

Barrier Functions Inspired Reward Shaping for Reinforcement Learning

Nilaksh^{1*}, Abhishek Ranjan^{2*}, Shreenabh Agrawal^{2*}, Aayush Jain¹, Pushpak Jagtap³, Shishir Kolathaya⁴

Abstract—Reinforcement Learning (RL) has progressed from simple control tasks to complex real-world challenges with large state spaces. While RL excels in these tasks, training time remains a limitation. Reward shaping is a popular solution, but existing methods often rely on value functions, which face scalability issues. This paper presents a novel safety-oriented reward-shaping framework inspired by barrier functions, offering simplicity and ease of implementation across various environments and tasks. To evaluate the effectiveness of the proposed reward formulations, we conduct simulation experiments on CartPole, Ant, and Humanoid environments, along with real-world deployment on the Unitree Go1 quadruped robot. Our results demonstrate that our method leads to 1.4-2.8 times faster convergence and as low as 50-60% actuation effort compared to the vanilla reward. In a sim-to-real experiment with the Go1 robot, we demonstrated better control and dynamics of the bot with our reward framework. We have open-sourced our code at https://github.com/Safe-RL-IISc/barrier_shaping.

I. INTRODUCTION

Reinforcement Learning (RL) has demonstrated substantial success in various domains, including gaming (e.g., Minecraft [1] and Atari [2]), language model optimization (e.g., Sparrow [3] and InstructGPT [4]), and robotics [5], [6]. However, all of these methods including RLHF [7], which addresses RL’s sample efficiency issues in other domains, can be costly when applied to robotics, including sim-to-real transfer scenarios. Despite having its limitations, reward shaping provides a simpler, more accessible and efficient alternative to address these issues even now. Well-crafted reward functions guide the agent’s behaviour towards desired outcomes, facilitating successful learning of the intended task. Previous works ([8], [9]) have established reward-shaping to accelerate algorithm convergence. Potential-based reward shaping [10] is a well-known work that suggests adding a potential function term initialized to the value function for reward-shaping. However, it needs a good estimate of the value function, which is challenging in the case of sparse reward models and in complex environments due to dimensionality [11], [12], [13].

With a view to improve upon existing works on reward shaping methods, we introduce Barrier Function (BF) inspired reward shaping, a simple, safety-focused framework

This work is supported in part by the Google Research Grant, the ART-PARK, and the SERB Grants SRG/2022/001807 and CRG/2021/008115.

¹Indian Institute of Technology (IIT), Kharagpur {nilaksh404, aayushjain}@kgpian.iitkgp.ac.in

²Indian Institute of Science (IISc), Bangalore {abhishekr, shreenabhm}@iisc.ac.in

³RBCCPS, IISc - pushpak@iisc.ac.in

⁴CSA & RBCCPS, IISc - shishirk@iisc.ac.in

*Equal Contribution

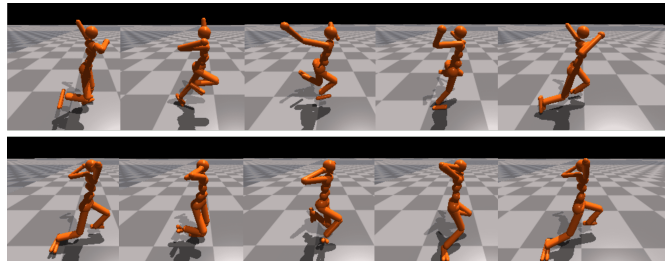


Fig. 1: An example run of Humanoid with vanilla reward (Top) and the exponential barrier reward (Bottom), trained for the same number of time steps. Unlike the vanilla reward, barrier function-based reward leads to more natural and less wasteful movements.

that enhances training efficiency and safety. Our approach uses BFs to supplement the base reward in an environment. Intuitively, when a BF is positive, the system’s state is within the safe region. Accordingly, we can construct an inequality by using the derivative of the barrier function, which is in turn encoded in the form of a reward. This reward-shaping term encourages the RL agent’s states to not only remain within the safe zone, but also ensure that undesirable behaviors are avoided at the limits. In this paper, we propose two BF-based reward formulations: the *exponential barrier* and the *quadratic barrier*. We assess our framework across various environments, from CartPole[14] to Humanoid[15]. Additionally, our approach improves the walking performance of agents, as evident in Fig 1. We also demonstrate sim-to-real transferability by applying our framework to the Unitree Go1 robot [16]. The key highlights of the proposed framework are:

- A safety-oriented, intuitive and easy-to-implement barrier function-inspired reward shaping framework.
- The framework leads to faster convergence towards the goal and efficient state exploration by enforcing the system within the safe set.
- It leads to lesser energy expenditure as the barrier function constrains the states within desired limits, thus avoiding extreme actions.

II. RELATED WORK

Reward Shaping : Positive linear transformation [17], [18], [19], leverages the principles of utility theory to enhance agent performance by adding rewards for state transitions as the difference between the values of the arbitrary potential functions applied to the respective states. In [20], a Bayesian approach to reward shaping is presented, which incorporates prior beliefs and adapts with experience. While effective, this method is model-based and may not generalize well across

different environments. Signal temporal logic, explored in [21] and [22], can serve as a formal specification for reward formulation but is also model-dependent. The safety and stability of the agent and environment remain unaddressed. To the best of our knowledge, for the first time, we explore the agent’s safety and stability using reward shaping that ensures faster convergence along with less energy consumption during training and testing.

It is worth mentioning that, in our formulation, we omit energy terms from the shaped reward as they often clash with the environment’s goals. For example, in CartPole, including an energy term would prioritize having the pendulum vertically down, which isn’t the intended objective. Adjusting energy terms requires extra tuning, complicating things for larger robot models. Similarly, alternative reward-shaping approaches like angle limits and velocity limits may lead to undesirable behavior, keeping the agent within bounds but failing to achieve the main objective. Intuitively, there is a relationship between the positions and velocities that must be respected at the limits. Violation of this can restrict exploration and learning, hindering effective policy development. The CBF based constraints encode this relationship, thereby ensuring that the position-velocity values are not in conflict.

Safety: [23] introduces a framework that combines model-free reinforcement learning with model-based CBF controllers and learned system dynamics to ensure safety during exploration. Since it is model-based, a good knowledge of the system dynamics is required. [24] proposes using reachability constraints to expand the feasible set, resulting in a less conservative policy compared to CBF-based approaches. [25] explores a more data-driven approach using two neural networks to learn the controller and safety barrier certificate simultaneously achieving a verification-in-the-loop synthesis, but requires formal verification (using SMT solvers) to ensure constraint satisfaction. Lastly, [26] discusses learning to restructure an MDP reward function for accelerated reinforcement learning, but it can be computationally intensive. Tan et al. [27] link CBF to a value function to enforce verifiable safety. Our approach aligns with the idea of leveraging model-based constraints for improved performance, but it doesn’t require learning system dynamics, making it more generalizable. We also allow for adjustable constraint levels (soft or hard) to prioritize returns or constraint satisfaction. Furthermore, we eliminate the need for formal verification, as our formulation inherently satisfies constraints.

III. PRELIMINARY

A. Reinforcement Learning

Reinforcement learning (RL) can be described as a discounted Markov decision process (MDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where \mathcal{S} is a set of states, \mathcal{A} is a set of actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the deterministic state transition function, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, and $\gamma \in (0,1)$ is the discount factor. In RL, the goal of the learning algorithm is to converge on a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the total (discounted) reward after performing actions on an MDP, i.e., the objective is

to maximize $\sum_0^\infty \gamma^t r_t$, where r_t is the output of the reward function r for the sample at instance t . Since a policy maps a state to an action, the value of a policy is evaluated according to the discounted cumulative reward. Given this MDP, we are interested in shaping the rewards by using barrier functions, described in the next section.

B. Barrier Functions

For practical applications, we often want the system state s to stay within a safe region, denoted as a set \mathcal{C} . The set \mathcal{C} is defined as the *super-level set* of a continuously differentiable function $h : \mathcal{S} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ satisfying,

$$\mathcal{C} = \{s \in \mathcal{S} : h(s) \geq 0\} \quad (1)$$

$$\partial\mathcal{C} = \{s \in \mathcal{S} : h(s) = 0\} \quad (2)$$

$$\text{Int}(\mathcal{C}) = \{s \in \mathcal{S} : h(s) > 0\}, \quad (3)$$

where $\text{Int}(\mathcal{C})$ and $\partial\mathcal{C}$ denote the interior and the boundary of the set \mathcal{C} , respectively. It is assumed that $\text{Int}(\mathcal{C})$ is non-empty and \mathcal{C} has no isolated points, i.e. $\text{Int}(\mathcal{C}) \neq \phi$ and $\text{Int}(\mathcal{C}) = \mathcal{C}$. The set \mathcal{C} is said to be forward invariant (safe) if $\forall s(0) \in \mathcal{C} \implies s(t) \in \mathcal{C} \quad \forall t \geq 0$. We can mathematically verify the safety of the set \mathcal{C} by establishing the existence of a barrier function. We have the following definition of a barrier function (BF) from [28].

Definition 3.1: Given the set \mathcal{C} defined by (1)-(3), the function h is called the barrier function (BF) defined on the set \mathcal{S} if there exists an extended class \mathcal{K} function κ such that for all $s \in \mathcal{S}$:

$$\dot{h}(s, \dot{s}) + \kappa(h(s)) \geq 0. \quad (4)$$

Here $\kappa : (-\infty, \infty) \rightarrow (-\infty, \infty)$ is a strictly increasing continuous function, with $\kappa(0) = 0$. κ is widely called an extended class \mathcal{K} function. Note that $\dot{h}(s, \dot{s}) := \frac{\partial h}{\partial s} \dot{s}$, where \dot{s} is the time derivative of s . Even if the MDP is in discrete time, \dot{s} , and consequently \dot{h} , can be calculated approximately as $\dot{h} \approx \frac{\partial h}{\partial s} (s_t - s_{t-1}) / \Delta t$, where s_t, s_{t-1} are the samples of the states obtained at time steps t and $t-1$, and Δt is the time interval between two samples. It is worth mentioning that the classical definition has the actions or controls u as one of the arguments along with model in (4), thereby making it a *control barrier function* (CBF). For this paper, we avoid the use of inputs and make estimates of the derivative of h by using \dot{s} . This allows us to use barrier functions in a model-free way and is sufficient for the presented work as our focus is on reward shaping.

If we are able to restrict the states of the system s, \dot{s} in such a way that the inequality (4) is satisfied, then we know that the set \mathcal{C} is forward invariant (safe). We can use this idea to shape the reward functions in such a way that any violation of safety causes a loss of reward. We will formally show the reward-shaping methodology in the next section.

IV. REWARD SHAPING METHODOLOGY

A. Reward Shaping using Barrier Functions

Having described BFs and their associated formal results, we now discuss BF-inspired reward shaping in the context

of RL. Reward shaping is a method for engineering a reward function to provide more frequent feedback on appropriate behaviours. To illustrate our framework, we propose the shaped reward r' in (5), which we obtain by adding a term to the traditional vanilla reward r .

$$r'(s, \dot{s}) = r(s) + \underbrace{r^{\text{BF}}(s, \dot{s})}_{\text{additional reward}}, \quad (5)$$

$r^{\text{BF}}(s, \dot{s})$ is our barrier function inspired reward shaping term. We emphasise that the new reward r' depends on \dot{s} . From Definition 3.1, \mathcal{C} is forward invariant if and only if there exists a barrier function h such that it satisfies (4). Thus we define r^{BF} as

$$r^{\text{BF}}(s, \dot{s}) = \dot{h}(s, \dot{s}) + \gamma h(s). \quad (6)$$

In order to satisfy (4) we have taken the extended class \mathcal{K} function κ as $\kappa(m) = \gamma m$, $\gamma \in \mathbb{R}_{>0}$. This gives our shaping term r^{BF} the desirable property of positively rewarding the agent when it is in a set of safe states \mathcal{C} while negatively rewarding otherwise.

The BF $h: \mathcal{S} \rightarrow \mathbb{R}$ is chosen to be a suitable function that constrains specific quantities in \mathcal{S} , such that it would lead to desirable properties like actuation safety and training efficiency. The following section provides specific examples of BF-based reward shaping.

Remark 4.1: Since, in RL task, our goal is to find a policy $\pi: \mathcal{S} \rightarrow \mathcal{A}$ that maximizes total (discounted) reward, as we maximise the proposed reward (5) it will also enforce the condition (4) in Definition 3.1 and thus implies safe execution of the task. However, it is essential to note that this approach does not provide a safety guarantee and does not imply safety during training.

B. Barrier Function Formulation

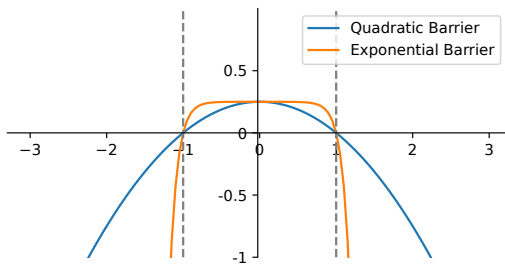


Fig. 2: Plots illustrating the proposed barrier functions (7)-(8). Dashed lines represent the constraint limits (-1,1).

In an RL task, some state variables should ideally lie between a safe range of values. Violating these bounds can result in undesirable behaviours. To constrain them within their safe bounds, we must use an appropriate barrier function, $h(s, \delta) \equiv h(s)$ parameterized by δ . We propose two barrier functions: a quadratic function h_{quad} (7) and an exponential function h_{exp} (8).

$$h_{\text{quad}}(s, \delta) = \sum_{l \in \mathcal{L}} \delta_a (s_l - s_l^{\text{max}}) (s_l^{\text{min}} - s_l) \quad (7)$$

$$h_{\text{exp}}(s, \delta) = \sum_{l \in \mathcal{L}} \delta_a \left[1 - \left(e^{\delta_b (s_l - s_l^{\text{max}})} + e^{\delta_b (s_l^{\text{min}} - s_l)} \right) \right] \quad (8)$$

where $\delta = [\delta_a, \delta_b]$, $\delta \in \mathbb{R}_{>0}^2$ is the vector parameter, and $\mathcal{L} = \{l_0, l_1, l_2, \dots\}$ is the set of indices for the elements of the state variables of the model whose values we want to constrain. In particular, $\{s_l\}$ is l^{th} element of state vector s , which is the value of the state variable upon which we wish to enforce the bounds $(s_l^{\text{min}}, s_l^{\text{max}})$. The choice of δ and the bounds $(s_l^{\text{min}}, s_l^{\text{max}})$ is elaborated in the following section.

Since $h(s, \delta)$ is positive for an $l \in \mathcal{L}$ only when $s_l \in (s_l^{\text{min}}, s_l^{\text{max}})$, and negative otherwise, $h(s, \delta)$ qualifies as a barrier function and r^{BF} can be computed by Eq. (6).

For appropriate values of δ , h_{exp} is flat within the bounds and tapers down sharply outside them, which allows for better exploration while staying within the bounds. In contrast, the shape of h_{quad} makes it more suitable for tasks that require s_l to be constrained near the central values (Fig. 2).

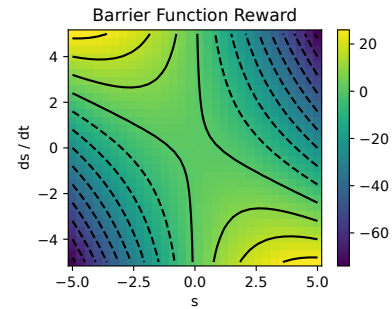


Fig. 3: Density plot of the quadratic BF reward, $r_{\text{quad}}^{\text{BF}}$ constructed using (6) and (7) with (-1,1) as the bounds on s . Notice that the reward depends on s and \dot{s} . The solid and dotted contour lines correspond to positive and negative values of the reward, respectively.

The reward r^{BF} is a function of both s and \dot{s} . Thus, it jointly constrains them depending on their values (Fig. 3). To get a more intuitive understanding, consider the following,

$$\begin{aligned} r^{\text{BF}}(s, \dot{s}) &= \dot{h}(s, \dot{s}) + \gamma h(s) \\ &= \frac{\partial h(s)}{\partial s} \frac{ds}{dt} + \gamma h(s) = \frac{\partial h(s)}{\partial s} \dot{s} + \gamma h(s) \end{aligned} \quad (9)$$

Taking the partial derivative of r^{BF} with respect to \dot{s} gives,

$$\frac{\partial r^{\text{BF}}(s, \dot{s})}{\partial \dot{s}} = \frac{\partial}{\partial \dot{s}} \frac{\partial h(s)}{\partial s} \dot{s} + \frac{\partial}{\partial \dot{s}} (\gamma h(s)) = \frac{\partial h(s)}{\partial s} = h'(s) \quad (10)$$

Now, if the barrier function $h(s)$ is taken to be a concave function, then there exists an $s_0 \in (s^{\text{min}}, s^{\text{max}})$ such that $h'(s) > 0$ for all $s < s_0$ and negative otherwise. Thus, from (10) we can infer that when $s < s_0$ i.e. it is near the lower bound, $(\partial/\partial \dot{s})r^{\text{BF}} = h'(s)$ is positive. As an RL policy wants to maximize r^{BF} , it promotes \dot{s} to be positive. Given that \dot{s} denotes the rate of change of s over time, a positive \dot{s} causes s to increase, moving it away from the lower bound s^{min} . Conversely, the opposite effect applies when s is near s^{max} . In both cases, s is promoted to move towards s_0 .

Thus, r^{BF} works with both s and \dot{s} to promote safety. This effect is lacking with other reward shaping methods that constrain only s within some bounds for safety. It can be verified that both of our proposed barriers h_{quad} and h_{exp} are concave functions with $s_0 = (s^{\text{max}} + s^{\text{min}})/2$.

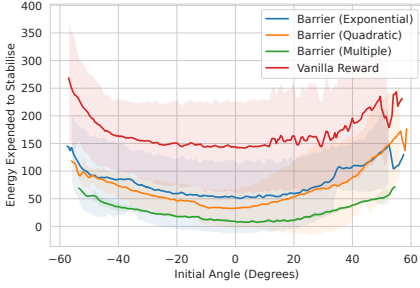


Fig. 4: Energy expended to stabilize initial angles for each reward formulation in the cartpole environment

V. SIMULATION EXPERIMENTS

We experimentally evaluate our BF-based reward-shaping formulation on OpenAI Gym’s Cartpole [29], and MuJoCo environments like Half-Cheetah [30], Humanoid [15], and Ant [14]. We used the Twin Delayed DDPG (TD3) [31] algorithm with two variants of the BF-based reward shaping: $r_{\text{quad}}^{\text{BF}}$ (7) and $r_{\text{exp}}^{\text{BF}}$ (8), which yield the $\pi_{\text{quad}}^{\text{BF}}$ and $\pi_{\text{exp}}^{\text{BF}}$ policies respectively. The environment’s *vanilla* reward is used as the baseline. All experiments were performed on a Ryzen Threadripper CPU, 64GB RAM, and an NVIDIA RTX 3080.

A. Cartpole

Reward Shaping: The state-space contains the pole angle θ_p , pole angular velocity ω_p , cart’s position x_c , and velocity v_c . The task consists of balancing a pole attached to a moving cart, i.e. we want $\theta_p = 0$. The vanilla reward is $r = +1$ for every step taken. We define $(\theta_p^{\min}, \theta_p^{\max}) \equiv (s_l^{\min}, s_l^{\max})$ as the pole angle’s desired threshold range for the goal ($\theta_p = 0$) around which we want the pendulum to stabilize, hence $h = h(\theta_p, \delta)$. This bound is taken as the maximum angle deviation ϕ from which the pole can balance itself. According (5)-(6), the quadratic BF (7) shaped reward is given as, (taking $\gamma = 1$, $\theta_p^{\max} = -\theta_p^{\min} = \phi$)

$$\begin{aligned} r'_{\text{quad}} &= 1 + \dot{h}_{\text{quad}}^{\text{cart}}(s, \dot{s}) + \gamma h_{\text{quad}}^{\text{cart}}(s) \\ &= 1 + \delta_a((\phi^2 - \theta^2) - 2\theta_p \omega_p), \delta_a \in \mathbb{R}_{>0} \end{aligned} \quad (11)$$

Results: Fig. 4 shows energy expended till stabilisation vs initial angle for a range of angles. As can be read, the performance of policy using vanilla reward is energy-expensive compared to that of with r^{BF} . For one unit of energy spent by the vanilla policy, $\pi_{\text{exp}}^{\text{BF}}$ policy spends 0.79 units, $\pi_{\text{quad}}^{\text{BF}}$ policy spends 0.59 units. Fig. 5 describes the control performance for the same initial state. In contrast to $\pi_{\text{quad}}^{\text{BF}}$ policy, the Vanilla policy exhibits chaotic behaviour with no convergence to the $\theta_p = 0$ value. Thus our $\pi_{\text{quad}}^{\text{BF}}$ policy maintains safety during deployment. In Fig. 4 multiple barrier refers to r^{BF} constructed on ω_p and v_c along with θ_p .

B. MuJoCo Walker Environments

Reward Shaping: For walkers like Half-Cheetah [30], Ant [14], and Humanoid [15], the task is to learn to run as fast as possible. We use the same r^{BF} formulation for all these environments as essentially they all have a similar state space \mathcal{S} containing x_{pos}^w , Θ^w and Ω^w , where Θ^w and Ω^w are the

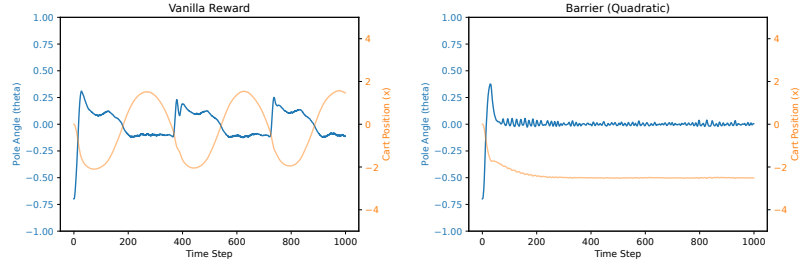


Fig. 5: θ_p and x -position vs time-step graph showing control flow to stabilize -40° pole angle in the cart pole environment. The policy clearly shows that the vanilla policy struggles to stabilize the pole angle close to zero, while the quadratic policy accomplishes this in a few time-steps.

set of agent’s joint angles and angular velocities, respectively. The vanilla reward r given by (12) has multiple terms that guide the agent to move forward (r_{forward}) without falling (r_{health}) and minimizing the contact forces (r_{contact}).

$$r = r_{\text{health}} + r_{\text{forward}} + r_{\text{contact}}. \quad (12)$$

Since the task involves running, we seek to constrain the agent’s joint angles $\theta_l \in \Theta^w$ within a range of angles $(\theta_l^{\min}, \theta_l^{\max}) \forall l \in \mathcal{L}$. In this case, \mathcal{L} is the set of all joint angles. Thus, θ_l is analogous to s_l . These bounds correspond to the range of angles within which a joint can safely turn, violating which could lead to damage to the actuators or collision with other parts of the body. These general bounds are specified in the robot description and do not require additional domain knowledge. The optimal parameters δ for (7)-(8) were found by performing a grid search.

Metrics: The metrics used to evaluate the policy performance are (i) *Actuation Coefficient*, computed from the episodic energy-velocity curve (Fig. 7), (ii) *Training Speed*, computed from the velocity-timestep curve (Fig. 6). All the experiments were repeated for ten random seeds.

The energy \mathcal{E}_w used by a walker w in a time-step is calculated by

$$\mathcal{E}_w = \sum_{j \in \mathcal{J}_w} \Delta\theta_j \cdot \tau_j \quad (13)$$

where \mathcal{J}_w is the set of all joints of the walker, τ_j is the torque exerted by the agent on the joint j , and $\Delta\theta_j$ is the change in the joint angle. The episodic energy is the sum of \mathcal{E}_w over an episode. Since the task of these environments is to maximise the velocity v_w , we define the *Actuation Coefficient* (given in Table. I) as

$$(\text{Actuation Coefficient})_w = \frac{\mathcal{E}_w^{\text{episodic}}}{(v_w^{\text{mean}})^2} \quad (14)$$

This term represents the energy expended or the effort made by the agent to achieve a certain kinetic energy.

Results: Table I shows that for Humanoid, $\pi_{\text{exp}}^{\text{BF}}$ policy takes only about 49% actuation energy to achieve the same kinetic energy as the vanilla policy. Fig. 7 (Left) shows that the π^{BF} policies reach a higher maximum velocity for the same training time-steps. Fig. 6 (Left) shows that $\pi_{\text{exp}}^{\text{BF}}$ policy converges to a higher velocity in much fewer training time-steps. Table I shows that $\pi_{\text{exp}}^{\text{BF}}$ policy converges 1.56 times faster, while $\pi_{\text{quad}}^{\text{BF}}$ policy is no worse than vanilla policy. Continuing the trends from Table I, we see that the π^{BF}

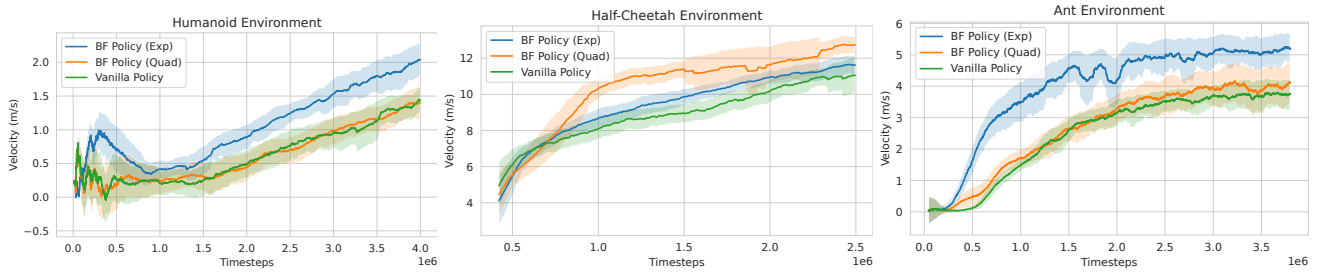


Fig. 6: The plots for each MuJoCo walker environment, averaged for ten random seeds, show the episodic velocity for each training time step. Since these environments aim to achieve as high a velocity as possible, these plots provide a good metric to judge the training speed.

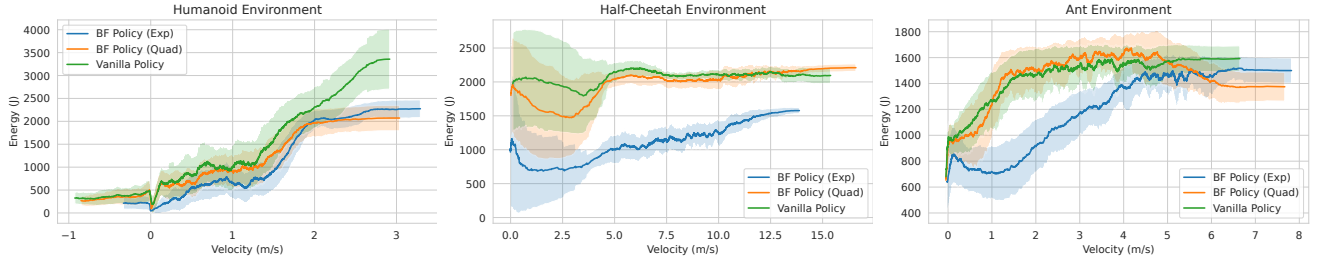


Fig. 7: The episodic energy (in joules) spent by the agent for achieving a particular velocity, averaged over ten random seeds. We see that the policy trained with exponential BF-based reward shaping performs the best across all the environments, consuming the least energy.

| Policy | Humanoid | | Half-Cheetah | | Ant | |
|----------------|-----------------------|--------------------|-----------------------|--------------------|-----------------------|--------------------|
| | Rel. Actuation Coeff. | Rel. Time | Rel. Actuation Coeff. | Rel. Time | Rel. Actuation Coeff. | Rel. Time |
| Vanilla | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Quadratic BF | 0.66 ± 0.08 | 0.89 ± 0.12 | 0.97 ± 0.08 | 0.68 ± 0.03 | 0.62 ± 0.05 | 0.76 ± 0.15 |
| Exponential BF | 0.49 ± 0.03 | 0.64 ± 0.06 | 0.92 ± 0.05 | 0.96 ± 0.02 | 0.63 ± 0.06 | 0.36 ± 0.05 |

TABLE I: Experiment results for different environments and reward formulations. The *Relative Actuation Coefficient* field is the episodic energy expended in actuation (in joules) divided by the squared velocity (m^2s^{-2}) relative to vanilla reward. The result is averaged over 100 best runs. *Relative Time* is the relative training time steps taken by a reward formulation to achieve the same maximum velocity as the Vanilla reward (averaged over 10 random seeds).

policies on Ant and Half-cheetah share similarities with the results in the Humanoid environment. These similarities include reduced kinetic energy, quicker convergence, and increased agent velocity within the same time-step frame.

VI. SIM-TO-REAL ON HARDWARE

A. Implementation Details

Hardware: We deploy our trained policies on the Unitree Go1 robot [16], a quadruped with 12 DOFs. We rely on the SDK provided by [32] to communicate between our code and the low-level control SDK provided by Unitree. The control frequency is 50Hz for both simulation and hardware.

Simulation and Training: We use an open source implementation of the Go1 in the Isaac Gym simulator [33]. We train the policies for 2500 episodes on 4000 parallel environments using PPO [34] with both the Vanilla Reward and r_{exp}^{BF} constructed using h_{exp} (8), having same formulations as given in V-B for other walkers. The joint angle bounds were taken from the physical specifications of the robot.

Domain Randomization: To enhance sim-to-real transfer, we train a policy that remains robust across variations in robot attributes such as body mass, motor strength, joint position calibration, ground friction, restitution, and gravity orientation and magnitude by varying these quantities.

B. Simulation and Training Results

The Vanilla reward is comprised of three components: target velocity tracking, target angle tracking, and target height jump. As depicted in Fig. 8 (Top Row), our proposed policy π^{BF} trained with the r_{exp}^{BF} shaping term added to the vanilla reward consistently outperforms the Vanilla policy across all three tasks. In Fig. 8 (Bottom Row), it’s evident that π^{BF} utilizes only 78% of the energy compared to Vanilla, as calculated using (13). Furthermore, π^{BF} exhibits improved action smoothness (negative of change in action) and significantly lower maximum action values, thereby enhancing safety for deployment.

C. Hardware Results

Vanilla Policy: The Vanilla policy (Fig. 9 Bottom) resulted in unstable movements and limb instability, particularly affecting the front right limb during basic commands like forward motion and re-direction (Fig. 9). This led to a fall, from which the robot autonomously recovered but at the cost of posing practical deployment risks. The robot’s joint movements lacked coordination and efficiency, frequently deviating from the intended path and causing the front hinge joints to approach dangerously close to the ground.

π^{BF} Policy: The policy trained with the new reward significantly enhanced the Go1 robot’s performance, with improved balanced and coordinated movements despite having half

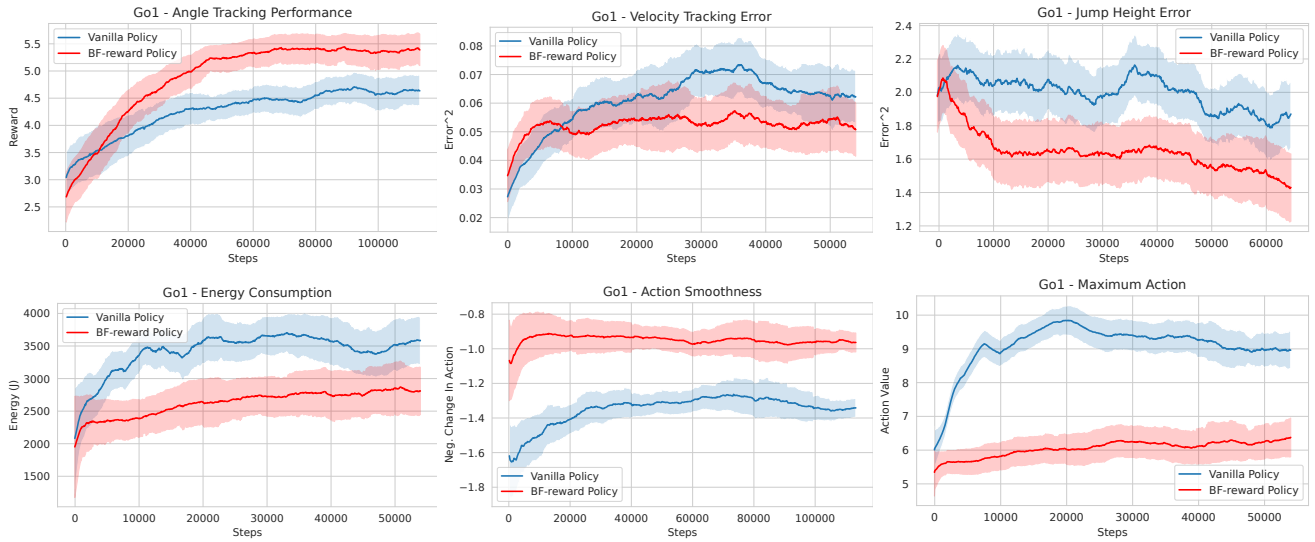


Fig. 8: Training plots for the Go1 robot in the Issac Gym simulation environment. The top row show that our policy π^{BF} trained using $r_{\text{exp}}^{\text{BF}}$ outperforms the policy trained without it (vanilla) in all three tasks - Velocity Tracking, Angle Tracking, and Jumping. The angle tracking reward is inversely proportional to the error in angle. Note that for both the policies we are comparing the same vanilla reward in the plot. Action smoothness is negative of change in action. The bottom row provides insights about the safety and efficiency of π^{BF} .

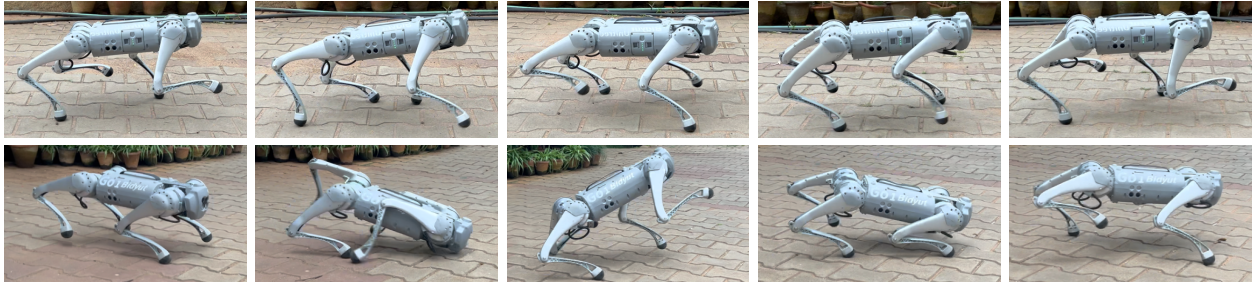


Fig. 9: Top row shows a sample run for the velocity tracking task using π^{BF} policy trained for 2.5k episodes. Bottom row shows the vanilla policy performance, trained for 5k episodes. Frames are 0.4 sec apart. The video can be found here: <http://www.stochlab.com/redirects/rewardshaping2023.html>.

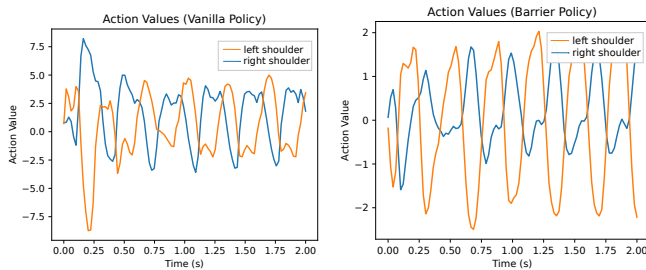


Fig. 10: Action values for two front shoulder flexion joints of Go1 for Vanilla Policy (left) and π^{BF} Policy (right) for a sample run of the velocity tracking task. The action values represent angle targets that are later clipped to a suitable range. It can be seen that the actions for π^{BF} are smoother, rhythmic and have lower action value.

the training duration as the Vanilla policy (Fig. 9 Top). It flawlessly followed RC controller commands, ensuring safety without falls or risky behaviour. The advantages of the π^{BF} policy are further affirmed by Fig. 10 where π^{BF} policy leads to more consistent and rhythmic action values. This characteristic enhances safety for the actuators and contributes to increased task efficiency.

VII. CONCLUSIONS

In this paper we propose barrier function (BF) inspired reward shaping, a safety-oriented, easy-to-implement reward

shaping formulation for robotic platforms. This approach is based on theoretical principles of safety provided by barrier functions. The shaping term aims to encourage agents to remain within predefined safe states during training. This enhances training efficiency and ensures safer exploration. To illustrate our formulation process, we proposed two barrier functions: *exponential* and *quadratic*.

While prior works, e.g. by Cheng et al. [23] achieved high training efficiency and state safety, their framework needs the system dynamics model. In contrast, our method eliminates this need, thus being easy to implement in complex environments. However, a limitation of our study is that we have only tested with barrier functions of joint angles. Further investigation into other quantities such as joint angular velocities could offer a more comprehensive understanding of our reward's effectiveness.

We employed the Unitree Go1 robot [16] as our hardware platform for sim-to-real experiments. Our reward-shaping methodology emerged superior through comparative analysis with [32], revealing smoother, more rhythmic control dynamics. The results indicate that our formulation is an easy way to introduce safety and efficiency in RL training.

REFERENCES

- [1] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.
- [2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [3] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker *et al.*, "Improving alignment of dialogue agents via targeted human judgements," *arXiv preprint arXiv:2209.14375*, 2022.
- [4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.
- [5] S. Arimoto, S. Kawamura, and F. Miyazaki, "Bettering operation of robots by learning," *Journal of Robotic systems*, vol. 1, no. 2, pp. 123–140, 1984.
- [6] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [7] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.
- [8] B. Marthi, "Automatic shaping and decomposition of reward functions," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 601–608.
- [9] M. Grzes and D. Kudenko, "Plan-based reward shaping for reinforcement learning," in *2008 4th International IEEE Conference Intelligent Systems*, vol. 2, 2008, pp. 10–22–10–29.
- [10] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99. Citeseer, 1999, pp. 278–287.
- [11] Y. Dong, X. Tang, and Y. Yuan, "Principled reward shaping for reinforcement learning via lyapunov stability theory," *Neurocomputing*, vol. 393, pp. 83–90, 2020.
- [12] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete event dynamic systems*, vol. 13, no. 1-2, pp. 41–77, 2003.
- [13] R. Bellman, "Dynamic programming," *science*, vol. 153, no. 3731, pp. 34–37, 1966.
- [14] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.
- [15] Y. Tassa, T. Erez, and E. Todorov, "Synthesis and stabilization of complex behaviors through online trajectory optimization," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 4906–4913.
- [16] Unitree, "Go1," 2022. [Online]. Available: <https://www.unitree.com/products/go1>
- [17] M. Grzes and D. Kudenko, "Plan-based reward shaping for reinforcement learning," in *2008 4th International IEEE Conference Intelligent Systems*, vol. 2. IEEE, 2008, pp. 10–22.
- [18] S. M. Devlin and D. Kudenko, "Dynamic potential-based reward shaping," in *Proceedings of the 11th international conference on autonomous agents and multiagent systems*. IFAAMAS, 2012, pp. 433–440.
- [19] M. Grzes, "Reward shaping in episodic reinforcement learning," 2017.
- [20] O. Marom and B. Rosman, "Belief reward shaping in reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [21] A. Balakrishnan and J. V. Deshmukh, "Structured reward shaping using signal temporal logic specifications," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3481–3486.
- [22] N. Saxena, G. Sandeep, and P. Jagtap, "Funnel-based reward shaping for signal temporal logic tasks in reinforcement learning," *IEEE Robotics and Automation Letters*, 2023.
- [23] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3387–3395, Jul. 2019.
- [24] D. Yu, H. Ma, S. E. Li, and J. Chen, "Reachability constrained reinforcement learning," 2022.
- [25] H. Zhao, X. Zeng, T. Chen, Z. Liu, and J. Woodcock, "Learning safe neural network controllers with barrier certificates," 2020.
- [26] B. Marthi, "Automatic shaping and decomposition of reward functions," in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 601–608.
- [27] D. C. Tan, F. Acero, R. McCarthy, D. Kanoulas, and Z. A. Li, "Your value function is a control barrier function: Verification of learned policies using control theory," *arXiv preprint arXiv:2306.04026*, 2023.
- [28] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European control conference (ECC)*. IEEE, 2019, pp. 3420–3431.
- [29] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE transactions on systems, man, and cybernetics*, no. 5, pp. 834–846, 1983.
- [30] P. Wawrzyński, "A cat-like robot real-time learning to run," in *Adaptive and Natural Computing Algorithms: 9th International Conference, ICANNGA 2009, Kuopio, Finland, April 23-25, 2009, Revised Selected Papers 9*. Springer, 2009, pp. 380–390.
- [31] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.
- [32] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," *Conference on Robot Learning*, 2022.
- [33] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.
- [34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.