Robust Reachability within Deep Reinforcement Learning Framework

Satya Marthi*, Mayank S. Jha, Didier Theilliol, Jean-Christophe Ponsart *

* *Universite de Lorraine, CNRS, CRAN, F-54000 Nancy, France. (e-mail: {satya.marthi, mayank-shekhar.jha, didier.theilliol, Jean-christophe.Ponsart}@univ-lorraine.fr)*

Abstract: We propose a novel deep reinforcement learning based approach to solve Hamilton–Jacobi reachability (HJ-R) problem for nonlinear control-affine systems with external disturbance. By recasting reachability as an optimal control problem, we solve it within a Deep Deterministic Policy Gradient (DDPG) framework: the critic learns the HJ-R value function and the actor synthesizes the optimal policy. We propose a Telescopic Incentive Reward Function that improves learning efficiency, promotes finite-time convergence to the target set, and reduces control oscillations near constraint boundaries. Disturbance is incorporated through agent-environment interaction, enabling robust optimal policy learning without an explicit disturbance model. The proposed approach fares well against classical grid-based dynamic programming approach and mitigates the curse of dimensionality through deep neural approximation, yielding scalability to higher-dimensional states. Numerical studies demonstrate target reach across diverse initial conditions, smooth control inputs relative to dynamic programming baselines, and resilience to worst-case disturbances. These results establish the proposed Robust Reachability-DDPG framework as an efficient, scalable, and robust alternative for HJ–R controller synthesis in continuous state–action spaces. The efficacy of the approach is assessed in simulation.

Keywords: Reinforcement Learning, Hamilton-Jacobi Reachability, Deep Deterministic Policy Gradient, Deep Neural Networks

1 INTRODUCTION

The Hamilton-Jacobi reachability (HJ-R) framework has become an important research field in the context of designing safe controllers for autonomous systems particularly for safety-critical systems (Mitchell et al., 2005; Bansal et al., 2017). The emerging research in HJ-R domain targets providing a holistic framework to solve the class of reach-avoid problems (Margellos and Lygeros, 2011). Typically, a reach-avoid problem (also liveness-safety problem) is formulated as one where the controller tries to reach the target in presence of an external disturbance, where the disturbance often acts as an adversary and opposes the controller action (Bansal et al., 2017). Historically, various approaches have been proposed to solve the HJ-R problem starting with the use of dynamic programming (DP) based approaches (Chen and Tomlin, 2018) where the HJ-R is solved as an optimal control problem. Subsequently, the field of differential games (Bansal et al., 2017) that provides a framework to solve the HJ reachability problem by treating them as a two-player (zero-sum) game by assigning the controller as a *reach player* and disturbance as *avoid player*, which are in conflict with one another (Bansal et al., 2017). It must be noted that the success of HJ-R framework is its capability to provide formal verification to the nonlinear systems with state and control constraints under adversarial disturbances (Chen and Tomlin, 2018). Another notable benefit of this framework is that the controller is designed for the worst-case scenario by considering all possible control inputs and trajectories. Hence, the recomputation of the safety-controller is avoided for any change in state or disturbance (Borquez et al., 2024). However, it must be noted that the optimal control policy (controller) synthesized by solving the reachability problem does not provide any performance guarantees (Borquez et al., 2024). Thus, there is a need to develop approaches that solve HJ-R problem whilst guaranteeing system performance.

Further, in the context of HJ-R problem, the traditional approaches typically solve the *backward reachable tube* (BRT), also called the *capture basin* (Mitchell et al., 2005; Aubin et al., 2011). Generally, the BRT for a dynamical system gives the set of initial states from which the system will eventually reach the target set despite the external (worst-case) disturbance (Mitchell et al., 2005; Aubin, 2001). The solution to the BRT can be obtained for any dynamical system by obtaining the viscosity solution to the Hamilton–Jacobi–Isaacs Variational Inequality (HJI-VI), which results in enforcing the reach and/or avoid sets as terminal constraints to the Hamilton–Jacobi–Bellman (HJB) partial differential equation (PDE) (Fisac et al., 2015; Bansal et al., 2017).

The HJI-VI solves a value function with minimal control effort under worst-case disturbance while minimizing the distance to the target set. The knowledge of the value function encodes the information of the set of initial states within the state-space through which a safe controller can be synthesized (Bansal et al., 2017). It is interesting to note that the safe controller computation is heavily reliant on the accuracy of the value function (Borquez et al., 2024).

However, most of the above mentioned approaches suffer with *curse of dimensionality*, notably the DP based approaches where the solutions also depend on the grid resolution (Chen and Tomlin, 2018). To alleviate problems arising from the curse of dimensionality and scale the HJ-R solutions to higher dimensional systems, recently, reinforcement learning (RL) based solutions have been proposed that utilize the power of neural networks (NN) to

⋆ This work is supported and funded by ANR (Agence nationale de la recherche) under the Projet ANR SOS (Self-organizing, smart and safe robots). ⋆Satya Marthi is the corresponding author.

act as universal function approximators (So et al., 2024). The review article (Bansal et al., 2017) and the works in (Ganai, 2024) provide a detailed summary on the use of RL techniques to solve the HJ-R problem and propose a new methodology called *Hamilton-Jacobi reachability estimation*. However these works consider the system dynamics without external disturbance. In the works (Kokolakis et al., 2023) the HJ-R problem is cast as a Mayer optimal control problem which provides convergence guarantees does not consider external disturbance. However, (Bansal and Tomlin, 2021) proposed deep neural networks (DNN) to approximate the HJI-VI PDE and demonstrated the scalability of HJ-R to higher dimensional systems while also considering the external disturbance. Therein, the authors have noted that in presence of disturbance, the training time is very high (16 hours for a *Air-3D*) and that the latter along with memory requirement does not scale with the dimension of the state-space, but rather depends on the signal complexity. Additionally, the work (Fisac et al., 2018) solve the safe reinforcement learning problem using HJ-R framework to synthesize safe controllers but do not treat the robustness in a formal manner.
In the light of such developments, we note the need for the development of a comprehensive approach to solve the HJ-R problem in presence of external disturbance, whilst synthesizing (learning) the optimal controller under worst-case scenario. To address the existing scientific gaps, this paper proposes a novel methodology by leveraging the deep deterministic policy gradient (DDPG) framework Lillicrap et al. (2015) that is robust against external disturbance. The contributions of this paper are as follows:

- a novel DDPG based robust approach for solving the HJ-R problem under worst-case disturbance, termed as Robust Reach DDPG (RR-DDPG) approach;
- a novel Telescopic Incentive Reward Function (TIRF) that is proposed to solve the robust reach problem, whilst learning robust optimal controller (optimal policy/law).

Section-II introduces the problem statement, section-III presents the formalisms necessary to solve HJ-R using DDPG. Section-IV presents a DDPG-based algorithm to solve the robust-reach problem. Section-V discusses the simulation results to demonstrate the efficacy of the proposed approach, while section-VI details the training and simulation parameters, followed by the conclusions of the work.

Notations The Euclidean norm is denoted by $||\cdot||$. The Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$ is denoted by $||A||_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *Lipschitz continuous* on a compact set $\mathcal{X} \subset \mathbb{R}^n$ if there exists a constant $L \in \mathbb{R}^+$ (set of positive real numbers) such that $||f(x_2) - f(x_1)|| \leq L||x_2 - x_1||, \forall x_2, x_1 \in \mathcal{X}$.

2 Background and Problem Formulation

2.1 System dynamics

Consider the following continuous-time nonlinear dynamics affine in control input and disturbance

$$\dot{x} = f(x) + g_u(x)u + g_d(x)d, \quad x(t_0) = x_0 \quad t \geq t_0 \tag{1}$$

where, for every $t \geq t_0$; $x(t) \in \mathscr{X} \subset \mathbb{R}^n$ is the state vector, $u(t) \in \mathscr{U} \subset \mathbb{R}^m$ is the control input vector and $d(t) \in \mathscr{D} \subset \mathbb{R}^p$ is the disturbance to the system. The disturbance considered is bounded within $||d(t)||_2 \in \rho$. We assume that the control input and disturbance are continuous and bounded. The drift dynamics $f(x) \in \mathbb{R}^n$, control input $g_u(x) \in \mathbb{R}^{n\times m}$ and disturbance $g_d(x) \in \mathbb{R}^{n\times p}$ are Lipschitz continuous. We consider that $g_u(x)$ and $g_d(x)$ satisfy $||g_u(x)||_{\mathrm{F}} \leq g_U$ and $||g_d(x)||_{\mathrm{F}} \leq g_D$. We also assume that the matrices $g_u(x)$ and $g_d(x)$ are invertible such that $g_u^{-1}(x)g_u(x) = \mathbb{I} \in \mathbb{R}^{m\times m}$. The system trajectory is denoted by $\eta_{x,t}^{u,d}(t)$ corresponding to (1) for a given $u(\cdot)$, $d(\cdot)$ signals at time t.
We assume that for a given initial condition, there exists a unique and continuous trajectory $\eta_{x,t}^{u,d}$ given input u and disturbance signals are piecewise continuous with time (Borquez et al., 2024). We also assume that there exists a set of admissible controls ($\mathscr{U}$) for (1). The safe controllers synthesized are $\mathscr{U}_s \subset \mathscr{U}$. The origin is assumed to be the equilibrium point and the system is fully *controllable* and fully *observable* and the system dynamics are completely known.

Definition 1. Target set (Evans, 2010)
A set $\mathscr{T} \subset \mathscr{X}$ is called a user-defined target set for the sub-zero level set of a continuous differentiable, bounded function $\ell(x)$ given by:

$$\mathscr{T} := \{x(t)|\ell(x) \leq 0\}. \tag{2}$$

Definition 2. Reach set (Evans, 2010)
A set $\mathscr{R} \subset \mathscr{X}$ is a reach set of a target set of the system (1) for a given initial condition x_0 and a time t_0 such that:

$$\mathscr{R}(t_0; t, \mathscr{T}) = \left\{ x_0 \middle| \max_{d(\cdot)\in\mathscr{D}} \min_{u(\cdot)\in\mathscr{U}} \ell(\eta(t; t_0, x_0, u, d)) \leq 0 \right\}. \tag{3}$$

Definition 3. Admissible control policy (Kokolakis et al., 2023)
A control policy $\pi(x) : \mathbb{R}^n \to \mathscr{U}$ is said to be admissible with respect to the target set $\mathscr{T}$ if the closed-loop trajectories for an initial state are unique, well-posed and bounded while satisfying the input constraints.

2.2 Reach problem using Hamilton-Jacobi Reachability

The primary objective of this paper is to synthesize controllers $\pi(x) : [0, T]\ \mathscr{X} \to \mathscr{U}$ that ensure that the system reaches a desired set of states $\mathscr{T}$ (see definition 1) in finite-time conditions (Fisac et al., 2015). Here the set $\mathscr{T}$ could represent the goal of the system (for example, the hovering destination of a drone in fire fighting situations). The objective of the system is to reach the target set in finite time under external disturbance and control input saturation (Borquez et al., 2024).

In the traditional HJ-R framework, the reach problem is cast as an optimal control problem. The optimization is performed to minimize the distance between the current system state and the target set over the finite horizon. For the reach case (also called the liveness problem) the optimization is defined as:

$$\max_{d(\cdot)} \min_{u(\cdot)} \min_{\tau\in[t,T]} \ell(\eta_{x,t}^{u,d}), \tag{4}$$

$$\text{s.t.} \quad \dot{x} = f(x) + g_u(x)u + g_d(x)d, \tag{5}$$

where the minimum cost over time is defined by,

$$J(x, t, u(\cdot), d(\cdot)) = \min_{\tau\in[t,T]} \ell(\eta_{x,t}^{u,d}). \tag{6}$$

Remark 2.1. The cost function computes the minimal distance along the entire trajectory $\eta_{x,t}^{u,d}$ to the target set given by the sub-zero level set of the function $\ell(x)$.

The minimum distance under the optimal control and external disturbance is considered as a value function given by:

$$V(x,t) = \max_{d(\cdot)} \min_{u(\cdot)} J(x,t,u(\cdot),d(\cdot)). \tag{7}$$

The viscosity solution of the Hamilton-Jacobi-Isaacs Variational inequality (HJI-VI) is the value function (7) (Borquez et al., 2024).

$$\max\left\{\frac{\partial V(x,t)}{\partial t} + H\left(x,t,\nabla V(x,t)\right), \ell(x) - V(x,t)\right\} = 0 \quad \text{for } t \in [0,T] \text{ and } V(x,T) = \ell(x). \tag{8}$$

The value function obtained by solving (8) can be used to obtain the Backward Reachable Set (BRS) for the system (1). BRS gives the initial set of states from which despite the worst-case disturbance there exists a control input such that the system can reach the target in time $(t-T)$. BRS is computed from the sub-zero level set of the value function solution given by (Borquez et al., 2024):

$$\mathscr{B}(t) = \{x : V(x,t) \leq 0\}. \tag{9}$$

Thus, computing the BRT by taking the sub-zero level set of the value function (7), gives the set of all possible initial states subject to various input and disturbance signals to drive the system to the target. The HJ reachability framework also enables the computation of optimal controllers and optimal disturbance that solve the reach problem (8). For a given state $x(t)$ the optimal control policy can be computed by:

$$\pi^\star(x,t) = \arg\min_{u\in\mathscr{U}} \max_{d\in\mathscr{D}} \nabla V(x,t)\left(f(x) + g_u(x)u + g_d(x)d\right), \tag{10}$$

and the optimal disturbance can be obtained using:

$$d^\star(x,t) = \min_{u\in\mathscr{U}} \arg\max_{d\in\mathscr{D}} \nabla V(x,t)\left(f(x) + g_u(x)u + g_d(x)d\right). \tag{11}$$

In essence, the optimal control policy steers the system towards the target whereas the optimal disturbance drives the system away from the target. The work (Borquez et al., 2024) presents a comprehensive discussion on the controller selection to solve the reach problem. This outcome follows the solution of (10), which does not consider any performance related constraints. In addition, several other studies (Bui et al., 2021) provide numerous approaches to solve both the reach problem and avoid problem and analyze the computational requirements to solve the reach-avoid problem (**?**). However, most of the works simplify the problem by neglecting the disturbance to the system. In contrast, this paper proposes a novel approach to address the reach problem while explicitly considering the disturbance with lower computational cost by leveraging the DDPG framework.

2.3 Deep Deterministic Policy Gradient (DDPG)

Introduced in (Lillicrap et al., 2015), DDPG is a robust off-policy reinforcement learning algorithm that solves the optimal control problems with continuous state-space. The DDPG framework uses a Markov decision process (MDP). The MDP is characterized by the tuple $\{\mathscr{X}, \mathscr{A}, \mathscr{R}, \mathscr{D}, \gamma\}$ where $\mathscr{X}$ is the state-space, $\mathscr{A}$ being the action-space and $\mathscr{R}$ as reward with the assumption that the rewards are positive $r \geq 0$ and $\mathscr{D}$ is the disturbance and thus implicitly includes the transition matrix (system dynamics) with $0 \leq \gamma \leq 1$ being the discount factor. DDPG adopts the actor-critic structure to solve the control problems by using the Q-value function also called the *state-action value function* $Q(x,u)$ using the recursive Bellman equation for a finite horizon problem is given by:

$$Q^\pi(x_k,u_k) = \mathbb{E}_{\tau\sim\pi,\mathscr{D}}\left[\sum_{k=0}^{T} \gamma^k r(x_k,u_k)\right]. \tag{12}$$

where $r(x,u)$ is the reward function. The optimal control policy can be computed by

$$u^\star(x) = \arg\min_{u(\cdot)} Q(x_k,u_k). \tag{13}$$

In DDPG approaches, the learning is driven by the choice of reward function which must be engineered with precision for successful completion of a task. The algorithm also employs the use of target networks and replay buffer to stabilize the learning under the nonexistence of convergence guarantees Lillicrap et al. (2015).

2.4 Problem Statement

For the system (1) with a given terminal cost condition $\phi(x)$, the value function to solve the finite horizon optimal control problem becomes:

$$V(x,t) = \min_{u(\cdot)} \max_{d(\cdot)} \int_t^T L(x(s),u(s,d(s)))\, ds + \phi(x(T)). \tag{14}$$

The HJI takes the form:

$$\frac{\partial V(x,t)}{\partial t} + \min_{u\in\mathscr{U}} \max_{d\in\mathscr{D}} \{\nabla V(x,t)^\top \left(f(x) + g_u(x)u + g_d(x)d\right) + L(x,u,d)\} = 0. \tag{15}$$

where $L(\cdot)$ is the running cost and the target set (see definition 1) becomes the terminal cost condition. Then, the goal is to employ the DDPG algorithm to approximate the value function (15) using the critic network and the corresponding optimal feedback policy (optimal control law) (10) using the actor network that is robust against disturbance. To this end, we propose a novel reward signal that incorporates worst-case disturbance considered through agent-environment interactions, leading to a novel DDPG-based robust-reachability algorithm.

3 Solving the Value function using DDPG

To solve the HJ-R optimal control problem using DDPG, it is necessary to define following key properties of the DDPG learning framework like, the environment, the reward function and the loss function that drives the learning procedure under finite time condition.

3.1 Designing the reward function

For the system (1), the signed distance function to the target set $\mathscr{T}$ is minimized with the terminal cost condition $\ell(x)$ which is encoded using the value function. Typically, the Q-value function computes the cumulative reward over a finite horizon. However, in our work, the optimality must be achieved over a finite time horizon. To achieve finite-time solutions to (15), in this paper, the following reward function is proposed:

$$r(x_k) = J_k - J_{k+1} \tag{16}$$

where J_k is the cost (6) computed at time-step k. The reward function (16) is now introduced in value function (12) with $\gamma = 1$ to give:

$$V(x_k) = \sum_{k=0}^{T-1} r(x_k) = \sum_{k=0}^{T-1} (J_k - J_{k+1}) = J_0 - J_T. \tag{17}$$

where J_0 is the cost at initial condition (x_0) and J_T is the terminal, finite horizon cost function. Given the finite-time nature of the value function, the reward function (16) becomes a telescopic reward function. The telescopic nature

of the reward function becomes crucial to solve the HJ-R problem using DDPG.

Lemma 1. For the system (1), the telescoping reward function (16) is considered within the value function (12) solves the HJ-R optimal control problem.

Proof. Consider a function $c(\cdot) : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ that computes the instantaneous distance to the target set at each time-step k which is also the instantaneous value of the cost function (6):

$$c(x_k) = \max\{0, \|x_k - x_T\| - r_T\} \geq 0 \tag{18}$$

where x_T is the center of the target set and r_T represents the size of the target set. Using the function $c(\cdot)$ the telescoping reward to compute the minimum distance is given by:

$$J_{k+1} = \min\{J_k, c(x_{k+1})\} \quad (k = 0, \dots T-1). \tag{19}$$

Now, the undiscounted (where $\gamma = 1$) cumulative reward function over a fixed time horizon T is:

$$\sum_{k=0}^{T-1} r(x_k) = \sum_{k=0}^{T-1} (J_k - J_{k+1}) = J_0 - J_T, \tag{20}$$

here, $J_0 = c(x_0)$ and J_T enforces the terminal condition. Therefore, under the condition that J_0 does not depend on the policy $\pi(x)$, (20) gives the same result as solving (6) thus concluding the proof. □

Using the aforementioned formulations, the telescopic reward function becomes:

$$r_t(x) = \begin{cases} \mathbf{w}, & \text{if } \ell(x_{k+1}) < J_k, \\ 0, & \text{otherwise,} \end{cases} \tag{21}$$

where $\mathbf{w}$ is a positive constant (hyper-parameter). At each time step, the telescopic reward function (21) is nonzero when the system trajectory moves close to the target but necessarily zero otherwise. Thus making the reward structure sparse. Learning algorithms like DDPG with sparse reward structures require high exploration noise or sometimes lead to suboptimal learning. As such, there is a need to mitigate the problem arising from the sparsity, whilst preserving the telescopic nature of the reward function (16) we propose an *Incentive function* that produces nonzero reward at each time-step.

Incentive Function (IF) To continuously guide the learning with dense reward, an Incentive function is added to the telescopic reward. The proposed *Incentive Function* (IF) is positive within the super-level set and negative within the sub-level set of the target set (2). Thus, we propose the following Incentive Function given by:

$$i(x) = \exp(\alpha\ x) - 1, \tag{22}$$

where α is a positive hyper-parameter. The Incentive function is an continuous, differentiable $\mathcal{C}^\infty$ and strictly convex function that is locally Lipschitz continuous and a strictly convex function. Incorporating IF within the telescopic reward function (16) offers two advantages. First, the function returns negative rewards when the agent is outside the target and returns positive rewards within the target set, thereby encouraging the actions towards the target set. Second, the IF also allows flexible adjustment of the function according to the boundaries of the target set. Leveraging the aforementioned incentive function, a novel *incentive reward function* (IRF) is proposed to 'incentivize' the agent's actions towards the target set. To this end, the instantaneous cost function (18) is reformulated into the following equation:

$$\delta(x) = \|x - x_T\| - r_T. \tag{23}$$

Now, the function (23), when introduced in IF (22) gives the IRF:

$$i(x) = \exp(\alpha\ \delta(x)) - 1, \tag{24}$$

Augmenting the proposed IRF (24) with the telescoping reward (21), the novel reward structure proposed for this study is called the *telescopic-incentive reward function* (TIRF):

$$r(x_k) := \begin{cases} \mathbf{w} + \mathbf{w_C}\ i(x_k), & \ell(x_{k+1}) < C_k, \\ \mathbf{w_C}\ i(x_k), & \text{otherwise,} \end{cases} \tag{25}$$

with $\mathbf{w_C}$ is a positive hyperparameter to dampen the effect of dense reward from the exponential function.

Remark 3.1. To learn the HJ-R value function using the DDPG learning framework, the use of telescoping reward function becomes critical to preserve the Markovian property that is critical in the learning paradigm. This originates from the original HJ-R cost function that computes the minimum distance over the entire trajectory which introduces a dependency over all the previous system states.

3.2 Environment

In order to leverage DDPG for learning optimal control policy, we consider the state space formulation of the system in (1) through a MDP that evolves (the transition is given by $\mathscr{A}, \mathscr{X}, \mathscr{D}$) and consider the discretized (in-time) equations in the following way:

$$x_{k+1} = f(x_k) + g_u(x_k)u_k + g_d(x_k)d_k, \quad \|d_k\|_2 \leq \rho. \tag{26}$$

Proposition 3.1. For the system (26), using the Definition 3, the functions $f(x)$, $g_u(x)$ and $g_d(x)$ are bounded and the instantaneous cost function (18) is convex in $\mathscr{D}$.

Proof. The control input and disturbance matrices are bounded under the Frobenius norm $\|g_u(x)\|_{\mathrm{F}} < g_U$ and $\|g_d(x)\|_{\mathrm{F}} < g_d$. Given the existence of set of admissible control policies to the system (26), the system dynamics are bounded. The reader is advised to see Lemma-1 in (Huang and Liu, 2014) for detailed proof. Thus, for the system (26) for a given state and control policy (x_r, u_r) the dynamics can be written as follows:

$$x(d) = \underbrace{f(x_r) + g_u(x_r)u_r}_{\text{bounded function}(F(x,u))} + \underbrace{g_d(x_r)}_{(G(x))}\, d. \tag{27}$$

The (27) can be reduced to into a dynamics that is affine in disturbance given by:

$$x(d) = F(x, u) + G(x)\ d. \tag{28}$$

The cost function according to the new affine dynamics (28) becomes:

$$C(d) = c(x(d)) \tag{29}$$

where $c(\cdot)$ is the cost function (18). This renders the $C(d)$ a convex function on $\mathscr{D}$. The instantaneous cost function (22) computes the unsigned distance (Hinge version) which gives a convex function. Thus, the system dynamics are bounded under the stated assumptions and cost function is convex which concludes the proof. □

Lemma 2. For the system (26), for a fixed $(x_k, u_k) \in \Omega$ the worst-case disturbance to the system is attained only at the boundary of the disturbance.

Proof. For a fixed $(x_k, u_k) \in \Omega$ to the system (28) the optimal disturbance (11) can be rewritten by considering the Lipschitz property as follows :

$$\max_{d_k \in \mathscr{D}} C(F(x,u)+G(x)d) \leq C(F(x,u)) + L_c \max_{d_k \in \mathscr{D}} ||G(x)d||_2. \tag{30}$$

where L_c is the Lipschitz constant for the function (18). The bounded disturbance to the system (26) can be considered as:

$$\mathscr{D} := \{d_k \in \mathbb{R}^m : \|d\|_2 \leq \rho\}. \tag{31}$$

Then we have,

$$\max_{d_k \in \mathscr{D}} ||G(x)d||_2 = \rho ||G(x)||_2 \tag{32}$$

Thus,

$$\max_{d_k \in \mathscr{D}} C(F(x,u) + G(x)d) \leq C(F(x,u)) + L_c \rho ||G(x)||_2 \tag{33}$$

As we consider, the Euclidean norm, the constant $L_c = 1$ (Boyd and Vandenberghe, 2004). Therefore, from (33) and proposition 3.1 we prove that the worst-case disturbance to the system is only attained at the boundary of the disturbance set. □

Using lemma 2, the computational resources required to solve the optimal disturbance (11) problem is reduced. This is due to the fact that the disturbance is not computed over the compact interval $d_k(t)$ rather over the endpoint.

3.3 Finite-horizon computation

To solve the reach problem within a finite-horizon we employ two methodologies, one is by the use of telescopic reward (20) within the Q-value function and the other is by imposing the final value function ($V_T(x) = 0$) by fixing the episodic time for each epoch as T.

Lemma 3. For the system (26), using lemma 1 and 2 with telescoping reward and environment with disturbance for a fixed episode length T during the implementation, enforces the finite-time condition through the terminal value function condition that solves the HJ-R finite horizon optimal control reach problem.

Proof. Using the cumulative minimum distance reward function

$$V_k^{DDPG} = \min_{\pi} \max_{\mathscr{D}} \sum_{k=0}^{T-1} r(x_k) = J_0 - J_T. \tag{34}$$

as $c(x_0)$ is fixed and equal to the episodic time during the learning procedure. Thus, the Bellman equation obtained becomes:

$$V_k^{\text{DDPG}}(x, J) = r_k + V_{k+1}^{\text{DDPG}}(x_{k+1}, J_{k+1}), \tag{35}$$

with terminal condition $V_T^{DDPG}(x, J) = 0$ which concludes the proof. □

The solution to HJ-R value function (15) is therefore equivalent to solution obtained by solving the (35). For the rest of the paper the $V_k^{\text{DDPG}}(x, t)$ is represented by $Q^{\pi}(x_k, u_k)$.

3.4 Loss function

The value function (35) along with the control policy are approximated using a deep neural networks (DNNs). The target equation is given by:

$$Q^{\pi}(x_k, u_k) = \mathbb{E}_{\mathcal{D}}[r(x_k) + Q^{\pi}(x_{k+1}, u_{k+1})]. \tag{36}$$

In the DDPG algorithm, the critic network updates the (36) to solve the Bellman equation and the actor network approximates $u_k = \pi(x_k)$ that maximizes $Q^{\pi}(x_k, u_k)$.

To this end, the critic loss function using the temporal difference (TD) target is given by:

$$\mathcal{L}_{\text{critic}}(\phi) = \mathbb{E}_{\mathcal{D}}[(y - Q_{\phi}(x_k, u_k))^2], \tag{37}$$

with the Bellman target y is:

$$y = r(x_k) + \gamma Q_{\phi}(x_{k+1}, u_{k+1}), \tag{38}$$

where, the critic network is parameterized by ϕ.

Remark 3.2. We assume that the data collected using the replay buffer $\mathcal{B}$ is sufficiently rich to represent the value function and the optimal control policy. With this assumption, after T time steps for each episode the trained critic and actor networks are stable along with the target networks (Lillicrap et al., 2015).

Similarly, the actor loss function which is parameterized by θ is updated using the deterministic policy gradient (DPG) that becomes:

$$\mathcal{L}_{\text{actor}}(\theta) = \mathbb{E}_{\mathcal{D}}[Q_{\phi}(x, \pi_{\theta}(x))]. \tag{39}$$

The episodes terminate after T time steps or if the target is reached thus solving the reach problem in finite-horizon. The earlier-defined components like the environment, the reward function and the loss function are introduced within the DDPG framework for the finite-time, robust computation of optimal control policy that solves the reachability problem. Therefore, using all these formulations the following theorem can be established.

Theorem 1. For the control and disturbance affine system (1), the HJ reachability problem, which is formulated as an optimal control problem can be solved using the reward function (25) within the DDPG framework.

Proof. By combining the proofs of Lemmas 1, 2, and 3 with the loss functions of the actor and critic networks, the necessary conditions are established to approximately solve the HJ-R optimal control problem within the DDPG framework. However, similar to the standard DDPG algorithm, this formulation does not guarantee algorithmic convergence. Nonetheless, the training stability of the proposed approach can be demonstrated in the sense of (Lillicrap et al., 2015). □

4 Robust-reach algorithm

Following theorem 1, we propose the robust reach DDPG (RR-DDPG) algorithm that leverages the DDPG framework to solve the HJ-R optimal control problem (8) under worst-case disturbance. The proposed RR-DDPG algorithm adopts the architecture presented in (Lillicrap et al., 2015) while incorporating the proposed TIRF (25) and by interacting with the environment with disturbance. The resulting control policy obtained upon successful training ensures robustness and can reach the target within finite time horizon T. Although the controller synthesized is conservative in nature, such behavior remains acceptable while navigating an environment with external disturbance.

5 Simulation results

We study the efficacy of the proposed approach in simulation by considering an inverted pendulum system with external disturbance:

$$\begin{bmatrix} \dot{x_1} \\ \dot{x_2} \end{bmatrix} = \begin{bmatrix} x_2 \\ -\sin x_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0 \\ \cos x_1 \end{bmatrix} d, \tag{40}$$

where x_1 and x_2 are the angle (θ) and angular velocity (ω) of the pendulum respectively, u is the input torque applied and d is the external disturbance to the system. The target set considered is a circular zone with the origin as the center of the target set. The target set is defined by considering the sub-level set of the following function:

$$\ell(x) = x_1^2 + x_2^2 - 0.25.$$

The TIRF (25) is considered with $\mathbf{w} = 3.5$, $\mathbf{w_C} = 0.3$ with $\alpha = 0.45$ in the incentive reward function (24). Each training episode consists of simulation over 12s, which, under sampling period of 0.01s, translates to 1200 simulation steps. We train the system by running the DDPG over 250 episodes. The actor learning rate was set to $3e-4$, critic network is set to $2e-3$ with the discount factor 1. A L2 regularization was performed on both actor and critic with a factor of $3e-3$ and $2e-3$ respectively. The barrier-

like reward function $\alpha = 0.45$. After training, the critic network approximates the optimal value function (15) and the actor approximates the optimal control policy (10). The actor thus serves as a near-optimal feedback controller robust to worst-case disturbance. Figure 1 illustrates closed-loop trajectories from multiple initial conditions under the trained controller. Simulations of the system dynamics show that trajectories reach the target-set boundary within the finite horizon as intended. Figure 2

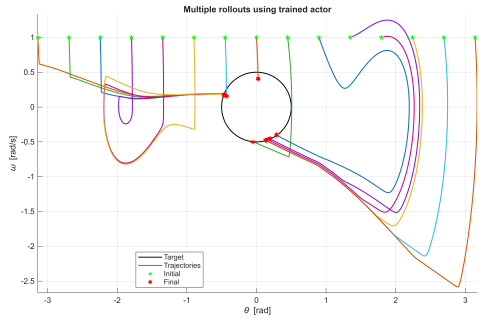

Fig. 1. System trajectory with multiple initial conditions

plots the state and control evolution under the trained controller alongside a grid-based dynamic programming baseline computed with the HelperOC toolbox (Chen and Herbert, 2019). The state trajectories align closely across the horizon, confirming consistency with the classical solution. In contrast, the DP control exhibits pronounced chattering effect, a well-known artifact of grid-based HJ/DP solvers that tend to produce bang–bang inputs in the absence of input regularization or control-effort penalties. By design, the proposed Telescopic Incentive Reward Function yields smooth, well-conditioned control while preserving reach performance. Finally, it is noted that trajectories generated by the proposed method terminate at the target-set boundary rather than at its center (the origin). This reflects the DDPG objective of boundary hitting within a finite horizon, whereas the HelperOC reference drives the state to the target center, explaining minor differences near the goal.

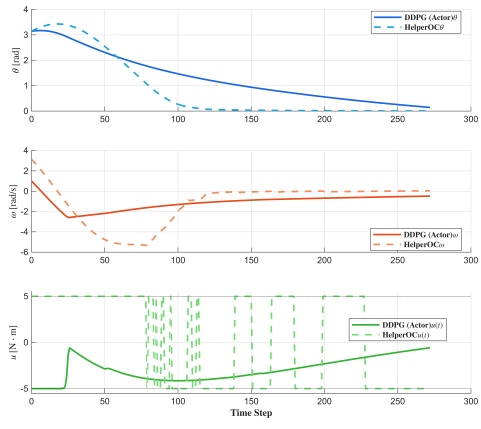

Fig. 2. System trajectory and control input

6 Conclusion

The work presents a novel approach to solve the robust reach problem using deep deterministic policy gradient (DDPG) algorithm for a nonlinear systems under external disturbance. We present a comprehensive approach by casting the Hamilton Jacobi -Reach (HJ-R) problem within DDPG framework learning both HJ-R value function and optimal control policy through critic and actor networks respectively. Simulations show that the proposed RR DDPG solves the reach task across diverse initial conditions, with trajectories reaching the target. Compared with the classical dynamic programming based approach (in HelperOC toolbox), our approach mitigates the undesirable chattering behavior and gives a smooth control input. These gains arise from the proposed telescopic incentive reward function, which drives the agent towards the target within a finite time horizon while regularizing the control effort. We also reduce computation by modeling the disturbance through agent-environment interaction. Because value and policy are represented by deep neural networks, the method scales to higher dimensional systems. In summary, RR DDPG provides an effective solution to HJ reachability for nonlinear systems with external disturbance under finite time conditions. Future work will address reach avoid and stay problems.

References

Aubin, J.P. (2001). Viability Kernels and Capture Basins of Sets under Differential Inclusions. *SIAM Journal on Control and Optimization*.

Aubin, J.P., Bayen, A.M., and Saint-Pierre, P. (2011). *Viability Theory: New Directions*. Springer.

Bansal, S., Chen, M., Herbert, S.L., and Tomlin, C.J. (2017). Hamilton–Jacobi Reachability: A Brief Overview and Recent Advances. In *Proceedings of the IEEE Conference on Decision and Control*.

Bansal, S. and Tomlin, C.J. (2021). DeepReach: A Deep Learning Approach to High-Dimensional Reachability. In *Proceedings of the IEEE International Conference on Robotics and Automation*.

Borquez, J., Chakraborty, K., Wang, H., and Bansal, S. (2024). On Safety and Liveness Filtering Using Hamilton–Jacobi Reachability Analysis. *IEEE Transactions on Robotics*.

Boyd, S.P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bui, M., Lu, M., Hojabr, R., and Chen, M. (2021). Real-Time Hamilton–Jacobi Reachability Analysis of Autonomous Systems with an FPGA. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Chen, M. and Herbert, S. (2019). HelperOC Library. `https://github.com/HJReachability/helperOC`.

Chen, M. and Tomlin, C.J. (2018). Hamilton–Jacobi Reachability: Recent Advances and Applications. *Annual Review of Control, Robotics, and Autonomous Systems*.

Evans, L.C. (2010). *An Introduction to Mathematical Optimal Control Theory*. Lecture Notes.

Fisac, J.F., Akametalu, A.K., Zeilinger, M.N., Kaynama, S., Gillula, J.H., and Tomlin, C.J. (2018). A General Safety Framework for Learning-Based Control. *IEEE Transactions on Automatic Control*.

Fisac, J.F., Chen, M., Akametalu, A.K., and Tomlin, C.J. (2015). Reach-Avoid Problems with Time-Varying Dynamics, Targets and Constraints. In *Proceedings of the IEEE Conference on Decision and Control*.

Ganai, M. (2024). *Hamilton–Jacobi Reachability Estimation in Reinforcement Learning*. Master's thesis, University of California, San Diego.

Huang, Y. and Liu, D. (2014). Neural-network-based optimal tracking control scheme for a class of unknown discrete-time nonlinear systems using iterative adp algorithm.

Kokolakis, N.M., Vamvoudakis, K.G., and Haddad, W. (2023). Reachability Analysis-Based Safety-Critical Control Using Online Fixed-Time Reinforcement Learning. In *Proceedings of the Learning for Dynamics and Control Conference*.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous Control with Deep Reinforcement Learning. *arXiv preprint arXiv:1509.02971*.

Margellos, K. and Lygeros, J. (2011). Hamilton–Jacobi Formulation for Reach-Avoid Differential Games. *IEEE Transactions on Automatic Control*.

Mitchell, I.M., Bayen, A.M., and Tomlin, C.J. (2005). A Time-Dependent Hamilton–Jacobi Formulation of Reachable Sets for Continuous Dynamic Games. *IEEE Transactions on Automatic Control*.

So, O., Ge, C., and Fan, C. (2024). Solving Minimum-Cost Reach-Avoid Using Reinforcement Learning. *Advances in Neural Information Processing Systems*.