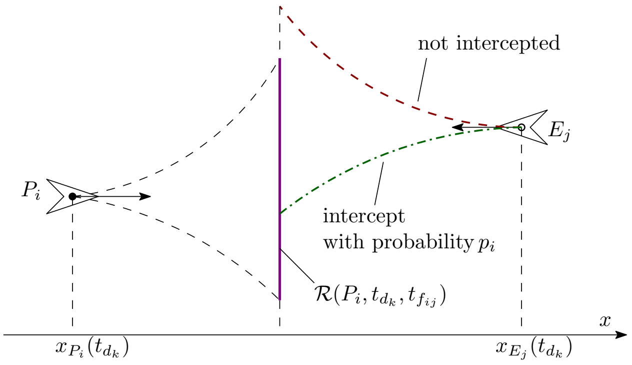


Fig. 6 Intercept model.

D. Cost Function

Our goal is to select for each pursuer P_i a virtual target VT_l or the evader E_j allocation at the decision times $\mathcal{T}_d$ such that the overall number of the real targets survived is minimal, i.e.

$$n(\mathcal{E}(T_f)) \longrightarrow \min \tag{13}$$

Next, we present the elements of the solution approach, including decentralized sequential decision-making (common for all algorithms), greedy allocation baseline, and the RL policy.

IV. Interceptor Decision Making

Each interceptor P_i at every decision time $t_{d_k} \in \mathcal{T}_d$ makes an allocation decision according to the scheme in Fig. 7. The pursuer may choose either one of L virtual targets from the set $\mathcal{V}(P_i, t_{d_k}, t_{d_{k+1}})$ or one of the remaining evaders from $\mathcal{E}(t_{d_k})$. If the evader is chosen, then this decision is fixed and cannot be changed. If a virtual target is chosen, then at the next step the pursuer may switch to another virtual target or an active evader that survived an intercept.

Interceptors have access to the following information

- Current decisions of other interceptors
- Relative locations of other interceptors and targets
- Relative velocities to other interceptors and targets

- Target maneuver
- Number of evaders that can be intercepted from the selected VT

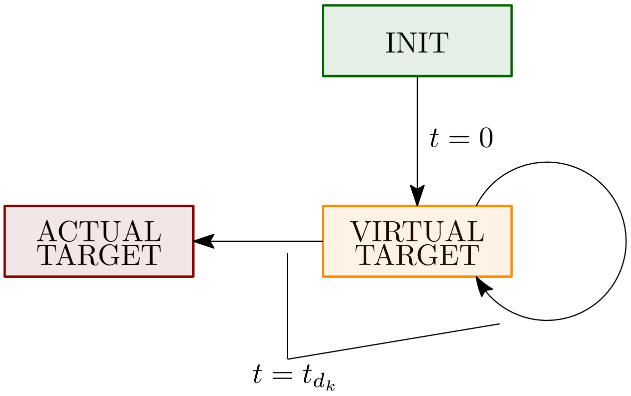


Fig. 7 Allocation decision scheme for trailing pursuer.

The decision-making at the decision times is decentralized and sequential. The order of sequence is fixed according to the number of the pursuer, i.e. in the example in Fig. 2, P_4 makes the allocation decision first; P_5 – second. More generally, the first pursuer in the sequence makes a pursuit decision. The next pursuer is aware of the decision of the previous pursuer and makes its decision based on that information and so on. At each decision time, there is one decision-making round, and the decisions do not change until the next decision time.

For simplicity, we assume that the first wave is allocated arbitrarily and we propose the next two decision algorithms for the interceptors of the second wave.

A. Greedy Algorithm

We first suggest a greedy allocation algorithm. The backup interceptors perform greedy sequential allocation decisions according to the following scheme described from the perspective of the decision time t_{d_k}.

1. If there is an unengaged evader E_j, then the first pursuer in the decision sequence that covers E_j is allocated against it.
2. If all evaders are engaged, then the current pursuer P_i under consideration is guided against a virtual target. The virtual target is selected as follows.
    2.1. Sample virtual target set $\mathcal{V}(P_i, t_{d_k}, t_{d_{k+1}})$.
    2.2. For each $VT_l(\cdot) \in \mathcal{V}(\cdot)$, $l = 1 \dots L$ compute the score S_{VT_l} as
        2.2.1. For each evader E_j that is covered from the current virtual target assign a score

$$S_j = \frac{1}{1 + q_j} \tag{14}$$

where q_j is the current number of backup pursuers that cover E_j before the allocation of P_i.

        2.2.2. Compute the updated score for each evader E_j considering coverage by P_i

$$S_j^{new} = \frac{1}{1 + q_j + c_i} \tag{15}$$

where $c_i = 1$ if P_i can intercept E_j from VT_l and $C_i = 0$, otherwise.

    2.3. Assign a virtual target score as a norm of “added allocation benefit”

$$S_{VT_l} = \left\| S_j^{new} - S_j \right\| \tag{16}$$

    2.4. Select the virtual target with the maximal score.

The idea behind the chosen “added allocation benefit” score is the assumption that the consistent distributed coverage of the evaders by the backup pursuers improves the number of available allocation decisions leading to a larger number of potentially successful intercepts.

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

B. RL-based Algorithm

We now present the RL-based WTA strategy for assigning the virtual targets for the backup pursuers. Let us describe the environment, the action space, the state representation, the network architecture, the reward and the training session.

Environment. Each episode starts with M pursuers assigned to M evaders. The RL agent considers only the excess of $N - M$ free interceptors. In each step t_{d_k} the agent sequentially assigns a virtual target to the next backup pursuer in the queue of free interceptors. When all free interceptors are assigned to virtual targets, the motion of pursuer and evaders is propagated until the next decision time $t_{d_{k+1}}$, which coincides with some engagement time $t_{f_{ij}}$. If the interception of E_j by P_i was successful, then new virtual targets are sampled for each backup pursuer, etc. Otherwise, the first pursuer from free interceptors that can reach the free evader is assigned to it. This interceptor is removed from the list of the free interceptors. If all virtual targets of a given backup pursuer can not cover any evader then this interceptor is removed from the list of the available free interceptors. The episode ends at the final time T_f.

Action Space. The action space is discrete with L actions, where L is the number of possible virtual targets for one interceptor. In each step, the current interceptor chooses one of its virtual targets.

State. The state includes the following:

- Coverage of current interceptor's virtual targets
- Virtual targets' coverage for all free interceptors currently not assigned to a virtual target
- Coverage of chosen virtual targets - for interceptors that already chose their virtual target
- Status of real targets. either free, occupied or intercepted.

The coverage is encoded as vectors with the number of elements equal to the number of evaders; the values are $\{-1, 0, 1\}$: -1 for intercepted evaders, 0 for not covered evaders and 1 for evaders that can be reached from the virtual target. The status is encoded using $\{-1, 0, 1\}$ as well.

Network Architecture. The state is fed into two fully connected neural networks. Each with 3 hidden layers of 512, 128 and 64 neurons respectively and RELU activation. One network for the Actor and the other for the Critic.

Reward. The selected reward given at each decision step is the same as in the greedy algorithm (16).

Remark 2. *A different reward was considered – a positive reward when a free interceptor is assigned. This reward provides good results for small scenarios, but for 20 evaders or more, it yields inferior results relative to the greedy score. We speculate that the reason is the sparsity of this reward.*

Training. Training was performed on an Azure *NC8as T4 v3* machine. PPO implementation of Stable-Baselines 3 [19] was utilized. After 15 million training steps (less than two hours of training) the RL cumulative reward was higher than the cumulative score of the greedy algorithm, but the expected number of ground hits was similar to the greedy algorithm. The training continued until 200 million steps where the RL reward improvement was negligible.

V. Results and Discussion

This Section presents a simulation example of a multiple vs multiple engagement with dynamic target allocation and the statistical performance analysis of the presented algorithms.

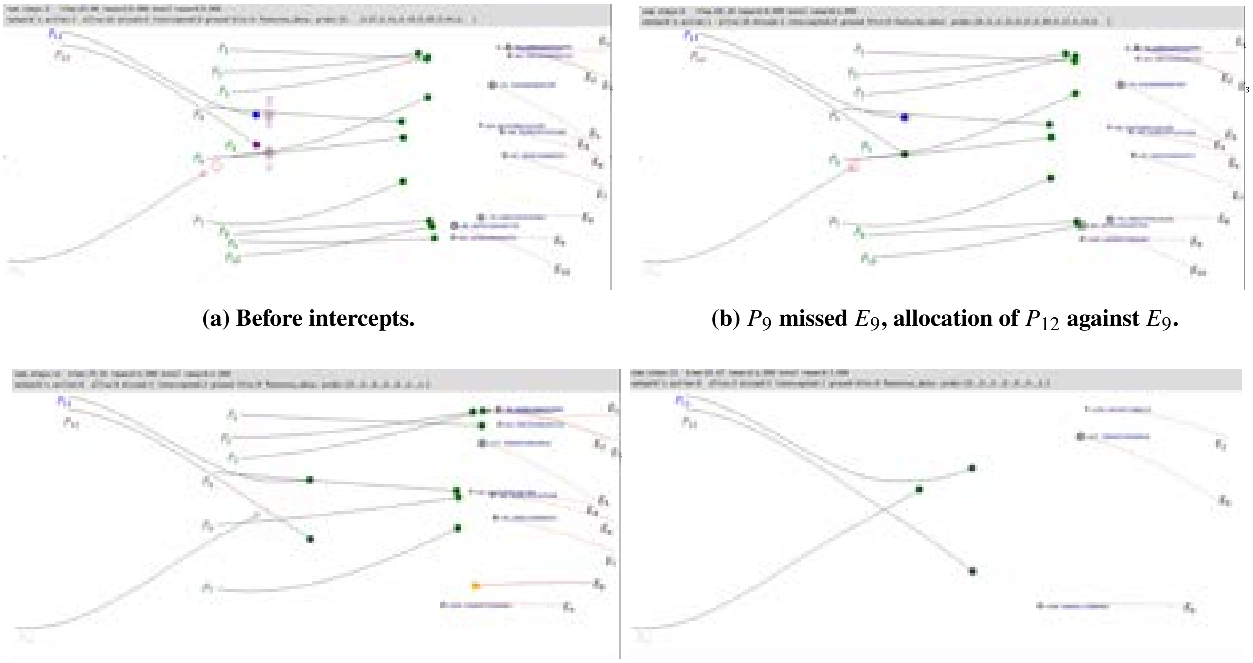
A. Example Simulation Scenario

Consider an engagement between 13 pursuers and 10 evaders as shown in the Fig. 8. The initial allocation of the first-wave pursuers to the evaders is according to their number, i.e. P_i is allocated against E_i for $i = 1 \dots 10$. The pursuers allocated against the evaders use augmented proportional navigation and the intercept success probability is $p = 0.8$. The backup pursuers are initially allocated against the virtual targets computed by the RL algorithm and use the trajectory shaping guidance to arrive at them. The initial phase of the scenario before any intercepts occur is shown in Fig. 8a. The pursuers allocated to real targets are shown in green, whereas backup pursuers have different colors to distinguish their corresponding virtual targets. The numbers near the evaders show the predicted engagement time

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

Table 1 Target allocation in example 13 vs. 10 scenario

Step	$\mathcal{T}_d$	Evader	First-wave pursuers										Backup pursuers		
			P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}
1	69.18	E_9									X			E_9	
2	69.89	E_{10}										V			
3	76.59	E_8								V					
4	79.33	E_5					X						E_5		
5	82.62	E_4				V									
6	82.89	E_3			V										
7	82.95	E_2		X											E_2
8	83.39	E_1	V												
9	86.91	E_6						V							
10	87.66	E_7							V						
11	108.54	E_9												V	
12	117.26	E_5											V		
13	130.32	E_2													V



(a) Before intercepts.

(b) P_9 missed E_9, allocation of P_{12} against E_9.

(c) P_5 missed E_5, allocation of P_{11} against E_5.

(d) P_2 missed E_2, allocation of P_{13} against E_2.

Fig. 8 Example simulation snapshots.

against the allocated pursuer. The full process of the target allocation is shown in Table 1 where V denotes a successful intercept and an X a unsuccessful one.

The first decision time is 69.18 [s], which corresponds to the engagement between P_9 and E_9. The result of this engagement is unsuccessful; therefore, the allocation algorithm assigned P_{12} to pursue E_9 and a new corresponding final time 108.54 was added to the decision times sequence $\mathcal{T}_d$. From this moment E_{12} cannot change its allocation; thus, P_{12} is colored in green. The simulation snapshot corresponding to the scenario state right after this reallocation is shown in Fig. 8b. The following intercepts of E_{10} and E_8 are successful. The next miss occurs in the engagement between P_5

and E_5. The allocation algorithm assigns P_{11} to pursue E_5, and the new engagement time 117.26 is added into $\mathcal{T}_d$. The corresponding simulation snapshot is shown in Fig. 8c. The following intercepts by P_4 and P_3 are successful. Pursuer P_2 misses E_2; therefore, the only backup interceptor left P_{13} is assigned against E_2 with the engagement final time of 130.32 [s]. All following intercepts are successful; thus, all evaders are hit.

Note that the first virtual target allocation was most significant since in the small time windows between the first-wave intercepts the second-wave interceptors do not have sufficient time to alter the trajectory. This can be seen in Fig. 8b and 8c, in which the virtual target candidates are all close to the respective pursuer.

B. Performance Analysis

Following we perform a statistical performance analysis of the presented algorithms. The example scenario we consider contains 20 targets and 24 interceptors, i.e. there are 4 second wave interceptors. The hit probability for all engagements is $p = 0.8$. At the beginning of the engagement, each first-wave interceptor was allocated to an actual target, while the second-wave interceptors were assigned to virtual targets. We investigated 4 options for each virtual target sampling: 3, 5, 7, and 9. The checkpoints for RL agent state were collected every 1 million steps of training. Every checkpoint network was tested on 1500 scenarios, each of which had identical initial conditions. Such a number of scenarios is needed to approximate the performance distribution for different intercept outcomes. Then, for each of the 4 virtual target configurations, we compared the best checkpoint to the greedy algorithm (section IV.A) for over $100,000$ cases with the same initial positions and virtual targets. The performance analysis results are presented in Table 2. Both algorithms yield similar performance with the RL policy being better in terms of the expected number of targets survived by approx. 1%.

Table 2 RL vs. Greedy Comparison

	Ground Hits Mean (Std)				**Score Mean (Std)**			
	3 vts	**5 vts**	**7 vts**	**9 vts**	**3 vts**	**5 vts**	**7 vts**	**9 vts**
RL	2.054 (1.49)	1.899 (1.448)	1.86 (1.433)	1.843 (1.423)	97.559 (21.751)	106.589 (19.271)	109.395 (18.845)	110.855 (18.657)
Greedy	2.08 (1.503)	1.927 (1.458)	1.89 (1.449)	1.871 (1.436)	96.13 (21.831)	104.225 (19.726)	106.837 (19.323)	108.193 (19.103)

To investigate the proximity of the proposed solution to the optimal one, we derived an approximation of the theoretical lower bound on the expected number of surviving targets (see Appendix B). This lower bound ignores the kinematics influence and is independent of the number of virtual targets. It provides the best possible outcome, from the point of view of the interceptor team, provided there are no kinematic limits (e.g. acceleration constraints). For the investigated case of 24 interceptors, 20 targets, and a hit probability of 0.8, Equation B.3 yields a lower bound of 1.225 surviving targets (on average). Table 3 describes the improvement of the RL policy over the greedy algorithm, relative to the approximate theoretical lower bound.

Table 3 RL Improvement over the Greedy Algorithm

	3 vts	**5 vts**	**7 vts**	**9 vts**
RL over Lower Bound	0.829	0.674	0.635	0.618
Greedy over Lower Bound	0.855	0.702	0.665	0.646
Improvement	3.04%	3.98%	4.51%	4.33%

Remark 3. *The simulations showed that the RL-training convergence depends on the number of backup interceptors $N - M$. The number of options to choose virtual targets is exponential in $N - M$. Our simulations show that for $N - M \le 5$, RL is better than the greedy algorithm. However, for $N - M \ge 6$, RL training did not converge. Convergence can be improved by truncating the RL state to include a limited number of backup interceptors in the queue.*

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

VI. Conclusions

We posed and investigated an integrated WTA and Guidance problem in a shoot-shoot-look scenario, where there is an excess number of interceptors. The optimization objective is to minimize the number of targets that are not intercepted. We proposed a solution approach for the allocation and guidance of the backup excess interceptors based on the virtual target approach. As long as all targets in the evader group are engaged, the backup pursuers are guided to dynamically allocated virtual targets before the engagement of actual targets. Two solution algorithms are proposed: greedy and RL. Both algorithms use the target coverage prediction as the main decision variable. The results of the numerical study demonstrate the viability of both the greedy and RL approaches, providing performance that is close to the approximation of the theoretical lower bound obtained by neglecting the interceptors' kinematics. In the investigated scenarios the RL approach provided somewhat better performance.

Appendix

A. Generation of Pursuer Trajectories

For computational simplicity, we generate the pursuit trajectories against virtual targets and evaders using trajectory-shaping guidance [20, Chapter 25] and augmented proportional navigation [20, Chapter 8], respectively. The guidance laws have the following forms

$$a_{P_i}^{TSG} = \frac{6}{(t_{d_{k+1}} - t)^2}\left((y_{VT_l}) - y_{P_i}) - \dot{y}_{P_i}(t_{d_{k+1}} - t)\right) + \frac{2V_{P_i}}{t_{d_{k+1}} - t}\left(\gamma_{P_i} - \gamma_{VT_l}\right) \tag{A.1}$$

$$a_{P_i}^{APN} = \frac{3}{(t_{f_{ij}} - t)^2}\left(\Delta y_{ij} + \Delta\dot{y}_{ij}(t_{f_{ij}} - t) + \frac{1}{2}u_{E_{ij}}(t_{f_{ij}} - t)^2\right) \tag{A.2}$$

These guidance laws do not take explicitly into account the maneuver limitation. In this case, we can treat the maneuver limit as a "design parameter" that limits the control effort of the above guidance laws.

B. Benchmark Performance – Neglecting Kinematics

To provide performance bounds on for the proposed greedy and RL algorithms, we propose an approximate approach to evaluate the expected number of surviving targets neglecting the kinematics. To this end, we compute $Pr(\text{G.H.}) = k$ - the probability of having exactly k ground hits for $0 \leq k \leq M$.

Recall that we have N interceptors and M targets with $N > M$, and the interception probability of each engagement is p. While in most cases there are two waves of interceptors, it is possible to have only few targets in the second wave, such that some of the second-wave interceptors will be used for the third wave. To ease the computation we neglect the case of more than three waves since they are very rare and hence almost do not change the expected number of ground hits.

Let $s_1 \geq k$. Denote the probability that s_1 targets survive the first wave and k targets hits by $Pr(\text{G.H.} = k, s_1)$. To calculate $Pr(\text{G.H.} = k, s_1)$ we distinguish between the following two cases:

Case 1: $\mathbf{s_1 \geq N - M}$. In this case, all second wave interceptors are engaged to targets that survived the first wave i.e. there is no third wave. Let $b(r, n, p)$ be the standard probability mass function of the binomial distribution i.e. the probability to intercept r targets out of n engagements where each interception occurs with probability p. Then

$$Pr(\text{G.H.} = k, s_1) = b(M - s_1, M, p) \cdot b(s_1 - k, N - M, p) \tag{B.1}$$

The first multiplier is the probability to intercept $M - s_1$ targets in the first wave and the second multiplier is the probability for the $N - M$ second wave interceptors to intercept $s_1 - k$ targets.

Case 2: $\mathbf{s_1 < N - M}$. In this case, we use s_1 interceptors for the second wave and $N - M - s_1$ interceptors are left for the third wave. Let s_2 be the the number of targets that survived the second wave. Since s_2 may take any value between k and s_1, we have that

$$Pr(\text{G.H.} = k, s_1) = \sum_{s_2=k}^{s_1} b(M - s_1, M, p) \cdot b(s_1 - s_2, s_1, p) \cdot b(s_2 - k, min(s_2, N - M - s_1), p) \tag{B.2}$$

In the last multiplier of B.2 we consider the third wave in which the number of engagements is the minimum between the number of targets s_2 and the number of interceptors $N - M - s_1$

We can conclude that

$$Pr(\text{G.H.} = k) = \sum_{s_1=k}^{M} Pr(\text{G.H.} = k, s_1) \tag{B.3}$$

References

[1] Hosein, P., and Athans, M., "An asymptotic result for the multi-stage weapon-target allocation problem," *29th IEEE Conference on Decision and Control*, 1990, pp. 240–245 vol.1. https://doi.org/10.1109/CDC.1990.203589.

[2] Glazebrook, K., and Washburn, A., "Shoot-Look-Shoot: A Review and Extension." *Operations Research*, Vol. 52, No. 3, 2004, pp. 454 – 463. URL http://ezlibrary.technion.ac.il/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.30036595&site=eds-live&scope=site.

[3] Kline, A., Ahner, D., and Hill, R., "The Weapon-Target Assignment Problem," *Computers & Operations Research*, Vol. 105, 2019, pp. 226–236. https://doi.org/10.1016/j.cor.2018.10.015.

[4] Shalumov, V., and Shima, T., "Weapon–Target-Allocation Strategies in Multiagent Target–Missile–Defender Engagement," *Journal of Guidance, Control, and Dynamics*, Vol. 40, No. 10, 2017, pp. 2452–2464. https://doi.org/10.2514/1.G002598, URL https://doi.org/10.2514/1.G002598.

[5] Farrell, J., "Recent advances in linear programming applied to global defense," *Proceedings of the 27th IEEE Conference on Decision and Control*, 1988, pp. 2428–2430 vol.3. https://doi.org/10.1109/CDC.1988.194777.

[6] Vermeulen, A., and Savelsberg, R., "Optimal mid-course doctrine for multiple missile deployment," *AIAA Guidance, Navigation, and Control Conference*, American Institute of Aeronautics and Astronautics, 2012. https://doi.org/10.2514/6.2012-4912, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088755373&doi=10.2514%2f6.2012-4912&partnerID=40&md5=76d9027145df2314df6e8e0e36ad9526.

[7] Pryluk, R., Shima, T., and Golan, O. M., "Shoot–Shoot–Look for an Air Defense System," *IEEE Systems Journal*, Vol. 10, No. 1, 2016, pp. 151–161. https://doi.org/10.1109/JSYST.2014.2344755.

[8] Turetsky, V., Weiss, M., and Shima, T., "Minimum Effort Pursuit Guidance with Delayed Engagement Decision," *Journal of Guidance, Control, and Dynamics*, Vol. 42, No. 12, 2019, pp. 2664–2670. https://doi.org/10.2514/1.G004393, URL https://doi.org/10.2514/1.G004393.

[9] Weiss, M., Shalumov, V., and Shima, T., "Minimum Effort Pursuit Guidance with Multiple Delayed Engagement Decisions," *Journal of Guidance, Control, and Dynamics*, Vol. 45, No. 7, 2022, pp. 1310–1319. https://doi.org/10.2514/1.G006494, URL https://doi.org/10.2514/1.G006494.

[10] Merkulov, G., Weiss, M., and Shima, T., "Virtual Target Approach for Multi-Evader Intercept," *2022 European Control Conference (ECC)*, IEEE, 2022. https://doi.org/10.23919/ecc55457.2022.9838427.

[11] Ryoo, C.-K., Cho, H., and Tahk, M.-J., "Optimal Guidance Laws with Terminal Impact Angle Constraint," *Journal of Guidance, Control, and Dynamics*, Vol. 28, No. 4, 2005, pp. 724–732. https://doi.org/10.2514/1.8392, URL https://doi.org/10.2514/1.8392.

[12] Sutton, R. S., and Barto, A. G., *Reinforcement learning: An introduction*, MIT press, 2018.

[13] Mouton, H., Le Roux, H., and Roodt, J., "Applying reinforcement learning to the weapon assignment problem in air defence," *Scientia Militaria: South African Journal of Military Studies*, Vol. 39, No. 2, 2011, pp. 99–116.

[14] Na, H., Ahn, J., and Moon, I.-C., "Weapon–Target Assignment by Reinforcement Learning with Pointer Network," *Journal of Aerospace Information Systems*, Vol. 20, No. 1, 2023, pp. 53–59.

[15] Shokoohi, M., Afsharchi, M., and Shah-Hoseini, H., "Dynamic distributed constraint optimization using multi-agent reinforcement learning," *Soft Computing*, Vol. 26, No. 8, 2022, pp. 3601–3629.

[16] Bertram, J., and Wei, P., "An Efficient Algorithm for Multiple-Pursuer-Multiple-Evader Pursuit/Evasion Game," *AIAA Scitech 2021 Forum*, American Institute of Aeronautics and Astronautics, 2021. https://doi.org/10.2514/6.2021-1862, URL https://arc.aiaa.org/doi/abs/10.2514/6.2021-1862.

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125

[17] Asgharnia, A., Schwartz, H. M., and Atia, M., “Multi-Invader Multi-Defender Differential Game Using Reinforcement Learning,” *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2022, pp. 1–8.

[18] Lee, E. B., and Markus, L. L., *Foundations of optimal control theory / [by] E. B. Lee [and] L. Markus.*, SIAM series in applied mathematics, Wiley, New York, 1967.

[19] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N., “Stable-Baselines3: Reliable Reinforcement Learning Implementations,” *Journal of Machine Learning Research*, Vol. 22, No. 268, 2021, pp. 1–8. URL http://jmlr.org/papers/v22/20-1364.html.

[20] Zarchan, P., *Tactical and strategic missile guidance / Paul Zarchan.*, 6th ed., Progress in astronautics and aeronautics ; vol. 239, American Institute of Aeronautics and Astronautics, Reston, Va, 2012.

Downloaded by TECHNION - ISRAEL INST OF TECH on April 15, 2024 | http://arc.aiaa.org | DOI: 10.2514/6.2024-0125