Abstract: Reinforcement learning is generating considerable interest in terms of building guidance law and solving optimization problems that were previously difficult to solve. Since reinforcement learning-based guidance laws often show better robustness than a previously optimized algorithm, several studies have been carried out on the subject. This paper presents a new approach to training missile guidance law by reinforcement learning and introducing some notable characteristics. The novel missile guidance law shows better robustness to the controller-model compared to the proportional navigation guidance. The neural network in this paper has identical inputs with proportional navigation guidance, which makes the comparison fair, distinguishing it from other research. The proposed guidance law will be compared to the proportional navigation guidance, which is widely known as quasi-optimal of missile guidance law. Our work aims to find effective missile training methods through reinforcement learning, and how better the new method is. Additionally, with the derived policy, we contemplated which is better, and in which circumstances it is better. A novel methodology for the training will be proposed first, and the performance comparison results will be continued therefrom.

Keywords: missile guidance law; robustness; linear system; reinforcement learning; control model tolerance

1. Introduction

The proportional navigation guidance (PNG) is known as a quasi-optimal of interception guidance law. The simple concept of it can be found in nature. The authors of [1] show that it is possible to explain predatory flies’ pursuit trajectories with the PNG, in which a rotation rate of a pursuer can be modeled as a multiplication of specific constant and line-of-sight (LOS) rate. In [2], the authors also describe the terminal attack trajectories of peregrine falcons as PNG. The head of a falcon acts as a gimbal of missile seeker, guiding its body like a modern missile does. That shows centuries of training and evolution possibly brought those creatures to fly with quasi-optimal guidance and gives us an insight to see those natural phenomena with engineered mechanisms on the same boundary. Reinforcement learning (RL)-based training has a similar process. Each generation of an animal and an RL agent act stochastically, have some mutational elements, and follow what has a better reward, and aims for the optimal. Meanwhile, some studies on RL-based missile guidance law have and have carried on. Accordingly, Gaudet [3] had suspected an RL-based guidance law may help the logic itself to become robust. He proposed an RL framework that has the ability to build the guidance law for the homing phase. He trained Single-Layer Perceptron (SLP) by the stochastic search method to build a guidance law. Reinforcement learning-based missile guidance law surprisingly works well on the noisy stochastic world, possibly because its algorithm is based on probability. He brought PNG and Enhanced PNG as the comparison targets and showed that the RL generated guidance law has
better performance in terms of energy consumption and miss distance than the comparison targets concerning the stochastic environment. Additionally, in another paper [4], he proposed a framework for 3-dimensional RL-based guidance law design for an interceptor trained via the Proximal Policy Optimization (PPO) [5] algorithm, and showed by comparative analysis that the derived guidance law also has better performance and efficiency than the augmented Zero-Effort Miss policy [6]. However, previous research on RL-based missile guidance law compared policies that have a state of unidentical input, so the comparison is deemed unfair.

Optimization of RL policy depends on the reward, which is effective when the reward can indicate which policy and single-action is better than the others. Besides, the missile guidance environment we are dealing with makes it harder for RL agents to acquire an appropriate reward. For this kind of environment, called a sparse reward environment, some special techniques were introduced. The simplest way to deal with such a sparse reward environment is by giving an agent a shaped reward. While general rewards are formulated with a binary state, e.g., success or not, shaped reward implies the information of the goal, e.g., formulate reward as a function of the goal. The authors of [7] proposed an effective reward formulation methodology that is in accordance with the optimization delay due to the sparse reward. A sparse reward can hardly show what is better, so it makes the reward validity weak and training slow. Hindsight Experience Replay (HER) is a replay technique that makes the failed experience an accomplished one. DDPG [8] combined with this technique shows faster and more stable training convergence than the original DDPG training method. HER has pumped up the learning performance with binary reward by mimicking human learning that we learn not only from accomplishments but even from failures.

This paper presents a novel training framework of missile guidance law and compares trained guidance law with PNG. Additionally, a comparison of the control-system-applied environment will be shown since the overall performance or efficiency of a missile is not only affected by its guidance logic but is affected by the dynamic system as well. The control system which connects guidance command to reality always has a delay. In missile guidance, a system can possibly make the missile miss the target and consume much more energy than expected. There is research on missile guidance to make it robust on uncertainties and diverse parameters. In [9], Gurfil presents a guidance law that works well even in an environment with missile parametric uncertainties. He focused on building a universal missile guidance law that even covers missiles having different systems, i.e., the purpose is making missiles work well with zero miss distance guidance even if they have various flight conditions. The result he showed was assuring zero miss distance for uncertainties. Additionally, due to the simplicity of the controller that the guidance law needs, proposed guidance law has simplicity in implementation.

The paper will first, in the problem formulation, define environments and engagement scenarios, and overviews of PNG and training methods will be continued therefrom. Finally, simulation results and conclusion will follow.

2. Problem Formulation

2.1. Equations of Motion

The most considerable location for a homing missile is the inside of the threshold that switches the INS (Inertial Navigation System)—based-guidance phase into the homing phase, which is the inner space of the gradation-filled boundary in Figure 1. $\phi_0$ is the initial approaching angle, and the angular initial conditions for simulations will be set by using a randomly generated approaching angle as a constraint condition, which is as follows:

$$\lambda_0 = \begin{cases} 
\phi_0 + \pi, & (if \ \phi_0 < \pi) \\
\phi_0 - \pi, & (if \ \phi_0 \geq \pi)
\end{cases} 
$$

$$\psi_0 = \lambda_0 + L_0$$

(1)

(2)
The heading of the missile is assumed to be parallel to the velocity vector, and the acceleration axis is attached perpendicular to the missile velocity vector. It is assumed that there is no aerodynamic effect nor dynamic coupling. That is, the mass point dynamic model will be applied. The equations of motion are as follows:

\[ R = |\vec{R}_c|, \quad V_m = |\vec{V}_m| \tag{3} \]

\[ L = \lambda - \psi = \arctan \left( \frac{\det[\vec{V}_m; \vec{R}_c]}{\vec{V}_m \cdot \vec{R}_c} \right) \tag{4} \]

\[ \dot{\psi} = \frac{a_m}{V_m} \tag{5} \]

\[ \Omega = \frac{V_m}{R} \sin(\lambda - \psi) \tag{6} \]

where \( R \) is the relative distance from the missile to the target, \( \lambda \) is the line of sight (LOS) angle, \( \psi \) is the heading of the missile, \( L \) is look-angle, \( V_m \) is the speed of the missile, and \( a_m \) is the acceleration applied to the missile. The acceleration of a missile is assumed to have a reasonable limitation of \(-10g\) to \(10g\), where \( g \) is the gravitational acceleration. The symbol ‘\( \cdot \)’ between two column vectors indicates a concatenation that joins two vectors into a square matrix.

The Zero-Effort-Miss (ZEM) is an expected miss distance if there will be no further maneuver from the current location \([10]\). The concept of this will be used to interpolate the minimum miss distance at the terminal state to compensate for the discreteness of the simulation. The ZEM* is the ZEM when the last maneuver command is applied, and we defined it as shown in Figure 2.

\[ \vec{R}_d = \vec{R}_{f-1} - \vec{R}_f \tag{7} \]

\[ \text{ZEM}^* = |\vec{R}_d| \sin \arctan \left( \frac{\det[\vec{R}_{f-1}; \vec{R}_d]}{\vec{R}_{f-1} \cdot \vec{R}_d} \right) \tag{8} \]

where \( \vec{R}_f \) and \( \vec{R}_{f-1} \) are \( \vec{R}_c \) at the final state and \( \vec{R}_c \) at the very former state, respectively.
2.2. Engagement Scenario

This paper works with an enemy target ship and a friendly anti-ship missile. A scenario of two-dimensional planar engagement will be discussed. The guidance phase of the scenario is switched by the distance between the target and the missile. The missile launches at an arbitrary spot far away from the target, which is guided by the Inertial Navigation System (INS)-aided guidance until it crosses the switching threshold. When the missile approaches closer than a specific threshold (16,500 m in this scenario), the missile switches its guidance phase into a seeker-based homing phase. The conceptual scheme of phase switching is depicted in Figure 3. \(L_0\) is the look angle at the threshold.

The reason the missile needs its phase switched is because INS-aided guidance is not that accurate nor precise. It acts as to place the missile at the position that makes the homing guidance available. Thus, the zero-effort-miss (ZEM) can be quite big at the point on the switching threshold, so the wide range of initial conditions for the homing phase is needed. Additionally, generation criteria for initial conditions are described in Table 1.

| \(R_{\text{threshold}}\) (m) | \(L_0\) (rad) | \(\phi_0\) (rad) | \(V_m\) (m/s) |
|---|---|---|---|
| 16,500 | \([-\pi, \pi]\) | \([0, 2\pi]\) | \([200, 600]\) |

In practice, there is a delay that disturbs the derived action command which prevents the action from occurring immediately. Eventually, there will be a gap between guidance law and real actuation; the missile consumes much energy and has less accuracy. The scenario must contain this type of hindrance, and the simulations for that should be implemented. The details will be discussed in Section 2.3.

2.3. Controller Model

The controller model can be described as a differential equation of the 1st order system. We simply define them as follows:

\[ \tau a_m + a_m = a_c \]

where \(\tau\) is the time constant, \(a_m\) is the derived actuation, and \(a_c\) is the acceleration command, respectively.
3. Proportional Navigation Guidance Law

Proportional navigation guidance (PNG) is the most widely used guidance law on homing missiles with an active seeker \[2,11,12\]. PNG generates an acceleration command by multiplying the LOS rate and velocity with a predesigned gain. The mechanism reduces the look-angle and guides the missile to hit the target.

The acceleration command by PNG in the planar engagement scenario is as follows:

\[
a_c = NV_m \Omega
\]

(10)

where \(N\) is navigation constant (design factor), \(V_m\) is the speed of the missile, and \(\Omega\) is the LOS rate. Figure 4 shows the conceptual diagram of PNG.

![Conceptual diagram of proportional navigation guidance.](image)

Meanwhile, Pure PNG, which is one of the branches of PNG, is adopted in this paper. Pure PNG is used for a missile to hit a relatively slow target like a ship. The acceleration axis of the Pure PNG is attached perpendicular to the missile velocity vector as it is shown in Figure 1. Additionally, the velocity term in Equation (10) of Pure PNG is the inertial velocity of the missile. PNG is very useful, and it basically is even optimal for missile guidance. Technically, creating a missile guidance law that has better performance than PNG, despite sharing the same engagement scenario that PNG is specialized for, is not possible under ideal circumstances. However, PNG is not optimized for practice. Thus, we will figure out whether RL-based is able to exceed the performance of PNG when the simulation environment gets more practical, i.e., applying a controller model. Of course, we did not apply the practical model to training but applied them to the simulation part.

The overall structure of the PNG missile simulation is shown in Figure 5.

![Missile simulation schematic diagram with proportional navigation guidance.](image)

4. Reinforcement Learning

Prior research on RL-based missile guidance was studied under a very limited environment and did not show whether a single network can cover a wide environment and initial conditions which push the seeker to the hardware-limit or not. Additionally, the comparison had been implemented under an unfair condition for the comparison target, which each of them does not take identical states. In this paper, it is demonstrated that the RL-based guidance is able to cover a wide environment under plausible wide initial conditions, and with various system models. In this section, a novel structure and methodology for training will be proposed. We adopted a Deep Deterministic Policy Gradient (DDPG)-based algorithm \[8\]. DDPG is a model-free reinforcement learning algorithm that generates deterministic action in continuous space, which makes off-policy training available. The NN we used was optimized via gradient descent optimization. Since the environment we are dealing with is vast and can be described as a sparse reward environment, conventional binary rewards nor even a simple shaped reward do not really help to accelerate training. Thus, the algorithm we used in this paper was slightly modified by making every step of an episode get an equational final score of the episode.
as a reward. It helps the agent to effectively consider the reward of the far future. Figure 6 is an overview of the structure we designed. The agent feeds \( \Omega \) and \( V_m \) as input and stores a transition dataset, which consists of state, action, reward, and next state into a replay buffer.

![Figure 6. Structure of the training algorithm.](image)

Developed policy gradient algorithm using an artificial neural network (ANN) needs two separated ANNs which are the actor that estimates the action and the critic that estimates the value. Figure 7 shows each structure. DDPG, generally, calculates its critic loss from the output of the critic network and the reward of a randomly sampled minibatch, and updates the critic and the actor every single step. \( Q \) is the action value, and \( \sigma \) is the hyperbolic tangent activation function.

![Figure 7. Actor-network and critic-network.](image)

Both consist of three steps of hidden layers, and actor-network has \( \pm 10 \) g of clipping step as we set. We used the Huber function [13] as a loss function that shows high robustness to outlier and excellent optimization performance. The equation is as follows:

\[
L_\phi(y, f(x)) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \begin{array}{ll}
\frac{1}{2} (y_i - f(x_i))^2, & \text{if } |y_i - f(x_i)| \leq \delta \\
\delta (|y_i - f(x_i)| - \frac{1}{2} \delta), & \text{otherwise}
\end{array} \right. 
\]  

(11)

The training proceeds with a designed missile environment. The environment terminates the episode if it encounters one or more termination conditions as follows—Out of Range condition: target information unavailable since the target is outside of the available field of view of the missile seeker; Time Out condition: time exceeds the maximum timestep which is 400 s; there is also a Hit condition (if \( R < 5 \)): judged that the missile hits the target. However, the last condition is not for termination, but exists as an observable data that decides whether the policy is proper for the mission or not. As an episode ends, each of the transition datasets gets the corresponding reward which is...
the final integrated reward of the episode. We estimated the closest distance between a missile and a
target to interpolate the closest approach in a discrete environment as follows:

\[
R_{\text{impact}} = \begin{cases} 
ZEM^*, & \text{if Hit} \\
\frac{1}{R_f}, & \text{otherwise} 
\end{cases}
\] (12)

where Hit is when Hit condition is true, \(\vec{R}_f\) is \(\vec{R}_c\) at the final state, \(ZEM^*\) is defined in Equation (8).

The reward is designed with the concept that its maximization minimizes the closest distance
between the missile and the target and aims for saving energy as much as possible. To achieve that,
two branches of reward term should be considered. One of them is a range reward term. The range
reward term should get value from how the missile approaches close to the target. We used negative
log-scaled \(R^2_{\text{impact}}\) (the impact distance error of an episode) as a reward to make the reward overcome
the following problem. A simple linear scale reward may cause inconsistency and high variance of
reward. While an early training is proceeding in the environment, the expected bounds of \(R^2_{\text{impact}}\) is
from 0 to 16,500. This colossal scale disturbs the missile not to approach extremely close to the target
since the reward of the terminal stage only seemingly slightly varies the variance of reward even if
there are distinguishable good and bad actions. Besides, the log-scaled \(R^2_{\text{impact}}\) helps the missile get
extremely close to the target by amplifying the small value of \(R^2_{\text{impact}}\) at impact stages. We express the
range reward as \(r_R\) and is as follows:

\[
r_R = -\log(R^2_{\text{impact}} + \epsilon)
\] (13)

where \(\epsilon\) (tolerance term for log definition) is set to \(10^{-8}\). The second branch is an energy reward term. It
represents the energy consumption in an episode. The energy term should be distinguishable from
what is better or not regardless of the initial condition. Therefore, it must contain influences from the
flight range and the initial look-angle. The energy reward term also has a high variance between the
early stage and the terminal stage. For the earlier stage, the missile is prone to terminate its episode
earlier because the Out of Range termination condition can be satisfied early when the policy is not
that well established. In other words, the magnitude of the energy reward term can be too small. Thus,
the log-scaled reward was also applied in the energy reward term as follows:

\[
r_E = -\ln \left( \frac{\int_0^{t_f} a_m^2 dt}{V_m \cdot t_f \cdot \sqrt{|L_0|} + 0.5} \right)
\] (14)

where \(a_m\) is acceleration, \(t_f\) is the termination time.

Therefore, the total reward is a weighted sum of both (13) and (14) as follows

\[
r_{\text{total}} = \omega_1 \left( \frac{R^2_{\text{impact}} - \mu_{R^2}}{\sigma_{R^2}} \right) + \omega_2 \left( \frac{r_E - \mu_{r_E}}{\sigma_{r_E}} \right)
\] (15)

where \(\omega_1\) and \(\omega_2\) are weights for each reward that satisfy \(\omega_1 + \omega_2 = 1\), \(\mu\), and \(\sigma\) are the means and
standard deviations for normalization, respectively, and subscripts represent the properties of each
reward. Table 2 shows the termination conditions and corresponding rewards.

Table 2. Definition of conditions and transition rewards under corresponding conditions.

| Condition Name | Conditions | Rewards |
|----------------|------------|---------|
| Out of range   | if \(|L| < \frac{\pi}{2}\) | \(r_{\text{total}} = \omega_1 \left( \frac{r^2_{\text{impact}} - \mu_{r^2}}{\sigma_{r^2}} \right) + \omega_2 \left( \frac{r_E - \mu_r}{\sigma_r} \right)\) |
| Time out       | if \(t \geq t_{\text{max}}\) | \(r_{\text{total}} = -1\) |
Our rewarding method makes the agent trained faster than the general DDPG algorithm and makes the training stable. Clearly, our rewarding method works better for the sparse reward and vast environment of our missile, and the learning curves for them are shown in Figure 9.

The full lines represent the reward trend and color-filled areas represent local mins and maxes. Additionally, all data was plotted by applying a low pass filter for the intuitive plot.

After the training of 20,000 episodes, we implemented the utility evaluation of each trained weight. Specifically, the training started to converge at about the 70th episode via the training structure we designed. Yet, for finding the most optimized weight, we ran the test algorithm with each trained weight by feeding 100 of the uniformly random predefined initial conditions dataset. The dataset was set to range from 30,000 m to 50,000 m as depicted in Figure 10. It is because we wanted a weight that covers universal conditions. The ZEM* was not considered to decide the utility as long as the ZEM* was less than 5 for all predefined initial conditions. In addition, the weight having the best utility was decided by selecting a weight that has a minimum average $\int_0^{T_f} a^2_{m} \, dt$ for all predefined initial conditions. In Figure 10, the vectors and green dots represent the initial headings and the launching position, respectively.

Figure 8. Missile simulation schematic diagram with reinforcement learning (RL).

Figure 9. Learning curves comparison for rewarding methods.

Figure 10. Missile simulation schematic diagram with reinforcement learning (RL).
performance. It is intriguing because the training optimization resembles the algebraic quasi-optimal,
predefined initial conditions. Comparison of PNG with various navigation coefficients and the
RL-based guidance is as follows. Under the ideal condition, RL-based guidance law in Table 3 shows it
works well even if the missile launches far from the trained initial conditions. In Figure 10, the
vectors and green dots represent the initial headings and the launching position, respectively.

The flow chart for overview of this study is shown in Figure 11.

5. Simulation

The simulation was implemented under conditions that acceleration command bypasses or passes
through the 1st order controller model. The test dataset was distinguished from the utility evaluation
dataset and contains 100 of uniformly random initial conditions that satisfy the initial condition of the
engagement scenario which is described in Table 2.

5.1. Bypassing Controller Model

We used the weight with the highest utility. This was obtained through the simulation with
predefined initial conditions. Comparison of PNG with various navigation coefficients and the
RL-based guidance is as follows. Under the ideal condition, RL-based guidance law in Table 3 shows it
consumes similar energy to PNG that has navigation constant 3–4 under ideal condition. Even though
the training was implemented on a single range parameter of 16,500 m, the RL-based algorithm shows
it works well even if the missile launches far from the trained initial conditions.

| N  | PNG       | RL-Based |
|----|-----------|----------|
| 2  | 3901.99   | 3206.764 |
| 3  | 3142.81   |          |
| 4  | 3424.72   |          |
| 5  | 3881.95   |          |

As shown in Table 3, RL-based missile guidance law has a reasonable and appropriately optimized
performance. It is intriguing because the training optimization resembles the algebraic quasi-optimal,
just like the predatory flies and peregrine falcons of [1,2]. Additionally, the trajectories of both have similar aspects as follows.

Figure 12 shows trajectories for all simulation datasets of RL and PNG. The target is at the coordinate of (0, 0) and the starting point is the other end of the trajectory spline. The graph plotted on the right shows trajectories of a magnified scale. As it is shown in the trajectories in Figure 12, every trajectory pierces the central area of the target and unquestionably satisfies 5 m accuracy. The simulation result here implies that the RL-based missile guidance law is able to replace the PNG.

5. Simulation

The simulation of this part is implemented via the acceleration command passing through the controller model. As it is mentioned in Section 2.3 the controller is simply modeled with 1st order differential equation i.e., (9). For the 1st order, controller model is formulated with condition of $\tau \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.75\}$. We carried out a simulation with those predefined parameters and graphed its results including the trend lines. The miss cost is the average ZEM* for the 100 initial conditions, and the energy cost is the average of integral values for the square of derived acceleration with each initial condition. The simulation results are as follows

Note that the RL-based is trained without any controller model and has an equivalent energy cost of PNG that has navigation constant 3–4. In terms of the miss cost, RL-based seems less robust to the system. Yet, the time constant of 0.75 is quite a harsh condition. Nevertheless, they have a hit rate of 100% for all initial conditions. RL-based and PNG both far exceed the requirement even if they have been engaged with a system that we had intended to push them to the limit. Therefore, the robustness on ZEM* does not matter. On the other hand, the energy consumption of RL-based seems more robust to the system than the PNG is. As shown in Figures 13–15, the energy cost trend line of PNG with respect to time constants shows a larger gradient than the energy cost trend line of RL-based. We consider a robustness criterion as follows:

$$\eta[\%] = \frac{\nabla f_{PNG} - \nabla f_{RL}}{\nabla f_{PNG} + \nabla f_{RL}} \times 100$$

where $f$ is the linear trend line of each cost for each guidance law that their subscript denotes, and $\nabla$ indicates the gradient of each trend line. With the criterion of the robustness, RL-based shows 64%, 12%, and 12% better robustness than PNG with $N = 2$, $N = 3$, and $N = 4$ respectively. We suspect that it is because the RL is trained with uncertainties and may bring policies that are robust to the system. Contrary to ZEM* robustness, energy robustness matters in the engagement scenario, because less energy consumption suggests farther radius of action of a missile.
where we formulated PNG with the input of the LOS rate only by merging previous navigation constant and expectation of velocity into a single navigation constant, which is as follows.

$$\eta [\%] = \frac{\omega_0 + \omega}{\omega_0} \times 100$$

We found that it is possible to make a LOS-rate-only-input missile guidance law via RL. We trained it with the same method and network, but without the V input. It is useful under circumstances that V cannot be fed anymore due to the sensor fault and unknown hindrances. To make a comparison target, we formulated PNG with the input of the LOS rate only by merging previous navigation constant and expectation of velocity into a single navigation constant, which is as follows.

$$\bar{a}_c = \bar{N}\Omega$$  \hspace{1cm} (17)

5.3. Additional Experiment

We found that it is possible to make a LOS-rate-only-input missile guidance law via RL. We trained it with the same method and network, but without the V input. It is useful under circumstances that V cannot be fed anymore due to the sensor fault and unknown hindrances. To make a comparison target, we formulated PNG with the input of the LOS rate only by merging previous navigation constant and expectation of velocity into a single navigation constant, which is as follows.

$$\bar{a}_c = \bar{N}\Omega$$

where $\bar{N}$ is the navigation constant for LOS-rate-only guidance. The performance comparison run with the simulation dataset of Figure 10a is summarized in Table 4 and Figure 16.

**Table 4. Energy consumption for each guidance law.**

| PNG N=3   | PNG $\bar{N}$=1200 | RL-Based | PNG N=4   | PNG $\bar{N}$=1600 |
|------------|---------------------|----------|------------|---------------------|
| $\Omega, V_m$ | $\Omega$ only | $\Omega, V_m$ | $\Omega$ only | $\Omega, V_m$ |
| 3142.81    | 3240.3              | 3213.42  | 3280.5     | 3424.72            |
| 3429.86    |                     |          |            |                     |
Every single simulation easily satisfies the hit condition. Additionally, understandably, they show poorer energy performance than intact guidance laws do. Meanwhile, even with this scenario, the RL-based algorithm is more robust to the controller model. In the same manner as the criterion of robustness described in Equation (16), LOS-rate-only RL-based shows 44% and 10%, respectively, better than LOS-rate-only PNG with $N = 1200$ and $N = 1600$. The results of the simulation suggest that the RL-based missile guidance law works better than the PNG when the actuation delay goes past a certain threshold point.

6. Conclusions

The novel rewarding method, performance, and special features of RL-based missile guidance were presented. We showed that a wide environment of missile training can be trained faster with our rewarding method of Monte-Carlo based shaped reward rather than temporal difference training methodology. Additionally, the performance comparison was presented. This showed that the energy-wise performance of our RL-based guidance law and PNG with navigation constant of 3 to 4 were similar, when using a system that immediately acts. Yet, when it comes to the delayed 1st order system, RL-based shows better energy-wise system robustness than PNG. Moreover, we showed that the RL-based missile guidance algorithm could be formulated without the input state of V. Additionally, this algorithm has better robustness to the controller models than PNG without V. Its significant expandability is that it can replace PNG and has good robustness to the controller model without designing a specific guidance law for the controller model of a missile itself.

This paper showed that RL-based not only can easily replace the existing missile guidance law but is also able to work more efficiently in various circumstances. These special features give insights into the RL based guidance algorithm which can be expanded for future work.

Author Contributions: Methodology, D.H., S.P.; software and simulation, D.H., M.K. and S.P.; writing—original draft preparation, D.H. and M.K.; writing—review and editing, D.H., M.K. and S.P.; supervision, S.P.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019R1F1A1060749).

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Fabian, S.T.; Sumner, M.E.; Wardill, T.J.; Rossoni, S.; Gonzalez-Bellido, P.T. Interception by two predatory fly species is explained by a proportional navigation feedback controller. *J. R. Soc. Interface* **2018**, *15*. [CrossRef] [PubMed]
2. Brighton, C.H.; Thomas, A.L.R.; Taylor, G.K. Terminal attack trajectories of peregrine falcons are described by the proportional navigation guidance law of missiles. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 201714532. [CrossRef] [PubMed]
3. Gaudet, B.; Furfaro, R. Missile Homing-Phase Guidance Law Design Using Reinforcement Learning. In Proceedings of the AIAA Guidance, Navigation, and Control Conference 2012, Minneapolis, MN, USA, 13–16 August 2012. [CrossRef]
4. Gaudet, B.; Furfaro, R.; Linares, R. Reinforcement Learning for Angle-Only Intercept Guidance of Maneuvering Targets. *Aerosp. Sci. Technol.* 2020, *99*, 105746. [CrossRef]

5. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* 2017, arXiv:1707.06347v2.

6. Zarchan, P. *Tactical and Strategic Missile Guidance*, 6th ed.; American Institute of Aeronautics and Astronautics, Inc.: Reston, VA, USA, 2013. [CrossRef]

7. Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; MacGrew, B.; Tobin, J.; Abbeel, P.; Zaremba, W. Hindsight Experience Replay. *arXiv* 2018, arXiv:1707.01495v3.

8. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.P.; Heess, N.H.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. *arXiv* 2019, arXiv:1509.02971v6.

9. Gurfil, P. Robust Zero Miss Distance Guidance for Missiles with Parametric Uncertainties. In Proceedings of the 2001 American Control Conference, Arlington, VA, USA, 25–27 June 2001. [CrossRef]

10. Li, H.; Li, H.; Cai, Y. Efficient and accurate online estimation algorithm for zero-effort-miss and time-to-go based on data driven method. *Chin. J. Aeronaut.* 2019, *32*, 2311–2323. [CrossRef]

11. Madany, Y.M.; El-Badawy, E.A.; Soliman, A.M. Optimal Proportional Navigation Guidance Using Pseudo Sensor Enhancement Method (PSEM) for Flexible Interceptor Applications. In Proceedings of the 2016 UKSim-AMSS 18th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, UK, 6–8 April 2016. [CrossRef]

12. Raj, K.D.S. Performance Evaluation of Proportional Navigation Guidance for Low-Maneuvering Targets. *Int. J. Sci. Eng. Res.* 2014, *5*, 93–99.

13. Huber, P.J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* 1964, *35*, 73–101. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).