IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 9, 2020, accepted August 14, 2020, date of publication August 18, 2020, date of current version August 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017480

Range-Aware Impact Angle Guidance Law With Deep Reinforcement Meta-Learning

CHEN LIANG[1], WEIHONG WANG[1], ZHENGHUA LIU[1], CHAO LAI[2], AND SEN WANG[2]
[1]School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
[2]Navigation and Control Technology Research Institute of China North Industries Group Corporation, Beijing 100089, China

Corresponding author: Chen Liang (tccliangchen@hotmail.com)

This work was supported by the Defense Industrial Technology Development Program under Grant JCKY2018601B101.

ABSTRACT In this article, a new guidance law is proposed for impact angle constrained missile with time-varying velocity against a maneuvering target. The proposed guidance law is based on model-based deep reinforcement learning (RL) technique where a deep neural network is trained to be a predictive model used in model predictive path integral (MPPI) control. Tube-MPPI, a robust approach utilizing ancillary controller for disturbance rejection, is introduced in guidance law design in this work to deal with the MPPI degradation of robustness when the deep predictive model differs with actual environment. To further improve the performance, meta-learning is utilized to enable the deep neural dynamics adapt to environment changes online. With this approach the model mismatch of the nominal controller is reduced to improve tube-MPPI performance. Furthermore, a range-aware hyperbolic function is proposed as an adaptive function in the MPPI performance index design. Thus, reduced initial acceleration command and increased terminal velocity benefit guidance performance. Numerical simulations under various conditions demonstrate the effectiveness of proposed guidance law.

INDEX TERMS Missile guidance, tube model predictive control, meta-learning, deep reinforcement learning, impact angle constraint.

I. INTRODUCTION

Interception at a desired intercept angle help missile in increasing penetration capability, warhead effectiveness and reduce collateral damage. It may be necessary for modern missile to intercept target not only at a small miss distance, but also at a desired intercept angle. When facing these new requirements, conventional guidance law design method faced elevated difficulty, and deep reinforcement learning is a powerful tool in tackling these problems.

Rising interest has been witnessed on the application of deep RL in guidance design, with great potential shown by deep reinforcement learning. Compared with guidance designed using traditional control theory [1], deep RL is a data driven method. Many recent works has utilized deep RL in guidance law design for performance enhancement [2], or for requirement traditional control theory hard to satisfy [3], [4]. Many works in deep RL guidance laws utilize model-free deep reinforcement learning. However, model-free RL lacks sample efficiency compared with model-based methods, thus require large number of interactions with environment. Deep model-based reinforcement learning utilizes deep neural network as model, is generally considered data efficient [5] and is welcomed in many real-world control tasks. MPPI is one of the typical methods of model-based deep reinforcement learning which utilizes a deep neural network as the dynamics model to obtain the optimal control solution to the Hamilton-Jacobi-Bellman (HJB) equation via Monte-Carlo sampling of path integrals, thus solves the optimal control problem and is widely used in many real-world tasks [6]. Ref [7] utilizes MPPI method to solve the guidance problem under impact angle constraint. However, MPPI sometimes suffer from degradation of robustness when the deep neural dynamics differs with the real environment. Many work try to robustified the MPPI method by ensemble models [8], $\mathcal{L}_1$ adaptive control [9], and so on. Tube-MPPI method is also proposed by combining an ancillary controller to keep the system states in the tube centered at nominal state computed using MPPI as nominal controller [10]. In this work, the deep neural dynamics

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

mismatch problem in Tube-MPPI is further improved using meta-learning to adapt the deep neural dynamics to environment changes online. Meta-learning provides learning-to-learn capability to deep neural network and is thus essential to real-world application of deep reinforcement learning to adapt to changes in environment online. This is critical in guidance problems since the target maneuver pose a large perturbation to the engagement dynamics. Thus in this work a Meta-learning Tube-MPPI method is proposed to tackle the impact angle guidance problem intercepting a strong maneuvering target.

Impact angle constrained guidance law helps to increase the missile capability, however in most of the existing guidance laws, high acceleration is needed at the beginning of the flight. The large acceleration command consumes excessive missile momentum energy, and will result control saturation. In [11], a hyperbolic tangent function weighted guidance law is proposed trying to tackle this problem. However, the value of proposed function grows exponentially with time, thus penalizing mostly impact angle and miss distance error while neglecting acceleration loss. Thus, utilizing the elegant saturation property of hyperbolic tangent function, a variant of hyperbolic function is proposed in this article that has an adjustable value during final phase of guidance. Different from [11] which use time as decision variable, range-to-go is used in this article since it provides more accurate information about current stage of guidance. Thus, in this work, a novel range-aware hyperbolic tangent function is proposed to reduce input saturation at the initial phase of guidance.

In this article, we develop a new range-aware meta-learning tube-MPPI guidance law with impact angle constraint. Given the limitations of prior work, the proposed approach is more sample efficient and impact angle constraint when compared with existing deep RL guidance laws. It also improves ancillary controller robustified MPPI method by reduced model mismatch using meta-learning, and benefit guidance performance by range-aware adaptive weighting compared with existing error shaping guidance laws. The main contribution of our work is as follows: 1) A meta-learning tube-MPPI control method is proposed. With this approach, the tube-MPPI performance is improved through reduced model mismatch of nominal controller using meta-learning model adaption. 2) A range-aware hyperbolic function is designed as an adaptive error shaping function in guidance law performance index design. This method benefits guidance performance by reduced initial acceleration and increased terminal velocity. 3) A new guidance scheme is formulated with aforementioned techniques for a varying velocity interceptor intercepting maneuvering target with desired terminal impact angle.

This article is organized as follows. Section II reviews existing works on deep RL guidance laws, weighted optimal guidance laws, MPPI and meta-learning. Section III details a novel guidance scheme based on model-based RL and meta-learning tube-MPPI. Numerical simulations are conducted to show the effectiveness of the proposed method in Section IV. Finally, conclusion is offered in section V.

II. RELATED WORK

A. DEEP RL IN GUIDANCE LAW DESIGN

Deep RL has proven to be successful in many control tasks. With fast evolving capability and good performance of deep RL, a growing trend emerges that modern guidance strategy incorporated deep RL framework to tackle guidance problem. Both model-free and model-based methods are incorporated in guidance design. For model-free methods, in [2], RL is used to design a missile guidance law in homing-phase, and it gives superior performance compared with guidance law design using Lyapunov theory. To tackle challenging environment and unknown highly variable dynamics, an adaptive guidance law and integrated navigation is proposed in [3] with deep meta-RL. Meta-learning can provide adaption to unforeseen environment changes through online learning while most traditional adaptive guidance is limited to specific faults [1], [14]. In Ref. [4], [12], Deep RL is also used to design a novel guidance law with solely seeker LOS angle and angular rate measurement for a mid-course exo-atmospheric interception. In [13], a deep RL based guidance law with missile attitude loop is proposed using PPO. Our work utilizes model-based RL techniques, thus has higher sample efficiency than these model-free RL guidance laws, and also achieves impact angle constrained guidance. Model-based RL is also used in guidance law design. In [7], a novel adaptive intercept angle guidance law with deep meta-RL is proposed for missile with actuator failures. Our work differs with [7] in the tube-MPPI approach and range aware hyperbolic functions that enhance the guidance performance.

B. OPTIMAL ERROR SHAPING GUIDANCE LAWS

Over the years, various efforts have been made to design improved optimal guidance law using different performance index [15]. Weighted cost function is utilized to shape the missile trajectory and distribute acceleration command during the engagement. Time-to-go [16], range-to-go [17] and generalized formulation [18] of weighted cost are used to alleviate initial high acceleration and highly curved trajectory problems in impact angle constraint guidance law. Other functions such as sinusoidal function [15], Gaussian function [19] have also been utilized in designing weighted optimal guidance laws. Ref. [11] employ hyperbolic tangent function as weighting in guidance design. However, the value of this variant of hyperbolic function in [11] grows exponentially with time. A range-aware hyperbolic tangent function is designed in this work to tackle this problem. Recent work in [20] also use error shaping to trade off acceleration against rate of error convergence. Our method inspired by this trade-off but differs by the range-aware weighting function that is adaptive respect to different stage of engagement, which benefit guidance performance.

C. MODEL-BASED RL

Model-based reinforcement learning is welcomed in many real-world control tasks for its high efficiency since a deep neural networks model is learned to solve the control task. MPPI is a typical method of model-based reinforcement learning to solve the control problem using the deep system model. MPPI is firstly used on real hardware in aggressive driving of rally vehicles in Ref. [22], and is implemented in a wide range of real control tasks including complex robot manipulation [23], missile guidance [7] and so on. Many attempts have been made to robustifies MPPI method. Ref [8] utilized model ensemble to tackle this problem, however, the ensemble of models normally deteriorate computation speed and may be inefficiency for real-time system. In [9], a $\mathcal{L}_1$ adaptive control method is combined with MPPI to address this problem and validated in multirotor racing. The Tube-MPPI in [10] utilize tube-based model predictive framework and robustifies MPPI by combine an ancillary controller as the tracking controller of nominal MPPI controller. Still, large difference between deep neural dynamic and true environment will impact central path and deteriorate ancillary controller tracking performance. Thus, meta-learning Tube-MPPI method is proposed in this work to address this problem. By utilizing meta-learning deep network dynamics is able to learn changes in environment online via learning to learn [24]. This is usually done by an update rule to the learner [25]. In this work, the tube-MPPI performance is improved through reduced model mismatch through meta-learning constantly adapting neural dynamics to changes in environment.

III. PROBLEM FORMULATION

The missile-target engagement dynamics is established for the purpose of guidance law development. Consider skid-to-turn roll-stabilized missile, the three dimensional missile target engagement geometry between missile M and target T in the inertial coordinate frame $O_I X_I Y_I Z_I$ is shown in Fig. 1, where the missile M has a velocity V_M, with direction defined by θ_m and ϕ_m; the target has a velocity V_T, with direction defined by θ_t and ϕ_t; line-of-sight (LOS) angles are denoted by θ_L and ϕ_L and the relative range is denoted by R.

Then the three-dimensional relative kinematic dynamics between missile and target can be expressed as follows [27]:

$$\dot{R} = V_T cos\theta_t cos\phi_t - V_M cos\theta_m cos\phi_m, \tag{1}$$

$$R\dot{\theta}_L = V_T sin\theta_t - V_M sin\theta_m, \tag{2}$$

$$R\dot{\phi}_L cos\theta_L = V_T cos\theta_t sin\phi_t - V_M cos\theta_m sin\phi_m. \tag{3}$$

The maneuver dynamics for target can be expressed as:

$$\dot{\theta}_t = \frac{a_{zt}}{V_T} - \dot{\phi}_L sin\theta_L sin\phi_t - \dot{\theta}_L cos\phi_t, \tag{4}$$

$$\dot{\phi}_t = \frac{a_{yt}}{V_T cos\theta_t} + \dot{\phi}_L tan\theta_t cos\phi_t sin\theta_L - \dot{\theta}_L tan\theta_t sin\phi_t - \dot{\phi}_L cos\theta_L, \tag{5}$$

where a_{yt} and a_{zt} are target accelerations. The forces acting on missile include thrust T, drag D, zero-lift drag D_0 and induced drag D_i. With missile mass denoted by m, and missile acceleration denoted by a_{ym} and a_{zm}, the dynamics of missile motion can be expressed as follows [28]:

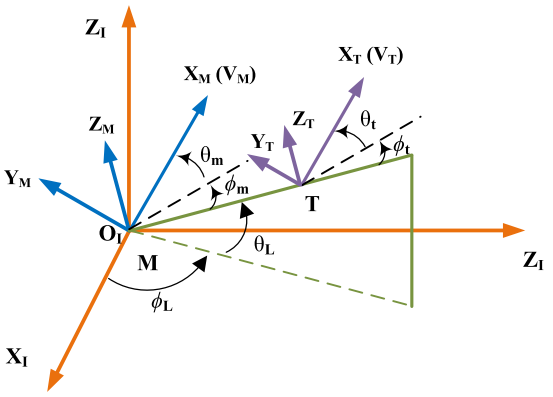

FIGURE 1. Missile-target interception geometry.

$$\dot{V}_M = \frac{T - D}{m} - g\,(cos\phi_m cos\theta_m sin\theta_L + sin\theta_m cos\theta_L), \tag{6}$$

$$\dot{\theta}_m = \frac{a_{zm} + g(cos\phi_m sin\theta_m sin\theta_L - cos\theta_m cos\theta_L)}{V_M} - \dot{\phi}_L sin\theta_L sin\phi_m - \dot{\theta}_L cos\phi_m, \tag{7}$$

$$\dot{\phi}_m = \frac{a_{ym} + g sin\phi_m sin\theta_m}{V_M cos\theta_m} - \dot{\theta}_L tan\theta_m sin\phi_m + \dot{\phi}_L tan\theta_m cos\phi_m sin\theta_L - \dot{\phi}_L cos\theta_L. \tag{8}$$

The equations of the forces can be expressed as:

$$D = D_0 + D_i; \quad D_0 = C_{D0} Qs, \tag{9}$$

$$D_i = \frac{Km^2(a_{ym}^2 + a_{zm}^2)}{Qs}, \tag{10}$$

$$K = \frac{1}{\pi A_r e}, \tag{11}$$

$$Q = \frac{1}{2}\rho V_M^2, \tag{12}$$

where C_{D0}, K, A_r, e, ρ, s and Q are zero-lift drag coefficient, interceptor, induced frag coefficient, aspect ratio, efficient factor, atmosphere density, reference area and dynamic pressure.

The objective of the guidance law is to achieve the interception of missile and target, with desired impact angle θ_{LD} and ϕ_{LD}. The impact angle is defined to be the LOS angles as in [27], [28]. As nullifying the LOS angular rates $\dot{\theta}_L$ and $\dot{\phi}_L$ can lead to the interception, the solution to this problem is to design the missile accelerations to guarantee the following equations:

$$\dot{\theta}_L = \dot{\phi}_L = 0, \quad \theta_L = \theta_{LD}, \ \phi_L = \phi_{LD}. \tag{13}$$

Thus we can see the problem of guidance law with desired impact angle can be reduced to the problem of controlling the LOS angles and angular rates as described in the above equation.

IV. DESIGN OF PROPOSED GUIDANCE LAW

In this section, a range aware impact angle guidance law is proposed with model-based RL and range-aware hyperbolic tangent function. By utilizing model-based RL and meta-learning, better data efficiency and online adaption capability is achieved. Meta-learning tube-MPPI approach which combine online adaption sampling based RL and disturbance rejection ancillary controller is constructed as model-based RL approach. Range-aware hyperbolic tangent function is then constructed as adaptive function used in performance index design to alleviate large initial acceleration command and highly curved trajectory problem in impact angle constrained guidance law. The schematic diagram of proposed guidance scheme with meta-learning tube-MPPI is shown in Fig.2.

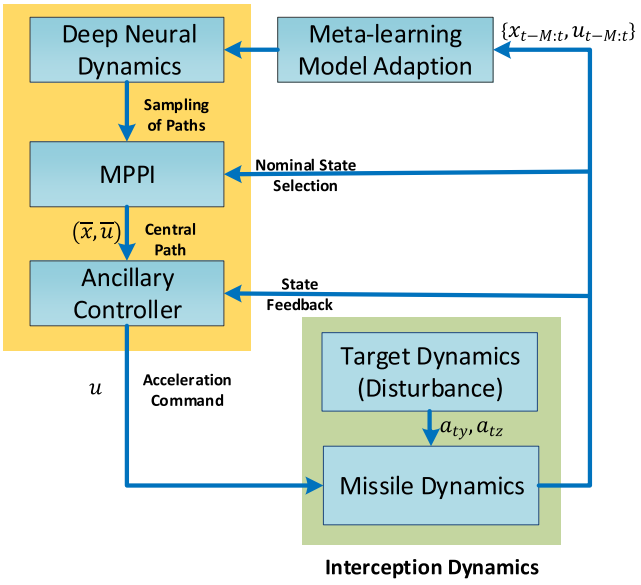

FIGURE 2. Schematic diagram of the proposed approach.

A. META-LEARNING NEURAL NETWORK DYNAMICS MODEL

A deep neural network dynamic model is built to be the predictive system dynamics model in model-based RL. Such neural dynamic model can be learned from observation data from real system. The neural network dynamics is noted as $\hat{\boldsymbol{x}}_{\boldsymbol{t+1}} = f_\theta(\boldsymbol{x_t}, \boldsymbol{u_t})$, where $\boldsymbol{x_t}$ and $\boldsymbol{u_t}$ are system state and input at time t, θ is the weight coefficients in the neural network, and $\hat{\boldsymbol{x}}_{\boldsymbol{t+1}}$ is the predicted system state at time t+1. The neural network utilized is a multi-layer dense network with ReLU activations. This deep neural dynamic is verified in [7] to have a neglectable prediction error, which will make failure of proposed deep RL controller unlikely.

Using meta-learning, deep neural dynamics can adapt to changes in environment online which solve the changing environment problem deep model-based RL facing. The meta-learning approach we adopted from [26] has two phases to make the neural dynamics optimized to the training dataset and also adapted to environment online. These two phases, the meta-training step and online adaption step are reviewed in the rest of this session.

In the meta-training step, the optimized deep neural dynamic model parameter θ_* is trained to be further adapted online. The model is trained using normalized training dataset by minimizing the mean square error of the prediction and actual value with 12 hours of offline training. The data is normalized to help the gradient flow in the training. Adam optimizer [29], a stochastic gradient descent optimization method, is employed to tackle this optimization problem.

In the online adaption phase, the meta-trained model $f_{\theta_*}(\boldsymbol{x_t}, \boldsymbol{u_t})$ is adapted using recent experience $\tau_\varepsilon(t-M, t-1)$ gained through environment to be a more accurate predictor. The adaption rule is selected to be gradient ascent of the likelihood of mean square error between prediction and ground truth using the recent experience:

$$\begin{aligned}&\mathcal{N}_\psi\left(\tau\left(t-M, t-1\right), \theta\right)\\&= \theta + \alpha\nabla_\theta\Big(\frac{1}{M}\sum_{m=t-1}^{t-M}\Big\|\left(\dot{q}_{m+1}-\dot{q}_m\right)-\hat{f}_\theta\left(\boldsymbol{x_m}, \boldsymbol{u_m}\right)\Big\|^2\Big),\end{aligned} \quad (14)$$

where α is the learning rate.

B. TUBE-MPPI CONTROLLER

Based on the meta-learning neural dynamic model trained above, a tube-MPPI controller can be built for the guidance problem. Tube-MPPI is a variant of tube-MPC which consist of a nominal controller and an ancillary controller [10]. The nominal considers general costs and generates nominal state: the central path, while the ancillary controller tracks the actual system state in a tube centered at the central path. The actual state is bound in a tube centered at the central path. Since in this guidance law we are more concerned with the robust ability of tube-MPPI and there are no other state constraints, therefor we do not concern with the computation of this bound in this work.

The nominal system can ignore system disturbances, like in [10], [30], two copies of the nominal controller are run with one from the actual state and the other one from nominal state. The mechanism accepts the MPPI solution of actual state if its cost is lower than the cost of solution from nominal state plus some threshold. In this way disturbances that are not catastrophic are feedback to the nominal controller to replan.

1) NOMINAL CONTROLLER

The MPPI controller, which is a sampling based optimal control method solving the stochastic HJB equation [22], is used as the nominal controller and is given in Alg. 1. In MPPI, we consider the optimal control problem to find a control sequence that minimize the cost functions:

$$U^* = \underset{U}{\operatorname{argmin}}\,\mathbb{E}_{\mathbb{P}}[\phi\left(\boldsymbol{x_T}\right) + \sum_{t=0}^{T-1} c_t\left(\boldsymbol{x_t}, \boldsymbol{u_t}\right)], \quad (15)$$

where the c and ϕ are positive definite running and terminal cost function respectively, and $\mathbb{P}$ denote the distribution

Algorithm 1 Nominal Controller (MPPI) at Each Time Stamp

Given: $\hat{f}_\theta(\boldsymbol{x}_t, \boldsymbol{u}_t)$: transition model with parameter θ of a prior;
$\mathcal{N}_\psi$: update rule of parameter θ;
$\tau\,(t-M, t-1)$: experience;
N: Number of samples;
T: Horizon;
Σ, r_t, ϕ: Control hyper-parameters

$\theta'_* \leftarrow \mathcal{N}_\psi\,(\tau\,(t-M, t-1), \theta_*)$
$\overline{\boldsymbol{x}_0} \leftarrow$ SetNominalState()
for $n = 0, \ldots, N-1$ **do**
  $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$
  Sample $\mathcal{E}_t^n = \{\delta\boldsymbol{u}_0^n, \delta\boldsymbol{u}_1^n, \ldots, \delta\boldsymbol{u}_{T-1}^n\}$
  for $t = 1, \ldots, T$ **do**
    $\boldsymbol{x}_t \leftarrow \hat{f}_{\theta'_*}(\boldsymbol{x}_{t-1}, \boldsymbol{u}_{t-1} + \delta\boldsymbol{u}_{t-1}^n)$
    $S\left(\mathcal{E}_t^n\right) \mathrel{+}= c_t\,(\boldsymbol{x}_t, \boldsymbol{u}_t) + \lambda \boldsymbol{u}_{t-1}^T \Sigma^{-1} \delta\boldsymbol{u}_{t-1}^n$
  end for
  $S\left(\mathcal{E}_t^n\right) \mathrel{+}= \phi\,(\boldsymbol{x}_T)$
end for
$S'\left(\mathcal{E}_t^n\right) = S\left(\mathcal{E}_t^n\right) - min_n[S\left(\mathcal{E}_t^n\right)]$
$\lambda = \sigma(S'\left(\mathcal{E}_t^n\right))$
$\eta = \sum_{n=0}^{N-1} \exp(-\frac{1}{\lambda}(S'\left(\mathcal{E}_t^n\right)))$
for $n = 0, \ldots, N-1$ **do**
  $w_t^n \leftarrow \frac{1}{\eta} exp(-\frac{1}{\lambda}(S'\left(\mathcal{E}_t^n\right)))$
end for
for $t = 0, \ldots, T-1$ **do**
  $\overline{\boldsymbol{u}_t} \mathrel{+}= \sum_{n=1}^{N} w_t^n \delta\boldsymbol{u}_t^n$
end for
$\overline{\boldsymbol{X}} \leftarrow$ Simulate($\overline{\boldsymbol{x}_0}, \overline{\boldsymbol{U}}$)
PublishSolution($\overline{\boldsymbol{X}}, \overline{\boldsymbol{U}}$)
for $t = 0, \ldots, T-1$ **do**
  $\boldsymbol{u}_{t-1} \leftarrow \boldsymbol{u}_t$
end for
$\boldsymbol{u}_T \leftarrow Initialize(\boldsymbol{u}_{T-1})$

corresponding to the dynamics $F\,(\boldsymbol{x}, \boldsymbol{u} + \delta\boldsymbol{u})$, $\delta\boldsymbol{u}$ is a Gaussian noise vector. The noise is essential to use sampling method originated from stochastic optimal control and also as a way of exploration. Denote V as perturbed input into the system, $\boldsymbol{h}$ as the input sequence of uncontrolled system and $\mathbf{p}$ as the open-loop control sequence, the free energy of the dynamic system is defined as follows:

$$\mathcal{F}\,(V) = -\lambda \log(\mathbb{E}_{\mathbb{P}}[\exp(-\frac{1}{\lambda} S(V))]), \tag{16}$$

where λ is a positive scalar. According to [31], the cost of optimal control problem is bounded below from this free energy. Further derivation using Jensen's inequality can get the optimal distribution of the control sequence:

$$\boldsymbol{p}^*\,(V) = \frac{1}{\eta} \exp\left(-\frac{1}{\lambda} S\,(V)\right) \boldsymbol{h}\,(V), \tag{17}$$

where η is the normalizing factor. Then we can get the optimal control solution by minimizing the gap measured by Kullback-Leibler divergence:

$$U^*\,(V) = \underset{U}{\operatorname{argmin}}\, \mathbb{D}_{KL}(\mathbb{P}^* || \mathbb{P}). \tag{18}$$

After expanding out the KL divergence, and analyzing the concave result, the optimal control sequence can be derived as follows. Since the optimal distribution $\mathbb{Q}^*$ cannot be directly sampled, importance sampling technique is taken to get the sequence

$$\boldsymbol{u}_t^* = \int \boldsymbol{p}^*\,(V)\, \boldsymbol{v}_t dV, \tag{19}$$

$$= \mathbb{E}_P\,[w\,(V)\, \boldsymbol{v}_t], \tag{20}$$

where the importance sampling weight $w\,(V)$ is:

$$w_t(V) = \frac{\boldsymbol{p}^*\,(V)}{\boldsymbol{h}\,(V)} \exp(\sum_{t=0}^{T-1} (-\boldsymbol{v}_t^T \Sigma^{-1} \boldsymbol{u}_t + \frac{1}{2} \boldsymbol{u}_t^T \Sigma^{-1} \boldsymbol{u}_t)),$$
$$= \frac{1}{\eta} exp\left(-\frac{1}{\lambda}\left(S(U+\mathcal{E}) + \lambda \sum_{t=0}^{T-1} \frac{1}{2} \boldsymbol{u}_t^T \Sigma^{-1} (\boldsymbol{u}_t + 2\delta\boldsymbol{u}_t)\right)\right). \tag{21}$$

The temperature coefficient λ of this softmax distribution is designed to be as follows to normalize the cost function distribution as in [7]:

$$\lambda = \lambda^* \sigma\,(S\,(V)), \tag{22}$$

Then the control sequence of MPPI is updated using N samples as iterative update law:

$$\boldsymbol{u}_t^{i+1} = \boldsymbol{u}_t^i + \sum_{n=1}^{N} w_t^n \delta\boldsymbol{u}_t^n. \tag{23}$$

Remark 1: The MPPI framework can be viewed as a stochastic optimal control (SOC) approach. With inspiration from [32]–[34], the stability is discussed as follows. If we denote the corresponding continuous dynamical system which is the guidance dynamics as:

$$d\boldsymbol{x}_t = (\boldsymbol{f}\,(\boldsymbol{x}_t) + \boldsymbol{G}\,(\boldsymbol{x}_t)\,\boldsymbol{u}_t)\,dt + Bd\boldsymbol{w}, \tag{24}$$

where B defines the covariance of the system, and $\boldsymbol{w}$ is a Brownian disturbance. If we denote the value function as $\mathcal{V}$, then the continuous time value function is:

$$\mathcal{V}\,(\boldsymbol{x}_t) = \min_u \mathbb{E}_{\boldsymbol{w}}\left[\phi\,(\boldsymbol{x}_T) + \int_t^T \left(q\,(\boldsymbol{x}_t) + \frac{1}{2} \boldsymbol{u}_t^T R \boldsymbol{u}_t\right) dt\right],$$

then the stochastic HJB equation is given as:

$$-\partial_t \mathcal{V} = c\,(\boldsymbol{x}_t) + \mathcal{V}_x^T f\,(\boldsymbol{x}_t) - \frac{1}{2} \mathcal{V}_x^T G\,(\boldsymbol{x}_t)\, R^{-1} G^T\,(\boldsymbol{x}_t)\, \mathcal{V}_x$$
$$+ \frac{1}{2} Tr\left(BB^T \mathcal{V}_{xx}\right) \tag{25}$$

with the boundary condition $\mathcal{V}\,(\boldsymbol{x}) = \phi\,(\boldsymbol{x})$ and optimal control expressed as:

$$\boldsymbol{u}^*\,(\boldsymbol{x}_t, t) = -R^{-1} G^T\,(\boldsymbol{x}_t)\, \mathcal{V}_x. \tag{26}$$

According to [21], this control can be computed using Feynman-Kac formula to the transformed Chapman-Kolmogorov equation, and (26) can also be expressed as:

$$\boldsymbol{u}^* dt = \mathcal{G}(\boldsymbol{x}_t, t) B \frac{\mathbb{E}_{\mathbb{P}}[\exp(-\frac{1}{\lambda} S(V)) d\boldsymbol{w}]}{\mathbb{E}_{\mathbb{P}}[\exp(-\frac{1}{\lambda} S(V))]}, \tag{27}$$

where $\mathcal{G}(\boldsymbol{x}_t, t) = R^{-1} G_c^T(\boldsymbol{x}_t) \left(G_c(\boldsymbol{x}_t) R^{-1} G_c^T(\boldsymbol{x}_t)\right)^{-1}$.

As the importance sampling is an unbiased estimate [37] and the universal approximation theorem of the feed-forward neural network used in our work, the noise adopted from the MPPI control method is zero mean. If we denote the variance of noise profile as Σ, the noise enters the system through control is $B = G\sqrt{\Sigma}$. Thus after discretization, and set Δt as the unit time, the control command of (27) can be derived as:

$$\begin{aligned} \boldsymbol{u}^* &= \mathcal{G}(\boldsymbol{x}_t, t) B \frac{\mathbb{E}_{\mathbb{P}}[\exp(-\frac{1}{\lambda} S(V)) \frac{\epsilon}{\sqrt{\Delta t}}]}{\mathbb{E}_{\mathbb{P}}[\exp(-\frac{1}{\lambda} S(V))]} \\ &= \frac{\mathbb{E}_{\mathbb{P}}[\exp(-\frac{1}{\lambda} S(V)) \sqrt{\Sigma} \epsilon]}{\mathbb{E}_{\mathbb{P}}[\exp(-\frac{1}{\lambda} S(V))]}, \end{aligned} \tag{28}$$

where ϵ is the Gaussian noise vector. Thus this is equivalent to the solution of the path integral approach in (20) and the resulting control command is the solution to the stochastic HJB equation, which can also be expressed in (26), which is also shown in [38].

If we choose the value function $\mathcal{V}$ as the stochastic control Lyapunov function (SCLF) [35], according to proof of Lemma 3.14 in A.1.4 in [39], $\mathcal{V}$ is positive definite in Lyapunov sense such that $\mathcal{V}(0, t) = 0, \mathcal{V}(e, t) \geq \mu(|e|)$ $\forall t > 0, \mu \in \mathcal{K}$. Then recall (25) and (26),

$$\begin{aligned} L(\mathcal{V}) &= \partial_t \mathcal{V} + \mathcal{V}_x^T (f(\boldsymbol{x}_t) + G(\boldsymbol{x}_t) \boldsymbol{u}_t) + \frac{1}{2} Tr(BB^T \mathcal{V}_{xx}), \\ &= -\left(\frac{1}{2} \mathcal{V}_x^T G(\boldsymbol{x}_t) R^{-1} G^T(\boldsymbol{x}_t) \mathcal{V}_x + c(\boldsymbol{x}_t)\right). \end{aligned} \tag{29}$$

Since by definition, c is positive definite and R is positive definite matrix, thus the value function is a strict SCLF with $L(\mathcal{V}(\boldsymbol{x}, t)) \leq 0$. According to theorem 5.3 in [36], the corresponding system (24) is stable in probability and the MPPI controller is a stabilizing controller.

2) ANCILLARY CONTROLLER

The ancillary controller acts as a tracking controller, which keep the actual system state in the tube, centered at the central path computed by the nominal controller. This is a standard tracking problem with small initial error and quadratic cost. With many solution exist, a nonlinear MPC utilizing iLQG is selected as the ancillary controller as in [10], [30], with the state convergence in finite time shown in [41], this widely used control method provide relative good performance.

C. RANGE-AWARE HYPERBOLIC TANGENT FUNCTION

The hyperbolic tangent function can be expressed as follows:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{30}$$

and its figure is drawn in fig.3.

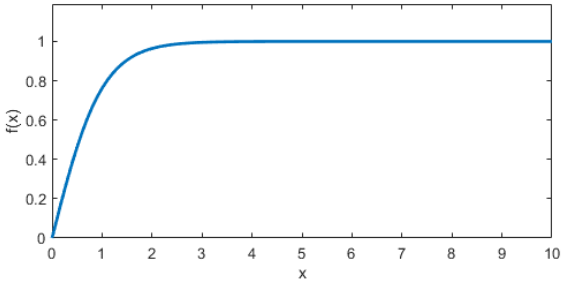

FIGURE 3. Hyperbolic tangent function.

Different from time, range-to-go provide more accurate information about the current stage of guidance, thus relative range is used as the decision variable of the adaptive function. We also want an adjustable saturated value at the end of guidance, thus the elegant property of hyperbolic function is necessary. Utilizing the above analysis, a variant of hyperbolic tangent function is designed as:

$$f_{RA-tanh}(R) = K_{RA} tanh\left(\frac{\sigma}{R + \varphi}\right), \tag{31}$$

where K_{RA}, φ, σ are positive coefficient. If we select them as 1, 1, 300 respectively, its figure can be drawn below:

From the figure we can see the function value is low when the interceptor is far from the target. The value then increases faster and faster as range closes and saturate at K_{RA} at the end. The saturated value, initial value and increasing speed can be adjusted using K_{RA}, φ, σ. Thus using this function, we can adjust the trading off between acceleration command and error approaching rate, and thus achieve error shaping in the objective function.

As in [7], based on (13), the derivative of LOS angular rates reference planning algorithm is defined as:

$$\boldsymbol{e}_0 = [\theta_L - \theta_{LD}, \phi_L - \phi_{LD}]^T, \tag{32}$$

$$\boldsymbol{e}_1 = \dot{\boldsymbol{e}}_0 + K_1 \boldsymbol{e}_0, \tag{33}$$

where $K_1 > 0, K_2 > 0$, and the planned reference is:

$$\dot{\boldsymbol{e}}_1 = -K_2 \boldsymbol{e}_1. \tag{34}$$

In this way, adjustable approaching rate make the error variables change more smoothly, and the convergence of the error variable to zero can be proved by selecting a Lyapunov function $V = \frac{1}{2} \boldsymbol{e}_1^2$, then:

$$\dot{V} = \boldsymbol{e}_1 \dot{\boldsymbol{e}}_1, \tag{35}$$

$$= -K_2 \boldsymbol{e}_1^2 \leq 0. \tag{36}$$

Thus the error variable will converge to zero and control objective is satisfied.

As we know, $\dot{\boldsymbol{e}}_1$ is the derivate of the LOS angular rate that is proportional to the acceleration command. Thus the trade-off between error approaching rate and acceleration command can be utilized using proposed range-aware hyperbolic function. The state dependent cost function of MPPI is then selected as:

$$\dot{\boldsymbol{e}}_2 = \dot{\boldsymbol{e}}_1 + f_{RA-tanh}(R)\,\boldsymbol{e}_1, \tag{37}$$

$$c(\boldsymbol{x}) = \| \boldsymbol{e}_2 \|^2. \tag{38}$$

V. NUMERICAL SIMULATION

In this section, numerical simulations for the proposed meta-learning tube-MPPI guidance law and other two guidance laws are conducted for comparison. The other two guidance laws taken into account for comparison are the meta-MPPI guidance law from [7] and a guidance law constructed using tube-MPPI in [10]. Monte Carlo simulation of the proposed guidance law is also utilized to verify the robustness and effectiveness of proposed guidance law.

The interceptor has acceleration limit $A_{M-max} = 200\text{m}/s^2$ in both directions, intercepting a strong maneuvering target with and the initial conditions are listed in table 1. In the simulation, a realistic interceptor velocity model from [28] is used, where the value of zero-lift drag coefficient C_{D0}, induced frag coefficient K, aspect ratio A_r, efficient factor e, atmosphere density ρ, reference area s can be found. The interceptor thrust and mass are considered as:

$$T = \begin{cases} 7500 & 0 \le t \le T_B, \\ 0 & t > T_B, \end{cases} \tag{39}$$

$$m = \begin{cases} 90.035 + 3.31\,(T_B - t) & 0 \le t \le T_B, \\ 90.035 & t > T_B. \end{cases} \tag{40}$$

TABLE 1. Case parameters.

Dataset	Case 1	Case 2	Monte Carlo
R(0) (m)	4300	4300	4500
$\theta_L(0)$ (rad)	-0.4	-0.8	$Unif(-0.3,-0.9)$
$\phi_L(0)$ (rad)	0.8	0.6	$Unif(0.5,1.1)$
$\theta_m(0)$ (rad)	0.1	-0.2	$Unif(-0.3,0.3)$
$\phi_m(0)$ (rad)	0.2	-0.2	$Unif(-0.3,0.3)$
$V_M(0)$ (m/s)	800	800	800
$\theta_t(0)$ (rad)	0.0	0.1	0.0
$\phi_t(0)$ (rad)	-0.1	0.0	0.0
V_T (m/s)	270	270	270
θ_{LD} (rad)	-0.6	-0.6	-0.6
ϕ_{LD} (rad)	0.8	0.8	0.8
a_{yt} (m/s^2)	80sin(t)	80sin(t)	60sin(0.2t)
a_{zt} (m/s^2)	80sin(t)	80sin(t)	60sin(0.2t)
T_B(s)	3.0	3.0	4.0

The initial engagement parameters in the simulation is given in table 1, where *Unif* means a uniform distribution, and the interceptors have a max 200 m/s^2 acceleration limit in these cases.

The MPPI controller used as nominal controller has a horizon 3 with control cycle 5ms, 1000 trajectories drawn, temperature coefficient λ^* set to 1. Ancillary controller has control cycle of 2ms. The meta-learning neural network dynamics has two hidden layers with 512 neurons, ReLU activation, and is trained using twenty minutes of data. The online learning rate α for meta-learning is set to 0.001. The step size of simulation integration is adaptive and less than 0.01ms for the environment.

Table 2 shows the simulation result of case 1, where ϕ_{LT} and ϕ_{LT} is terminal LOS angles respectively. The miss distance of proposed guidance law is smaller than the other guidance laws. Which indicates a better LOS angular rate tracking performance at the end of guidance. The error in terminal impact angle is also smaller with the proposed guidance law. With range-aware hyperbolic function, the proposed guidance law has a smaller cumulative control effort that results quicker impact and increased terminal velocity. Thus the results demonstrate the proposed guidance law has better performance than the meta-MPPI guidance law and the proposed meta-learning MPPI method has better performance than tube-MPPI in [10].

TABLE 2. Case 1 simulation result.

	Proposed	Meta-MPPI	Tube-MPPI
Miss Distance(m)	1.50×10^{-5}	2.46×10^{-4}	4.31×10^{-5}
θ_{LT} (rad)	-0.6019	-0.5882	-0.5960
ϕ_{LT} (rad)	0.7979	0.8180	0.8031
Impact Time(s)	10.92	12.51	10.93
V_{MT}(m/s)	447.06	376.23	425.64
$\int_{t_0}^{t_f} \lvert u_t \rvert\, dt$	1411.3	2084.9	1600.1

In case 1, the scenario setting is as listed in table 1. A comparison demonstration of the proposed guidance, meta-MPPI law and tube-MPPI guidance law is conducted. As the simulation result in table 2 shows, the proposed guidance achieves better outcome than the other guidance law in miss distance, terminal angle error and impact time. Fig. 4 shows the trajectories of interceptor and target in these guidance laws, all guidance laws drive the interceptor to interception, but quicker impact is achieved through shaping of the trajectory in the proposed guidance law. The LOS angles and angular rates during the interception in this case are shown in Fig. 6. The LOS angles of meta-MPPI guidance law diverge at the end of the interception, due to the fluctuate in the LOS angular rates which shows the guidance law has difficulty in tracking the desired LOS angle and angular rates. The proposed method achieves better tracking performance with ancillary controller and tube method. LOS angles and angular rates of tube-MPPI guidance law also fluctuate since nominal controller diverge greatly with strong target maneuver

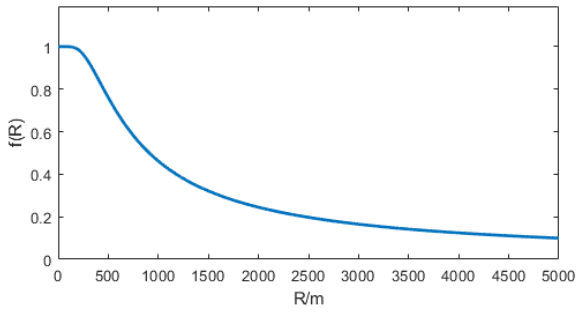

FIGURE 4. Range-aware hyperbolic tangent function.

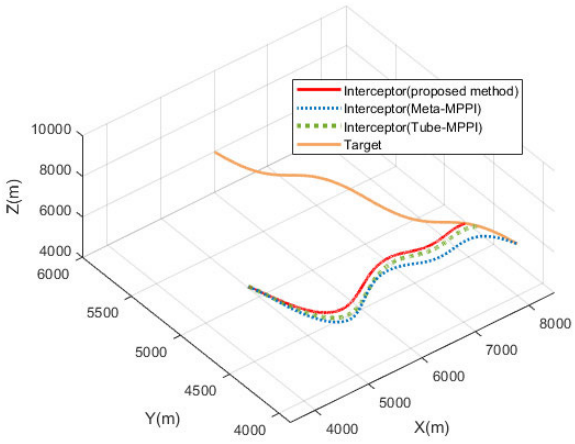

FIGURE 5. Trajectories of interceptors and target in case 1.

disturbance, and we can see the proposed method has better tracking ability since online adaption to environment changes is done by meta-learning. Better LOS angle and angular rates tracking performance also result in better terminal angle and miss distance as shown in table 2. Fig. 7 shows the acceleration profile of the guidance law during engagement, it can be seen that the acceleration command of meta-MPPI guidance law chatters at the terminal phase of the guidance law. From Fig. 6-8, the proposed guidance law consumes less energy and poses larger terminal velocity than meta-MPPI guidance law due to range-aware hyperbolic tangent function as weighting function in guidance law design. Thus the proposed guidance law achieves satisfactory performance with range-aware hyperbolic tangent function, tube-MPPI method and meta-learning.

Simulation using different setting in engagement scenario is conducted in case 2 to further demonstrate the comparative performance of the proposed guidance law. The simulation results are listed in table 3. The results shows the proposed guidance law achieves better performance than comparative guidance laws. We can see the tube-MPPI method has a worsen result than case 1. This is partly because the LOS angles happens to fluctuate near the desired LOS angles in case 1 which can be seen in Fig. 7. The range-aware hyperbolic tangent function result in a reduced cumulative control

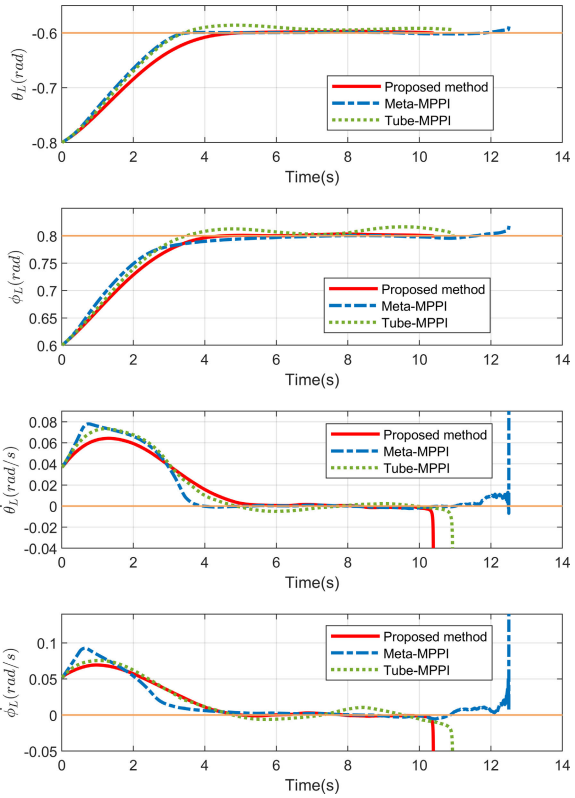

FIGURE 6. LOS angles and angular rates in case 1.

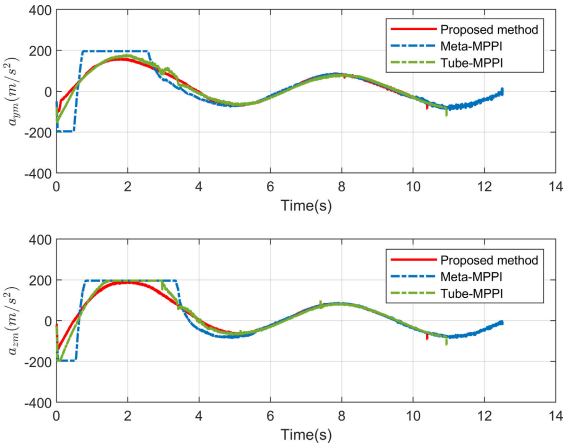

FIGURE 7. Interceptor accelerations command in case 1.

TABLE 3. Case 2 simulation result.

	Proposed	Meta-MPPI	Tube-MPPI
Miss Distance(m)	1.34×10^{-5}	3.41×10^{-4}	3.81×10^{-5}
θ_{LT} (rad)	-0.6018	-0.5901	-0.5675
ϕ_{LT} (rad)	0.7975	0.7911	0.8357
Impact Time(s)	10.60	11.53	10.61
V_{MT}(m/s)	427.22	391.99	426.75
$\int_{t_0}^{t_f} \lvert u_t \rvert \, dt$	1267.8	1671.0	1277.2

effort that also result in a reduced impact time and increased terminal velocity than meta-MPPI guidance law.

The trajectories of interceptors and target, LOS angle and angular rates, acceleration profile and interceptor velocity

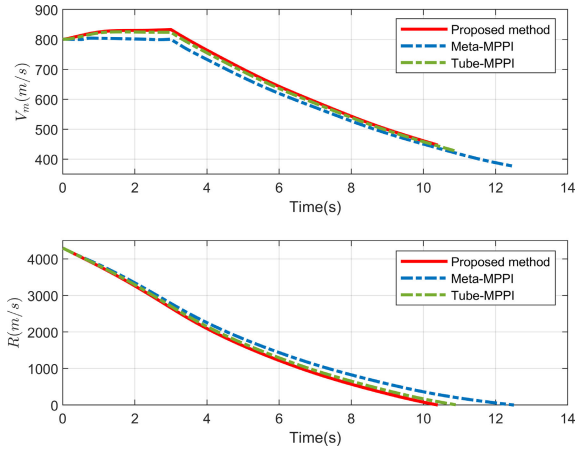

FIGURE 8. Interceptor velocity and relative distance in case 1.

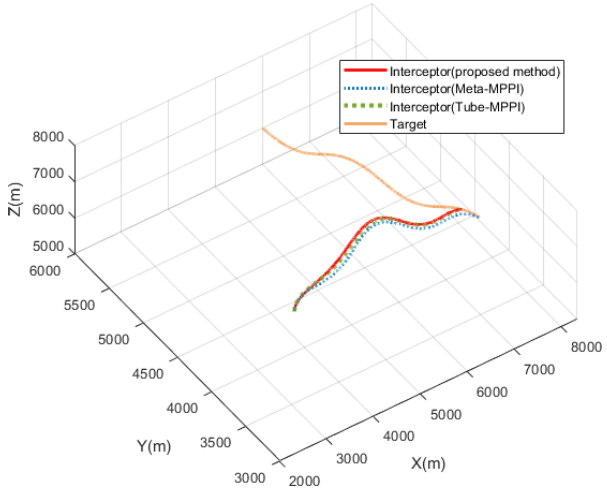

FIGURE 9. Trajectories of interceptors and target in case 2.

is shown in Fig. 9-12. We can see from these figures that guidance laws provide similar simulation result in case 1. From Fig. 11, the control command of meta-MPPI guidance law fluctuate at end, causing the LOS angular rate to diverge. Control command of guidance law jump at 3s which is caused by thrust burnout causing perturbation to the system. Fig.11 also shows the tube-MPPI control command sometimes deviate due to tracking performance, and cause a slight increase in cumulative control command between proposed control approach and tube-MPPI control method. With range-aware adaptive function, the control command in the initial of interception is reduced, which is shown in Fig.11 when the meta-MPPI command saturate. The reduced in initial command results increased terminal velocity and quicker impact time as in Fig.12. Thus the proposed guidance law achieves better LOS angle, angular tracking performance and consumes less energy which result in better miss distance, terminal angle error, impact and terminal velocity.

As the Monte Carlo method is powerful in analyzing effectiveness and robustness, 5000 rollouts are conducted

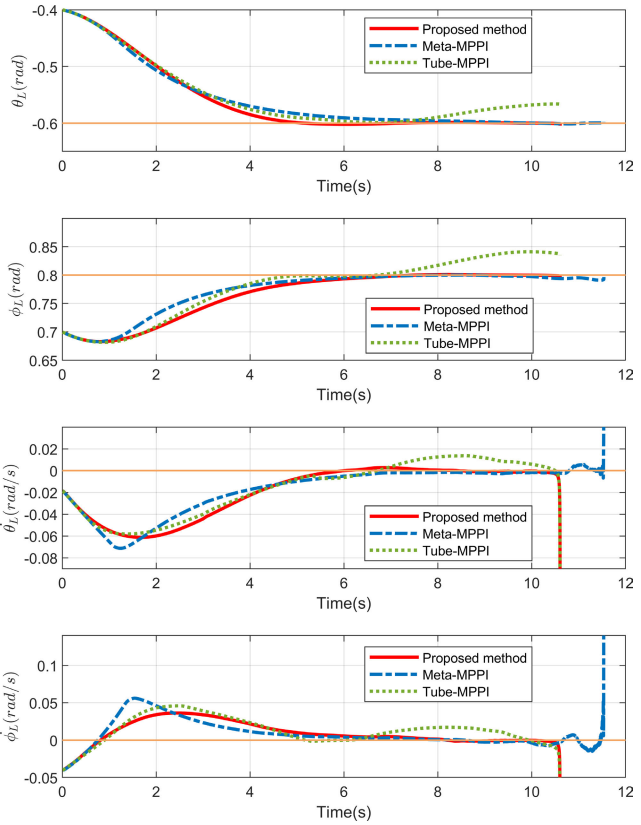

FIGURE 10. LOS angles and angular rates in case 2.

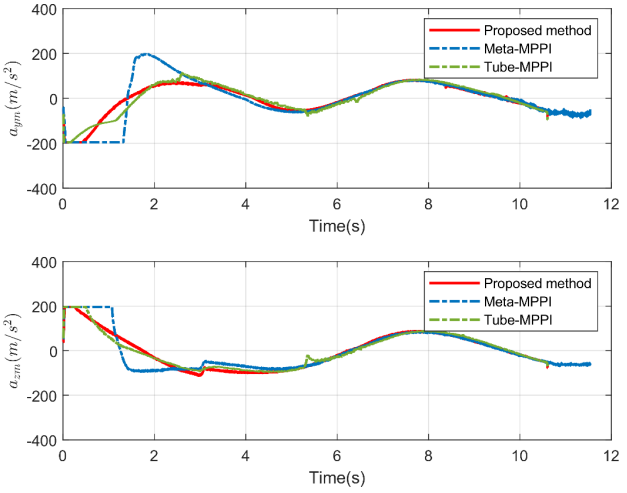

FIGURE 11. Interceptor accelerations command in case 2.

to further verify the performance of proposed method. The initial condition is listed in table 1, the initial LOS angle and missile heading is set to be in a uniform distribution ranging 0.6 rad to cover operating initial condition range. The target acceleration changing rate is reduced to get a clearer picture of result, and has 60 m/s^2 max value. The Gaussian measurement noise of LOS angle is set to zero mean, 8mrad standard deviation. And standard deviation of LOS angular

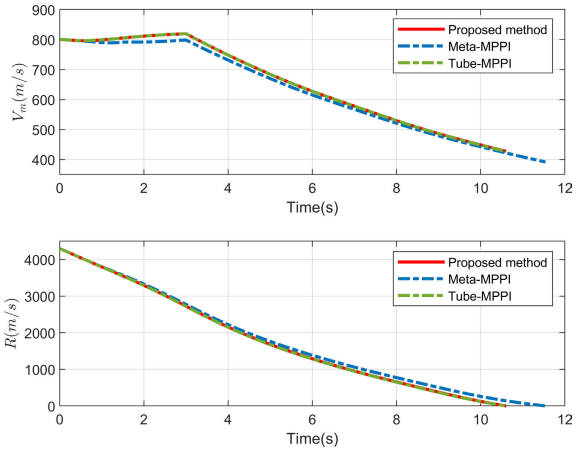

FIGURE 12. Interceptor velocity and relative distance in case 2.

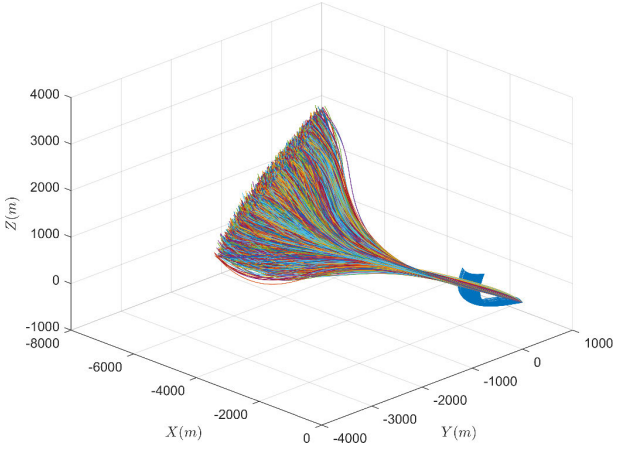

FIGURE 13. Trajectories of interceptors (multi-color) and target (blue) in Monte Carlo case.

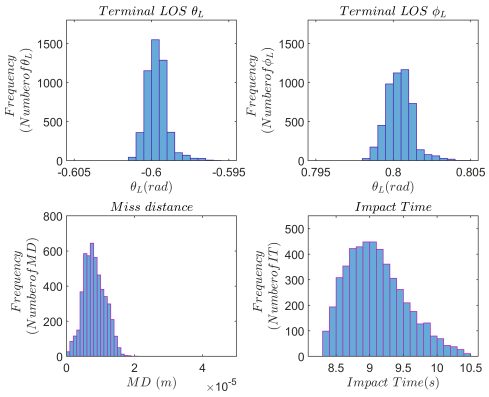

FIGURE 14. Histogram of terminal LOS angles, miss distance, and impact time in Monte Carlo case.

rates Gaussian noise is set to one percent its current measurement value. The results are shown in Fig.13-16. In Fig.13, all trajectories of interceptors and targets are transformed into origin at interception point to make them easy to see. The trajectories show all rollouts has successful hit. The histogram of miss distance, terminal impact angle and impact

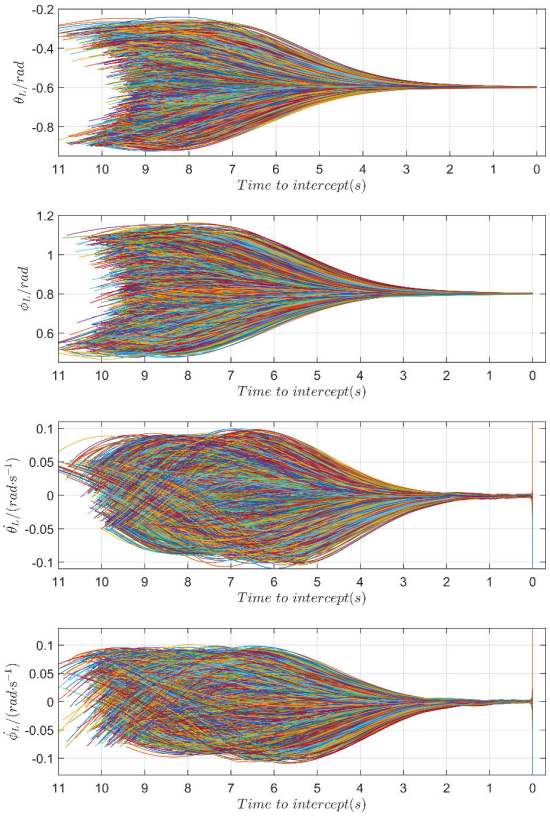

FIGURE 15. LOS angles and angular rates in Monte Carlo case.

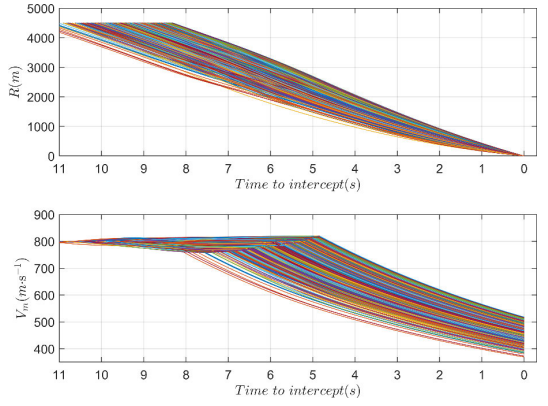

FIGURE 16. Miss distance and missile velocity in Monte Carlo case.

time is shown in Fig.14, and we can see majority has a small error. The deviation in mean is due to the consistency in target maneuver. As LOS angles and angular rates of interceptors shown in Fig.15, all LOS angles and angular rates converge to desired value. In Fig.16, we can see the different terminal velocity of missile caused by different engagement condition.

VI. CONCLUSION

In this article, we present a new range-aware impact guidance law using model-based RL technique for a varying velocity interceptor intercepting a maneuvering target with desired impact angle. Model-based deep RL method is used in guidance law design and a deep neural dynamic capable of online

adapting to environment change via meta-learning is used as predictive model. The predictive model is then utilized in MPPI to solve the optimal control problem via importance sampling of path integrals to compute the nominal state and control as central path. An ancillary controller tracks the central path to keep system states in the tube. The benefit of combing meta-learning and tube-MPPI method is that the model mismatch of the nominal controller is reduced to improve overall control performance. The benefit is verified in simulation which shows the proposed approach achieves better tracking performance than tube-MPPI method. Numerical simulation clearly indicates the proposed method can reduce acceleration at initial phase and increase terminal velocity. Compared with meta-MPPI guidance which acceleration command and LOS angular rate chatters at end, the proposed guidance law clearly shows more robust capability in disturbance rejection. Monte Carlo simulation result verifies the effectiveness and robustness of proposed guidance law under operating conditions.

REFERENCES

[1] D. Lin, Y. Ji, W. Wang, Y. Wang, H. Wang, and F. Zhang, "Three–dimensional impact angle–constrained adaptive guidance law considering autopilot lag and input saturation," *Int. J. Robust Nonlinear Control*, vol. 30, no. 9, pp. 3653–3671, Jun. 2020.

[2] B. Gaudet and R. Furfaro, "Missile homing-phase guidance law design using reinforcement learning," in *Proc. AIAA Guid., Navigat., Control Conf.*, Minneapolis, MN, USA, Aug. 2012, p. 4470.

[3] B. Gaudet, R. Linares, and R. Furfaro, "Adaptive guidance and integrated navigation with reinforcement meta-learning," *Acta Astronautica*, vol. 169, pp. 180–190, Apr. 2020.

[4] B. Gaudet, R. Furfaro, and R. Linares, "A guidance law for terminal phase exo-atmospheric interception against a maneuvering target using angle-only measurements optimized using reinforcement meta-learning," in *Proc. AIAA Scitech Forum*, Orlando, FL, USA Jan. 2020, p. 609.

[5] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *J. Intell. Robotic Syst.*, vol. 86, no. 2, pp. 153–173, May 2017.

[6] I. S. Mohamed, G. Allibert, and P. Martinet, "Model predictive path integral control framework for partially observable navigation: A quadrotor case study," 2020, *arXiv:2004.08641*. [Online]. Available: http://arxiv.org/abs/2004.08641

[7] C. Liang, W. Wang, Z. Liu, C. Lai, and B. Zhou, "Learning to guide: Guidance law based on deep meta-learning and model predictive path integral control," *IEEE Access*, vol. 7, pp. 47353–47365, 2019.

[8] I. Abraham, A. Handa, N. Ratliff, K. Lowrey, T. D. Murphey, and D. Fox, "Model-based generalization under parameter uncertainty using path integral control," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2864–2871, Apr. 2020.

[9] J. Pravitra, K. A. Ackerman, C. Cao, N. Hovakimyan, and E. A. Theodorou, "L1-adaptive MPPI architecture for robust and agile control of multirotors," 2020, *arXiv:2004.00152*. [Online]. Available: http://arxiv.org/abs/2004.00152

[10] G. Williams, B. Goldfain, P. Drews, K. Saigol, and E. Theodorou, "Robust sampling based model predictive control with sparse objective information," in *Proc. Robot., Sci. Syst.*, Pittsburgh, PA, USA, Jun. 2018, pp. 1–7.

[11] S. Xiong, M. Wei, M. Zhao, H. Xiong, W. Wang, and B. Zhou, "Hyperbolic tangent function weighted optimal intercept angle guidance law," *Aerosp. Sci. Technol.*, vol. 78, pp. 604–619, Jul. 2018.

[12] B. Gaudet, R. Furfaro, and R. Linares, "Reinforcement learning for angle-only intercept guidance of maneuvering targets," *Aerosp. Sci. Technol.*, vol. 99, Apr. 2020, Art. no. 105746.

[13] B. Gaudet, R. Furfaro, R. Linares, and A. Scorsoglio, "Reinforcement meta-learning for interception of maneuvering exoatmospheric targets with parasitic attitude loop," 2020, *arXiv:2004.09978*. [Online]. Available: http://arxiv.org/abs/2004.09978

[14] L. Fei and J. Haibo, "Guidance laws with input constraints and actuator failures," *Asian J. Control*, vol. 18, no. 3, pp. 1165–1172, May 2016.

[15] C.-H. Lee, J.-I. Lee, and M.-J. Tahk, "Sinusoidal function weighted optimal guidance laws," *Proc. Inst. Mech. Eng., G, J. Aerosp. Eng.*, vol. 229, no. 3, pp. 534–542, Mar. 2015.

[16] C.-K. Ryoo, H. Cho, and M.-J. Tahk, "Time-to-go weighted optimal guidance with impact angle constraints," *IEEE Trans. Control Syst. Technol.*, vol. 14, no. 3, pp. 483–492, May 2006.

[17] B.-G. Park, T.-H. Kim, and M.-J. Tahk, "Range-to-go weighted optimal guidance with impact angle constraint and seeker's look angle limits," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 3, pp. 1241–1256, Jun. 2016.

[18] C.-H. Lee, M.-J. Tahk, and J.-I. Lee, "Generalized formulation of weighted optimal guidance laws with impact angle constraint," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 2, pp. 1317–1322, Apr. 2013.

[19] J.-I. Lee, I.-S. Jeon, and C.-H. Lee, "Command-shaping guidance law based on a Gaussian weighting function," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 1, pp. 772–777, Jan. 2014.

[20] N. Cho, Y. Kim, and H.-S. Shin, "Generalized formulation of linear nonquadratic weighted optimal error shaping guidance laws," *J. Guid., Control, Dyn.*, vol. 43, no. 6, pp. 1143–1153, Jun. 2020.

[21] G. Williams, A. Aldrich, and E. A. Theodorou, "Model predictive path integral control: From theory to parallel computation," *J. Guid., Control, Dyn.*, vol. 40, no. 2, pp. 344–357, Feb. 2017.

[22] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Stockholm, Sweden, May 2016, pp. 1433–1440.

[23] E. Arruda, M. J. Mathew, M. Kopicki, M. Mistry, M. Azad, and J. L. Wyatt, "Uncertainty averse pushing with model predictive path integral control," in *Proc. IEEE-RAS Int. Conf. Humanoid Robot.*, Jan. 2017, pp. 497–502.

[24] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to Learn*. Boston, MA, USA: Springer, 1998.

[25] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017, *arXiv:1703.03400*. [Online]. Available: http://arxiv.org/abs/1703.03400

[26] A. Nagabandi, I. Clavera, R. S. Fearing, P. Abbeel, S. Levine, and C. Finn, "Learning to adapt in dynamic, real-world environments through meta-reinforcement learning," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–9.

[27] J. Gao and Y.-L. Cai, "Three-dimensional impact angle constrained guidance laws with fixed-time convergence," *Asian J. Control*, vol. 19, no. 6, pp. 2240–2254, Nov. 2017.

[28] S. R. Kumar and D. Ghose, "Three-dimensional impact angle guidance with coupled engagement dynamics," *Proc. Inst. Mech. Eng., G, J. Aerosp. Eng.*, vol. 231, no. 4, pp. 621–641, Mar. 2017.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[30] D. Q. Mayne, E. C. Kerrigan, and P. Falugi, "Robust model predictive control: Advantages and disadvantages of tube-based methods," *IFAC Proc. Volumes*, vol. 44, no. 1, pp. 191–196, Jan. 2011.

[31] E. A. Theodorou and E. Todorov, "Relative entropy and free energy dualities: Connections to path integral and KL control," in *Proc. IEEE 51st IEEE Conf. Decision Control (CDC)*, Dec. 2012, pp. 1466–1473.

[32] Z. Yao, J. Yao, and W. Sun, "Adaptive RISE control of hydraulic systems with multilayer neural-networks," *IEEE Trans. Ind. Electron.*, vol. 66, no. 11, pp. 8638–8647, Nov. 2019.

[33] F. Liu, J. Sun, J. Si, W. Guo, and S. Mei, "A boundedness result for the direct heuristic dynamic programming," *Neural Netw.*, vol. 32, pp. 229–235, Aug. 2012.

[34] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-Learning," 2019, *arXiv:1901.00137*. [Online]. Available: http://arxiv.org/abs/1901.00137

[35] Y. P. Leong, M. B. Horowitz, and J. W. Burdick, "Linearly solvable stochastic control Lyapunov functions," *SIAM J. Control Optim.*, vol. 54, no. 6, pp. 3106–3125, Jan. 2016.

[36] R. Khasminskii, *Stochastic Stability of Differential Equations*, 2nd ed. Berlin, Germany: Springer-Verlag, 2011.

[37] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Cham, Switzerland: Springer, 2001.

[38] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Information-theoretic model predictive control: Theory and applications to autonomous driving," *IEEE Trans. Robot.*, vol. 34, no. 6, pp. 1603–1622, Dec. 2018.
[39] Kamalapurkar, Rushikesh, *Reinforcement Learning for Optimal Feedback Control*. Cham, Switzerland: Springer, 2018.
[40] E. Todorov and W. Li, "A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems," in *Proc. Amer. Control Conf.*, 2005, pp. 1–7.
[41] J. A. Paulson, T. A. N. Heirung, and A. Mesbah, "Fault-tolerant tube-based robust nonlinear model predictive control," in *Proc. Amer. Control Conf. (ACC)*, Philadelphia, PA, USA, Jul. 2019, pp. 1648–1654.

CHEN LIANG received the B.E. degree from the Harbin Institute of Technology, China, in 2012, and the M.S. degree in electrical and computer engineering from Rutgers, The State University of New Jersey, New Brunswick, NJ, USA, in 2014. He is currently pursuing the Ph.D. degree in navigation, guidance, and control with Beihang University, China. His research interests include nonlinear control, machine vision, deep learning, and guidance.

WEIHONG WANG received the B.E., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1990, 1993, and 1996, respectively. She is currently a Professor with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. Her research interests include computer simulation, computer control and simulation, guidance, intelligent control, servo control, and simulation technique.

ZHENGHUA LIU received the B.E. and M.S. degrees from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Beihang University, Beijing, China, in 2004. He is currently an Associate Professor with the School of Automation Science and Electrical Engineering, Beihang University. His research interests include high-precision servo control, robotic systems, and flight control.

CHAO LAI was born in Shandong, China, in 1990. He received the B.E. degree from Dalian Maritime University and the Ph.D. degree from Beihang University, Beijing, China, in 2019. He is currently with the Navigation and Control Technology Research Institute, China North Industries Group Corporation (NORINCO). His current research interests include nonlinear control, servo control, adaptive control, and integrated guidance and control.

SEN WANG received the B.E. degree from Xi'an Jiaotong University, China, in 2011, and the Ph.D. degree from Beihang University, China, in 2017. He is currently involved in research on missile guidance and control with the Navigation and Control Technology Research Institute, NORINCO Group, China.