Reinforcement Learning Based Decentralized Weapon-Target Assignment and Guidance

Gleb Merkulov*, Eran Iceland°, Shay Michaeli*, Yosef Riechkind†,
Oren Gal*, Ariel Barel*, and Tal Shima*

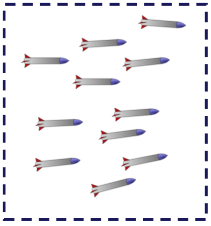
* Technion – Israel Institute of Technology
° Hebrew University of Jerusalem
† The Open University of Israel

2024 AIAA SciTech Forum

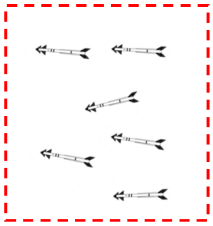
08/01/2024

Scenario and Objective I

Swarm Attack Scenario



Interceptor
single wave



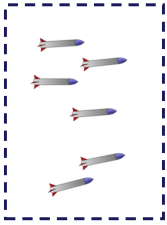
Swarm attack

Solution Approaches

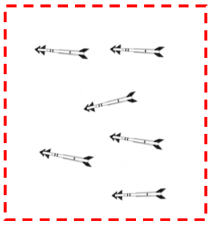
- **Single Shot**
  - Improved miss probability
  - Bad resource management

Scenario and Objective I

Swarm Attack Scenario



Interceptor
first wave



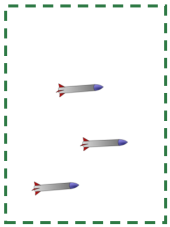
Swarm attack

Solution Approaches

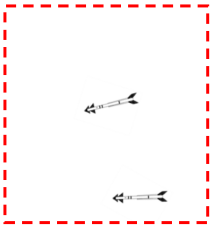
- **Single Shot**
  - Improved miss probability
  - Bad resource management
- **Shoot-Look-Shoot**
  - Better resource management
  - Time constraints, reallocation

Scenario and Objective I

Swarm Attack Scenario



Interceptor
second wave



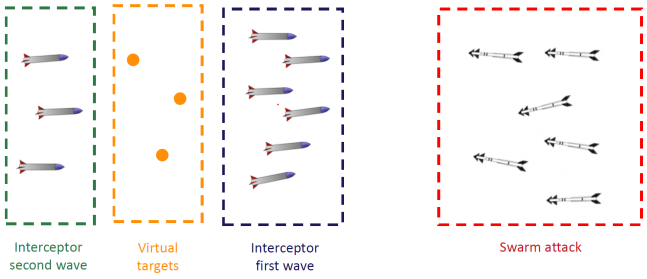
Remaining targets

Solution Approaches

- **Single Shot**
  - Improved miss probability
  - Bad resource management
- **Shoot-Look-Shoot**
  - Better resource management
  - Time constraints, reallocation

Scenario and Objective I

Swarm Attack Scenario



Solution Approaches

- **Single Shot**
  - Improved miss probability
  - Bad resource management
- **Shoot-Look-Shoot**
  - Better resource management
  - Time constraints, reallocation
- **Shoot-Shoot-Look**
  - Dynamic allocation
  - Highest complexity

Challenges

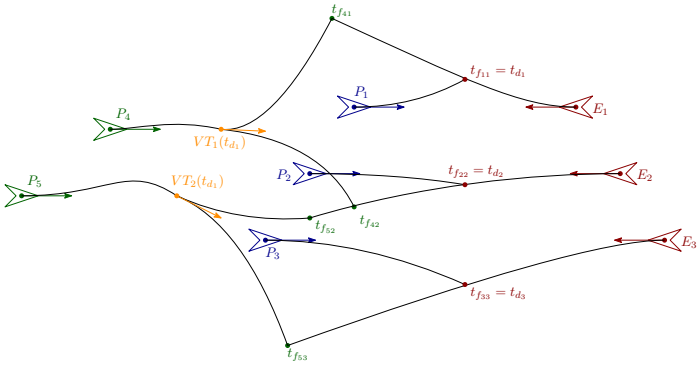
- Combinatorics make WTA computationally hard
- With the additional complexity of
  - Need to recompute, not just one shot
  - Incorporation of virtual target placement and selection

Objective

Dynamic WTA strategy for Shoot-Shoot-Look scenario

Scenario and Objective II

Engagement Example: 5 vs. 3



Solution approach – RL with decentralized decision making

DWTA – at each time instance, choose next backup interceptor allocation (VT or target) to **eliminate maximal number of real targets**.

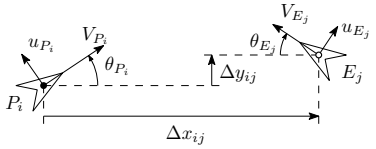
VT = Position + Heading

Assumptions:

- Linearized engagement
- Decision times known a-priori
- Perfect information
- **Bounded interceptor maneuver**
- Predictable target motion
- Sequential decision making
- Intercept probabilities are fixed (initially identical)

Engagement Kinematics

Linearized Kinematics

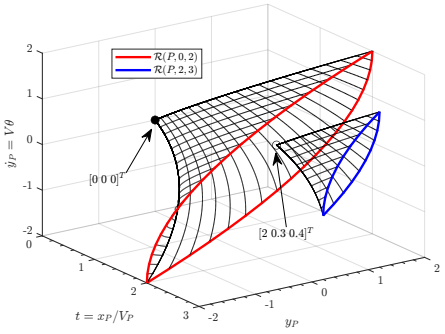


$$\dot{x}_{P_i} = V_{P_i} \qquad \dot{x}_{E_j} = -V_{E_j}$$
$$\dot{y}_{P_i} = V_{P_i}\theta_{P_i} \qquad \dot{y}_{E_j} = V_{E_j}\theta_{E_j}$$
$$\dot{\theta}_{P_i} = u_{P_i}/V_{P_i} \qquad \dot{\theta}_{E_j} = u_{E_j}/V_{E_j}$$

$\rightarrow$

analytical solution

Reachability Sets

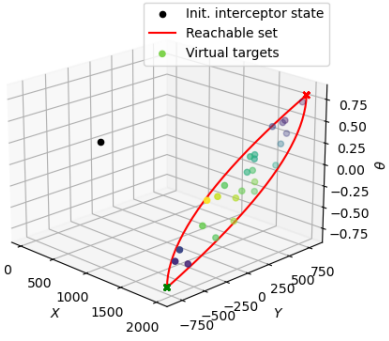


Definition

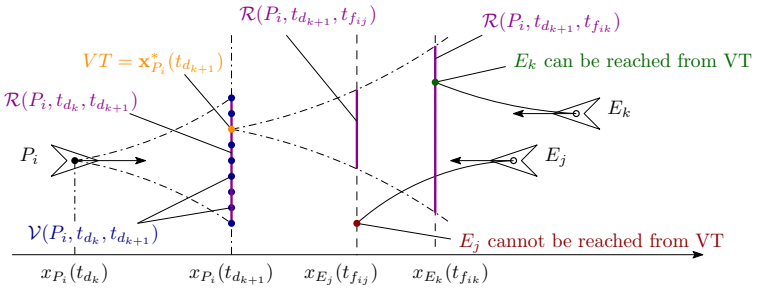
Reachable set at t is the set of all states $[x_k(t)\ \ y_k(t)\ \ \theta_k(t)]^T$ that can be achieved from the initial state $[x_k(t_0)\ \ y_k(t_0)\ \ \theta_k(t_0)]^T$ using PWC control $u_k < |u_k^{max}|$

VT Selection

VT Reachability



VT Sampling

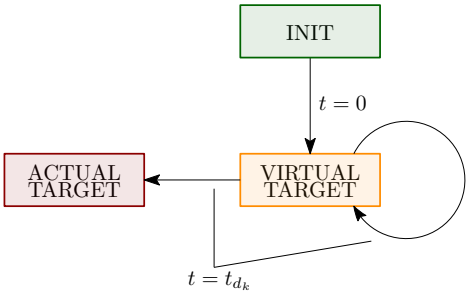


Definition

The evader is **covered** from the VT if its future position is inside the interceptor reachable set associated with this VT.

Interceptor Decision Making

Decision Flowchart



Information Available to Interceptor

- Kinematics → VT coverage
  - VT's coverage for current interceptor
  - Coverage of VT's for previous interceptors
  - Coverage of all VT's of next interceptors
- Target status
  - Free
  - Engaged
  - Destroyed

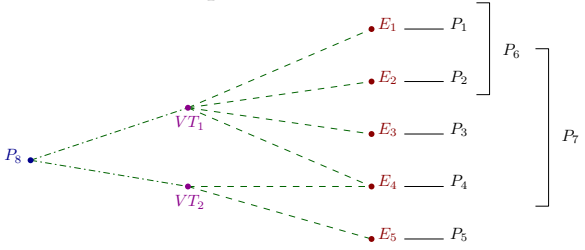
Greedy Heuristic Algorithm

Heuristic Idea

Added benefit

Choose VT that has largest coverage addition

Example VT Evaluation



Greedy Algorithm

1. If there is an unengaged evader – select free pursuer by reachability and queue position
2. If all evaders engaged
   - 2.1 Assign score to each evader before allocation of current interceptor

$$S_j = \frac{1}{1+q}, \quad q - \#\text{ pursuers covering } E_j$$

   - 2.2 Compute updated score with the current interceptor

$$S_j^k = \frac{1}{1+q+c}, \quad c \in \{0,1\}$$

   - 2.3 Assign score (**added benefit**) to each virtual target

$$\boxed{S_{VT_i} = \left\| \mathbf{S}^k - \mathbf{S} \right\|}$$

   - 2.4 Select VT with highest score

RL Algorithm

Algorithm steps:

1. If there is an unengaged evader – select free pursuer by reachability and queue position
2. Otherwise select VT as RL action

- **Environment** – N vs. M linearized engagement
- **Action Space** – L VT choices
- **State**:
  - Current interceptor VT coverage
  - VT coverage for all free interceptors
  - Coverage of selected virtual targets for interceptors prior in queue
  - Status of real targets: free, occupied, or intercepted
- **Network Architecture**: Fully connected Actor-Critic each with hidden layers of 512, 128 and 64 neurons respectively and RELU activation
- **Reward** – added allocation benefit (same as Greedy)
- **Training** – $2 \cdot 10^8$ steps

Scenario Example

- # Pursuers = 8, 6 – first wave, 2 – backup
- # Evaders = 6
- # VT = 4

Statistical Analysis

- # Pursuers = 24, 20 – first wave, 4 – backup
- # Evaders = 20
- Intercept probability $p = 0.8$

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)

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%

Conclusions

- **Formulation of dynamic WTA problem in Shoot-Shoot-Look scenario with excess of interceptors**
- Derived lower bound on performance
- Proposed two algorithms:
  - Greedy coverage heuristic
  - RL algorithm
- Viable performace for both algorithms
- RL slightly better than Greedy in investigated scenarios

Thank you for your attention!

Performance Lower Bound

Assumptions:

1. No reachability constraints
2. 3 intercept waves approximation

Let s_1 – # targets survived first wave. Calculate $Pr(\text{G.H.} = k)$ – ?

Case 1: $s_1 \geq N - M$ (all second-wave interceptors engage targets)

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

Case 2: $s_1 < N - M$ ($N - M - s_1$ second-wave interceptors engage targets)

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

Result: $\Pr(\text{G.H.} = k) = \sum_{s_1=k}^{M} Pr(\text{G.H.} = k, s_1)$

Interceptor Guidance Laws

Note: Guidance laws used only for trajectory generation.

Trajectory-shaping guidance:

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

Augmented proportional navigation (APN):

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