Adaptive Safety Shielding for Goal-Directed Navigation under Exogenous Disturbance

Debajyoti Chakrabarti   Shawn Yuxuan Dong

Abstract— **Learning-based navigation policies can be efficient but unsafe under deployment shifts such as time-varying wind and uncertain obstacle placement. This project combines a nominal Proximal Policy Optimization (PPO) policy with a dynamic-programming-based Hamilton-Jacobi backward reachable tube (BRT) shield. Offline, we solve reach-avoid dynamic programming problems for multiple disturbance bounds and store the resulting BRT certificates. Online, a moving-window disturbance estimator selects the least conservative valid certificate, while a hierarchical fallback rule recovers when the selected safe set does not contain the current state. Under shifted obstacles and time-varying disturbance, vanilla PPO achieves 58% success with 42% collision, whereas the proposed hierarchical adaptive BRT shield achieves 100% success and 0% collision. The results show that integrating offline dynamic programming with online disturbance-aware switching can significantly improve learned navigation safety.**

I. Introduction

Autonomous aerial vehicles must navigate safely even when wind, obstacle locations, and model conditions differ from training. A learned policy may reach goals efficiently in nominal settings, but the same policy can collide when disturbances push the vehicle toward obstacles. This motivates a controller that preserves the efficiency of learning while adding a safety certificate during deployment.

Hamilton-Jacobi (HJ) reachability provides a dynamic-programming framework for robust safety certification. Given bounded control and disturbance, it computes value functions whose level sets represent states from which safety can be maintained under worst-case disturbance. However, recomputing an HJ partial differential equation online whenever the disturbance bound changes is too expensive for real-time use.

This project asks:

> *Can we adapt HJ safety certificates online under changing disturbance without solving a new HJ problem at every time step?*

Our answer is an adaptive BRT filter bank. Offline, we solve multiple HJ reach-avoid dynamic programming problems for different disturbance bounds. Online, the system estimates the active disturbance, selects a precomputed certificate, and overrides the PPO action only near the certified boundary.

The main contributions are:

- a concrete dynamic-programming implementation of HJ BRT safety certificates for a planar double-integrator reach-avoid task;
- an offline BRT filter bank indexed by disturbance magnitude;
- online integration of disturbance estimation, certificate selection, and safety filtering;
- a hierarchical fallback rule that improves practical recovery when conservative estimates make the selected certificate infeasible;
- quantitative comparison against vanilla PPO, fixed BRT shields, and direct adaptive switching.

UCLA ECE M237 Spring 2026 Course Project Final Report. Instructor: Jason J. Choi.

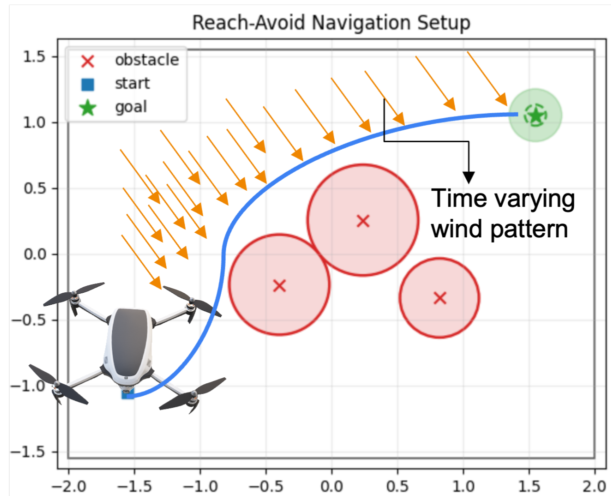

Fig. 1: Motivating reach-avoid navigation problem. The learned policy drives the vehicle toward the goal, while the shield prevents unsafe actions under disturbance.

II. Related Work

HJ reachability computes robust safety and reachability certificates for nonlinear dynamical systems under worst-case disturbance [1]. The resulting value function can be used as a safety certificate and its gradient gives a safety-preserving control direction. Recent work connects reachability value functions to real-time filtering through robust control barrier-value functions [2], and scalable safety learning methods have been proposed for autonomous systems with changing conditions [3]. Online shielding for learned policies has also been studied through dynamic model predictive shielding [4].

Our method is closest to adaptive HJ-based safety certification, but avoids solving new HJ equations online. Instead, the expensive dynamic programming computation is moved offline into a disturbance-indexed certificate bank, and online computation is reduced to estimation, lookup, interpolation, and action replacement.

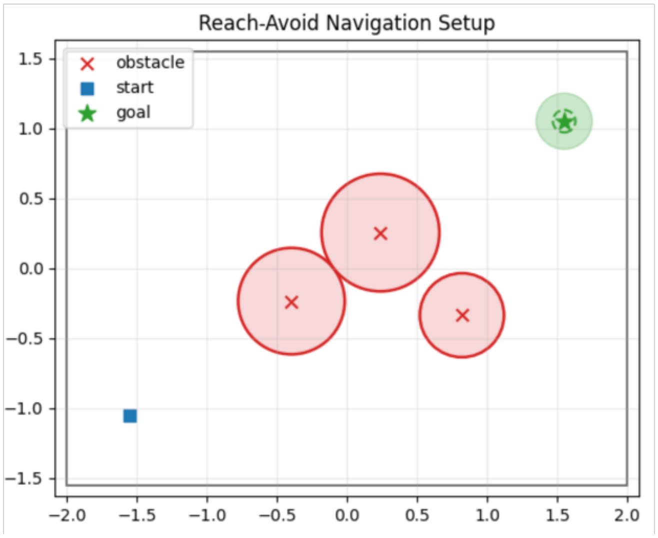

Fig. 2: Planar reach-avoid navigation environment used for safety shielding experiments.

III. Problem Formulation

We model the vehicle as a planar double integrator with state

$$x = [p_x, p_y, v_x, v_y]^\top, \tag{1}$$

and dynamics

$$\dot{p}_x = v_x, \qquad \dot{p}_y = v_y, \tag{2}$$
$$\dot{v}_x = u_x + d_x, \qquad \dot{v}_y = u_y + d_y, \tag{3}$$

where $u = [u_x, u_y]^\top$ is commanded acceleration and $d = [d_x, d_y]^\top$ is wind disturbance. The input satisfies

$$|u_x| \leq u_{\max}, \qquad |u_y| \leq u_{\max}, \qquad u_{\max} = 1.0. \tag{4}$$

The disturbance is bounded by

$$d(t) \in \mathcal{D}_{\bar{d}(t)} = \{d \in \mathbb{R}^2 : \|d\|_\infty \leq \bar{d}(t)\}, \tag{5}$$

where $\bar{d}(t)$ varies over time.

The objective is to reach a target set $\mathcal{T}$ while avoiding an obstacle/failure set $\mathcal{A}$. The nominal controller is a PPO policy,

$$u_{\text{nom}} = \pi_\theta(x), \tag{6}$$

trained for goal reaching and obstacle avoidance. For a fixed disturbance bound $\bar{d}_i$, the HJ computation returns a BRT value function $V_i(x)$ and safe set

$$\mathcal{S}_i = \{x : V_i(x) \leq 0\}. \tag{7}$$

States in $\mathcal{S}_i$ are certified for the corresponding disturbance bound.

IV. Proposed Method

The proposed controller has an offline dynamic programming phase and an online adaptation phase.

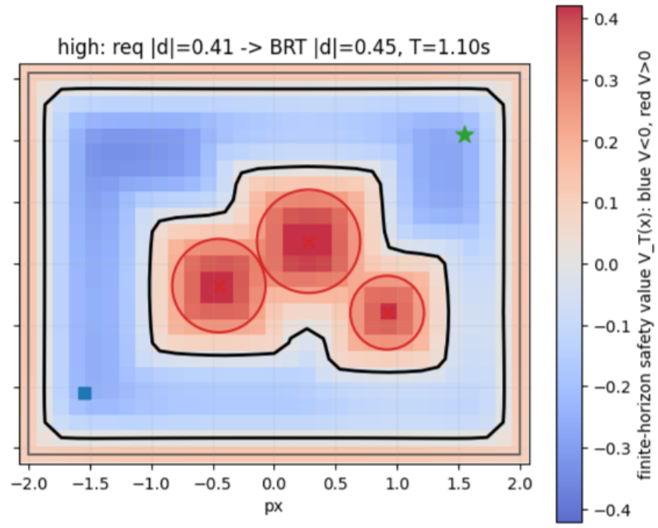

Fig. 3: Offline BRT filter bank. Larger disturbance bounds produce smaller and more conservative safe sets.

A. Offline HJ Dynamic Programming

For each disturbance bound $\bar{d}_i$, we solve an HJ reachability problem on a grid over (p_x, p_y, v_x, v_y). The value function $V_i(x, t)$ evolves according to

$$\frac{\partial V_i}{\partial t}(x, t) + H_i(x, \nabla V_i(x, t)) = 0, \tag{8}$$

with Hamiltonian

$$H_i(x, p) = \min_{u \in \mathcal{U}} \max_{d \in \mathcal{D}_{\bar{d}_i}} p^\top f(x, u, d). \tag{9}$$

The maximization represents worst-case disturbance and the minimization gives the safety controller. We propagate the value function until it converges to an approximately infinite-horizon certificate,

$$\mathcal{S}_i^\infty = \{x : V_i^\infty(x) \leq 0\}, \qquad \|V_i^{k+1} - V_i^k\|_\infty \leq \delta_V. \tag{10}$$

The corresponding safety action is

$$u_{S,i}^*(x) = \arg\min_{u \in \mathcal{U}} \max_{d \in \mathcal{D}_{\bar{d}_i}} \nabla V_i(x)^\top f(x, u, d). \tag{11}$$

Gradients are computed from the stored value table using finite differences and interpolation.

B. BRT Filter Bank

We precompute certificates for ordered disturbance bounds

$$0 < \bar{d}_1 < \bar{d}_2 < \cdots < \bar{d}_M. \tag{12}$$

For each $\bar{d}_i$, we store V_i^∞, ∇V_i, the safety controller $u_{S,i}^*$, and metadata. Larger disturbance bounds produce more robust but smaller safe sets; therefore, the online controller should choose the smallest certificate that covers the estimated disturbance.

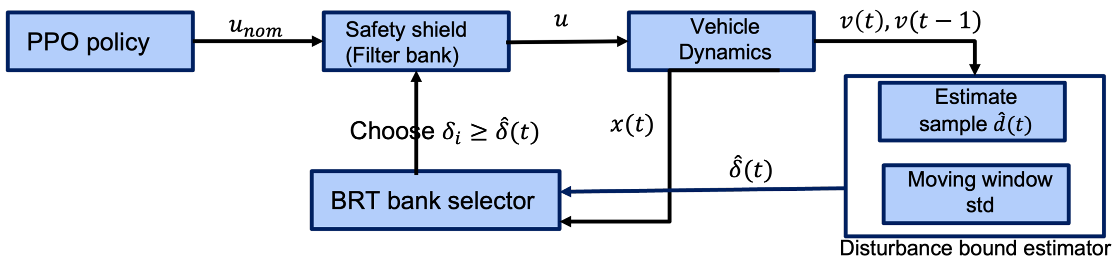

Fig. 4: Runtime architecture of the adaptive BRT shield. PPO proposes a nominal action, the disturbance estimator selects a compatible BRT table, and the shield overrides only near the certified boundary.

C. Online Estimation, Selection, and Shielding

At runtime, we estimate the disturbance bound from a moving window of recent disturbance samples $\{d_{k-W+1}, \dots, d_k\}$. Assuming that disturbances are sampled from a uniform distribution, we use the running standard deviation $\hat{\sigma}_k$ to estimate the disturbance bound $\hat{d}_k$. The controller selects the least conservative certificate satisfying

$$i_k = \min\{i : \bar{d}_i \geq \hat{d}_k\}. \tag{13}$$

The PPO policy proposes $u_{\text{nom},k} = \pi_\theta(x_k)$. The shield uses the PPO action inside the safe set and overrides near the boundary:

$$u_k = \begin{cases} u_{\text{nom},k}, & V_{i_k}(x_k) < -\epsilon, \\ u^*_{S,i_k}(x_k), & V_{i_k}(x_k) \geq -\epsilon. \end{cases} \tag{14}$$

D. Hierarchical Fallback

A conservative disturbance estimate can select a BRT whose safe set does not contain the current state. To avoid immediate infeasibility, we use a hierarchical fallback rule:

$$j_k = \max\{j \leq i_k : V_j(x_k) \leq 0\}. \tag{15}$$

The shield then uses certificate j_k. This relaxes the strict guarantee for the conservative estimate, but it gives a feasible recovery controller and substantially improves empirical performance.

E. Offline-Online Integration

The expensive HJ dynamic programming is performed offline. Online computation only requires disturbance estimation, table lookup, interpolation, certificate selection, and action replacement. This division is the key implementation advantage: the method adapts to changing disturbance without solving a new HJ PDE during deployment.

V. Experiment Results

A. Simulation Setup

We evaluate the planar double-integrator system with a PPO nominal policy. Deployment introduces shifted obstacle placement and time-varying disturbance. A representative disturbance schedule and its estimation is shown in Fig. 7. We compare vanilla PPO, fixed low BRT, fixed high BRT, switched BRT, and hierarchical BRT. The metrics are success rate, collision rate, workspace exit rate, filter intervention rate, and fraction of time outside the selected safe set. Overall, we design the experiments to verify that our adaptive safety shielding can enable the policy to safely handle varying disturbance bounds while successfully reaching the target.

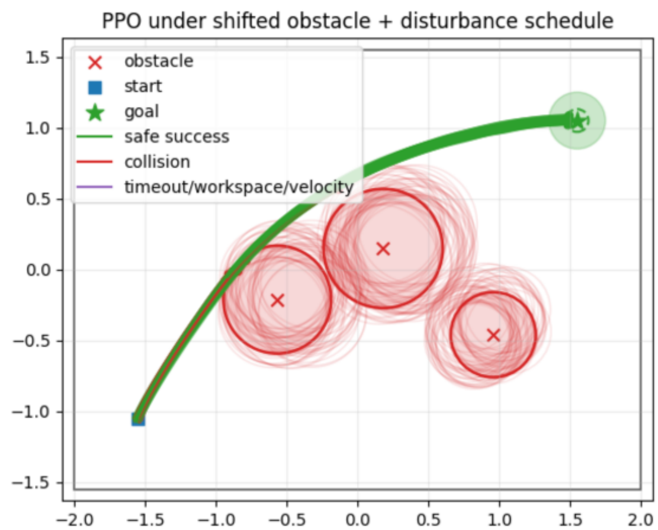

Fig. 5: PPO-alone baseline under deployment shift. Without a safety certificate, the learned policy collides more frequently when obstacle placement and disturbance differ from training.

B. Vanilla PPO Baseline

The PPO-alone baseline motivates the need for shielding. Under the training distribution, PPO achieves 86% success and 14% collision. Under shifted obstacles and time-varying disturbance, success drops to 58% and collision rises to 42%, as illustrated in Fig. 5. Thus, the learned policy is efficient but not safety-certified under deployment shift.

C. Quantitative Comparison

Table I shows the training-distribution setting. All BRT-shielded methods remove collisions with only 12% intervention, indicating that the shield does not dominate the PPO policy when the policy is already safe.

Table II and Fig. 6 show the main stress test with shifted obstacles and time-varying disturbance. Vanilla PPO suffers 42% collision. Fixed low BRT improves safety but still has 4% collision because the certificate does not always match

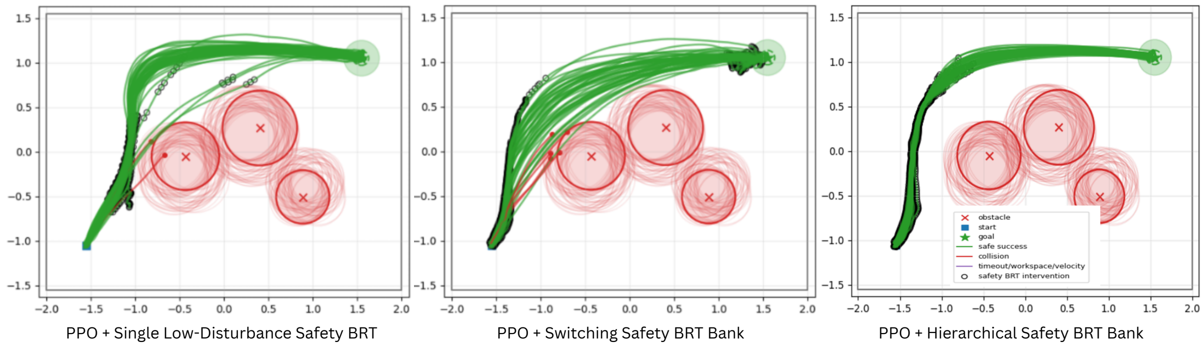

Fig. 6: Representative rollouts under shifted obstacles and time-varying disturbance. The hierarchical fallback method gives the most reliable recovery by avoiding infeasible conservative certificate selections.

TABLE I: Experiment 1: Training obstacle distribution and nominal disturbance.

Method	Succ.	Coll.	Exit	Interv.	Outside
Vanilla PPO	82%	18%	0%	–	–
PPO + fixed low BRT	100%	0%	0%	12%	0%
PPO + switched BRT	100%	0%	0%	12%	0%
PPO + hierarchical BRT	100%	0%	0%	12%	0%

TABLE II: Experiment 2: Shifted obstacles and time-varying disturbance.

Method	Succ.	Coll.	Exit	Interv.	Outside
Vanilla PPO	58%	42%	0%	–	–
PPO + fixed low BRT	96%	4%	0%	12%	3%
PPO + fixed high BRT	58%	42%	0%	0%	100%
PPO + switched BRT	90%	10%	0%	27%	51%
PPO + hierarchical BRT	100%	0%	0%	54%	14%

the active disturbance. Fixed high BRT is too conservative: the system is outside the selected safe set for the whole rollout. Direct switching improves over vanilla PPO, but estimator conservatism causes frequent infeasible selections. Hierarchical BRT performs best, reaching 100% success and 0% collision.

D. Discussion

The results show three main lessons. First, PPO is effective for goal reaching but fragile under deployment shift: collision rises to 42% without shielding. Second, BRT shielding improves safety by replacing unsafe nominal actions near the certified boundary; in the nominal setting, safety improves to 100% success with only 12% intervention. Third, robustness alone is not enough. The fixed high-disturbance certificate is theoretically conservative, but practically unusable when the current state lies outside its safe set. The hierarchical method resolves this by choosing the most conservative feasible certificate.

We also evaluated bank sizes $M = 6$, $M = 10$, and $M = 19$ and observed qualitatively consistent behavior, suggesting that a moderate-sized bank can be sufficient. The key tradeoff is safety versus autonomy: hierarchical BRT achieves the best safety but intervenes more often. This is acceptable for safety-critical navigation, but future work should reduce unnecessary intervention through value-function interpolation or smoother QP-style filtering.

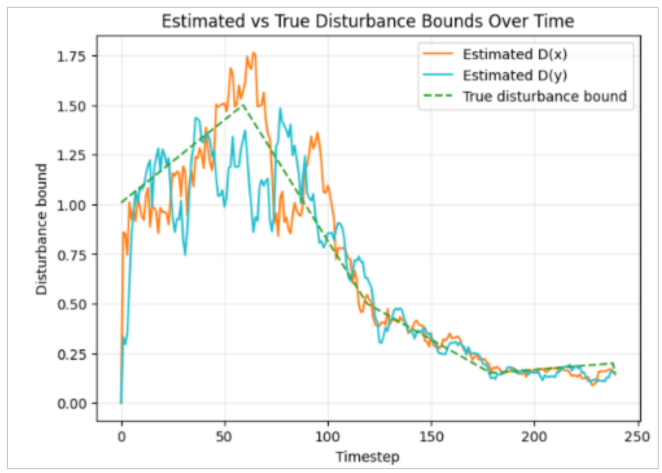

Fig. 7: Moving-window estimator used for online BRT certificate selection.

VI. Conclusion and Future Work

This project implemented adaptive safety shielding for goal-directed navigation under exogenous disturbance. The proposed method combines PPO with an offline bank of dynamic-programming-based HJ BRT certificates. Online, a disturbance estimator selects the active certificate, and a safety shield overrides the PPO action near the boundary. A hierarchical fallback mechanism improves practical recovery when the selected certificate is infeasible.

In the most challenging setting, vanilla PPO achieved 58% success and 42% collision, while hierarchical BRT achieved 100% success and 0% collision. The main lesson is that adaptive shielding is not only about reducing collisions; it is about matching the safety certificate to the active disturbance while keeping the certificate feasible for the current state. Future work will study interpolation between adjacent BRTs, improved disturbance estimation, and smoother action-projection filters.

References

[1] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin, “Hamilton-Jacobi reachability: A brief overview and recent advances,” in *Proc. IEEE Conf. Decision and Control (CDC)*, 2017, pp. 2242–2253.

[2] J. J. Choi, D. Lee, K. Sreenath, C. J. Tomlin, and S. L. Herbert, “Robust control barrier-value functions for safety-critical control,” in *Proc. IEEE Conf. Decision and Control (CDC)*, 2021, pp. 6814–6821.

[3] S. Herbert, J. J. Choi, S. Sanjeev, M. Gibson, K. Sreenath, and C. J. Tomlin, “Scalable learning of safety guarantees for autonomous systems using Hamilton-Jacobi reachability,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2021, pp. 5914–5920.

[4] A. Banerjee, K. Rahmani, J. Biswas, and I. Dillig, “Dynamic model predictive shielding for provably safe reinforcement learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.