DOI: 10.13009/EUCASS2023-518
Aerospace Europe Conference 2023 – 10TH EUCASS – 9TH CEAS

Deep reinforcement learning based integrated guidance and control for a longitudinal missile system

Jeongsu Ahn[⋆]*, Jongho Shin*[⋆†] *and Hyeong-Geun Kim*[⋆⋆]
[⋆]*School of Mechanical Engineering, Chungbuk National University*
[⋆] *Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644, Korea*
[⋆⋆]*School of Mechanical Engineering, Incheon National University*
[⋆⋆] *Academy-ro 119, Yeonsu-gu, Incheon, 22012, Korea*
hoeunsu3@cbnu.ac.kr · jshin@cbnu.ac.kr · hgkim@inu.ac.kr
[†]Corresponding author

Abstract

In this paper, we propose deep reinforcement learning (DRL)-based robust integrated guidance and control (IGC) law for longitudinal missile dynamics corrupted with noise. The longitudinal missile model is defined as the Markov decision process (MDP) and the IGC law is described as a policy network. The actor networks are trained using soft actor-critic (SAC) method and the output of the proposed method is determined as the normalized action. Then, the action multiplied by a scale factor becomes tail fin deflection command of the missile. Finally, numerical simulations are performed in learning and evaluation phases, and the results are analyzed and compared with those of sliding mode guidance control (SMGC)-based simulations.

1. Introduction

In general, the missile guidance and control systems have been developed using various methods such as nonlinear control and optimal control. They consist of guidance and control and have been developed individually. The previous studies were carried out under the premise that there is no coupling between the guidance loop and the control loop. In ref [1], a three-loop structure was designed for missile control, and control gains were derived through linear quadratic regulator. Ref [2] uses the backstepping technique and incorporates state reconstruction and a neural network to enhance robustness. Ref [3] uses the nonlinear sliding mode control (SMC) technique to avoid the chattering problem and analyzes the effect according to the boundary layer thickness. Although the performance of the previous studies is satisfactory, designing and integrating guidance and control separately is complex and costly. In addition the controller cannot follow the acceleration command due to the rapid geometrical change or the stability of the system cannot be guaranteed.

To address these problems, an integrated guidance and control (IGC) method that handles the guidance and control simultaneously has been developed. Ref. [4, 5] defined the dynamics of missiles and conducted the IGC research based on model predictive control (MPC). Ref. [6] designed the SMC to minimize zero-effort-miss (ZEM) on the premise that the maneuvering acceleration of the target is known. Ref. [7] developed the IGC system that is robust to disturbances by combining the SMC technique with a robust disturbance observer. Ref. [8] considered the field-of-view of a strap-down seeker that observes the state of the target. Ref. [9] considered the terminal impact angle to enhance the effectiveness of the intercept. Ref. [10] conducted a study to respond to rapid geometric changes by using two controllers, fast and slow. Even though the overall researches generated satisfactory performance, they did not consider corrupted observation with noise.

To alleviate this problem, deep reinforcement learning (DRL) is attracting attention as a new approach. The DRL is a field of reinforcement learning, which combines deep neural networks and reinforcement learning algorithms so that an agent interacts with the environment and learns a policy with maximal rewards. This approach demonstrates great potential in solving problems without predefined solutions, and has been utilized for the missile guidance and control systems. Ref [11] conducted a study to replace the missile attitude controller using the deep deterministic policy gradient (DDPG) technique. Ref. [12] attempted to replace the existing guidance technique using the DDPG technique in 2D kinematics. However, the DRL-based studies are not being actively conducted in the IGC system.

In this study, to overcome the limitations of the above studies, we propose a DRL-based integrated guidance and control law. This method proceeds by integrating guidance and control into the policy network. For this purpose, the missile

DOI: 10.13009/EUCASS2023-518

longitudinal dynamics and engagement kinematics are defined as the markov decision process (MDP). The defined MDP is designed using OpenAI's Gym environment [13], and the policy network is trained using soft actor-critic (SAC) method [14]. In addition, to improve the convergence of the policy to the optimal policy, the input and output are normalized between -1.0 and +1.0. The output is multiplied by the scale factor so that the driving range of the tail fin ranged from $-30°$ to $30°$. In order to verify the performance of the proposed method, numerical simulations are performed in an environment with seen/unseen observations and noise, and the results are analyzed. Additionally, the performance of the proposed method is compared with sliding mode guidance and control (SMGC) [6].
The remainder of this paper is organized as follows: Sec. 2 defines the longitudinal engagement kinematics and dynamics of the missile. In Sec. 3, the missile longitudinal engagement problem is defined as the MDP, and the DRL-based integrated guidance and control framework is proposed. In Sec. 4, numerical simulations composed of the learning and evaluation phases are performed and the results are analyzed. Sec. 5 describes the conclusion of this paper.

2. Model formulation

In this chapter, the longitudinal engagement kinematics and dynamics of missiles are considered and defined as MDP to perform the DRL-based integrated guidance and control.

2.1 Engagement kinematics

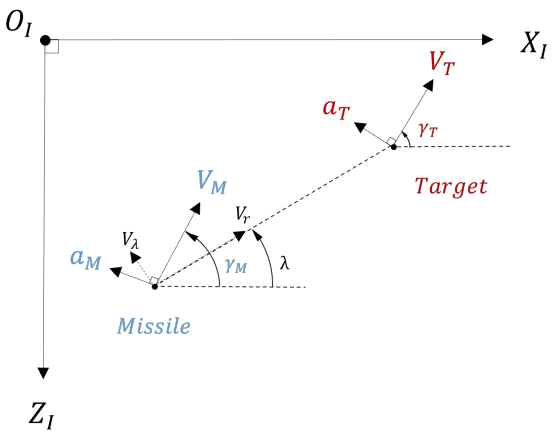

Figure 1: Longitudinal engagement geometry.

As shown in Fig. 1, the longitudinal engagement kinematics of the missile and target are as follows:

$$\dot{R} = V_r = V_T \cos(\gamma_T - \lambda) - V_M \cos(\gamma_M - \lambda) \tag{1a}$$

$$R\dot{\lambda} = V_\lambda = V_T \sin(\gamma_T - \lambda) - V_M \sin(\gamma_M - \lambda) \tag{1b}$$

$$\dot{\gamma}_T = \frac{a_T}{V_T} \tag{1c}$$

where (X_I, O_I, Z_I) is Cartesian inertial reference frame. The variables D, λ, V_r and V_λ are relative distance, line-of-sight (LOS) angle, speed in the approach direction, and speed perpendicular to the LOS, respectively. The variables V_i, a_i and γ_i are velocity, acceleration perpendicular to the velocity vector, and flight-path angle$(i = M, T)$, respectively. The missiles and targets are denoted by subscripts M and T, respectively. The missile moves at a constant speed of 500 m/s, while the target has the initial attitude of γ_T and moves at a speed of 300 m/s. The acceleration of the target is a constant, a_T is applied by $-5g$

2.2 Missile dynamics

Fig. 2 shows the coordinate systems of the longitudinal dynamics of the missile. (X_{bf}, O_M, Z_{bf}) is parallel to the inertial frame and is at the center of gravity of the missile. The variable α is the angle-of-attack, θ is the pitch angle of the

DOI: 10.13009/EUCASS2023-518

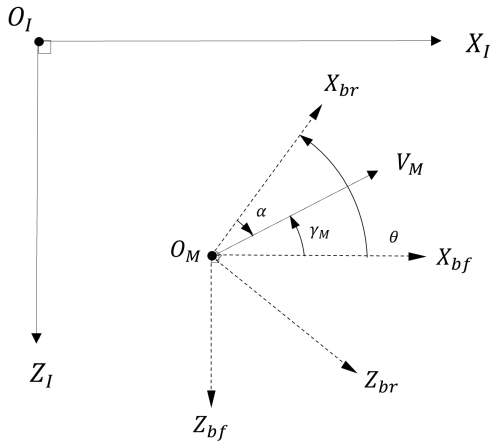

Figure 2: Missile coordinate systems.

missile. Based on the previous definition, the dynamics of the missile can be expressed as:

$$\gamma_M = \theta - \alpha \tag{2a}$$

$$\dot{\alpha} = q - \frac{\mathbf{L}(\alpha, \delta)}{mV_M} \tag{2b}$$

$$\dot{q} = \frac{\mathbf{M}(\alpha, q, \delta)}{\mathbf{I}} \tag{2c}$$

$$\dot{\theta} = q \tag{2d}$$

$$\dot{\delta} = \frac{\delta_c - \delta}{\tau_s} \tag{2e}$$

where q represents the pitch rate of the missile. $\mathbf{M}$ and $\mathbf{L}$ are the pitch moment and the lift force, respectively. $\mathbf{I}$ refers to the moment of inertia, and m is the mass. The angle of the missile's tail fin, denoted as δ, is modeled as a first-order dynamic system that tracks the desired command, δ_c, with a time constant of τ_s. The lift and pitch moment of Eqs. (2) are expressed as follows:

$$\mathbf{L}(\alpha, \delta)/m = \mathbf{L}_\alpha^B f_1(\alpha) + \mathbf{L}_\delta f_2(\alpha + \delta) \tag{3a}$$

$$\mathbf{M}(\alpha, q, \delta)/\mathbf{I} = \mathbf{M}_\alpha^B f_3(\alpha) + \mathbf{M}_q q + \mathbf{M}_\delta f_4(\alpha + \delta) \tag{3b}$$

$$\mathbf{L}_\alpha^B = \mathbf{L}_\alpha - \mathbf{L}_\delta \tag{3c}$$

$$\mathbf{M}_\alpha^B = \mathbf{M}_\alpha - \mathbf{M}_\delta \tag{3d}$$

$f_i(\cdot), i = 1, 2, 3, 4$, where f_i represents a bounded function, describe the nonlinear aerodynamic characteristics of the missile. The range of the boundary condition was set from $-30°$ to $30°$, and the missile model parameters used are shown in Table 1.

Table 1: Missile characteristics

Variable	Value	Unit
$\mathbf{L}_\alpha^B$	1190	m/s^2
$\mathbf{L}_\delta$	80	m/s^2
$\mathbf{M}_\alpha^B$	-234	s^{-2}
$\mathbf{M}_\delta$	160	s^{-2}
$\mathbf{M}_q$	-5	s^{-1}
τ_s	0.02	s

DOI: 10.13009/EUCASS2023-518

3. DRL-based integrated guidance and control system

In this chapter, we propose the DRL-based integrated guidance and control law. To implement this, the kinematics and dynamics model of the missile and target are defined as the MDP, and the policy network is trained using the SAC method.

3.1 Soft actor-critic

The Soft Actor-Critic (SAC) algorithm was developed as a solution to address the limitations of policy-based approaches in the reinforcement learning. The SAC introduces an off-policy technique that enhances sample efficiency by storing and reusing previous experiences stored in a replay buffer. Moreover, the SAC is well-suited for continuous action spaces as it leverages the maximum entropy model, leading to improved convergence speed towards the optimal policy.

The SAC algorithm adopts an actor-critic architecture comprising an actor and a critic. The actor generates actions based on the current observation, while the critic evaluates these actions and provides feedback to the actor to update the policy accordingly. The critic plays a crucial role in estimating the Q-function, which assesses the value of observed state-action pairs based on the current policy. By doing so, the actor learns to optimize its policy by seeking the maximum return within the given environment, thereby facilitating effective learning.

3.2 SAC-based integrated guidance and control law

Fig. 3 shows the overall conceptual diagram of the proposed integrated guidance and control system. Since the actual missile system observes the state information of the target through a sensor such as a strap-down seeker, it is assumed that the target's LOS angle, LOS angular rate, and missile state can be used.

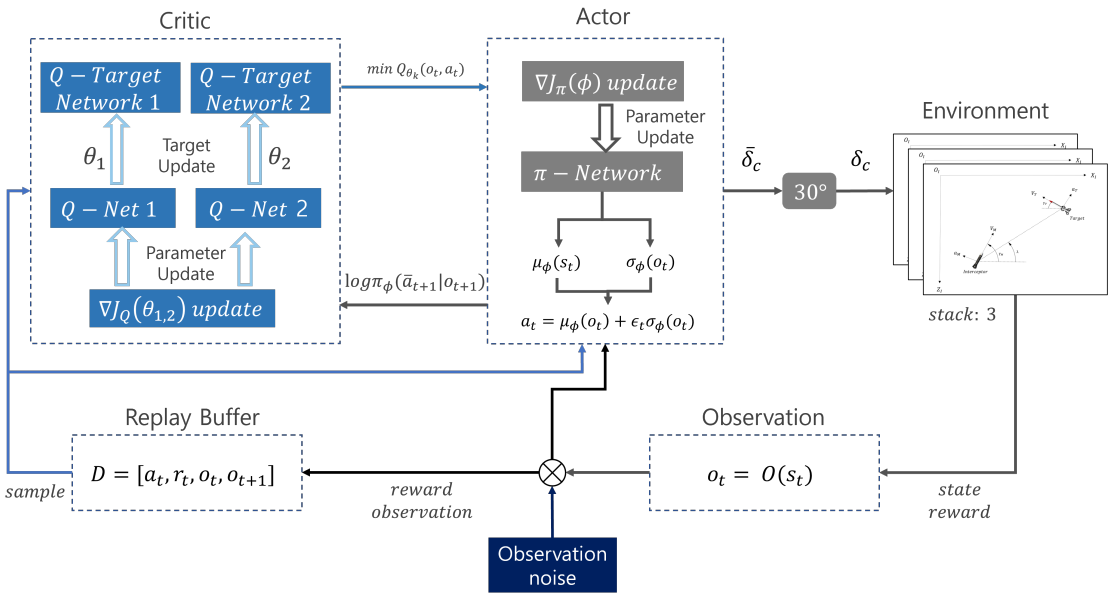

Figure 3: Conceptual diagram of the proposed SAC-based IGC framework

Using the two-dimensional longitudinal kinematics and dynamics model of the missile and target, the following MDP is constructed.

DOI: 10.13009/EUCASS2023-518

$$
\begin{aligned}
&\mathcal{S} = [\mathcal{S}_M, \mathcal{S}_T] \\
&\quad (\mathcal{S}_M = [x_M, z_M, \alpha_M, q_M, \theta_M, \delta_M], \mathcal{S}_T = [x_T, z_T, \gamma_T]) \\
&\mathcal{O} = \left[\sin(\lambda - \gamma_M), \cos(\lambda - \gamma_M), \dot{\lambda}, q, \delta\right] \\
&\mathcal{A} = \bar{\delta_c} \\
&\mathcal{R} = \begin{cases} -100, & \text{if } R > R_{prev} \\ -(\text{ZEM}/1e4)^2 & \text{otherwise} \end{cases}
\end{aligned} \tag{4}
$$

The state vector ($\mathcal{S}$) consists of six missile states and three target states. To ensure the continuity of the observation vector ($\mathcal{O}$), the relationship between λ and γ_M is represented using trigonometric functions, specifically the sin and cos functions. This representation allows for maintaining smooth transitions in the target state information even when it exceeds the range of -π to π, while also facilitating the normalization effect. Furthermore, to enforce continuity in the state information for the policy network, three observation vectors are stacked and used as input. The action ($\mathcal{A}$) represents the output of the policy network and ranges from -1 to 1. The final control command for the missile's control system is obtained by multiplying the output by a scale factor of 30 degrees. The reward function ($\mathcal{R}$) is designed to enable the missile to guide itself towards the target. R is a relative distance between the missile and the target, and the relative distance at the previous time step, denoted as R_{prev}. If R is greater than R_{prev}, it is considered a failure, while if R becomes smaller than 3, it is considered a success. The value ZEM in the reward function is defined as follows:

$$
\text{ZEM} = \frac{RV_r}{\sqrt{V_r^2 + V_\lambda^2}} \tag{5}
$$

4. Validation

In this chapter, numerical simulations are conducted to validate the performance of the proposed integrated guidance and control system. The results are then analyzed to evaluate its effectiveness. Furthermore, a comparison is made with the sliding mode guidance and control (SMGC) system to provide a comparative analysis of the outcomes.

4.1 Experiment setup

The structure of the actor and critic network of The SAC is shown in Table. 2. All networks have an input layer, two hidden layers with 256 nodes, and an output layer. The well-known function ReLU was utilized for the activation function. The actor network receives the latest three stacked observations to create an action. The critic network creates a Q value through stacked observation and actions. The actor network takes the three latest observations as input and generates an action. The critic network takes the three observations and actions as input and evaluates the current actor network.
Table. 3 shows the hyperparameters of the SAC for the learning. The learning process begins after 100 episodes have been performed to build up the replay buffer. Each episode ends when the end terminal condition is stisfied or 5000 timestep is exceeded. The timestep required for the entire learning was set to 2e7.

Table 2: SAC network architecture

	actor network		**critic network**	
layers	unit	activation	unit	activation
input	obs dim	Linear	obs dim + act dim	Linear
hidden	256	ReLU	256	ReLU
hidden	256	ReLU	256	ReLU
output	act dim	Linear	1	Linear

The missile longitudinal engagement environment was created with the OpenAI Gym python module. When deriving the next state from the environment, it was integrated at 100 Hz using the fourth-order Runge-Kutta method. The overall simulation is composed of the learning and evaluation phases which were determined as follows:

DOI: 10.13009/EUCASS2023-518

Table 3: Hyperparameter of SAC

Variable	Value
learning rate	0.0003
discounting factor	0.99
replay buffer size	1e6
learning start	100
batch experience	256
episode timestep	5e3
total timestep	2e7

- **Learning phase**
  - The initial positions of the missile and target are $(0, 0)$ m and $(5000, 0)$ m, respectively.
  - The velocities of the missile and target are assigned as 500 m/s and 300 m/s, respectively.
  - The missile's initial angle-of-attack is randomly selected in 1° intervals within the range of 0° to 5°.
  - The missile's initial pitch angle is randomly determined in 10° intervals ranging from 0° to 50°.
  - The target's initial pitch angle is randomly determined in 10° intervals between 110° and 150°.
- **Evaluation phase**
  - The initial positions, velocities, and angle-of-attack of the missile are the same as those in the learning phase.
  - The initial positions and velocities of the target are the same as those in the learning phase.
  - The missile's initial pitch angle is randomly determined in 1° intervals ranging from 0° to 70°.
  - The target's initial pitch angle is randomly determined in 1° intervals between 110° and 170°.

In addition, in order to implement the noise generated in the environment, the observations were contaminated with noise. During the learning phase, a relatively low noise level ($\mu = 1, \sigma = 0.001$) was used, while during the evaluation phase, a higher noise level ($\mu = 1, \sigma = 0.01$) was employed.

4.2 Results

In this section, the performance of the proposed method is analyzed. The SAC was trained for 2e7 timesteps in the learning phase. In order to evaluate the performance of the policy network after learning was completed, 1000 samples of Monte Carlo simulation were performed in the evaluation phase with noise $N(1, 0.001/0.01)$. Fig. (4-7) show the results of 100 out of 1,000 verifications. Fig. 4 represents the flight trajectory of the missile and target. Fig. 5 is the angle-of-attack, pitch angles, and pitch angular rates among missile states, respectively. Fig. 6 shows the vertical acceleration of the missile due to the tail fin. The vertical acceleration is the lift force divided by the mass in Eqs. (3). Fig. 7 shows the control command generated by the policy network with a control gain of 30°. In the evaluation phase, the policy network encounters situations not experienced in the learning phase. Nevertheless, the performance of the policy network is similar to the results in the learning phase.

In order to further prove the validity of the proposed method, a comparative analysis was conducted with the SMGC method. That is, the success rate was analyzed by changing the magnitude of the acceleration applied to the target. In addition, the corrupted λ and $\dot{\lambda}$ with noise were utilized in the SMGC method, while the observation were corrupted with noise in the proposed method. Table. 4 shows the success rate according to the target's maneuvering acceleration and noise. As shown in the table, the proposed method shows a consistent performance according to the variation of the target's maneuvering acceleration and noise magnitude while the success rate of the SMGC method is not maintained similarly depending on the noise and acceleration.

In summary, it can be confirmed that the proposed method in this study guarantees consistent performance, while the SMGC techniques have difficulty maintaining performance in a situation where the size of noise increases and the acceleration of the target changes.

DOI: 10.13009/EUCASS2023-518

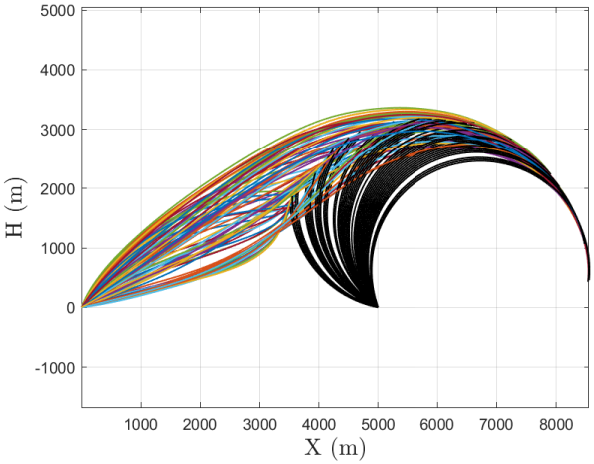

Figure 4: Trajectories of missiles.

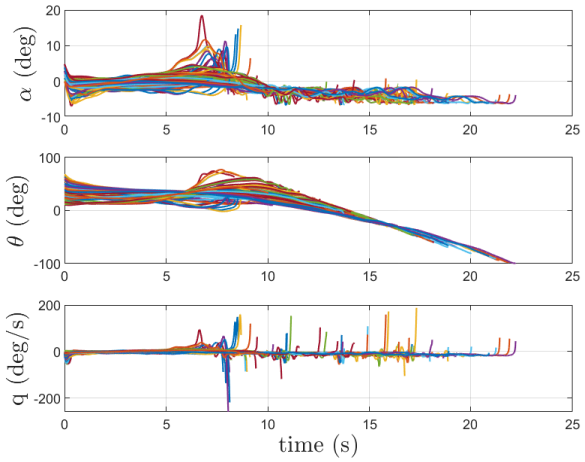

Figure 5: Angular states of the missiles.

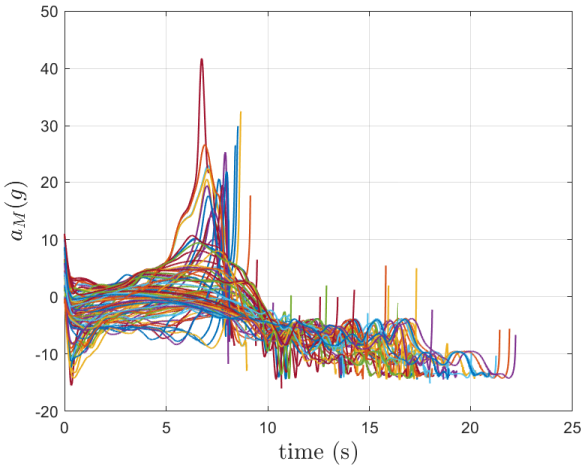

Figure 6: Vertical acceleration of the missiles.

DOI: 10.13009/EUCASS2023-518

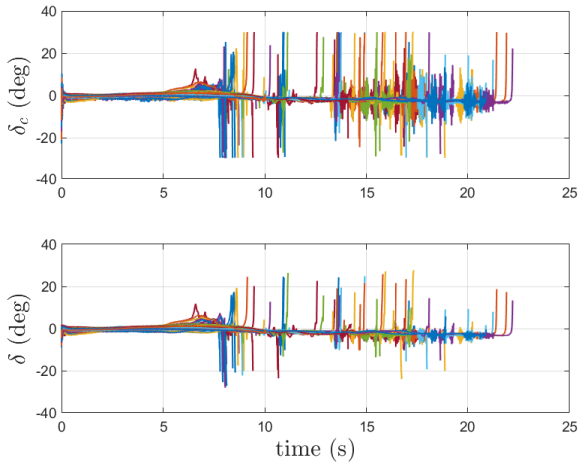

Figure 7: DRL-based tail fin control command and response.

Table 4: Success rate of the proposed method and SMGC

	DRL(our method)		**SMGC [6]**	
Target a_T	$N(1, 0.001)$ [%]	$N(1, 0.01)$ [%]	$N(1, 0.001)$ [%]	$N(1, 0.01)$ [%]
$-7g$	**100**	**100**	82.9	60.8
$-6g$	**100**	**100**	81.8	60.7
$-5g$	**100**	**99.8**	98.0	97.4
$-4g$	99.2	99.3	**99.3**	**99.9**
$-3g$	92.8	92.2	**97.9**	**97.4**

5. Conclusion

In this study, we proposed the integrated guidance and control law that is robust to noise based on the SAC method. To this end, the kinematics and dynamics models of the missile and target were integrated and defined as the MDP, and the reinforcement learning environment was created using OpenAI Gym. The proposed method consists of the learning phase and the evaluation phase. In the learning phase, it proceeded with limited initial conditions. In the evaluation phase, it was verified with the initial conditions and increased noise that were not encountered in the learning phase. Additionally, a simulation using the SMGC method was performed and the results were compared and analyzed. In the case of the SMGC method, the performance degradation was observed to be significant and dependent on the presence of noise and the acceleration variation of the target. In contrast, the SAC-based IGC exhibited a remarkably consistent and high success rate, even when confronted with observations and noise that were not encountered during the learning phase. Based on the obtained results, it is anticipated that an integrated guidance and control system considering not only longitudinal but also lateral motion can be developed. The proposed method can operate only with the information that the sensor can acquire. Therefore, if the stability of the learning results can be ensured, it is expected that it can be applied to the real-world environment.

6. Acknowledgments

This work was supported by Artificial Intelligence Research Laboratory for Flight Control funded by Agency for Defense Development and Defense Acquisition Program Administration under Grant UD230014SD and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.2021R1A4A1033141).

References

[1] D Ridgely, Yung Lee, and Todd Fanciullo. Dual aero/propulsive missile control-optimal control and control allocation. In *AIAA guidance, navigation, and control conference and exhibit*, page 6570, 2006.

DOI: 10.13009/EUCASS2023-518

[2] Liu Zhong and Jia Xiaohong. Novel backstepping design for blended aero and reaction-jet missile autopilot. *Journal of Systems Engineering and Electronics*, 19(1):148–153, 2008.

[3] Jingxian Liao and Hyochoong Bang. Precise missile autopilot design using nonlinear sliding mode control. In *2022 13th Asian Control Conference (ASCC)*, pages 1330–1335, 2022.

[4] Saeed Shamaghdari, SKY Nikravesh, and Mohammad Haeri. Integrated guidance and control of elastic flight vehicle based on robust mpc. *International Journal of Robust and Nonlinear Control*, 25(15):2608–2630, 2015.

[5] Prashant Kumar, Sarvesh Sonkar, AK Ghosh, and Deepu Philip. Nonlinear model-predictive integrated guidance and control scheme applied for missile-on-missile interception. In *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 318–324. IEEE, 2020.

[6] Tal Shima, Moshe Idan, and Oded M Golan. Sliding-mode control for integrated missile autopilot guidance. *Journal of guidance, control, and dynamics*, 29(2):250–260, 2006.

[7] Tae-Won Hwang and Min-Jea Tahk. Integrated backstepping design of missile guidance and control with robust disturbance observer. In *2006 SICE-ICASE International Joint Conference*, pages 4911–4915. IEEE, 2006.

[8] Xingwei Li, Bin Zhao, Jun Zhou, and Zhenxin Feng. Integrated guidance and control for missiles with strap-down seeker. In *2017 36th Chinese Control Conference (CCC)*, pages 6208–6212. IEEE, 2017.

[9] Peng Wu and Ming Yang. Integrated guidance and control design for missile with terminal impact angle constraint based on sliding mode control. *Journal of Systems Engineering and Electronics*, 21(4):623–628, 2010.

[10] Hyeong-Geun Kim and H Jin Kim. Integrated guidance and control of dual missiles considering trade-off between input usage and response speed. In *2013 13th International Conference on Control, Automation and Systems (ICCAS 2013)*, pages 55–60. IEEE, 2013.

[11] Angelo Candeli, Gianmaria De Tommasi, Dario Giuseppe Lui, Adriano Mele, Stefania Santini, and Gaetano Tartaglione. A deep deterministic policy gradient learning approach to missile autopilot design. *IEEE Access*, 10:19685–19696, 2022.

[12] Shaoming He, Hyo-Sang Shin, and Antonios Tsourdos. Computational missile guidance: A deep reinforcement learning approach. *Journal of Aerospace Information Systems*, 18(8):571–582, 2021.

[13] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[14] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.