Control-theoretic Evaluation of Policy in Sequential Decision Making via Data-driven Differential Game

Geunseob (GS) Oh*, Nauman Sohani*, Huei Peng
University of Michigan, Ann Arbor

***Abstract*—We propose a novel approach to the evaluation of agent policies in uncertain sequential decision making problems. We study a model-based two-player differential game with the first player being the agent of interest and the other player being a disturbance player may act against the agent. In particular, we focus on the problems where tail events are critical. Here, robustness of the policy against the disturbance actions must be guaranteed. We present a framework which relies upon backward reachable sets computed by solving the differential game with respect to the disturbance player. The disturbance action is modeled and learned as a set-valued mapping, rather than a deterministic or probabilistic policy. The solution is disturbance winning set (B) where a predefined metric is violated under all possible policies. By sampling test cases from the complement of B, we obtain challenging scenarios that can help evaluating robustness of policies. We demonstrate our framework in a simple autonomous driving example where an adaptive cruise control policy in a car-following scenario is evaluated. Our approach to the synthesis of realistic and challenging test cases can help to systematically evaluate the robustness and safety of policies.**

I. Introduction and Related Works

While there is significant optimism surrounding artificial intelligence systems, questions persist with respect to their readiness for wide-scale deployment and robustness under uncertainty. In order to solve complex sequential decision making problems, they rely on learning-based methods [1, 2, 3] and/or solutions to optimization problems where the models are optimized over specific cost functions [4, 5, 6]. They offer compelling sub-optimal solutions, however, the robustness and safety against various disturbances are often elusive. An important question follows: how can we evaluate if the agents act in desirable and robust manners under various scenarios?

In this work, we consider a strategy to systematically assess policies of agents in uncertain sequential decision making problems with the presence of a disturbance player. Specifically, we focus on the problems where rare events that have lower probabilities in the distributions are critical, either for the safety and/or performance requirements. The idea is to find meaningful test cases by first answering which test cases should not be considered for the evaluation. A test case of a sequential decision making problem is defined to be a tuple consisting of an initial condition and a sequence of disturbance actions. We specifically address cases challenging, yet where a solution is guaranteed to exist for the agent, meaning that given a test case, the agent can safely maneuver through the sequential decision making problem provided the agent made the right choices. Otherwise, a test may be wasted due to ambiguity in interpreting outcomes where the agent fails: did the policy fail because of the agent's policy or was the test simply too difficult? In this sense, we aim to synthesize the test cases certifying solution guarantees through principles of robust control leveraging game theoretic concepts. In philosophy, this approach is most similar to work done in the space of corner case generation and falsification [7, 8, 9].

In the aforementioned works, the modeling of the disturbance player has been done either using some polytopic representation or by specifying classes of template signals constructed based on our knowledge on the physics and behaviors of the disturbance players. These approaches involve simplifying assumptions. If the disturbance player always acts mildly, it may results in test cases with trivially easy disturbances and fails to capture challenging test cases. If we assume that the disturbance player always acts extremely, we may obtain completely unrealistic test cases of which only conservative policies might be able to pass. To address these limitations, we leverage the insight that the disturbance player reacts to the environment and such reactions are captured in the real-world data. This admits a *data-driven* and *state-dependent* disturbance model, which is then used in a differential game to find the set of *realistic* critical test cases.

As one of the building blocks, we utilize backward reachability [10, 11] in our differential game to answer the following: from which initial states is it possible to drive the agent to final states provided that the control and disturbance sets? This question has found particular consequence in safety-critical applications where the disturbance is conceived as an adversarial player trying to drive the system to an unsafe configuration. In this regard, we present results of our approach using a safety-critical scenario for autonomous vehicles (AV).

II. Solution Approach

A. Problem Formulation

In particular, we consider a two-player sequential decision making problem with agent state x, known state-transition model f of the form: $\dot{x} = f(x,u,w)$, where agent action $u(\cdot)$ and disturbance action $w(\cdot)$ are constrained to membership in sets U and W respectively. The goal is to systematically identify critical test cases to determine whether the agent policy $\pi : X \to 2^U$ results in a violation of a predefined metric of interest $g(x)$. We define the *collision set*, denoted C, as the set of states where $g(x)$ is violated. If an initial state lies outside of C, but there exists t when the state belongs to C

Email: {gsoh, nsohani, hpeng}@umich.edu, *Equal contributions.

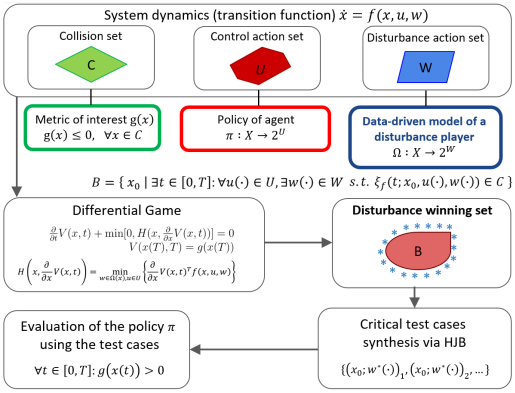

Fig. 1. Description of the solution approach to the systematic evaluation.

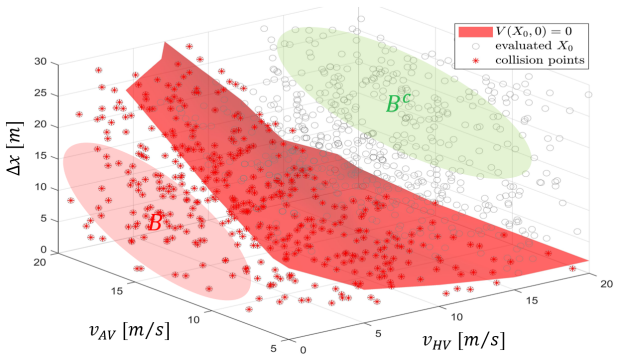

Fig. 2. Critical tests synthesized for the evaluation of a policy for AV.

during the test duration T, i.e. $\exists t \in (0,T] : x(t) \in C$, it is said that the agent fails the test.

We seek critical test cases, i.e. tuples of an initial state x_0 and a sequence of disturbance actions $w(\cdot)$ such that the agent gets close to C at some point of the test but where avoiding a collision is possible.

B. Data-driven Disturbance Modeling

We model the *disturbance* player action as a set-valued mapping $\Omega(\cdot) : X \to 2^W$ where $\Omega(x) \subset W$, rather than a deterministic or a probabilistic function that are more common approach to learning problems. This is because that we focus on the problems where the rare events are likely to jeopardize the safety or performance of the agent (e.g., autonomous vehicle planners). In these problems, it is critical to validate the robustness of the agent against as much scenarios as possible. In this regard, $\Omega(x)$ admits *all* probable disturbances at x. It is a realistic subset of W, where W represents the entire action space of the disturbance player. We model Ω using a single-class classifier $f_{\text{classifier}}(x,w)$ trained on tuples of states and disturbance actions. As all training samples are drawn from the real-world observations, we utilize a support vector data description (SVDD) [12, 13, 14] to learn the (state-dependent) smallest hyper-spheres $\Omega(x)$ for the disturbances [12].

C. Two-player Differential Game

Given the problem specified, a differential game between the agent and disturbance player is established as follows.

$$\begin{aligned} \text{given} \quad & T, U, W, \Omega, f, g \\ \text{minimize}_w \quad & \inf_{t\in[0,T]} \left[g(x(t)) \right] \\ \text{s.t.} \quad & \dot{x} = f(x,u,w), u(\cdot) \in U, w(\cdot) \in \Omega(x(\cdot)) \end{aligned}$$

where the minimization is performed with respect to w. The objective of the disturbance is to drive the agent to C. In reachability literature, such a disturbance is found by minimizing the distance to the boundary of C by defining a function $g(x)$ which satisfies: $g(x) > 0, x \notin C$. Accordingly, the cost function is the infimum of the distance of $x(t)$ to the boundary of C over the time span $t = [0:T]$ and measures how close the agent gets to C. For problems that exhibit optimal substructures, Bellman's Principle of Optimality has shown that the minimum values of time-dependent optimization are the solutions to Hamilton-Jacobi-Bellman PDE [10, 11, 15]:

$$\begin{aligned} \text{solve} \quad & \frac{\partial}{\partial t} V(x,t) + \min\left[0, H(x, \frac{\partial}{\partial x} V(x,t))\right] = 0. \\ \text{s.t.} \quad & V(x(T),T) = g(x(T)), \end{aligned}$$

where $V(x_0,0) = \inf_{t\in[0,T]} g(x(t))$ and H is a Hamiltonian. In our framework, H is a state-dependent Hamiltonian:

$$H\left(x, \frac{\partial}{\partial x} V(x,t)\right) = \min_{w\in\Omega(x), u\in U} \left(\frac{\partial}{\partial x} V(x,t)^T f(x,u,w) \right)$$

where $V(x,t)$ is called the value function of the state x at time t. We solve the above using the level set toolbox [16].

The outcome is *disturbance winning set* (B), which is a set of initial states x_0 where the disturbance player wins the differential game (i.e., $x \in C$ at any time of the game) against *all* possible $u \in U$. To simply put, if $x_0 \in B$, then there exists no policy can avoid the collision. Hence, test cases from B are considered too difficult. The complement of B defined as $B^c = \{x_0 \mid \forall w(\cdot) \in W, \forall t \in [0,T] : \exists u(\cdot) \in U, x(t) \notin C\}$ is utilized for evaluation; A test case consists of x_0 sampled in B^c and $w(\cdot)$.

D. Test Cases Synthesis and Evaluation

By setting the value of $g(x(t))$ which provides a quantitative measure of how close the agent gets to C, we can generate challenging test cases of various difficulty. The synthesis of test cases $(x_0, w(\cdot))$ is a two-step process with the first step being the selection of x_0 in B^c so that $0 < \varepsilon_1 < V(x_0,0) < \varepsilon_2$ and the second solving the PDE with respect to the x_0. As a result, we obtain $w^\star(\cdot)$ that minimizes the objective function.

Lastly, the evaluation of policies with the synthesized test cases is done by simulating the agent policy in the test cases to validate whether it drives the agent to C by checking $\forall t \in [0,T] : g(x(t)) > 0$ or $\min_{t\in[0,T]} g(x(t))$.

III. Preliminary Results and Conclusion

We demonstrate our framework using a safety-critical scenario for AV: we evaluate the robustness of adaptive cruise control with automatic emergency braking algorithm in car-following event with the preceding human-driven car being the disturbance. Fig. 2 describes the resulting B, B^c, evaluated test cases (empty circles), and test cases that had the agent collided with the preceding car. Our approach is a pragmatic mix of theories drawn from robotics, learning, game-theory, and optimization and it is a novel systematic approach to the evaluation through challenging yet realistic test cases.

References

[1] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3389–3396.

[2] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[3] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1343–1350.

[4] G. Oh and H. Peng, "Eco-driving at signalized intersections: What is possible in the real world?" *The 21st IEEE International Conference on Intelligent Transportation Systems*, 2018.

[5] M. P. Vitus and C. J. Tomlin, "A probabilistic approach to planning and control in autonomous urban driving," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 2459–2464.

[6] J. F. Fisac, E. Bronstein, E. Stefansson, D. Sadigh, S. S. Sastry, and A. D. Dragan, "Hierarchical game-theoretic planning for autonomous vehicles," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9590–9596.

[7] G. Chou, Y. E. Sahin, L. Yang, K. J. Rutledge, P. Nilsson, and N. Ozay, "Using control synthesis to generate corner cases: A case study on autonomous driving," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, 2018.

[8] X. Jin, A. Donzé, J. V. Deshmukh, and S. A. Seshia, "Mining requirements from closed-loop control models," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 11, p. 1704–1717, 2015.

[9] N. Sohani, G. Oh, and X. Wang, "A data-driven, falsification-based model of human driver behavior," *arXiv preprint arXiv:1912.08361*, 2019.

[10] L. C. Evans and P. E. Souganidis, "Differential games and representation formulas for solutions of hamilton-jacobi-isaacs equations," *Indiana University mathematics journal*, vol. 33, no. 5, pp. 773–797, 1984.

[11] I. Mitchell, A. Bayen, and C. Tomlin, "A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games," *IEEE Transactions on Automatic Control*, vol. 50, no. 7, p. 947–957, 2005.

[12] D. M. J. Tax, "One-class classification: concept-learning in the absence of counter-examples," Ph.D. dissertation, 2001.

[13] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[14] R. Sadeghi and J. Hamidzadeh, "Automatic support vector data description," *Soft Computing*, vol. 22, no. 1, pp. 147–158, 2018.

[15] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin, "Hamilton-jacobi reachability: A brief overview and recent advances," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 2242–2253.

[16] I. M. Mitchell, "The flexible, extensible and efficient toolbox of level set methods," *Journal of Scientific Computing*, vol. 35, no. 2-3, p. 300–329, 2007.