Proceedings of AIAA SCITECH 2022 Forum, 3-7 January, 2022, San Diego, CA, USA and Virtual Event
Paper number AIAA 2022-2104, DOI:10.2514/6.2022-2104

Development of Reinforcement Learning Based Mission Planning Method for Active Off-board Decoys on Naval Platforms

Enver Bildik*, Burak Yuksek†, Antonios Tsourdos‡, Gokhan Inalhan§
School of Aerospace, Transport and Manufacturing
Cranfield University, United Kingdom, MK43 0AL

Abstract

In this paper, reinforcement learning based decoy deployment strategy is proposed to protect naval platforms against radar seeker equipped anti-ship missiles. Decoy system consists of a rotary-wing unmanned aerial vehicle (UAV) and an integrated onboard jammer. This decoy concept enables agility which is quite critical for jamming operations against a high-speed anti-ship missile. There are two main purposes of the developed jamming strategy; a) flying in the field of view of the anti-ship missile to conceal the naval platform, and b) flying away from the target ship to increase the miss distance between the anti-ship missile and naval platform. Here, it is aimed to meet these requirements simultaneously. Kinematics models are used to represent missile, decoy UAV and target motion. Jammer and seeker signal strengths are modeled and radar-cross section of a frigate is utilized to increase the realism of the simulation environment. Deep Deterministic Policy Gradient (DDPG) algorithm is applied to train an actor-critic agent which maps the observation parameters to decoy's lateral acceleration. A heuristic way is chosen to create appropriate reward function to solve the decoy guidance problem. Finally, simulations studies are performed to evaluate the system performance.

I. Introduction

Naval platforms can be considered as a portable space for war vehicles, and they are moving slowly due to their heavy structure. Because of housing significant assets, the safety of these platforms becomes a major issue in modern warfare. During the maritime missions, naval platforms operated in combat area are under multi-directional anti-ship missile threats equipped with advanced radar seekers. Therefore, an effective defence strategy should be carried out to enhance the survival probability of target ships from approaching threats. Many concepts have been designed to develop a decoy strategy that performs a countermeasure/soft kill mission against missile threats. On-board countermeasures are one of the first attempts to conceal main platform from approaching threats. However, due to intelligent tracking algorithms executed by the radar seeker technology degrade the effectiveness of this technique.

Protecting of these naval platforms has been studied for years because of their importance in the warfare. Several methods have been developed in literature based on active and passive protection. Passive systems that are applied as decoys are not reactive to a specific threat, and their structure are utilized to mimic radar cross section signature, such as floating decoy system. In [1], a towed-decoy launched from an aircraft is analysed to evaluate its performance against an anti-air missile which is equipped with a monopulse radar seeker. Relative distance between missile and target is called as a miss distance, and it is an important metric to evaluate the effectiveness of decoys. The signal strength, tether length, and release direction of the decoy are parameterised to be utilized for different scenarios. However, because of the tether between towed-decoy and mother platform, manoeuvre capability of the aircraft has been restricted. In [2], the author examines the use of off-board active decoys against anti-ship missiles. These decoys can be deployed by a helicopter or launched by a rocket and then kept in air by a parachute system. They emit radiation that can jam the radar signal of the seeker and direct the missile away from the ship. The decoy deployment time and decoy launch direction are critical parameters, so they are examined to determine the success of decoy mission. The authors in [3] evaluate the effectiveness of jamming strategy carrying out by a group of unmanned air vehicles (UAVs) flying based on a close formation. UAVs move cooperatively to expand the jammed area by superimposing their jamming signals. Moreover, a methodology is proposed for UAVs to follow a minimum-risk path planning within the zone where surface-to-air missile radars exist. In [4], feasibility of electronic attack executed by multiple autonomous vehicles against integrated air defence systems (IADS) is discussed. Resource allocation and cooperative path planning are highlighted problems to be formulated in this

*PhD. Student, e.bildik@cranfield.ac.uk
†Postdoctoral Research Fellow, burak.yuksek@cranfield.ac.uk, AIAA Professional Member.
‡Head of Centre for Autonomous and Cyber-physical Systems, a.tsourdos@cranfield.ac.uk.
§BAE Systems Chair, Professor of Autonomous Systems and Artificial Intelligence, inalhan@cranfield.ac.uk, AIAA Associate Fellow.

context. A collaborative decoy jamming strategy is proposed in [5] to degrade the performance of the inverse synthetic aperture radar (ISAR) by utilizing a group of small-scale UAVs. Also, a cooperative decision-making algorithm is applied for coordination of multiple UAVs. A same-side-deployment and zig-zag deployment strategies are proposed in [6] to evaluate the efficiency of a single and multiple decoys. For aircraft targets, an analytical expression is developed in [7] to perform the optimal deployment of decoys and vertical-S manoeuvre strategy, simultaneously. The survival probabilities of the radar against anti-radiation missiles are evaluated for the cases in which decoys are positioned in an appropriate quadrangular topology [8]. Effectiveness of decoys in the evader-pursuer engagement scenario is analysed in [9] for various decoy launch angles and launch time. The range of launch angles and launch time which ensure that the decoy remains within the radar seekers' field-of-view (FOV) are derived in [10] based on the intersection point between FOV's boundaries and the loci of decoy position. A method to simulate echo signal at the tracking radar is proposed in [11], and the effects of circular and linear polarization of signals are analysed for repeater-type active decoys against ground tracking radar. Jamming performance of an active repeater decoy is evaluated in [12] based on RF specifications of the decoy, such as antenna patterns, and amplifier gains. In [13] , a ducted-fan flight array system is allocated for the role of decoy to guarantee the protection of the target ship against anti-ship missiles (ASMs). A decoy deployment strategy is developed by means of a sequential logic algorithm. Q-learning based reinforcement learning algorithm is designed in [14] for decoy guidance to direct the allocated decoy to the optimal direction. In [15], an auction-based task assignment algorithm is applied to effectively manage the decoy mission conducted by multiple UAVs against anti-ship missiles.

In this study, main aim is to protect the ship from an approaching missile threat by deploying a decoy which gets the attention of the missile seeker. The mission success depends on both signal power and path of the decoy. Here, path planning plays a crucial role to settle the decoy to an appropriate location. An idea is proposed to localize the decoy to the right position where the decoy stays within the field of view of radar seeker. For this purpose, Deep deterministic policy gradient algorithm is applied to train an agent which predicts the acceleration of the decoy with respect to changing dynamics environment. In each simulation case, decoy deployment angle is set randomly from a predefined angle interval. Based on the assigned deployment angle, the agent can predict the best action for optimal solution.

The rest of this paper is organized as follows: Section II formulates and explains kinematics models of the ship, the missile, and the decoy, respectively. Section III provides fundamental information about reinforcement learning, and then gives details about deep deterministic policy gradient algorithm applied in this study. Section IV describes the implementation stage of the proposed algorithm and shares results about the effectiveness of decoying strategy. Concluding remarks and future works are given in section V.

II. Problem Definition

A mission is envisioned in which an off-board decoy deployed from the target ship aims to lure the anti-ship missile away from the main platform. Models utilized in the scenario has a substantial impact on the accomplishment of the mission. That is why, firstly, kinematics models for the target ship, anti-ship missile, and decoy are introduced. In the case of multiple targets, simply, a radar seeker tends to move towards the direction of the centroid of signal power source of targets. The main idea of this work is to demonstrate the power of reinforcement learning based approach in terms of decoying strategy. As illustrated in the Figure 1, the scenario is that as soon as approaching missile is detected by the ship sensor, decoy mission is executed, and decoy is ejected from the ship to the appropriate position that ensures the protection of the ship.

A. Target ship kinematics model

It is assumed that the ship moves in 2D environment, along East and North axis. A simple point mass kinematics equation is described as,

$$\begin{bmatrix} \dot{X}_s \\ \dot{Y}_s \\ \dot{\phi}_s \end{bmatrix} = \begin{bmatrix} V_s cos\phi_s \\ V_s sin\phi_s \\ \omega_s \end{bmatrix} \tag{1}$$

where X and Y indicate the position of the ship in the East (x) and North (y) axis respectively. V_s and ω_s are ground velocity of the ship and the turning rate of the ship respectively. During simulation V_s is fixed as 15 m/sec. Another important parameter for target ship is Radar Cross Section (RCS) which is used to calculate the reflected signal from the target to the radar seeker. RCS is a plenty significant design factor for stealth of a naval ship, so as to decrease

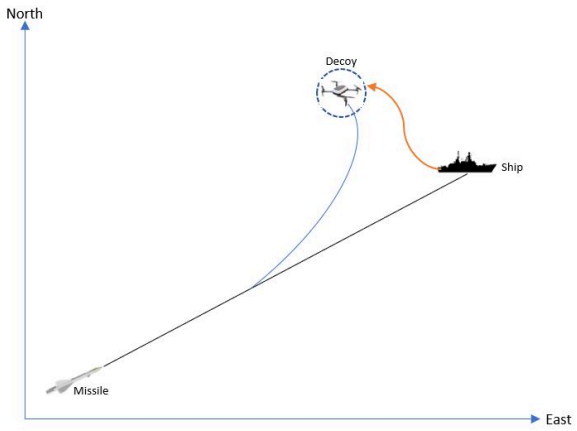


Fig. 1 Missile Target Decoy engagement

the detectability of naval platforms ship designers perform some methods. However, due to huge structure of naval platforms, it may not be possible to make them full stealth. The radar cross section of the ship depicted in the Figure 2 is computed by means of toolkit created by Pofacets [16]. After this section, the ship will be named as a target.

B. Missile kinematics model

In this study, it is assumed that missile updates its information about the target ship by using a radar seeker. For the kinematics model of the anti-ship missile (ASM), a point mass model utilized. The below equation is given to express an ASM model as;

$$\dot{X}_m = V_m cos\phi_m \tag{2}$$

$$\dot{Y}_m = V_m sin\phi_m \tag{3}$$

where X_m, and Y_m depict the positions of the missile in the inertial frame. V_m and ϕ_m are the velocity and heading angle of the ASM respectively. A 2D planar missile-target engagement geometry [17] is illustrated in Figure 3 to better figure out the conceptual approach of PNG. Subscripts M and T give information about missile and target respectively. The V_M and V_T represent the magnitude of velocity of missile and target respectively. It can be seen in the figure, the missile moves, with a velocity denoted as V_M, at an angle of L+HE with respect to the line-of-sight (LOS). The missile lead angle is defined as the angle L, and the angle HE illustrates the heading error which is known as an initial deviation angle of missile. The line linking the missile and target during the engagement is termed as line of sight. This line creates an angle of lambda (λ) with respect to reference point.The mission allocated to the missile is to catch the target as short time as possible. Miss distance is well-known as the closest point between missile and target. The expected miss distance to hit a target should be as close to zero as possible, but in reality, impossible. Some basic mathematical equations are given below to derive the required parameters employed for lateral acceleration. The R_{TM} is a relative separation between target and missile during engagement, and it is calculated as;

$$R_{TM} = \sqrt{(R_{TMe}^2 + R_{TMn}^2)} \tag{4}$$

As seen in figure 3, by means of trigonometry, it is straightforward to find the line-of-sight angle (λ) as;

$$\lambda = \arctan\left(\frac{R_{TMe}}{R_{TMn}}\right) \tag{5}$$

The relative velocity components of the target and missile are calculated separately for each axis as below;

$$V_{TMe} = V_{Te} - V_{Me} \tag{6}$$

$$V_{TMn} = V_{Tn} - V_{Mn} \tag{7}$$

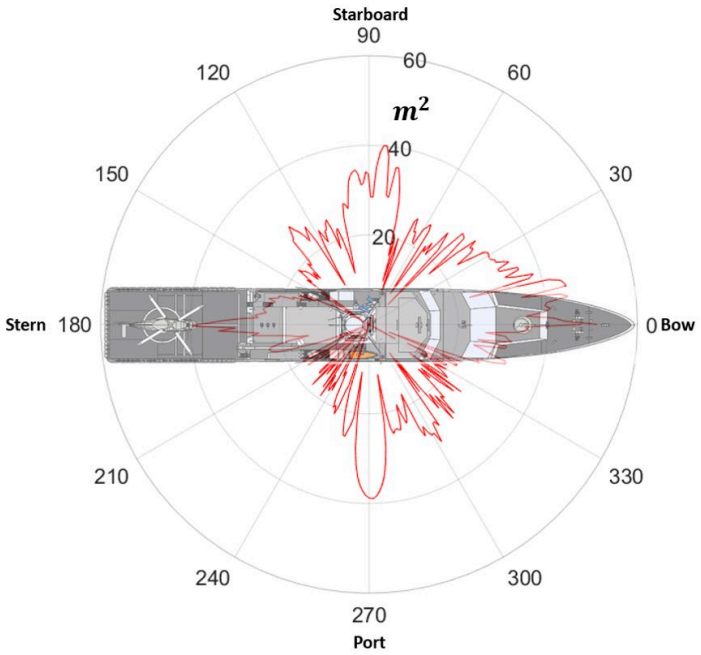


Fig. 2 Radar Cross Section of the frigate

The line-of-sight rate is derived by the differentiation of line-of-sight angle equation. After some simplifications the value is obtained as;

$$\dot{\lambda} = \frac{R_{TMe}V_{TMn} - R_{TMn}V_{TMe}}{R_{TM}^2} \tag{8}$$

The closing velocity is described as the negative rate of change of the missile target separation i.e R_{TM}, by applying differential equations law, it can be easily calculated as

$$V_c = \frac{-(R_{TMe}V_{TMe} + R_{TMn}V_{TMn})}{R_{TM}} \tag{9}$$

Lastly, by the definition of proportional navigation guidance law, the magnitude of the missile guidance command n_c is received as;

$$n_c = NV_c\dot{\lambda} \tag{10}$$

Here, N is constant and its value is between 3 and 5.Usually, the identification of the target by the radar seeker is accomplished with respect to the reflected radar signal from the target to the seeker. The back scattered signal to the seeker is formalized as

$$S = \frac{PG_tG_r\lambda^2\sigma}{(4\pi)^3R^4} \tag{11}$$

where P is the output power of the seeker, and G_t,and G_r, are the transmitter and receiver gains of the seeker, respectively. The λ is the wavelength of the seeker, and σ is equal to the radar cross section (RCS) of the target ship.

C. Decoy kinematics model and jammer

A UAV equipped with a jammer is considered as a decoy to lure the approaching missile threat away from the ship. It is ejected from the ship and continue its motion in 2D environment. The motion equation is calculated based on the point mass model.

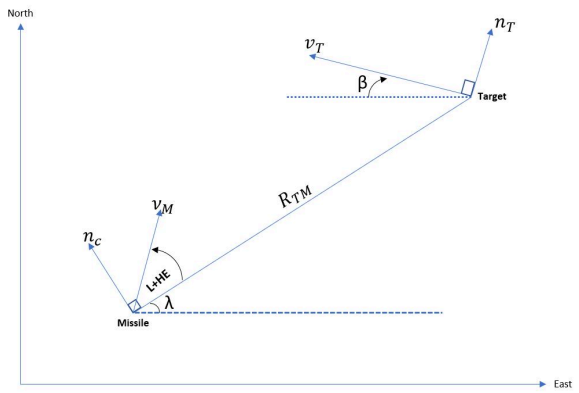


Fig. 3 Missile target engagement

$$\begin{bmatrix} \dot{X}_d \\ \dot{Y}_d \\ \dot{\phi}_d \end{bmatrix} = \begin{bmatrix} V_d cos\phi_d \\ V_d sin\phi_d \\ \omega_d \end{bmatrix} \tag{12}$$

where $[X_d\ Y_d]$, ϕ_d and V_d represent coordinates of the decoy, heading angle, and the velocity of the decoy respectively. The signal strength of the decoy during engagement is calculated as [2]

$$J = \frac{P_d G_d G_r \lambda^2}{(4\pi)^2 R^2} \tag{13}$$

where P_d, G_d and G_r symbolize the jammer's output power, the transmitter gain, and receiver gain of the decoy jammer respectively. In this study, values used for the calculation of the seeker and decoy signal power are taken from this paper [15], and are seen in the table 1.

Table 1 Seeker/decoy signal parameters.

Parameters	Values	Units
P_k	200	kw
G_t	35	dB
G_r	35	dB
λ	0.003	m
σ	50	m^2
P_d	1	kW
G_d	35	dB

III. Reinforcement learning Approach for Decoy Guidance

Reinforcement learning is a promising sub-field of machine learning, and its learning ability is dependent on the interaction between the agent and the environment [18]. The aim of reinforcement learning is to maximize the cumulative long-term reward in a Markov Decision Process (MDP). An MDP is formulated by a five-element tuple (S, O, A, R, P), where S represents the set of states, O the set of observations, A the set of actions, R the reward function (R: S×A → R) and P the state transition probability (P: S×A×A→ [0,1]). The fundamental working principle of reinforcement learning is that agent takes an action based on the current observations from the environment to move

to the next state, and simultaneously, the agent receives reward or penalty in return for the goodness/badness of taken action. Observation/states are the information received from the environment during interaction. There is a minor distinction between observation and state, in short, state covers observation. While state is a full description of the world, observation is a partial description of the world. The action space is defined as a set of all possible actions in an environment. In discrete space case, the number of actions is finite while in continuous space case it is infinite. In reinforcement learning, after taking an action reward (positive or negative) is given as feedback to evaluate whether taken action is good or not. The cumulative reward at each timestep t can be computed by means of the equation as:

$$R(\tau) = r_{t+1} + r_{t+2} + r_{t+3} + r_{t+4} + \dots . \tag{14}$$

However, directly summation of rewards is not applicable in reality, so a new term is added to the equation which is named as discount. The future reward is discounted by the exponent of the time step to be comparable with the near reward. The near rewards are more probable to occur because they are more presumable than the long-term future reward. The modified cumulative reward equation is rewritten as:

$$R(\tau) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \dots . \tag{15}$$

which is equivalent to:

$$R(\tau) = \sum_{n=1}^{\infty} \gamma^k r_{t+k+1} \tag{16}$$

γ (gamma) is a discount rate which is bounded between 0 and 1. The bigger the gamma the lesser the discount, which means the agent focuses more about the long-term reward. On the other hand, the lesser the gamma the bigger discount which means the agent considers more about the short term reward. The exploration/exploitation trade-off is a crucial part of learning process in reinforcement learning. Exploration means trying random actions in order to discover new information about the world, meantime exploitation means using experienced information about the world to take action. There are two methods to train an RL agent, first is policy-based method, and other is value-based method. In the policy-based method, learning takes place directly, that directs the agent to find the most appropriate action which maximizes the cumulative expected return. In deterministic policy case, policy function tries to match each state to the best action, while in stochastic case, policy function outputs a probability distribution over the set of possible action at that state. In Value based methods, training takes place based on a Value function which map a state to the expected value of being at that state. The value of a state is the expected discounted return the agent can receive during interaction. Value Function is a very indirect process to decide the best action. Instead, "Action–Value Function", denoted by Q (s, a) is widely used for this purpose. One of the most popular methods to compute the Q-value is Q-learning.

A. Deep Deterministic Policy Gradient

The Deep Deterministic Policy Gradient is a model-free, online, and off-policy algorithm in reinforcement learning family. It is a conceptual integration of well-known Deep Q Networks (DQN) and Deterministic Policy Gradient (DPG) algorithms to learn a deterministic policy which outputs a continuous action in an unknown environment. The structure of DDPG consists of actor-critic learning algorithm, which is a combination of policy gradient function and value function. Like DQN, DDPG employs target networks to untangle instability during training caused due to the use of the Bellman Error. In this case, four neural networks are required namely actor, target actor, critic, and target critic to run this DDPG algorithm. The μ, Q, $\mu\prime$ and Q' represent actor, critic, target actor, and target critic respectively. The critic takes current state and the action predicted by the actor, and the output layer of the critic network with a single neuron generates the Q-value of the given state-action pair. As seen in the Figure 4, the task of the critic is to critique the actor regarding by determining the quality of the predicted action given the current state. Learning for the critic happens by using critic loss function derived below as [18],

$$y_i = r_i + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1} \mid \theta^{\mu'}\right) \mid \theta^{Q'}\right) \tag{17}$$

$$L = \frac{1}{N}\sum_i \left(y_i - Q\left(s_i, a_i \mid \theta^{Q}\right)\right)^2 \tag{18}$$

Thus, the loss depicted as L is calculated as the Mean-Squared Error between the TD-Target and the Q-value estimated for current state and corresponding action pair. The TD-Target y employs the target networks. The next state s' and the

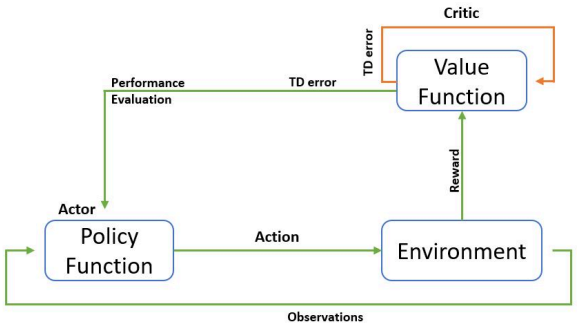


Fig. 4 Actor-Critic Algorithm

associated action predicted by the target actor $\mu\prime$ are provided as input to the target critic Q'. As the name suggests, in this algorithm the policy is deterministic, this refers that the actor network learns to map a given state to a specific action, instead of a probability distribution on the actions. The actor-network takes as input the state vector and generates the action to be executed based on the size of the action space. During training, a small amount of noise is added as a role of exploration to the action predicted by the actor to guarantee that the network does not get stuck in a local minimum. In this algorithm, a time-corelated noise depicted as ε formed by the Ornstein-Uhlenbeck process is utilized.

$$a = \mu\left(s\right) + \epsilon \tag{19}$$

The actor network is updated by executed a gradient ascent with respect to the policy, along the direction pointed out by the critic. Target network parameters are updated smoothly based on the value of τ (tau) parameter. The update equations are indicated below.The τ is between 0 and 1.

$$\theta' \leftarrow \tau \cdot \theta + (1-\tau) \cdot \theta' \tag{20}$$

$$\phi' \leftarrow \tau \cdot \phi + (1-\tau) \cdot \phi' \tag{21}$$

Here, θ', and ϕ' represent target critic and target actor respectively.

B. Training DDPG agent

In this section, the aim is to propose a DDPG algorithm that teaches the agent to execute the best action based on the given current observations. Here, the action is allocated as the acceleration of the decoy to create a path planning in which decoy is effective. The agent is trained with respect to the observation parameters, reward function and Isdone statue of the simulation environment. Observation parameters are related with missile states and decoy states, and namely they are missile positions (2x1), decoy positions (2x1), missile path angle (1x1) and decoy path angle (1x1).

$$\left[O\right] = \left[X_m, Y_m, X_d, Y_d, \phi_m, \phi_d\right] \tag{22}$$

The most challenging part of this study is to create an appropriate reward function which can guarantee to carry out the mission successfully. After some trial and error, it is observed that the below reward function can converge to an acceptable point. Reward function consists of three different terms with their coefficients, and it is formulated as

$$r_t = \omega_1 r_1 + \omega_2 r_2 + \omega_3 r_3 \tag{23}$$

where

$$r_1 = \frac{RTM}{RDM} \tag{24}$$

$$r_2 = 1 \tag{25}$$

$$r_3 = \begin{cases} 1, & \text{if } DFOV. \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Table 2 Initial parameters of missile, ship and decoy.

-	Position	Velocity	Flight path angle
Missile	[600;800] m	300 m/sec	85 (deg)
Ship	[3000;4000] m	15 m/sec	0 (deg)
Decoy	[3000;4000] m	15 m/sec	[165 180 195] (deg)

Algorithm 1 DDPG algorithm

Initialization of critic network $Q\left(s, a \mid \theta^{Q}\right)$ and actor network $\mu\left(s \mid \theta^{\mu}\right)$ with weights θ^{Q} and θ^{μ} respectively.

Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^{Q}, \theta^{\mu'} \leftarrow \theta^{\mu}$

Initialize replay buffer R

For each episode start with 1 , until the M

Initialize a random process $\mathcal{N}$ for action exploration

Receive initial observation state s_1

for t = 1, T do

Select action $a_t = \mu\left(s_t \mid \theta^{\mu}\right) + \mathcal{N}_t$ according to the current policy and exploration noise

Execute action a_t and observe reward r_t and observe new state s_{t+1}

Store transition $\left(s_t, a_t, r_t, s_{t+1}\right)$ in R

Sample a random minibatch of N transitions $\left(s_i, a_i, r_i, s_{i+1}\right)$ from R

Set $y_i = r_i + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1} \mid \theta^{\mu'}\right) \mid \theta^{Q'}\right)$

Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i \left(y_i - Q\left(s_i, a_i \mid \theta^{Q}\right)\right)^2$

Update the actor policy using the sampled policy gradient:

$\nabla_{\theta^{\mu}} J \approx \frac{1}{N}\sum_i \nabla_a Q\left(s, a \mid \theta^{Q}\right)\big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^{\mu}} \mu\left(s \mid \theta^{\mu}\right)\Big|_{s_i}$

Target critic update:

$\theta' \leftarrow \tau \cdot \theta + (1-\tau) \cdot \theta'$

Target actor update:

$\mu' \leftarrow \tau \cdot \mu + (1-\tau) \cdot \mu'$

end for

end for

Coefficients ω_1, ω_2, and ω_3 are three constant to shape the reward function, and they are given in the Table 3. The notation RTM and RDM denotes the range between the target and missile, and the range between the decoy and missile respectively, and DFOV represents that decoy is within the field of view. The first term r_1 relates the relative distance between the missile and the target, and relative distance between the missile and the decoy. This rate indicates that whether the missile is more close to the decoy than the target. That means the r_1 is greater in the case that the decoy is hit by the missile instead of the target. The second term r_2 is a constant value and it is used as a penalty to complete the mission as short time as possible. The appointed value is depicted in the Table 3, and it is given to the agent for each time step during training. The third term r_3 is encouraging the decoy to stay within the field of view of radar seeker during the mission. Because, if the decoy moves out of the field of view, the mission will be unsuccessful. Isdone statue is used to terminate the training session for each episode if any of the situations where RDM and RTM are smaller than the fuze range occurs. 40m is assigned to fuse range.

$$Isdone = \begin{cases} 1, & \text{if RDM} < \text{Fuse range.} \\ 1, & \text{if RTM} < \text{Fuse range} \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Table 3 Reward weights

ω_1	ω_2	ω_3
1	-0.1	0.2

The training scenario is run in which the missile firstly moves towards the target. After decoy deployment, based on the signal powers coming from the ship and the decoy, missile choose a target to hit. Before training the agent, some required specifications are carried out. The observation and action are 6 by 1 dimensional and 1 by 1 dimensional respectively.In observation, state parameters have different scale and units therefore for training efficiency it is required to normalize these parameters. The action represents the decoy's acceleration to create an optimal path trajectory, and it is a continuous value bounded between $-1.5\,g$ and $1.5\,g$. By an optimum action value, the cumulative reward for each episode will be maximized. Initialization is done in each episode, an angle randomly chosen from angle data given above is assigned as a decoy deployment angle, and missile and target positions are randomly chosen from bounded values given in the Table 2 as well. In the initial stage, the decoy and the ship is collocated. As soon as simulation starts, the decoy launched from the ship with a heading angle to move towards a point that can guarantee the survivability of the ship. The information regarding these layers are given in the Table 4. A rectified linear units (Relu) function formulated in Eq. 28 is applied as an activation function for each neuron in layers, except for that in the actor output layer, tanh function 29 is employed.

$$g(z) = \begin{cases} z, & \text{if } z >= 0. \\ 0, & \text{if } z < 0. \end{cases} \tag{28}$$

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{29}$$

After several trial-and-error tests, it is decided to assign numeric values given in the Table 5 as hyperparameters.

Table 4 Network layer properties

Layer	Critic Network	Actor Network
Input layer	7 (dimension of observation + action)	6 (dimension of observations)
Hidden layer 1	400	400
Hidden layer 2	300	300
Hidden layer 3	300	-
Output layer	1 (dimension of Action-Value function)	1 (dimension of action)

IV. Simulations Results

Extensive numerical simulations are carried out to obtain the preliminary results and training results. In the preliminary results, we observed in which conditions the decoy can execute the mission successfully. The expectetion from the trained agent which guides the decoy is that the decoy move towards to the optimal point where the mission can be achieved.

A. Preliminary Simulations

Before mentioning training results, it is worth to talk about the simulation environment that is a testbed of the proposed approach in this study. Initially, the missile locks-on the target and the PNG law which guides the missile during engagement is applied based on target ship parameters. After decoy is activated, the radar seeker resolves two targets, and based on the signal powers returned from them to radar seeker, the missile chooses one to hit. For simplicity,

Table 5 Hyperparameters settings

Parameters	Value
Maximum episodes number	4000
Maximum steps number	2000
Critic learning rate	1e-3
Actor learning rate	1e-4
Experience Buffer Length	1e6
Discount Factor	0.99
Mini Batch Size	256
Sample Time	0.01
Target Smooth Factor	1e-3
Noise standard deviation	0.9
Noise standard deviation decay rate	1e-6

the signal power computation of seekers and jammers have been converted into log scale (dBm). Consequently, 11 and 13 can be rewritten as

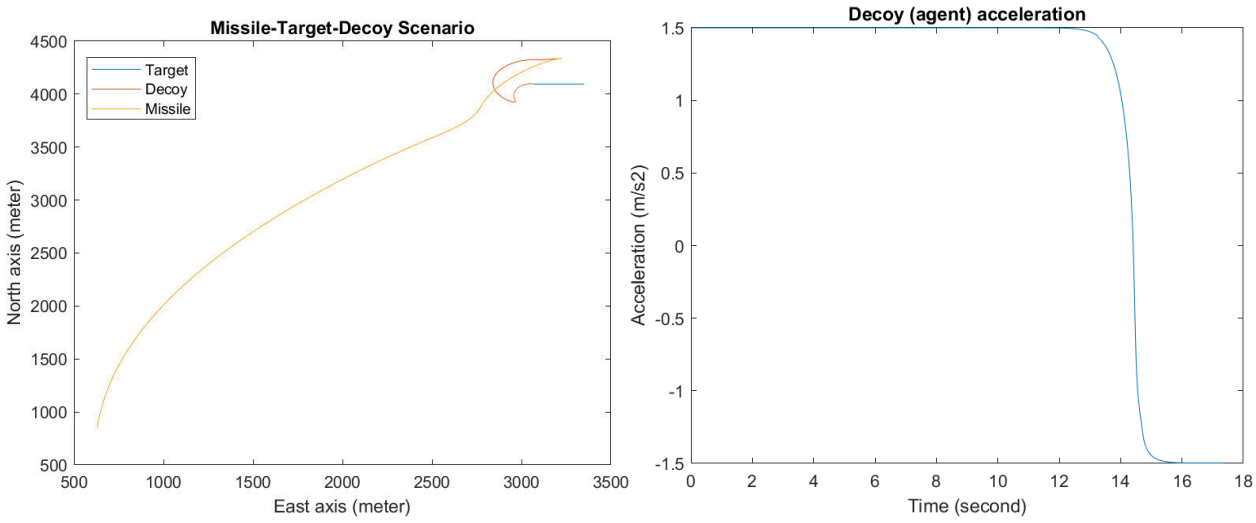
$$S = 10\log(P_S) + G_{S,t} + G_{S,r} + 10\log(\sigma) - 40\log(R) - 20\log(F) - 163.4 \tag{30}$$

$$J = 10\log(P_D) + G_{D,t} + G_{D,r} - 20\log(R) - 20\log(F) - 92.45 \tag{31}$$

These notations are explained above, so we don't need mentioning about it again.In each sample time, signal strength calculations are done.In each sample time, signal strength calculations are done based on the given values in the Table 1. As long as the ship and the decoy are within the field of view of radar seeker, their signal powers are taken into account, otherwise they will be eliminated. A task is accepted to switch seen target by the radar seeker between the ship and the decoy based on the their power signal and field of view statue. The backscattered signal to the seeker from the target ship vary with respect to the angle between the missile and the ship, and this case is taken into account in this study.

B. Training Results

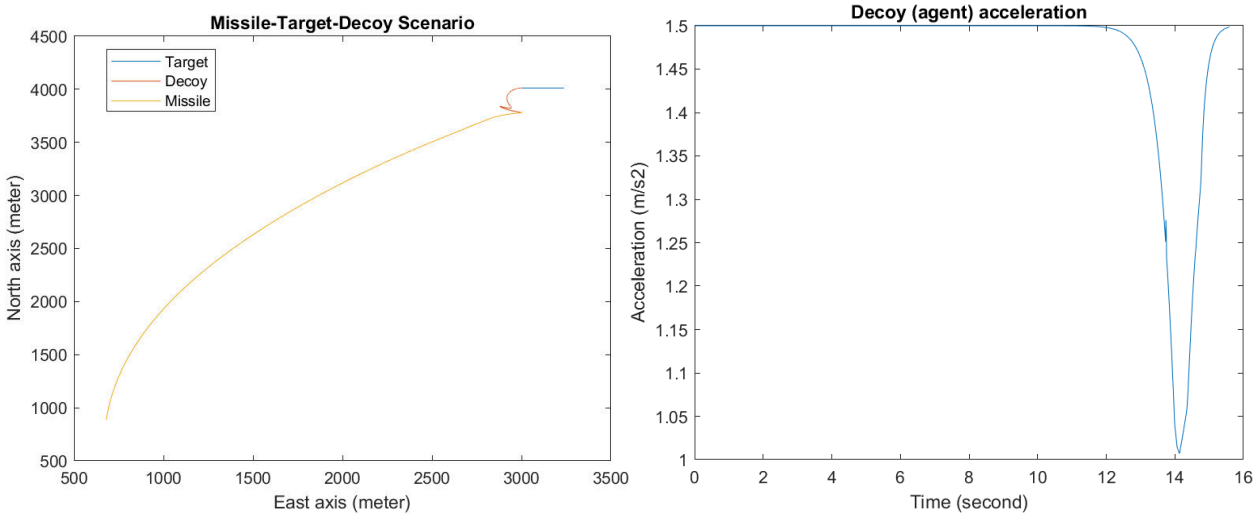
Training process of the agent is started as soon as requirements are completed, and it continues until the maximum number of episodes reach the 4000. After thousands of episodes, the curve of the average reward begins converging. The agent tries to learn to take the best action which navigates the decoy to move toward the optimal position in which miss distance can be maximum. Miss distance which is the range between the missile and the ship is a metric to evaluate the performance of the proposed approach. In each episode, initially the ship and the decoy are co-located, and the positions of the missile and the target are randomly chosen from the minimum and maximum values. The Figure 5 is obtained in the case decoy deployment angle is equal to 165° degree. The Figure 5b depicts the change in the acceleration of the decoy during the implementation of the scenario under dynamics environment. As can be seen that the agent learned taking reasonable action for each observation situation.The Figure 5a depicts the positions of the missile, the ship, and decoy during the simulation. The ship moves align east axis, and decoy is ejected from the stern part of the ship. As seen is the Figure 5b, the decoy's acceleration value start with a positive value, and after a while the sign change. After deployment, the decoy followed the created path to attract the missile which locks-on the target. The obtained miss distance in this case is 270 meter, but it can be claimed that at that point even though the radar seeker realizes the decoy, the re-lock-on the target can be difficult because of deceived heading angle.



(a) Path trajectories **(b) The change in acceleration**

Fig. 5 Decoy deployment angle = 165°

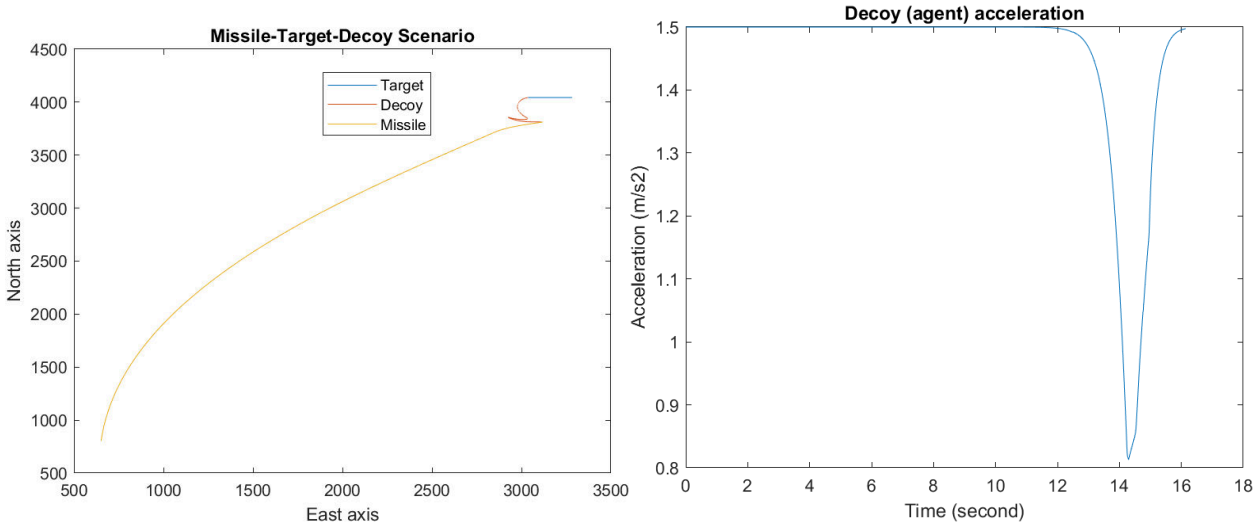
In the case when decoy deployment angle is equal to 180 ° which means opposite direction of the ship motion trajectory, results are presented in Figure 6. As can be seen in the Figure 6b, the agent provides acceleration value between 1.5 and 1. After 12th seconds, the agent gave unexpected acceleration value for a very short time, but this situation did not affect the path trajectory of the decoy. The decoy moves towards the port side of the ship to lure the approaching missile threat as demonstrated in the Figure 6a. The obtained miss distance in this case is 340 meter, but the target can be on the verge of danger. Because after the missile arrives the decoy, it still has enough time to re-scan the real target.



(a) Path trajectories **(b) The change in acceleration**

Fig. 6 Decoy deployment angle = 180°

Figure 7 provides information about the instant position of the missile, target and decoy during engagement, also agent's response based on the given current observations. The acceleration indicated in the Figure 7b takes a value between 0.8 and 1.5, and the path planning for the decoy is created based on the this value. The decoy achieves staying within the field of view of the radar seeker during the simulation time, and this will make the mission a success. A 330 meter miss distance is obtained when the decoy deployment angle is 195°. However, as the previous case, the target can



(a) Path trajectories **(b) The change in acceleration**

Fig. 7 Decoy deployment angle = 195°

be on the verge of danger in this case too because of that the missile having sufficient time to maneuver towards to the real target.

Table 6 Miss distance between Missile and Ship

Decoy deployment angle (degree)	165°	180°	195°
Miss distance (meter)	270	340	330

V. Conclusion

In this paper, a decoy deployment strategy is proposed to enhance the survival probability of the naval platform under one-decoy/one-missile threat case. A single UAV equipped with a jammer is considered as a decoy. 2D point mass kinematics equation was employed to model the ship, the missile, and the decoy respectively. Also, the strength of the back scattered signal from the target and jammer signal is calculated. Furthermore, general concept of the Deep deterministic policy gradient reinforcement learning algorithm is given and then observation and reward function utilized during training process are created to provide an optimal path planning to the decoy. To evaluate the performance of the proposed AI-based approach, a numerical simulation is carried out, and miss distance between the naval platform and anti-ship missile is utilized as a performance metric. Decoy deployment angle is critical parameter to accomplish the mission, a limited range is created to randomly assign a value to the angle. It is seen that to extend the range of assigned deployment angle interval, a huge number of episodes is required and it is necessary to perform long training processes. This part is also considered as future work. Moreover, future works include developing multi-agent algorithms to create an optimal path for multiple decoys against multiple missile threats.

Acknowledgments

This work is supported, in parts, by the Engineering and Physical Sciences Research Council [Grant number: EP/V026763/1].

References

[1] Yeh, J.-H., "Effects of towed-decoys against an anti-air missile with a monopulse seeker," Ph.D. thesis, Monterey, California. Naval Postgraduate School, 1995.

[2] Tan, T.-H., "Effectiveness of Off-Board Active Decoys Against Anti-Shipping Missiles." Tech. rep., NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 1996.

[3] Kim, J., and Hespanha, J. P., "Cooperative radar jamming for groups of unmanned air vehicles," *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, Vol. 1, IEEE, 2004, pp. 632–637.

[4] Mears, M. J., "Cooperative electronic attack using unmanned air vehicles," *Proceedings of the 2005, American Control Conference, 2005.*, IEEE, 2005, pp. 3339–3347.

[5] Ilaya, O., Bil, C., and Evans, M., "Distributed and Cooperative Decision Making for Multi-UAV Systems with Applications to Collaborative Electronic Warfare," *7th AIAA ATIO Conf, 2nd CEIAT Int'l Conf on Innov and Integr in Aero Sciences, 17th LTA Systems Tech Conf; followed by 2nd TEOS Forum*, 2007, p. 7885.

[6] Akhil, K., Ghose, D., and Rao, S. K., "Optimizing deployment of multiple decoys to enhance ship survivability," *2008 American Control Conference*, IEEE, 2008, pp. 1812–1817.

[7] Vermeulen, A., and Maes, G., "Missile avoidance maneuvres with simultaneous decoy deployment," *AIAA Guidance, Navigation, and Control Conference*, 2009, p. 6277.

[8] Zhou, W., Luo, J., Jia, Y., and Wang, H., "Performance evaluation of radar and decoy system counteracting antiradiation missile," *IEEE transactions on aerospace and electronic systems*, Vol. 47, No. 3, 2011, pp. 2026–2036.

[9] Ragesh, R., Ratnoo, A., and Ghose, D., "Analysis of evader survivability enhancement by decoy deployment," *2014 American Control Conference*, IEEE, 2014, pp. 4735–4740.

[10] Ragesh, R., Ratnoo, A., and Ghose, D., "Decoy Launch Envelopes for Survivability in an Interceptor–Target Engagement," *Journal of Guidance, Control, and Dynamics*, Vol. 39, No. 3, 2016, pp. 667–676.

[11] Rim, J.-W., Koh, I.-S., and Choi, S.-H., "Jamming performance analysis for repeater-type active decoy against ground tracking radar considering dynamics of platform and decoy," *2017 18th International Radar Symposium (IRS)*, IEEE, 2017, pp. 1–9.

[12] Rim, J.-W., and Koh, I.-S., "Effect of beam pattern and amplifier gain of repeater-type active decoy on jamming to active RF seeker system based on proportional navigation law," *2018 19th International Radar Symposium (IRS)*, IEEE, 2018, pp. 1–9.

[13] Jeong, J., Yu, B., Kim, T., Kim, S., Suk, J., and Oh, H., "Maritime application of ducted-fan flight array system: Decoy for anti-ship missile," *2017 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS)*, IEEE, 2017, pp. 72–77.

[14] Rajagopalan, A., "Active Protection System Soft-Kill Using Q-Learning," *International Conference on Science and Innovation for Land Power, Australia Defence Science and Technology*, 2018.

[15] Dileep, M., Yu, B., Kim, S., and Oh, H., "Task Assignment for Deploying Unmanned Aircraft as Decoys," *International Journal of Control, Automation and Systems*, Vol. 18, No. 12, 2020, pp. 3204–3217.

[16] Chatzigeorgiadis, F., and Jenn, D., "A MATLAB physical-optics RCS prediction code," Vol. 46, 2004, pp. 137–139.

[17] Zarchan, P., *Tactical and strategic missile guidance*, American Institute of Aeronautics and Astronautics, Inc., 2012.

[18] Sewak, M., *Deep reinforcement learning*, Springer, 2019.

2021-12-29

Development of reinforcement learning based mission planning method for active off-board decoys on naval platforms

Bildik, Enver

AIAA

Bildik E, Yuksek B, Tsourdos A, Inalhan G. (2021) Development of reinforcement learning based mission planning method for active off-board decoys on naval platforms. In: AIAA SciTech 2022 Forum, 3-7 January 2022, San Diego, CA, USA and Virtual Event, Paper number AIAA 2022-2104

https://doi.org/10.2514/6.2022-2104