

# Learning Optimal Strategies for Temporal Tasks in Stochastic Games

Alper Kamil Bozkurt<sup>1</sup>, Yu Wang<sup>1</sup>, Senior Member, IEEE, Michael M. Zavlanos<sup>1</sup>, Senior Member, IEEE, and Miroslav Pajic<sup>1</sup>

**Abstract**—Synthesis from linear temporal logic (LTL) specifications provides assured controllers for systems operating in stochastic and potentially adversarial environments. Automatic synthesis tools, however, require a model of the environment to construct controllers. In this work, we introduce a *model-free* reinforcement learning (RL) approach to derive controllers from given LTL specifications even when the environment is completely unknown. We model the problem as a stochastic game (SG) between the controller and the environment; we then learn optimal strategies that maximize the probability of satisfying the LTL specifications against the *worst-case* environment behavior. We first construct a product game using the deterministic parity automaton (DPA) translated from the given LTL specification. By deriving distinct rewards and discount factors from the acceptance condition of the DPA, we reduce the maximization of the worst-case probability of satisfying the LTL specification into the maximization of a discounted reward objective in the product game; this enables the use of model-free RL algorithms to learn an optimal controller strategy. To address the scalability issues arising when the number of sets defining the acceptance condition of the DPA, usually referred to as colors, is large; we propose a lazy color generation method where distinct rewards and discount factors are utilized only when needed, and an approximate method where the controller eventually focuses on only one color. In several case studies, we show that our approach is scalable to a wide range of LTL formulas, significantly outperforming existing methods that learn controllers from LTL specifications in SGs.

**Index Terms**—Controller synthesis, linear temporal logic (LTL), reinforcement learning (RL), stochastic games (SGs).

Manuscript received 30 August 2023; revised 8 February 2024; accepted 28 March 2024. Date of publication 18 April 2024; date of current version 25 October 2024. This work was supported in part by ONR under Grant N00014-17-1-2504 and Grant N00014-20-1-2745, in part by NSF under Grant CNS-1652544 award, in part by AFOSR under Grant FA9550-19-1-016, and in part by the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks under Grant CNS-2112562. Recommended by Associate Editor A. Mahajan. (Corresponding author: Alper Kamil Bozkurt.)

Alper Kamil Bozkurt, Michael M. Zavlanos, and Miroslav Pajic are with the Department of Computer Science, Mechanical Engineering and Material Science, and Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: alper.bozkurt@duke.edu; mz61@duke.edu; miroslav.pajic@duke.edu).

Yu Wang is with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: yuwang1@ufl.edu).

Digital Object Identifier 10.1109/TAC.2024.3390848

## I. INTRODUCTION

LINEAR temporal logic (LTL) [1] provides a formal specification language that can be used to express many control tasks with temporal objectives, such as safety, sequencing, conditioning, and repetition. For instance, the majority of robotics tasks can be expressed as an LTL formula (e.g., [2], [3], [4], [5], [6], and [7]) such as a path planning task where the objective includes “first go to the entrance and then go to the workspace,” “if the battery is low, go to the charger,” and “continuously monitor a particular region while avoiding the danger zones.” Due to this expressiveness, automatic controller synthesis from LTL specifications has been widely studied in the control community (e.g., [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], and [15]). However, such synthesis approaches require a model of the operational environment a priori, which is mostly impracticable as modern autonomous systems are being deployed in more complex and stochastic environments.

The common situations where a model of the environment is not available necessitate the use of learning-based methods to design controllers for LTL specifications. A growing body of literature has investigated learning controllers when the interaction with stochastic environments can be modeled as Markov decision processes (MDPs). Several studies (e.g., [16], [17], and [18]) have introduced model-based probably approximately correct (PAC) methods that construct a product MDP using the deterministic Rabin automata (DRAs) derived from the LTL specifications and detect the product MDP components where the specifications are satisfied. However, to detect such components, these methods learn and store the transition structure of the MDP, potentially resulting in large memory requirements.

Recently, model-free reinforcement learning (RL) methods have been proposed to mitigate this problem (e.g., [19] and [20]). With such methods, the LTL specifications are translated into suitable limit-deterministic Büchi automata (LDBAs) [21] to construct product MDPs. Rewards are then derived from the Büchi (i.e., repeated reachability) acceptance condition such that controller strategies maximizing sums of discounted rewards maximize the satisfaction probabilities; then, off-the-shelf RL algorithms are used to learn such strategies. However, all these approaches focus on learning for MDPs and do not consider *non-deterministic* adversarial environments that can take nonrandom actions to disrupt performing the given tasks.

A key challenge in enabling autonomous systems to operate in unpredictable environments is to learn controller strategies robust to adversarial inputs. The adversarial inputs usually cannot be restricted to simple perturbations, and thus, they

need to be considered as separate unpredictable environment actions (e.g., [14] and [22]). The design of controllers resilient against the worst-case environment behavior is often crucial for autonomy when safety and security are of the utmost importance. Though such controllers might be conservative, they provide assurance against any potentially adversarial environment behaviors.

Formally, the interaction between controllers and potentially adversarial environments can be modeled as zero-sum turn-based stochastic games (SGs) [23], where the objective is to learn optimal controller strategies to perform the tasks specified as LTL formulas. SGs are natural extensions of MDPs to systems where some states are under the control of an adversary whose objective is to prevent performing the given tasks. Although more restrictive than the general-sum concurrent games, the zero-sum turn-based games are used to model many control and sequential decision-making problems where tasks need to be successfully carried out no matter how the adversarial environment acts [24], [25]; thus, are widely adopted by the control community (e.g., [26], [27], [28], [29], [30], [31], and [32]). For example, the problem of security-aware motion planning against stealthy attacks can be expressed as an SG [33], [34], [35].

There have been studies investigating controller synthesis from LTL specifications in SGs [32] including synthesis from the generalized reactivity(1) fragment of LTL [36] and adaptive approaches [37]; yet, only few have investigated the use of learning for LTL specifications in SGs. A model-based PAC method from [26] precomputes the winning states with respect to the LTL specification based on the assumption that the given LTL specifications can be translated into deterministic Büchi automata (DBAs) and the transition structure of the SG is available a priori. Another model-based PAC method from [38] learns the winning states for reachability specifications, a limited fragment of LTL that cannot be used to specify continuous tasks. As model-based methods, these are inefficient in terms of memory requirements when the number of different states that actions can lead to is large.

More recently, two automata-based model-free learning methods have been proposed to learn controller strategies for LTL specifications in SGs [39], [40]. Our preliminary work [40] has introduced a method that translates the LTL specifications into DRAs, and provides reward and discount factors for each Rabin pair in the acceptance conditions. This method, however, does not guarantee convergence to the optimal strategies when there is more than one Rabin pair. In [39], the LTL specifications are translated to deterministic parity automata (DPAs), whose acceptance conditions are more intricate than the Büchi condition. The problem of satisfying the parity condition is then reduced to a reachability problem by extending the product game constructed using DPAs with terminal states where the probabilities of transitions to these states are determined by the sets in the parity condition. Yet, the episode lengths required to learn optimal strategies in this method *grow exponentially with the number of sets*; thus, substantially limiting its scalability as we demonstrate in our case studies.

Consequently, in this work, we introduce an approach to learn optimal strategies for *any* desired LTL specifications in SGs while significantly improving scalability compared to the existing methods. Our approach is *model-free* as it does not use

or construct transition models of SGs. We begin by composing an SG where the transition structure is completely unknown with a DPA that is automatically obtained from the given LTL specification. We then derive novel reward and discount factors from the parity acceptance condition such that a strategy maximizing the sum of the discounted rewards also *maximizes the probability of satisfying the LTL specification in the worst case*, i.e., for the worst (most-damaging) adversary actions from the controller's perspective.

Unlike the existing automata-based methods (discussed above), our approach provides dense rewards that significantly accelerate learning while being capable of capturing all possible LTL specifications. Also, to overcome the scalability issues when the number of sets defining the acceptance condition of the DPA, referred to as *colors*, is large; we present a model-free method that utilizes these distinct rewards and discount factors *only when needed* via lazy color generation; thus, our method facilitates efficient learning of the optimal strategies in the derived product SGs. Moreover, we provide a scalable approximate method for the scenarios where the controller can perform the task by eventually focusing on only one color.

We compare our methods with each other and with the existing methods in several robot navigation case studies with the LTL specifications, and we show that our lazy color generation and approximate methods outperform the others. Finally, we demonstrate the applicability of our methods in several robotic arm tasks.

The rest of this article is organized as follows. Section II reviews the necessary preliminaries before providing a reduction from LTL specifications to reward returns in Section III. To enhance learning scalability, we introduce the lazy color generation framework in Section IV, and evaluate our approach in Section VI. Finally, Section VII concludes this article.

## II. PRELIMINARIES AND PROBLEM FORMULATION

### A. Stochastic Games

We use SGs to model the problem of performing a given task by the controller (Player 1) against an adversary (Player 2) in a stochastic environment.

*Definition 1:* A (labeled fully observable turn-based two-player) SG is a tuple  $\mathcal{G} = (S, (S_\mu, S_\nu), s_0, A, P, AP, L)$  where  $S = S_\mu \cup S_\nu$  is a finite set of states,  $S_\mu$  is the set of states in which the controller takes action,  $S_\nu$  is the set of states under the control of the adversary, and  $s_0$  is the initial state;  $A$  is a finite set of actions and  $A(s)$  denotes the set of actions that can be taken in a state  $s$ ;  $P : S \times A \times S \mapsto [0, 1]$  is a probabilistic transition function such that  $\sum_{s' \in S} P(s, a, s') = 1$  if  $a \in A(s)$  and 0 otherwise; finally,  $AP$  is a finite set of atomic propositions and  $L : S \mapsto 2^{AP}$  is a labeling function.

SGs can be considered as games played by the controller and the adversary for infinitely many time steps on finite directed graphs consisting of state and state-action nodes. The state nodes are divided into two distinct subsets reflecting the turns of the controller and the adversary. The game starts in the initial state and moves between the state nodes as the controller and the adversary take actions. In each state node, only the owner of the state observes the state information and chooses one of the

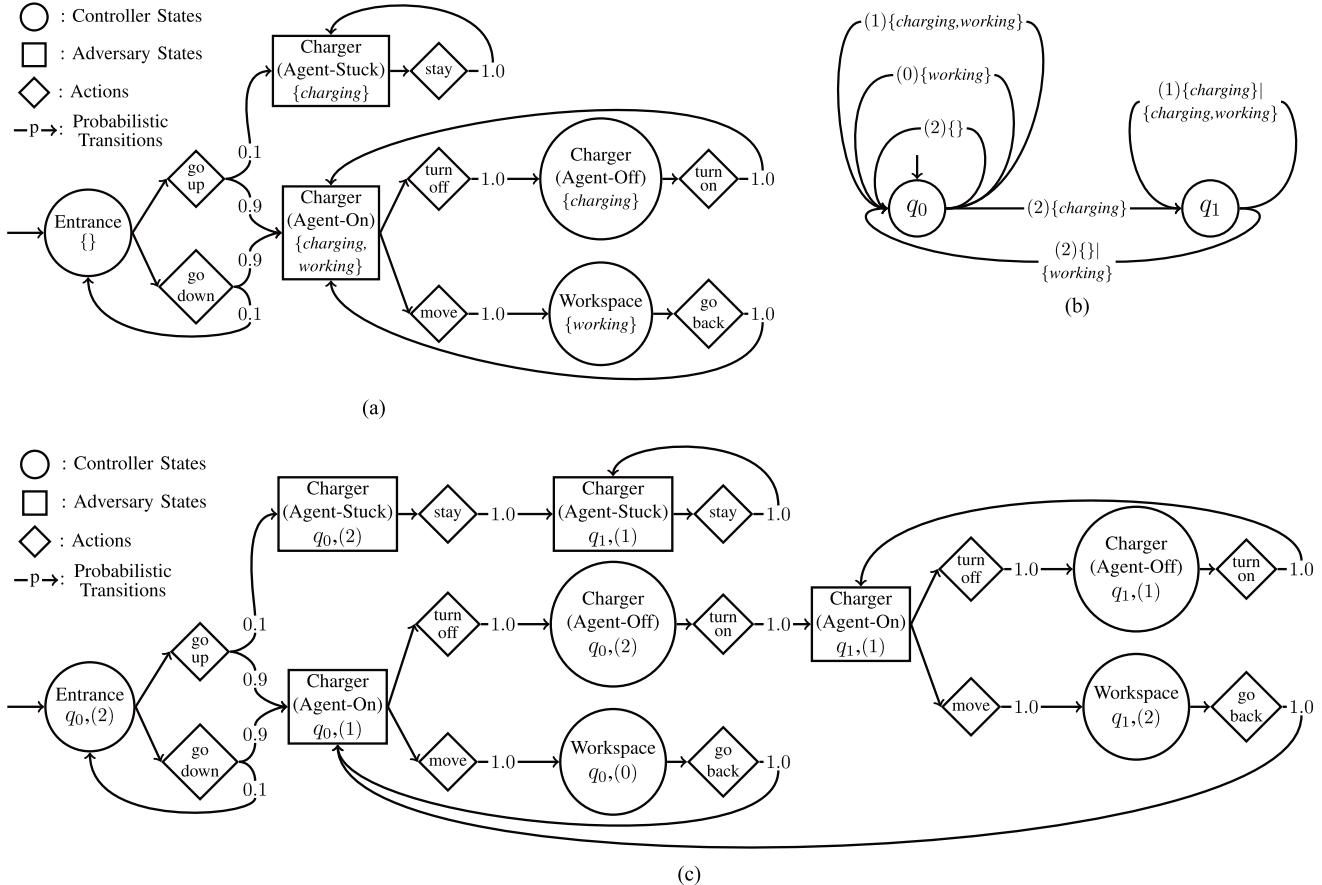


Fig. 1. Product game construction for the system from Example 1. (a) Example SG. The sets of words under the state names are the labels. (b) DPA derived from LTL formula  $\varphi = (\Diamond \square \text{working} \wedge \square \Diamond \text{charging}) \vee \Diamond \square \text{charging}$ . The numbers within parentheses and the set of words on the transitions are the colors and the labels of the transitions, respectively. (c) Product game of the SG in (a) and the DPA in (b). The numbers in the parentheses under the state names are the corresponding colors.

state-action nodes of the state and the game then probabilistically transitions to one of the successors of the node according to the given transition function. We illustrate this in the following example.

*Example 1:* An example SG of a robotics environment is shown in Fig. 1(a). A robotic agent (i.e., the controller) starts in “Entrance,” where an empty label {} is received. In “Entrance,” the agent can choose either “go up” or “go down.” With “go down,” the agent starts working and moves to the charging station with probability (w.p.) 0.9 (i.e., the game makes a transition to “Charger (Agent-On)” and a label of {charging, working} is received). If the agent keeps taking “go down,” it eventually moves to “Charger (Agent-On)” w.p. 1. With “go up,” the agent moves to the charging station but could get stuck w.p. 0.1. In “Charger (Agent-On),” the adversary can prevent the agent from working by turning it OFF (“Charger(Agent-Off)”), however, the agent can turn itself ON again. In addition, the adversary can move the agent from the charging station to “Workspace,” where the agent cannot get charged but can work or go back to the charging station.

We call  $\pi := s_0 s_1 \dots$  an infinite path (i.e., execution) of the SG  $\mathcal{G}$  if for all  $t \geq 0$  there exists an action  $a \in A(s_t)$  such that  $P(s_t, a, s_{t+1}) > 0$ . We denote the state  $s_t$  and the suffix  $s_t s_{t+1} \dots$  by  $\pi[t]$  and  $\pi[t:]$ , respectively. The behaviors of the

players can be specified by strategy functions mapping the history of the visited states to an action. We focus on finite-memory strategies since they suffice for the LTL tasks (see [23] and references therein).

*Definition 2:* A finite-memory strategy for an SG  $\mathcal{G}$  is a tuple  $\sigma = (M, m_0, T, \alpha)$  where  $M$  is a finite set of modes;  $m_0$  is the initial mode;  $T : M \times S \mapsto \mathcal{D}(M)$  is the transition function that maps the current mode and state to a distribution over the next modes;  $\alpha : M \times S \mapsto \mathcal{D}(A)$  is a function that maps a given mode  $m \in M$  and a state  $s \in S$  to a discrete distribution over  $A(s)$ . A controller strategy  $\mu$  is a finite-memory strategy that maps only the controller states to distributions over the actions. Similarly, an adversary strategy  $\nu$  is a finite-memory strategy mapping the adversary states to distributions over the actions. A finite-memory strategy is called pure memoryless if there is only one mode ( $|M| = 1$ ) and  $\alpha(m_0, s)$  is a point distribution assigning a probability of 1 to a single action for all  $s \in S$ .

Intuitively, a finite-memory strategy is a finite state machine moving from one mode (memory state) to another as the SG states are visited, outputting a distribution over the actions in each state. Unlike the standard definition of finite-memory strategies (e.g., [1] and [23]) where transitions among the modes are all deterministic, Definition 2 allows probabilistic transitions; this will later enable modeling of probabilistic transitions

between different levels of the derived product games where pure memoryless strategies induce finite-memory strategies with probabilistic mode transitions in the original SGs.

For a given pair of a finite-memory controller strategy  $\mu$  and an adversary strategy  $\nu$  in an SG  $\mathcal{G}$ , we denote the resulting *induced Markov chain* (MC) as  $\mathcal{G}_{\mu,\nu}$ . We use  $\pi \sim \mathcal{G}_{\mu,\nu}$  and  $\pi_s \sim \mathcal{G}_{\mu,\nu}^s$  to denote paths drawn from  $\mathcal{G}_{\mu,\nu}$  and  $\mathcal{G}_{\mu,\nu}^s$ , respectively, where  $\mathcal{G}_{\mu,\nu}^s$  is same as  $\mathcal{G}_{\mu,\nu}$  except that the state  $s$  is designated as the initial state instead of  $s_0$ . Finally, a *bottom strongly connected component* (BSCC) of an MC is a set of states such that there is a path from each state to any other state in the set without any outgoing transitions. We denote the set of all BSCCs of the MC  $\mathcal{G}_{\mu,\nu}$  by  $\mathcal{B}(\mathcal{G}_{\mu,\nu})$ .

### B. Linear Temporal Logic

LTL provides a high-level formalism to specify tasks with temporal properties by placing requirements for infinite paths [1]. LTL specifications consist of nested combinations of Boolean and temporal operators according to the grammar

$$\varphi := \text{true} \mid a \mid \varphi_1 \wedge \varphi_2 \mid \neg \varphi \mid \bigcirc \varphi \mid \varphi_1 \mathbf{U} \varphi_2, a \in \text{AP}.$$

The other Boolean operators can be defined via the standard equivalences (e.g.,  $\varphi_1 \vee \varphi_2 := \neg(\neg \varphi_1 \wedge \neg \varphi_2)$ ,  $\varphi \rightarrow \varphi' := \neg \varphi \vee \varphi'$ ). A path  $\pi$  of an SG  $\mathcal{G}$  satisfies an LTL specification  $\varphi$ , denoted by  $\pi \models \varphi$ , if the following recursively hold:

- 1) if  $\varphi = a$  and  $L(\pi[0]) = a$  (i.e.,  $a$  immediately holds);
- 2) if  $\varphi = \varphi_1 \wedge \varphi_2$ ,  $\pi \models \varphi_1$ , and  $\pi \models \varphi_2$ ;
- 3) if  $\varphi = \neg \varphi'$  and  $\pi \not\models \varphi'$ ;
- 4) if  $\varphi = \bigcirc \varphi'$  (called *next*  $\varphi'$ ) and  $\pi[1:] \models \varphi'$ ;
- 5) if  $\varphi = \varphi_1 \mathbf{U} \varphi_2$  (called  $\varphi_1$  *until*  $\varphi_2$ ) and there exists  $t \geq 0$  such that  $\pi[t:] \models \varphi_2$  and for all  $0 \leq i < t$ ,  $\pi[i:] \models \varphi_1$ .

Intuitively, the temporal operator  $\bigcirc \varphi$  expresses that  $\varphi$  needs to hold in the next time step, whereas  $\varphi_1 \mathbf{U} \varphi_2$  specifies that  $\varphi_1$  needs to hold until  $\varphi_2$  holds. Other commonly used temporal operators such as *eventually* ( $\diamond$ ) and *always* ( $\square$ ) can be derived as:  $\diamond \varphi := \text{true} \mathbf{U} \varphi$  (i.e.,  $\varphi$  eventually holds) and  $\square \varphi := \neg(\diamond \neg \varphi)$  (i.e.,  $\varphi$  always holds).

LTL can be used to specify a wide range of robotics tasks such as sequencing, surveillance, persistence, and avoidance. In the following example, we illustrate how a persistence task can be specified as an LTL formula, and we discuss the optimal controller and adversary strategies.

*Example 2:* Consider a simple persistence task “ $\varphi = \diamond \square \text{charging}$ ” in the SG presented in Fig. 1(a). The robot controller can ensure that “Charger(Agent-On)” is eventually reached by taking “go down” in “Entrance” persistently. However, the adversary can move the robot to “Workspace,” a state that is not labeled with “charging,” thereby making the probability of satisfying the specification 0. To avoid this, the controller can instead take “go up,” which w.p. 0.1 leads to “Charger(Stuck),” a trap cell labeled with “charging,” where the specification is guaranteed to be satisfied.

### C. Deterministic Parity Automata

Any LTL task can be translated to a DPA that accepts an infinite path satisfying the LTL task [41].

*Definition 3:* A DPA is a tuple  $\mathcal{A} = (Q, q_0, \Sigma, \delta, \kappa, C)$  such that  $Q$  is the finite set of automaton states;  $q_0 \in Q$  is the initial

automaton state;  $\Sigma = 2^{\text{AP}}$  is a finite alphabet;  $\delta : Q \times \Sigma \mapsto Q$  is the transition function;  $\kappa$  is the number of colors;  $C : Q \times \Sigma \mapsto \{0, \dots, \kappa-1\}$  is the coloring function. A path  $\pi$  of  $\mathcal{G}$  induces an execution  $\rho_\pi = \langle q_0, L(\pi[0]) \rangle \langle q_1, L(\pi[1]) \rangle \dots$  such that for all  $t \geq 0$ ,  $\delta(q_t, L(\pi[t])) = q_{t+1}$ . Let  $\text{Inf}(\rho_\pi)$  denote the set of the transitions  $\langle q, l \rangle \in Q \times \Sigma$  made infinitely often by  $\rho_\pi$ ; then, a path  $\pi$  is accepted by a DPA if  $\max\{C(\langle q, l \rangle) \mid \langle q, l \rangle \in \text{Inf}(\rho_\pi)\}$  is an odd number.

For a given LTL task  $\varphi$ , we use  $\mathcal{A}_\varphi$  to denote a DPA derived from  $\varphi$ . DPAs provide a systematic way to evaluate the satisfaction of any LTL specification, which can be expressed by the satisfaction of the parity condition of a constructed DPA. The parity condition is satisfied simply when the largest color among the colors repeatedly visited is an odd number. This provides a natural framework to reason about LTL tasks in SGs; the controller tries to visit the states triggering the odd-colored transitions as often as possible, while the adversary tries to do the opposite (i.e., even-colored transitions).

*Example 3:* Fig. 1(b) shows a DPA derived from the LTL formula “ $\varphi = (\diamond \square \text{working} \wedge \square \diamond \text{charging}) \vee \diamond \square \text{charging}$ .” The executions visiting  $q_0$  and  $q_1$  infinitely many times make the interstate transitions colored with 2 infinitely many times. Since 2 is an even number and the largest color in this DPA, regardless of how many times the transitions colored with 1 or 0 are made, these executions would not be accepted. Therefore, any accepting execution, after some finite time steps, must stay forever in either  $q_0$  or  $q_1$ . The executions staying in  $q_0$  are accepting only if they receive the label {charging, working} infinitely many times and the label {} finitely many times, hence satisfying  $\diamond \square \text{working} \wedge \square \diamond \text{charging}$ . In addition, the accepting executions staying in  $q_0$  receive the labels {charging} or {charging, working}, thus satisfying  $\diamond \square \text{charging}$ .

### D. Problem Statement

We can now formalize the problem considered in this work.

*Problem 1:* For a given LTL task specification  $\varphi$  and an SG  $\mathcal{G}$  where the transition probabilities are *completely unknown*, design a model-free RL approach to learn an optimal controller strategy under which the given LTL tasks are performed successfully with the highest probability in the worst case (i.e., against an optimal adversary).

We consider control strategies maximizing the worst-case performance, in order to provide robust performance guarantees against unpredictable and potentially adversarial environmental behaviors. When the environment does not act as an optimal adversary, the learned controller strategy can exhibit an even higher probability of successfully performing the task. On the other hand, while there may exist superior controller strategies against a suboptimal adversary, the tradeoff is that adopting such strategies could lead to even zero satisfaction probability as soon as the adversary becomes optimal. Therefore, considering the worst-case scenario may be necessary, especially in the context of safety-critical systems.

Formally, our objective is to learn a controller strategy  $\mu_\varphi$  for the SG  $\mathcal{G}$  such that under  $\mu_\varphi$ , the probability that a path  $\pi$  satisfies the LTL specification  $\varphi$  of the given task is maximized in the worst case, i.e.,

$$\mu_\varphi := \arg \max_{\mu} \min_{\nu} \Pr_{\pi \sim \mathcal{G}_{\mu,\nu}} \{\pi \mid \pi \models \varphi\} \quad (1)$$

where  $\mu$  and  $\nu$  are finite-memory controller and adversarial strategies from Definition 2, and  $Pr_{\pi \sim \mathcal{G}_{\mu, \nu}} \{\pi \mid \pi \models \varphi\}$  denotes the probability that a path drawn from  $\mathcal{G}_{\mu, \nu}$  satisfies  $\varphi$ .

Model-free RL algorithms such as minimax-Q [42] require a discount factor, which can be considered as the probability of continuing the game, and a reward function providing a scalar reward after each transition for guidance, and therefore cannot learn directly from the LTL specifications. Hence, we solve Problem 1 by crafting rewards and discount factors from the given LTL specification in a way that a model-free RL algorithm can efficiently learn an optimal strategy from (1) by maximizing the minimum expected value of the return, which is the sum of the discounted rewards.

### III. REDUCTION FROM LTL SPECIFICATIONS TO REWARD RETURNS

In this section, we introduce a reduction from Problem 1 to the problem of learning a controller strategy maximizing the worst-case return. We start by constructing a product game by composing the given SG with the DPA derived from the given LTL specification of the desired task. We then introduce *novel rewards and discount factors* for each set in the acceptance condition of the DPA, and show that for any memoryless strategy pair, the expected sum of the discounted rewards approaches the probability of satisfying the parity condition as rewards approach zero.

#### A. Product Game Construction

By constructing a product game, the problem of learning an optimal controller strategy for an LTL task  $\varphi$  in an SG  $\mathcal{G}$  is reduced to meeting the parity condition of the derived DPA  $\mathcal{A}_\varphi$ .

**Definition 4:** A product game of an SG  $\mathcal{G}$  and a DPA  $\mathcal{A}_\varphi$  is a tuple  $\mathcal{G}^\times = (S^\times, (S_\mu^\times, S_\nu^\times), s_0^\times, A^\times, P^\times, \kappa, C^\times)$  where  $S^\times = S \times Q$  is the set of product states;  $S_\mu^\times = S_\mu \times Q$  and  $S_\nu^\times = S_\nu \times Q$  are the controller and adversary product states, respectively;  $s_0^\times = \langle s_0, q_0 \rangle$  is the initial state;  $A^\times = A$  with  $A^\times(\langle s, q \rangle) = A(s)$  for all  $s \in S, q \in Q$ ;  $P^\times: S^\times \times A^\times \times S^\times \mapsto [0, 1]$  is the probabilistic transition function such that

$$P(\langle s, q \rangle, a, \langle s', q' \rangle) = \begin{cases} P(s, a, s'), & \text{if } q' = \delta(q, L(s)) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

and  $C^\times: S^\times \mapsto \{0, \dots, \kappa-1\}$  is the product coloring function such that  $C^\times(\langle s, q \rangle) = C(q, L(s))$ . A path  $\pi^\times = \langle s_0, q_0 \rangle \langle s_1, q_1 \rangle \dots$  in a product game meets the parity condition if it holds that

$$\varphi^\times := \text{“max } \{C^\times(s^\times) \mid s^\times \in \text{Inf}^\times(\pi^\times)\} \text{ is odd”} \quad (3)$$

where  $\text{Inf}^\times(\pi^\times)$  is the set of product states visited infinitely many times by  $\pi^\times$ .

The product game  $\mathcal{G}^\times$  effectively captures this synchronous execution of the SG  $\mathcal{G}$  and the DPA  $\mathcal{A}_\varphi$ ; they start in the initial states, and whenever  $\mathcal{G}$  moves to a state,  $\mathcal{A}_\varphi$  consumes the label of the state and makes a transition. For example, when  $\mathcal{G}$  is in state  $s$  and  $\mathcal{A}_\varphi$  is in  $q$ , the product game  $\mathcal{G}^\times$  is in  $\langle s, q \rangle$ . If the SG moves to  $s'$ , the DPA moves to the state  $q' = \delta(q, L(s'))$ , represented by the product game transition from  $\langle s, q \rangle$  to  $\langle s', q' \rangle$ .

This implies a one-to-one mapping between the paths of  $\mathcal{G}$  and  $\mathcal{G}^\times$ .

**Lemma 1:** For a given SG  $\mathcal{G}$ , a DPA  $\mathcal{A}_\varphi$  and their product game  $\mathcal{G}^\times$ , let  $\pi = s_0 s_1 \dots$  be a path in  $\mathcal{G}$  and  $\pi^\times = \langle s_0, q_0 \rangle \langle s_1, q_1 \rangle \dots$  be the corresponding path in  $\mathcal{G}^\times$ . Then

$$(\pi^\times \models \varphi^\times) \Leftrightarrow (\pi \models \varphi). \quad (4)$$

*Proof:* It holds from the definition of SGs, as the DPA  $\mathcal{A}_\varphi$  makes a transition colored with  $k$  if and only if a product state  $\langle s, q \rangle$  colored with  $k$  is visited in the product game  $\mathcal{G}^\times$ . ■

A strategy in  $\mathcal{G}^\times$  induces a finite-memory strategy in  $\mathcal{G}$  where the states of  $\mathcal{A}_\varphi$  act as the modes governed by the transition function of  $\mathcal{A}_\varphi$ . To illustrate this, let  $\mu^\times$  denote a *pure memoryless strategy* in  $\mathcal{G}^\times$  and  $\mu$  denote its *induced strategy* in  $\mathcal{G}$ . While  $\mu$  is operating in mode  $m$  corresponding to the DPA state  $q$ , if a state  $s$  is visited in  $\mathcal{G}$ ,  $\mu$  changes its mode from  $m$  to  $m'$ , corresponding to the DPA state  $q' = \delta(q, L(s'))$ ; as a result,  $\mu$  chooses the action that  $\mu^\times$  selects in  $\langle s', q' \rangle$ .

Hence, from Lemma 1, the probability of satisfying the *parity condition*  $\varphi^\times$  under a strategy pair  $(\mu^\times, \nu^\times)$  in the product game  $\mathcal{G}^\times$  is equal to the probability of satisfying the LTL specification  $\varphi$  in the SG  $\mathcal{G}$  under the induced strategy pair  $(\mu, \nu)$ ; this is formalized in the following lemma.

**Lemma 2:** For a given strategy pair  $(\mu^\times, \nu^\times)$  in the product game  $\mathcal{G}^\times$  and its induced strategy pair  $(\mu, \nu)$  in  $\mathcal{G}$ , let  $\pi$  and  $\pi^\times$  be random paths drawn from the MCs  $\mathcal{G}_{\mu, \nu}$  and  $\mathcal{G}_{\mu^\times, \nu^\times}^\times$ , respectively. Then

$$Pr \{\pi \mid \pi \models \varphi\} = Pr \{\pi^\times \mid \pi^\times \models \varphi^\times\}. \quad (5)$$

*Proof:* Follows from Lemma 1, since there is a one-to-one mapping between the paths of  $\mathcal{G}_{\mu, \nu}$  and  $\mathcal{G}_{\mu^\times, \nu^\times}^\times$ . ■

**Example 4:** Fig. 1(c) presents the product game  $\mathcal{G}^\times$  obtained from the SG  $\mathcal{G}$  in Fig. 1(a) and the DPA  $\mathcal{A}_\varphi$  of the LTL task  $\varphi$  in Fig. 1(b). In  $\mathcal{G}$ , if the adversary follows a pure memoryless strategy, the controller can almost surely win the game in the sense that  $\varphi$  is satisfied by taking “go down” in “Entrance.” This eventually leads to “Charger (Agent-On),” from which  $\mathcal{G}$  can alternate either between “Charger(Agent-On)” and “Workspace” (satisfying  $\Diamond \Box \text{working} \wedge \Box \Diamond \text{charging}$ ), or between “Charger(Agent-On)” and “Charger(Agent-Off)” (satisfying  $\Diamond \Box \text{charging}$ ); thus,  $\varphi$  is satisfied. However, the adversary can win in the  $\mathcal{G}^\times$  by following a pure memoryless strategy that chooses “turn OFF” in  $\langle \text{Charger(Agent-On)}, q_0 \rangle$  and “move” in  $\langle \text{Charger(Agent-On)}, q_1 \rangle$  as in this case, the largest color in the obtained infinite cycle  $\langle \text{Charger(Agent-On)}, q_0 \rangle \rightarrow \langle \text{Charger(Agent-Off)}, q_0 \rangle \rightarrow \langle \text{Charger(Agent-On)}, q_1 \rangle \rightarrow \langle \text{Workspace}, q_1 \rangle \rightarrow \langle \text{Charger(Agent-On)}, q_0 \rangle \rightarrow \dots$  is 2 (even). This pure memoryless strategy in  $\mathcal{G}^\times$  induces a finite-memory strategy in  $\mathcal{G}$  that alternates between the modes  $m_0$  and  $m_1$ , under which “turn OFF” and “move” are selected in “Charger(Agent-On)” respectively, where  $m_0$  and  $m_1$  correspond to the DPA states  $q_0$  and  $q_1$ .

We use  $Pr_{\mu^\times, \nu^\times}(s^\times \models \varphi^\times)$  to denote the probability that a state  $s^\times \in S^\times$  satisfies the parity condition  $\varphi^\times$  under strategies  $(\mu^\times, \nu^\times)$ ; i.e.,

$$Pr_{\mu^\times, \nu^\times}(s^\times \models \varphi^\times) := Pr \left\{ \pi^\times \in \Pi_{\mu^\times, \nu^\times}^{\times s^\times} \mid \pi^\times \models \varphi^\times \right\}$$

here,  $\Pi_{\mu^\times, \nu^\times}^{\times s^\times}$  denotes the set of all paths of the product MC  $\mathcal{G}_{\mu^\times, \nu^\times}^{\times s^\times}$  that is obtained from  $\mathcal{G}_{\mu^\times, \nu^\times}^\times$  by assigning  $s^\times$  as the initial state.

Therefore, from Lemma 2 and the fact that pure memoryless strategies suffice for the parity condition [23], our objective can be revised as learning a pure memoryless strategy defined as

$$\mu_{\varphi}^{\times} := \arg \max_{\mu^{\times}} \min_{\nu^{\times}} \Pr_{\mu^{\times}, \nu^{\times}}(s_0^{\times} \models \varphi^{\times}) \quad (6)$$

in the product game  $\mathcal{G}^{\times}$ . The strategy  $\mu_{\varphi}^{\times}$  is then used to induce the finite-memory strategy  $\mu_{\varphi}$  from (1) in the initial game  $\mathcal{G}$ .

### B. Reduction From Parity to Return

To obtain optimal strategies from (6) via model-free RL, we introduce a reward function  $R_{\varepsilon}^{\times} : S^{\times} \mapsto \mathbb{R}$  and a *state-dependent* discount function  $\gamma_{\varepsilon}^{\times} : S^{\times} \mapsto [0, 1)$ , parameterized by  $\varepsilon$  as

$$R_{\varepsilon}^{\times}(s^{\times}) := \begin{cases} \varepsilon^{\kappa - C^{\times}(s^{\times})}, & \text{if } C^{\times}(s^{\times}) \text{ is odd} \\ 0, & \text{if } C^{\times}(s^{\times}) \text{ is even} \end{cases} \quad (7)$$

$$\gamma_{\varepsilon}^{\times}(s^{\times}) := 1 - \varepsilon^{\kappa - C^{\times}(s^{\times})} \quad (8)$$

where  $\kappa$  is the number of colors. The idea behind the rewards structure is to encourage the agent to repeatedly visit a state colored with an odd number as large as possible by assigning a larger reward to the states with larger odd colors. Further, the rewards are discounted more in the states with larger colors to reflect the parity condition and to keep the return bounded. The parameter  $\varepsilon$  should be sufficiently small so that the effect of finitely visited states on the return becomes negligible.

We slightly extend the definition of the return of a suffix of a path  $\pi^{\times}[t:]$ , denoted by  $G_{\varepsilon}^{\times}(\pi^{\times}[t:])$ , for state-dependent discount factors as

$$G_{\varepsilon}^{\times}(\pi^{\times}[t:]) := \sum_{i=t}^{\infty} \left( \prod_{j=0}^{i-1} \gamma_{\varepsilon}^{\times}(\pi^{\times}[j]) \right) R_{\varepsilon}^{\times}(\pi^{\times}[i]). \quad (9)$$

Now, the return of the entire path is simply  $G_{\varepsilon}^{\times}(\pi^{\times}) := G_{\varepsilon}^{\times}(\pi^{\times}[0:])$ . We further use  $v_{\varepsilon, \mu^{\times}, \nu^{\times}}^{\times}(s^{\times})$  to denote the expected return of the paths starting from a state  $s^{\times}$  under a strategy pair  $(\mu^{\times}, \nu^{\times})$ , which is formally defined as

$$v_{\varepsilon, \mu^{\times}, \nu^{\times}}^{\times}(s^{\times}) := \mathbb{E}_{\pi_{s^{\times}}^{\times} \sim G_{\mu^{\times}, \nu^{\times}}^{\times}} [G_{\varepsilon}^{\times}(\pi_{s^{\times}}^{\times})]. \quad (10)$$

We now establish a connection similar to the Blackwell optimality [43] between the values and satisfaction probabilities, and show that as  $\varepsilon$  approaches 0, the value of each state approaches the probability of satisfying  $\varphi$  starting from that state.

*Theorem 1:* For a given product game  $\mathcal{G}^{\times}$  and a pure memoryless strategy pair  $(\mu^{\times}, \nu^{\times})$ , it holds for any  $s^{\times} \in S^{\times}$  that

$$\lim_{\varepsilon \rightarrow 0^+} v_{\varepsilon, \mu^{\times}, \nu^{\times}}^{\times}(s^{\times}) = \Pr_{\mu^{\times}, \nu^{\times}}(s^{\times} \models \varphi^{\times}). \quad (11)$$

Before proving Theorem 1, we provide several lemmas. Throughout this section,  $(\mu^{\times}, \nu^{\times})$  denotes an arbitrary but fixed strategy pair in a given product game  $\mathcal{G}^{\times}$ ; and we omit the subscripts from  $\mathbb{E}_{\pi_{s^{\times}}^{\times} \sim G_{\mu^{\times}, \nu^{\times}}^{\times}}$ ,  $\Pr_{\mu^{\times}, \nu^{\times}}$ ,  $v_{\varepsilon, \mu^{\times}, \nu^{\times}}^{\times}$  and use  $\mathbb{E}$ ,  $\Pr$ ,  $v_{\varepsilon}^{\times}$  instead to simplify our notation.

We first establish bounds on the returns of the paths of  $\mathcal{G}^{\times}$ . In the following lemma, the first inequality (12) states that the returns and thereby the values are always between 0 and 1. The second inequality (13) states that changing the color of a state on a path to a larger odd number increases the return of the

path. Similarly, the third inequality (14) states that changing the color to a larger even number decreases the return.

*Lemma 3:* For a given product game  $\mathcal{G}^{\times}$ , the following hold for any path  $\pi^{\times}$  and any time step  $t \geq 0$

$$0 \leq G_{\varepsilon}^{\times}(\pi^{\times}[t:]) \leq 1 \quad (12)$$

$$G_{\varepsilon}^{\times}(\pi^{\times}[t:]) \leq \varepsilon^{\kappa-k} + (1 - \varepsilon^{\kappa-k}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \quad (13)$$

$$G_{\varepsilon}^{\times}(\pi^{\times}[t:]) \geq (1 - \varepsilon^{\kappa-k}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \quad (14)$$

where  $k = \min\{\kappa, C^{\times}(\pi^{\times}[t]) + 1\}$ .

*Proof:* The lower bound in (12) holds since there is no negative reward. Now, assume that we can change the colors of the states at each time step. In this case, the maximum return can be obtained by assigning the same odd color to all the states, which results in a return  $\varepsilon^{\kappa-k} + \varepsilon^{\kappa-k}(1 - \varepsilon^{\kappa-k}) + \varepsilon^{\kappa-k}(1 - \varepsilon^{\kappa-k})^2 + \dots = 1$ , concluding the upper bound in (12) holds. By definition, for a given color  $k'$ , the return can be written recursively as

$$G_{\varepsilon}^{\times}(\pi^{\times}[t:]) = \begin{cases} \varepsilon^{\kappa-k'} + (1 - \varepsilon^{\kappa-k'}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]), & \text{if } k' \text{ is odd} \\ (1 - \varepsilon^{\kappa-k'}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]), & \text{if } k' \text{ is even.} \end{cases} \quad (15)$$

From (12), (15), and  $\varepsilon \in (0, 1]$ , we obtain (13) and (14) as

$$\begin{aligned} G_{\varepsilon}^{\times}(\pi^{\times}[t:]) &\leq \varepsilon^{\kappa-k'} + (1 - \varepsilon^{\kappa-k'}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \\ &= \varepsilon^{\kappa-k'}(1 - G_{\varepsilon}^{\times}(\pi^{\times}[t+1:])) + G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \\ &\leq \varepsilon^{\kappa-k}(1 - G_{\varepsilon}^{\times}(\pi^{\times}[t+1:])) + G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \\ &= \varepsilon^{\kappa-k} + (1 - \varepsilon^{\kappa-k}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \end{aligned} \quad (16)$$

$$\begin{aligned} G_{\varepsilon}^{\times}(\pi^{\times}[t:]) &\geq (1 - \varepsilon^{\kappa-k'}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \\ &\geq (1 - \varepsilon^{\kappa-k}) G_{\varepsilon}^{\times}(\pi^{\times}[t+1:]) \end{aligned} \quad (17)$$

since  $k' \leq k$ .  $\blacksquare$

The parity condition of a product game is defined over the recurrent states, i.e., the states that are visited infinitely many times, in a product game. In the MC induced by a strategy pair, a state is recurrent if and only if it belongs to a BSCC of the MC as the paths almost surely reach a BSCC and visit its states infinitely often. If the largest color among the colors of the states in a BSCC is an odd number, the BSCC is called *accepting* as all the paths reaching the BSCC satisfy the parity condition. Similarly, if the largest color is an even number, then the BSCC is called *rejecting*. Thus, the probability of satisfying the parity condition can be obtained by computing the probability of reaching an accepting BSCC (see the proofs provided for the Rabin condition in [1]).

In the following lemma, we show that the values of the states in a BSCC approach 1 if the BSCC is accepting and 0 otherwise. Using this lemma, we then prove Theorem 1 by showing that the values approach the probability of reaching accepting BSCCs.

*Lemma 4:* For a BSCC  $\mathcal{V}^{\times} \in \mathcal{B}(\mathcal{G}_{\mu^{\times}, \nu^{\times}}^{\times})$ , and for all  $s^{\times} \in \mathcal{V}^{\times}$ , it holds that

$$\lim_{\varepsilon \rightarrow 0^+} v_{\varepsilon, \mu^{\times}, \nu^{\times}}^{\times}(s^{\times}) = \begin{cases} 1, & \text{if } \mathcal{V}^{\times} \text{ is accepting} \\ 0, & \text{if } \mathcal{V}^{\times} \text{ is rejecting.} \end{cases} \quad (18)$$

*Proof:* To simplify our notation, for a BSCC  $\mathcal{V}^{\times} \in \mathcal{B}(\mathcal{G}_{\mu^{\times}, \nu^{\times}}^{\times})$ , we let  $k_{\mathcal{V}^{\times}}$  denote the largest color among the colors

of the states in  $\mathcal{V}^\times$ , i.e.,

$$k_{\mathcal{V}^\times} := \max \{C^\times(s^\times) \mid s^\times \in \mathcal{V}^\times\} \quad (19)$$

and we let  $r_{\mathcal{V}^\times}$  denote  $\varepsilon^{k_{\mathcal{V}^\times} - k_{\mathcal{V}^\times}}$ .

We first consider the case where  $k_{\mathcal{V}^\times}$  is an odd number; i.e.,  $\mathcal{V}^\times$  is accepting. Let  $s_*^\times \in \mathcal{V}^\times$  be a state colored with  $k_{\mathcal{V}^\times}$ . As  $\mathcal{V}^\times$  is a BSCC, a path  $\pi_{s_*^\times}^\times$ , starting from  $s_*^\times$ , almost surely visits  $s_*^\times$  and the other states in  $\mathcal{V}^\times$  infinitely many times. By Lemma 3, we can obtain a lower bound on  $v_\varepsilon^\times(s_*^\times)$ , the expected value of the return of  $\pi_{s_*^\times}^\times$ , by replacing the colors of the states other than  $s_*^\times$  by  $k_{\mathcal{V}^\times} - 1$ . Let  $N$  be the number of time steps between two consecutive visits to  $s_*^\times$ , then

$$v_\varepsilon^\times(s_*^\times) \geq r_{\mathcal{V}^\times} + (1 - r_{\mathcal{V}^\times}) \mathbb{E} \left[ (1 - \varepsilon r_{\mathcal{V}^\times})^N G_\varepsilon^\times(\pi_{s_*^\times}^\times[N:]) \right].$$

Now, by Jensen's inequality and the Markov property, we can simplify the bound as

$$v_\varepsilon^\times(s_*^\times) \geq r_{\mathcal{V}^\times} + (1 - r_{\mathcal{V}^\times})(1 - \varepsilon r_{\mathcal{V}^\times})^n v_\varepsilon^\times(s_*^\times) \quad (20)$$

where  $n := \mathbb{E}[N]$ . From this inequality and the fact that  $v_\varepsilon^\times(s_*^\times)$ ,  $r_{\mathcal{V}^\times}$ , and  $\varepsilon$  are all between 0 and 1, we further obtain

$$\begin{aligned} v_\varepsilon^\times(s_*^\times) &\geq \frac{r_{\mathcal{V}^\times}}{1 - (1 - r_{\mathcal{V}^\times})(1 - \varepsilon r_{\mathcal{V}^\times})^n} \\ &\geq \frac{r_{\mathcal{V}^\times}}{1 - (1 - r_{\mathcal{V}^\times})(1 - n\varepsilon r_{\mathcal{V}^\times})} \\ &= \frac{1}{1 - n\varepsilon(1 - r_{\mathcal{V}^\times})}. \end{aligned} \quad (21)$$

As  $\varepsilon$  goes to 0,  $n\varepsilon(1 - r_{\mathcal{V}^\times})$  goes to 0 as well, which makes  $v_\varepsilon^\times(s_*^\times)$  go to 1.

Furthermore, we can derive a lower bound on the value of any state  $s^\times \in \mathcal{V}^\times$  in a similar way as

$$v_\varepsilon^\times(s^\times) \geq (1 - \varepsilon r_{\mathcal{V}^\times})^{n'} v_\varepsilon^\times(s_*^\times) \quad (22)$$

where  $n'$  is the expected number of time steps between leaving  $s^\times$  and reaching  $s_*^\times$ . Since  $n'$  is a constant,  $(1 - \varepsilon r_{\mathcal{V}^\times})^{n'}$  along with  $v_\varepsilon^\times(s_*^\times)$  goes to 1 as  $\varepsilon$  goes to 0, making  $v_\varepsilon^\times(s^\times)$  goes to 1. Thus, since the lower bounds go to 1 when  $\varepsilon \rightarrow 0^+$  and the values are bounded above by 1 due to (12), we have that (18) holds when  $\mathcal{V}^\times$  is accepting.

We now consider the case where  $k_{\mathcal{V}^\times}$  is an even number (i.e.,  $\mathcal{V}^\times$  is rejecting). Similarly to the previous case, from Lemma 3, we obtain the following upper bound by changing the colors of the states other than  $s_*^\times$  with  $k_{\mathcal{V}^\times} - 1$ , i.e.,

$$\begin{aligned} v_\varepsilon^\times(s_*^\times) &\leq (1 - r_{\mathcal{V}^\times}) \mathbb{E} \left[ 1 - (1 - \varepsilon r_{\mathcal{V}^\times})^N \right. \\ &\quad \left. + (1 - \varepsilon r_{\mathcal{V}^\times})^N G_\varepsilon^\times(\pi_{s_*^\times}^\times[N:]) \right] \end{aligned} \quad (23)$$

where  $N$  is defined as in the previous case. Again, from Jensen's inequality, the Markov property, and the fact that  $1 - \varepsilon r_{\mathcal{V}^\times}$  and  $1 - r_{\mathcal{V}^\times}$  are between 0 and 1, we obtain that

$$v_\varepsilon^\times(s_*^\times) \leq 1 - (1 - \varepsilon r_{\mathcal{V}^\times})^n + (1 - r_{\mathcal{V}^\times}) v_\varepsilon^\times(s_*^\times) \quad (24)$$

where  $n'$  is expected value of  $N$ . Hence, it holds that

$$v_\varepsilon^\times(s_*^\times) \leq \frac{1 - (1 - \varepsilon r_{\mathcal{V}^\times})^n}{r_{\mathcal{V}^\times}} \leq \frac{n\varepsilon r_{\mathcal{V}^\times}}{r_{\mathcal{V}^\times}} = n\varepsilon. \quad (25)$$

Using this inequality, we obtain the upper bound

$$v_\varepsilon^\times(s^\times) \leq (1 - \varepsilon r_{\mathcal{V}^\times})^{n'} + v_\varepsilon^\times(s_*^\times) \leq (1 - \varepsilon r_{\mathcal{V}^\times})^{n'} + n\varepsilon$$

on any state  $s^\times \in \mathcal{V}^\times$  where  $n'$  is defined as in the previous case. Since  $n$  and  $n'$  are constant, this upper bound goes to 0 as  $\varepsilon$  goes to 0, which, by (12), concludes the proof.  $\blacksquare$

With the above results, we now prove Theorem 1.

*Proof. [Theorem 1]:* The probability that the parity condition is satisfied from a state  $s^\times \in S^\times$ , denoted by  $Pr(s^\times \models \varphi^\times)$ , is equivalent to the probability that a path  $\pi_{s^\times}^\times$ , which starts in  $s^\times$ , reaches an accepting BSCC. Let  $\mathcal{U}_+^\times$  and  $\mathcal{U}_-^\times$  denote the union of all the states belonging to accepting and rejecting BSCCs, respectively; then we can express the value as

$$\begin{aligned} v_\varepsilon^\times(s^\times) &= \mathbb{E} [G_\varepsilon^\times(\pi_{s^\times}^\times) \mid \pi_{s^\times}^\times \models \diamond \mathcal{U}_+^\times] Pr(s^\times \models \varphi^\times) \\ &\quad + \mathbb{E} [G_\varepsilon^\times(\pi_{s^\times}^\times) \mid \pi_{s^\times}^\times \models \diamond \mathcal{U}_-^\times] Pr(s^\times \not\models \varphi^\times). \end{aligned} \quad (26)$$

Let  $m_+$  denote the expected number of time steps until an accepting BSCC is reached. Using Lemma 3, we can obtain a lower bound on the value by changing the colors of all transient states to the largest even number; i.e.,

$$v_\varepsilon^\times(s^\times) \geq (1 - \varepsilon)^{m_+} v_\varepsilon^\times(\mathcal{U}_+^\times) Pr(s^\times \models \varphi^\times) \quad (27)$$

where  $v_\varepsilon^\times(\mathcal{U}_+^\times) := \min_{s^\times \in \mathcal{U}_+^\times} v_\varepsilon^\times(s^\times)$ . From Lemma 4, this lower bound goes to  $Pr(s^\times \models \varphi^\times)$  as  $\varepsilon$  goes to 0.

Similarly, let  $m_-$  denote the expected number of time steps until a rejecting BSCC is reached. As changing the colors of all transient states to the largest odd number increases the return due to Lemma 3, the following upper bound on the value:

$$\begin{aligned} v_\varepsilon^\times(s^\times) &\leq (1 - \varepsilon r_{\mathcal{V}^\times})^{m_-} \overline{v}_\varepsilon^\times(\mathcal{U}_-^\times) Pr(s^\times \not\models \varphi^\times) \\ &\quad + Pr(s^\times \models \varphi^\times) \end{aligned} \quad (28)$$

holds, where  $\overline{v}_\varepsilon^\times(\mathcal{U}_-^\times) := \max_{s^\times \in \mathcal{U}_-^\times} v_\varepsilon^\times(s^\times)$ ; this bound goes to 0 as  $\varepsilon$  goes to 0 due to Lemma 4.

Since both the lower bound (27) and the upper bound (28) go to  $Pr(s^\times \models \varphi^\times)$  as  $\varepsilon \rightarrow 0^+$ , it follows that  $v_\varepsilon^\times(s^\times)$  goes to  $Pr(s^\times \models \varphi^\times)$  as  $\varepsilon \rightarrow 0^+$ , concluding the proof.  $\blacksquare$

#### IV. LAZY COLOR GENERATION

In this section, we introduce a framework that efficiently uses the rewards and discount factors established in the previous section to learn optimal strategies. For a given number of colors  $\kappa$ , the functions in (7) and (8) provide  $\kappa$  distinct rewards and discount factors. Larger discount factors are used in the states with smaller colors so that visiting such states affects the return less than the ones with larger colors. However, if  $\kappa$  is large, the discount factors could be too large for the smaller colors. For example, for  $\varepsilon = 0.01$  and  $\kappa = 7$ , the discount factor for the color 0 is very close to 1 (i.e.,  $1 - 10^{-14}$ ), and thus, would extremely slowdown learning convergence.

To address this problem, we modify the design of the product game; we start with a product game having only a single color and increase the number of colors only when a new color is needed. The key idea is that if the large colors are not necessary to learn an optimal strategy, then there is no need to specify distinct reward and discount factors for them; thus, smaller discount factors can be also used for small colors.

*Definition 5:* A multilevel product game (MPG)  $\mathcal{G}^*$  of a product game  $\mathcal{G}^\times$  is a tuple  $\mathcal{G}^* = (\kappa, S^*, (S_\mu^*, S_\nu^*), s_0^*, A^*, P^*, R^*, \gamma^*)$  where  $\kappa$  is the number of colors (levels) and  $[\kappa] := \{1, \dots, \kappa\}$ ;  $S^* = S^\times \times [\kappa]$  is the set of multilevel

product states;  $S_\mu^* = S_\mu^{\times} \times [\kappa]$  and  $S_\nu^* = S_\nu^{\times} \times [\kappa]$  are the controller and adversary multilevel product states respectively;  $s_0^* = \langle s_0^{\times}, 1 \rangle$  is the initial state;  $A^* = A^{\times}$  is the set of actions;  $P^*: S^* \times A^* \times S^* \mapsto [0, 1]$  is the probabilistic transition function such that  $P^*(\langle s^{\times}, \bar{\kappa} \rangle, a^*, \langle s^{\times'}, \bar{\kappa}' \rangle) :=$

$$P^*(s^{\times}, a^*, s^{\times'}) \cdot \begin{cases} \tau_\varepsilon, & \text{if } \bar{\kappa}' > \bar{\kappa} \text{ and } C^*(s^{\times}) = \bar{\kappa}' - 1 \\ 1 - \tau_\varepsilon, & \text{if } \bar{\kappa}' = \bar{\kappa} \text{ and } C^*(s^{\times}) \geq \bar{\kappa} \\ 1, & \text{if } \bar{\kappa}' = \bar{\kappa} \text{ and } C^*(s^{\times}) < \bar{\kappa} \\ 0, & \text{otherwise} \end{cases} \quad (29)$$

where  $\tau_\varepsilon$  is a function of  $\varepsilon$  satisfying  $\lim_{\varepsilon \rightarrow 0^+} \frac{\varepsilon}{\tau_\varepsilon} = 0$  and  $R^*: S^* \mapsto [0, 1]$  and  $\gamma^*: S^* \mapsto [0, 1]$  are the reward and discount functions, respectively, such that

$$R_\varepsilon^*(\langle s^{\times}, \bar{\kappa} \rangle) := \begin{cases} \varepsilon^{\bar{\kappa} - C^*(\langle s^{\times}, \bar{\kappa} \rangle)}, & \text{if } C^*(\langle s^{\times}, \bar{\kappa} \rangle) \text{ is odd} \\ 0, & \text{if } C^*(\langle s^{\times}, \bar{\kappa} \rangle) \text{ is even} \end{cases}$$

$$\gamma_\varepsilon^*(\langle s^{\times}, \bar{\kappa} \rangle) := 1 - \varepsilon^{\bar{\kappa} - C^*(\langle s^{\times}, \bar{\kappa} \rangle)} \quad (30)$$

where  $C^*(\langle s^{\times}, \bar{\kappa} \rangle) := \min\{C^*(s^{\times}), \bar{\kappa} - 1\}$ .

The MPG  $\mathcal{G}^*$  of a product game  $\mathcal{G}^{\times}$  has a copy of  $\mathcal{G}^{\times}$  for each level  $\bar{\kappa} \in \{1, \dots, \kappa\}$ . In the  $\bar{\kappa}$ th level, the reward and discount factors are specified as if  $\bar{\kappa}$  is the total number of colors and  $\bar{\kappa} - 1$  is the largest color; thus, the colors greater than  $\bar{\kappa} - 1$  in the  $\bar{\kappa}$ th level copy of  $\mathcal{G}^{\times}$  are truncated to  $\bar{\kappa} - 1$ . This results in smaller discount factors and larger rewards for lower levels, thereby speeding up learning.

In addition, as long as the colors of the states visited infinitely often are less than  $\bar{\kappa}$  in the  $\bar{\kappa}$ th level, the return of a path approaches the probability of satisfying the parity condition as  $\varepsilon$  goes to 0. This is because the effect of the states visited only for a finite number of times on the return vanishes as  $\varepsilon$  goes to 0, as shown in the proof of Theorem 1. Therefore, truncating the larger colors to  $\bar{\kappa} - 1$  does not change the return in the limit. However, this does not hold if a state with a color  $k \geq \bar{\kappa}$  is visited infinitely often as it might alter the satisfaction of the parity condition. In such cases, a transition to the  $(k+1)$ th level needs to be made to reflect the parity condition.

The main challenge here is that since the transition model of the initial SG  $\mathcal{G}$  (and thus,  $\mathcal{G}^{\times}$  and  $\mathcal{G}^*$ ) is *unknown*, it is not possible to determine in advance whether a state with a color  $k \geq \bar{\kappa}$  is visited infinitely often under a strategy pair. Consequently, our key idea is to allow for probabilistic transitions between levels so that repeated visits can lead to upper levels. Specifically, the transitions of the multilevel game  $\mathcal{G}^*$  are constructed in (29) in a way that it is impossible to make a transition from the  $\bar{\kappa}$ th level to the  $\bar{\kappa}'$ th level if  $\bar{\kappa}' < \bar{\kappa}$ , and a transition from a state  $s^{\times}$  in the  $\bar{\kappa}$ th level to the  $\bar{\kappa}'$ th level can happen w.p.  $\tau_\varepsilon$  (i.e., stays in the same level w.p.  $1 - \tau_\varepsilon$ ) only if  $\bar{\kappa}' > \bar{\kappa}$  and  $C^*(s^{\times}) = \bar{\kappa}'$ . Such probabilistic transitions ensure that repeatedly visiting a state with a color  $k \geq \bar{\kappa}$  in the  $\bar{\kappa}$ th level almost surely results in a transition to the level where the color  $k$  is the largest color.

This captures our goal that, in order to accelerate learning, *the larger colors in the parity condition should be considered only when needed*. If the first  $\bar{\kappa}$  colors are sufficient to reflect the parity condition for the strategies being followed,  $\mathcal{G}^*$  stays in the  $\bar{\kappa}$ th level and uses only  $\bar{\kappa}$  distinct rewards and discount factors; and if they are not,  $\mathcal{G}^*$  eventually moves to an upper level where more colors are considered to provide a sufficient number of

distinct rewards and discount factors. However, the transitions to the lower levels are not allowed to avoid redundant circular transitions between the levels as they can possibly distort the returns.

Now, any pure memoryless strategy pair  $(\mu^*, \nu^*)$  in the MPG  $\mathcal{G}^*$  of a product game  $\mathcal{G}^{\times}$  induces a finite-memory strategy pair  $(\mu^{\times}, \nu^{\times})$  in  $\mathcal{G}^{\times}$ , where the levels of  $\mathcal{G}^*$  act as modes, i.e., whenever  $\mathcal{G}^*$  transitions to an upper level, the strategies  $\mu^{\times}$  and  $\nu^{\times}$  switch to the corresponding mode. Hence, we below capture the result that enables the use of MPGs; we show that under a strategy pair, the values of the states in  $\mathcal{G}^*$  approach the probability of satisfying the parity condition in  $\mathcal{G}^{\times}$  as  $\varepsilon$  goes to 0.

**Theorem 2:** For the MPG  $\mathcal{G}^*$  of a product game  $\mathcal{G}^{\times}$ , let  $(\mu^*, \nu^*)$  be an arbitrary pure memoryless strategy pair in  $\mathcal{G}^*$  and let  $(\mu^{\times}, \nu^{\times})$  be its induced strategy pair in  $\mathcal{G}^{\times}$ . Then, it holds that

$$\lim_{\varepsilon \rightarrow 0^+} v_{\varepsilon, \mu^*, \nu^*}^*(\langle s^{\times}, \bar{\kappa} \rangle) = \Pr_{\mu^{\times}, \nu^{\times}}(s^{\times} \models \varphi^{\times}) \quad (31)$$

for any  $s^{\times} \in S^{\times}$  and  $\bar{\kappa} \in \{1, \dots, \kappa\}$ .

**Proof:** A BSCC of the induced MC  $\mathcal{G}_{\mu^*, \nu^*}^*$  cannot contain states belonging to the different levels since transitions from an upper level to a lower level are not permitted. Moreover, a BSCC in the  $\bar{\kappa}$ th level cannot have a state whose color is greater than or equal to  $\bar{\kappa}$  because otherwise a transition to an upper level eventually happens, conflicting with the BSSC definition. This means that none of the colors in a BSCC is truncated, i.e., there is a distinct reward and a discount factor for each color. Thus, from Lemma 4, all the values in a BSCC approach either 1 or 0 depending on whether the BSCC is accepting or rejecting.

We can now complete the proof by following the same steps as in the proof of Theorem 1. The only difference is that the number of steps before reaching a BSCC might depend on the parameter  $\varepsilon$  and  $\kappa$  in  $\mathcal{G}_{\mu^*, \nu^*}^*$ . However, since  $\tau_\varepsilon$  approaches 0 slower than  $\varepsilon$ , both  $(1 - \varepsilon)^{\kappa m_+ / \tau_\varepsilon}$  and  $(1 - \varepsilon)^{\kappa m_- / \tau_\varepsilon}$  go to 1 as  $\varepsilon$  goes to 0, the same results hold. ■

In the MPGs, the probability of transitioning to an upper level is controlled by the parameter  $\tau_\varepsilon$  [see (29)]. A smaller  $\tau_\varepsilon$  encourages the controller and the adversary to learn optimal strategies utilizing a fewer number of colors. For many SGs for which such optimal strategies exist, this can result in faster convergence since the discount factors to be used are smaller, as we demonstrate in our case studies. In addition, the space complexity of the MPGs is *only linear* in the total number of colors, and any model-free RL algorithm is guaranteed to converge to an optimal controller strategy maximizing the minimum satisfaction probability for a sufficiently small  $\varepsilon$ , as formalized below.

**Theorem 3:** For the MPG  $\mathcal{G}^*$  constructed from the product game  $\mathcal{G}^*$  of a given SG  $\mathcal{G}$  and a given LTL specification  $\varphi$ , there exists  $\varepsilon' > 0$  such that for any  $\varepsilon \in (0, \varepsilon']$ , a model-free RL algorithm, such as minimax-Q, converges to a pure memoryless controller strategy  $\mu_*^*$  in  $\mathcal{G}^*$ , and its induced strategy  $\mu_\varphi$  maximizes the minimum probability of satisfying  $\varphi$  in  $\mathcal{G}$ , i.e.,  $\mu_\varphi$  satisfies (1).

**Proof:** Since there are only finitely many different pure memoryless strategies that can be followed in an MPG, the number of possible different satisfaction probabilities that can be obtained by them is also finite. Thus, the absolute differences between the

**Algorithm 1:** Model-free RL for LTL tasks.

---

**Input:** LTL formula  $\varphi$ , unknown SG  $\mathcal{G}$ , parameter  $\varepsilon$   
**Output:** optimal controller strategy  $\mu_\varphi$

Translate  $\varphi$  to a DPA  $\mathcal{A}_\varphi$   
Construct a product game  $\mathcal{G}^\times$  by composing  $\mathcal{G}$  and  $\mathcal{A}_\varphi$   
Construct the MPG  $\mathcal{G}^*$  from  $\mathcal{G}^\times$   
Initialize  $Q(s^*, a)$  for each state  $s^*$  and action  $a$   
**for** [ **do** # for each episode  $i = 0$  **to**  $\mathcal{T} - 1$   
  **for** [ **do** # for each time step  $t = 0$  **to**  $\mathcal{T} - 1$   
    Derive an  $e$ -greedy strategy pair  $(\mu^*, \nu^*)$  from  $Q$   
    Take the action  $a_t \leftarrow \begin{cases} \mu^*(s_t^*), & s_t^* \in S_\mu^* \\ \nu^*(s_t^*), & s_t^* \in S_\nu^* \end{cases}$   
    Observe the next state  $s_{t+1}^*$   
     $Q(s_t^*, a_t) \leftarrow (1 - \alpha_{i,t})Q(s_t^*, a_t) + \alpha_{i,t}R^*(s_t^*)$   
       $+ \alpha_{i,t} \gamma^*(s_t^*) \cdot \begin{cases} \max_{a'} Q(s_{t+1}^*, a'), & s_{t+1}^* \in S_\mu^* \\ \min_{a'} Q(s_{t+1}^*, a'), & s_{t+1}^* \in S_\nu^* \end{cases}$   
  **end for**  
**end for**  
Derive a greedy controller strategy  $\mu_*^*$  from  $Q$   
Obtain the induced controller strategy  $\mu_\varphi$  for  $\mathcal{G}$  from  $\mu_*^*$  from  $Q$

---

satisfaction probabilities for different strategy pairs are either 0 or greater than for some  $d > 0$ . Hence, from Theorem 2, for the MPG  $\mathcal{G}^*$  of the product game  $\mathcal{G}^\times$  constructed from an SG  $\mathcal{G}$  and an LTL task  $\varphi$ , there exists  $\varepsilon' > 0$  such that for all  $\varepsilon \in (0, \varepsilon']$ , it holds that

$$v_{\varepsilon, \mu_1^*, \nu_1^*}^*((s^\times, \bar{\kappa})) > v_{\varepsilon, \mu_2^*, \nu_2^*}^*((s^\times, \bar{\kappa})) \\ \iff Pr_{\mu_1^*, \nu_1^*}(s^\times \models \varphi^\times) > Pr_{\mu_2^*, \nu_2^*}(s^\times \models \varphi^\times) \quad (32)$$

for any two pure memoryless strategy pairs  $(\mu_1^*, \nu_1^*)$  and  $(\mu_2^*, \nu_2^*)$  in  $\mathcal{G}^*$  and their induced strategy pairs  $(\mu_1^\times, \nu_1^\times)$  and  $(\mu_2^\times, \nu_2^\times)$  in  $\mathcal{G}^\times$ .

As a result, a memoryless controller strategy  $\mu_*^*$  maximizing the minimum expected return in  $\mathcal{G}^*$  induces a finite-memory strategy  $\mu_\varphi^\times$  maximizing the minimum probability of satisfying the parity condition in  $\mathcal{G}^\times$ . Additionally, from Lemmas 1 and 2,  $\mu_\varphi^\times$  further induces a finite-memory strategy  $\mu_\varphi$  in  $\mathcal{G}$  that maximizes the probability of satisfying  $\varphi$  in the worst case. Finally, since all the discount factors are strictly less than 1 and all the rewards are Markovian, under regular conditions on the learning rate and the exploration strategy, a model-free RL algorithm such as minimax-Q [42] is guaranteed to converge to such optimal strategies. We note that using state-dependent discount factors does not violate the convergence conditions of minimax-Q. The largest discount factor in (30), i.e.,  $(1 - \varepsilon^{\kappa-1}) < 1$ , can be used as the contraction index in the convergence proofs (e.g., see [44] and [45]).

Our result also suggests a PAC algorithm based on the minimum transition probability as discussed in [46]. However, the convergence rate of such an algorithm could be arbitrarily slow depending on the minimum transition probability in the SG. In cases where multiple optimal controller strategies exist, the converged strategies tend to satisfy the specifications as quickly as possible due to discounting.  $\blacksquare$

The overall approach is summarized in Algorithm 1, which takes as input an LTL formula  $\varphi$ , an SG  $\mathcal{G}$ , and a parameter  $\varepsilon$ . The algorithm first translates  $\varphi$  into a DPA  $\mathcal{A}_\varphi$  and composes it with  $\mathcal{G}$  to construct the product game  $\mathcal{G}^\times$ . Then, it obtains the MPG  $\mathcal{G}^*$  of  $\mathcal{G}^\times$ . Since  $\mathcal{G}$  is unknown, the transition probabilities and the graph topologies of  $\mathcal{G}$ ,  $\mathcal{G}^\times$ , and  $\mathcal{G}^*$  are unknown.

In each episode, Algorithm 1 starts in the initial state and follows a  $e$ -greedy strategy pair, under which a random action is taken w.p.  $e$  and a greedy action is taken w.p.  $1 - e$ . At each time step  $t$ , after taking an action  $a$ , the next state  $s_{t+1}^*$  is observed and  $Q(s_t^*, a)$  is updated according to the minimax-Q algorithm [42] using a properly decreasing learning rate  $\alpha_{t,i}$ . Each episode terminates after  $\mathcal{T}$  time steps, and after  $\mathcal{T}$  episodes, a greedy controller strategy is derived from the Q-values, then used to induce the controller strategy for  $\mathcal{G}$ .

## V. THREE-COLOR APPROXIMATION

In this section, we provide a modified version of the MPGs for the scenarios where the controller can learn optimal strategies by eventually focusing only on a single odd color. The idea here is that controller nondeterministically decides to focus on an odd color  $k$ , and once the controller decides, the colors smaller than  $k$  are replaced with the color 0, the colors larger than  $k$  are replaced with the color 2, and finally, the color  $k$  is replaced with the color 1. This results in three distinct reward and discount factors regardless of the actual number of colors; thus, improving the convergence rate in the worst case compared to the lazy color generation.

Such an approach, however, may not yield optimal strategies if the adversary can observe and make use of the controller's decision. Nonetheless, in many scenarios, if the controller carefully chooses when to decide to ensure that, after the decision, the adversary cannot follow a strategy under which a state with a color larger than the focused color is repeatedly visited, the adversary cannot benefit from observing the decision.

To reflect this approach, we modify the MPGs as follows.

*Definition 6:* A three-color multilevel product game (TMPG)  $\mathcal{G}^*$  of a product game  $\mathcal{G}^\times$  is a tuple  $\mathcal{G}^* = (\kappa, S^*, (S_\mu^*, S_\nu^*), s_0^*, A^*, P^*, R^*, \gamma^*)$  where  $\kappa$  is the number of colors and  $[\kappa]_{\text{odd}} := \{k \in [\kappa] \mid k \text{ is odd}\}$ ;  $S^* = S^\times \times [\kappa]_{\text{odd}}$  is the set of multilevel product states;  $S_\mu^* = S_\mu^\times \times [\kappa]_{\text{odd}}$  and  $S_\nu^* = S_\nu^\times \times [\kappa]_{\text{odd}}$  are the controller and adversary multilevel product states respectively;  $s_0^* = \langle s_0^\times, 1 \rangle$  is the initial state;  $A^* = A^\times \cup \{\beta_i \mid i > 1 \text{ and } i \in [\kappa]_{\text{odd}}\}$  is the set of actions;  $P^*: S^* \times A^* \times S^* \mapsto [0, 1]$  is the probabilistic transition function such that  $P^*(\langle s^\times, \bar{\kappa} \rangle, a^*, \langle s^{\times'}, \bar{\kappa}' \rangle) :=$

$$\begin{cases} P^\times(s^\times, a^\times, s^{\times'}), & \text{if } \bar{\kappa}' = \bar{\kappa} \text{ and } a^\times \in A^\times(s^\times) \\ 1, & \text{if } \bar{\kappa}' > \bar{\kappa}, a^* = \beta_{\bar{\kappa}} \text{ and } s^{\times'} = s^\times \in S_\mu^* \\ 0, & \text{otherwise} \end{cases}$$

and  $R^*: S^* \mapsto [0, 1]$  and  $\gamma^*: S^* \mapsto [0, 1]$  are the reward and discount functions, respectively, such that

$$R_\varepsilon^*(\langle s^\times, \bar{\kappa} \rangle) := \begin{cases} \varepsilon^2, & \text{if } C^\times(s^\times) = \bar{\kappa} \\ 0, & \text{otherwise} \end{cases} \quad (33)$$

$$\hat{\gamma}_\varepsilon^*(\langle s^\times, \bar{\kappa} \rangle) := \begin{cases} 1 - \varepsilon, & \text{if } C^\times(s^\times) > \bar{\kappa} \\ 1 - \varepsilon^2, & \text{if } C^\times(s^\times) = \bar{\kappa} \\ 1 - \varepsilon^3, & \text{if } C^\times(s^\times) < \bar{\kappa}. \end{cases} \quad (34)$$

The TMPG  $\mathcal{G}^*$  of the product game contains a copy of the product game  $\mathcal{G}^\times$  for each odd color. The game starts in the first level, and in the  $\bar{\kappa}$ th level, the controller can nondeterministically take a  $\beta_{\bar{\kappa}'}$  action to transition to the  $\bar{\kappa}'$ th level where  $\bar{\kappa}'$  is an odd number larger than  $\bar{\kappa}$ . In the  $\bar{\kappa}$ th level, there are three distinct discount factors  $1 - \varepsilon$ ,  $1 - \varepsilon^2$ , and  $1 - \varepsilon^3$  for the colors larger than  $\bar{\kappa}$ , the color  $\bar{\kappa}$ , and the colors smaller than  $\bar{\kappa}$ , respectively. The controller's nondeterministic decision to focus on a particular odd color is represented by taking a  $\beta$ -action. The restriction in a TMPG is that after focusing on an odd color, the controller can only make a decision to focus on a larger color. This prevents cyclic transitions between the levels of the TMPG.

We now show that the values of the states in a TMPG provide a lower bound for the satisfaction probabilities.

**Theorem 4:** For the TMPG  $\mathcal{G}^*$  of a product game  $\mathcal{G}^\times$ , let  $(\mu^*, \nu^*)$  be an arbitrary pure memoryless strategy pair in  $\mathcal{G}^*$  and let  $(\mu^\times, \nu^\times)$  be its induced strategy pair in  $\mathcal{G}^\times$ . Then

$$\lim_{\varepsilon \rightarrow 0^+} \hat{v}_{\varepsilon, \mu^*, \nu^*}^*(\langle s^\times, \bar{\kappa} \rangle) \leq \Pr_{\mu^\times, \nu^\times}(s^\times \models \varphi^\times) \quad (35)$$

for any  $s^\times \in S^\times$  and  $\bar{\kappa} \in [\kappa]_{\text{odd}}$ .

**Proof:** Any path under any strategy pair in a TMPG eventually reaches a level where it stays forever because there is no transition from an upper level to a lower level. As  $\varepsilon$  goes to 0, the return of a path approaches 1 only if the states with the color  $\bar{\kappa}$  in the  $\bar{\kappa}$ th level are visited infinitely many times, and approaches 0 otherwise. This can be easily shown by following similar steps as in the proof of Theorem 1. Since visiting a state colored with an odd number infinitely often while visiting the states with larger colors only finitely many times satisfies the parity condition, the values of the states provide a lower bound of the satisfaction probabilities as  $\varepsilon$  goes to 0. ■

Theorem 4 implies that a controller strategy maximizing the minimum values in a TMPG achieves a lower bound on the satisfaction probabilities against an optimal adversary. This lower bound can be informally described as the probability of reaching a component in the game where the parity condition is satisfied, and the adversary cannot change the largest color among the colors of the states visited infinitely many times. In many SGs, such components exist and can be almost surely reached, making the lower bound 1, and thus, the optimal controller strategy for the TMPG induces a controller strategy in the original SG that is optimal for the given LTL specification as we show in a case study in Section VI.

## VI. CASE STUDIES

In this section, we illustrate the applicability of our approach in several case studies focused on robot navigation and control of a robotic arm. We consider different LTL tasks, larger grids, continuous state-spaces, continuous action spaces, different levels of stochasticity, and adversarial capabilities to evaluate the effectiveness and scalability of the methods. In particular, we compared the performance of our methods with the method from Hahn et al. [39], which we refer to as the augmented product game (APG) method; specifically, we evaluated our reduction

[referred to as product game (PG)], lazy color generation (referred to as MPG), and three-color approximation (referred to as TMPG) methods. We did not compare with our preliminary methods from [40] as those do not support general LTL formulas.

### A. Robot Navigation

We start with three path-planning case studies where a robot needs to learn to navigate to perform given tasks in planar environments.

**1) Active Adversaries:** In the first case study, the environments are modeled as grids. In these grids, each cell represents a state, and a robot can move from one cell to a neighbor cell by taking four actions: *up*, *down*, *right*, and *left*. An adversary can observe the actions the robot takes and can act to disturb the movement so that the robot may move in a perpendicular direction of the intended direction.

Specifically, there are four actions the adversary may choose *none*, *cw*, *ccw*, *both*; for *none*, the robot will move in the intended direction w.p. 1; for *cw* (or *ccw*), the robot will move in the intended direction w.p. 0.8, and move in the perpendicular direction that is 90° clockwise (or counter-clockwise, respectively) w.p. 0.2; for *both*, the robot will move in any of the perpendicular directions with w.p. 0.1. The robot cannot leave a trap cell and if the robot attempts to move towards an obstacle or a grid edge, the robot does not move and stays in the previous cell.

Fig. 2 captures the considered scenarios. The trap cells are represented as large empty circles, and the obstacles are represented as large circles filled with gray. The labels of the cells (i.e., states) are displayed as encircled letters filled with various colors.

The core navigation task of the robot is to enter the workspace (*w*) and stay there while repeatedly charging (*c*) and monitoring the assigned region (*r*). Alternatively, the robot can go to and stay in a charging station (*c*). In addition, the robot must always avoid the danger zone (*d*). This entire task can be formally captured as the LTL formula

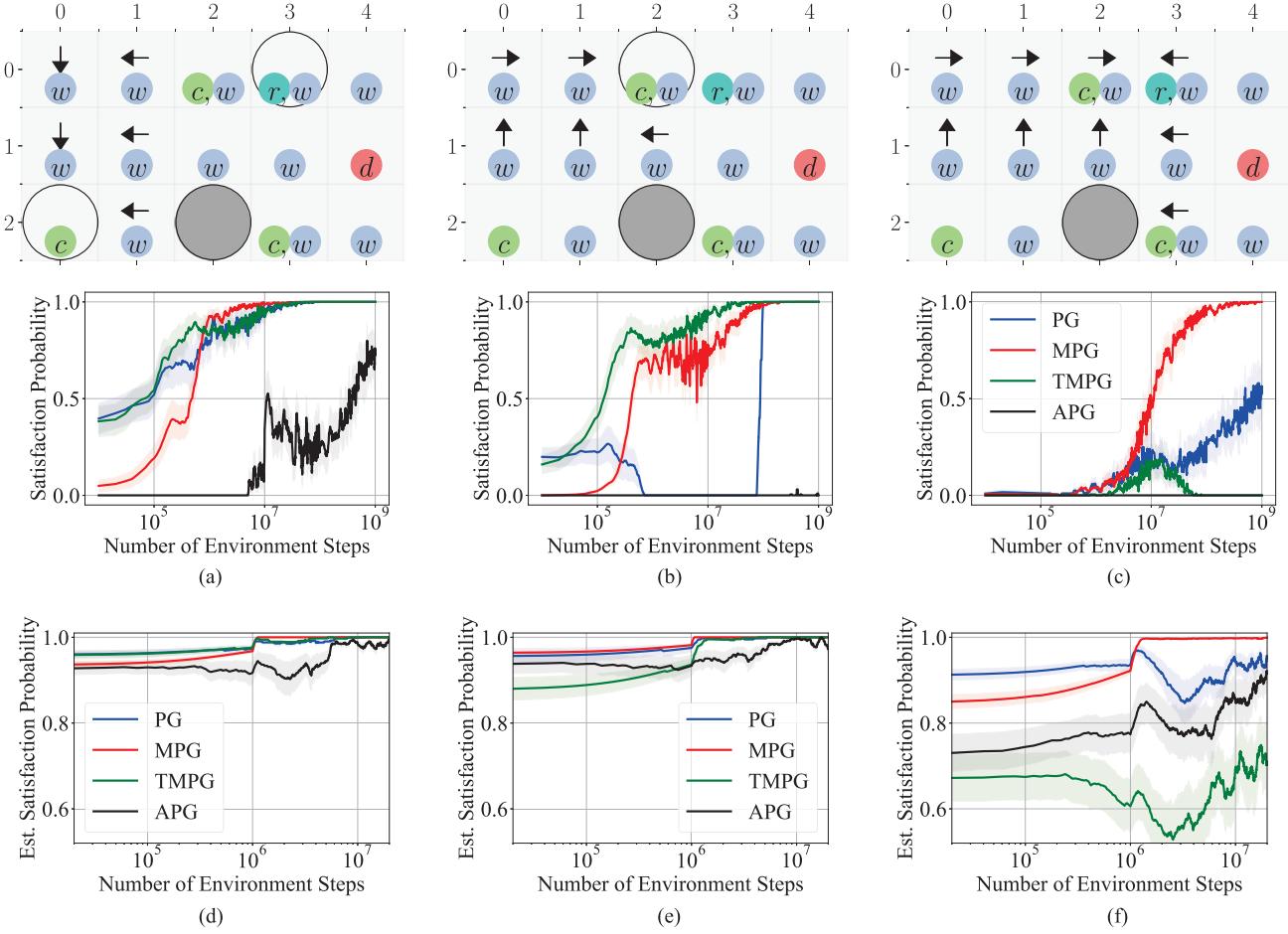
$$\varphi_1 := ((\Diamond \Box w \wedge \Box \Diamond c \wedge \Box \Diamond r) \vee \Diamond \Box c) \wedge \Box \neg d. \quad (36)$$

Fig. 2 illustrates the three ways of performing the task considered in this case study:

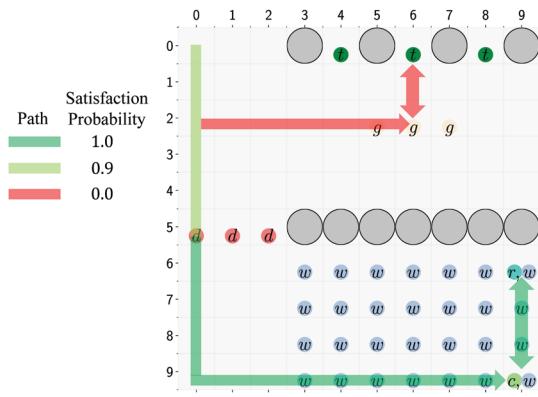
- I) reaching and staying in the charging station at (2, 0), which is not in the workspace;
- II) reaching and staying in the charging station at (0, 2) in the workspace;
- III) visiting repeatedly the charging station at (0, 2) and the assigned region at (0, 3) without leaving the workspace.

We used *Owl* [47] to automatically translate LTL specifications into DPAs. The DPA translated from  $\varphi_1$  has four states, and five colors (see Fig. 6). The largest colors among the colors of the DPA transitions made infinitely often while performing (I), (II), and (III) are 3, 1, and 3, respectively.

We used minimax-Q [42] to learn the optimal controller strategies. We set  $\varepsilon$  to 0.01 and  $\tau_\varepsilon$  to  $\sqrt{\varepsilon} = 0.1$ . During the learning phase, an *e*-greedy strategy was followed, which chooses random actions w.p. *e* and chooses greedily otherwise. We initially set *e* = 1 and gradually decrease it to 0.1 to encourage exploration as both the controller and the adversary need to actively search for optimal strategies against each other. We gradually decreased the learning rate  $\alpha$  with the number of steps from 0.5



**Fig. 2.** Environments and the obtained learning curves for the case studies in Sections VI-A1 and VI-A2. The shaded regions are the quarter of the standard deviations. The results are smoothed by moving averages for better visualization. In the environments, the encircled letters are labels, the empty circles are trap cells; the filled circles are obstacles, and the arrows are optimal controller actions. The blue, red, green, and black learning curves are for the PG, MPG, TMPG, and APG methods. (a) Environment I (Discrete). (b) Environment II (Discrete). (c) Environment III (Discrete). (d) Environment I (Continuous). (e) Environment II (Continuous). (f) Environment III (Continuous).



**Fig. 3.** Environment for the case study in Section VI-A3. The encircled letters are labels, the empty circles are trap cells; the filled circles are obstacles. The arrows represent two ways of satisfying the LTL task  $\varphi_2$  from (37) where the colors of the arrows represent the satisfaction probabilities.

to 0.05. Each episode started in the cell at  $(0, 0)$  and terminated after a  $\mathcal{T} = 10^3$  time steps. We evaluated the performance of the learned controller strategies against the optimal adversary strategies. We formally derived the optimal adversary strategies

and the associated minimum satisfaction probabilities from the full specification of the underlying SG using the PRISM model checking tool [48]. The learning curves of the methods for each environment are shown in the second row of Fig. 2 where the probabilities are averaged over 10 simulations.

In the first environment, learning to perform (I) is a simple task as the cell  $(2, 0)$  containing the charging station aimed to be reached is a trap cell, and the corresponding color 3 is large (yielding smaller discount factors) relative to the largest color 4. Thus, MPG and TMPG do not have any advantage over PG. However, the performance of MPG and TMPG is not significantly less than the one of PG despite the increased space complexity as shown in Fig. 2(a). APG performs poorly compared to these methods as the rewards provided by APG are very sparse.

In the second environment, learning to perform (II) is harder since the corresponding color is 1, resulting in large discount factors. TMPG outperforms the other methods as (II) can be performed by merely focusing on the color 3. The performance of MPG is significantly better than the ones of PG and APG as the task can be performed in the second level. However, the performance is slightly less than TMPG because of the transitions to the upper levels that can happen with a small

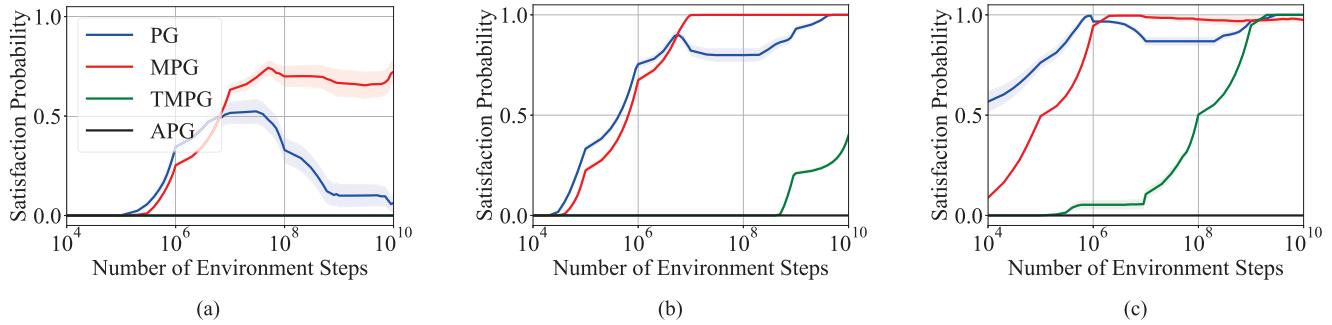


Fig. 4. Learning curves for the case study in Section VI-A3. The shaded regions are the quarter of the standard deviations, and the results are smoothed by moving averages for better visualization. (a) Start Cell: (0, 0). (b) Start Cell: (9, 0). (c) Start Cell: (9, 9).

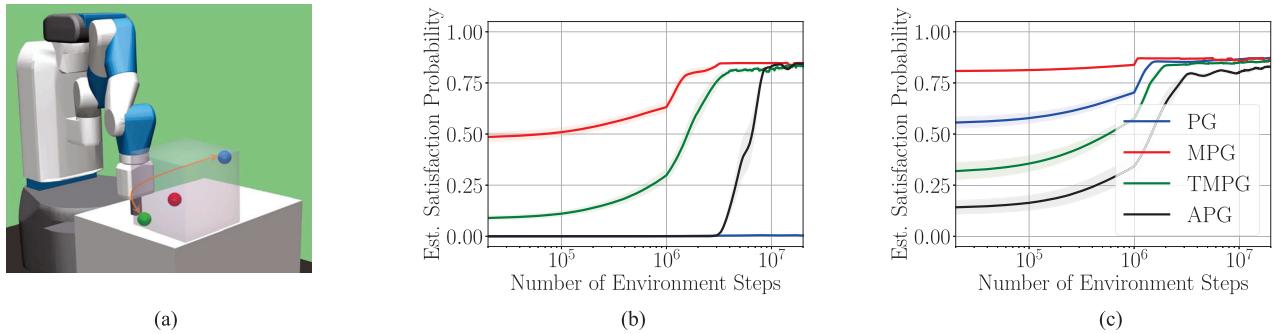


Fig. 5. Environment and the obtained learning curves for the robotic arm tasks in Section VI-B. The shaded regions are the quarter of the standard deviations, and the results are smoothed by moving averages for better visualization. (a) Robot trying to repeatedly reach the green and the blue balls while staying in the rectangular boundary and avoiding the red ball placed by the adversary. (b) Task 1. (c) Task 2.

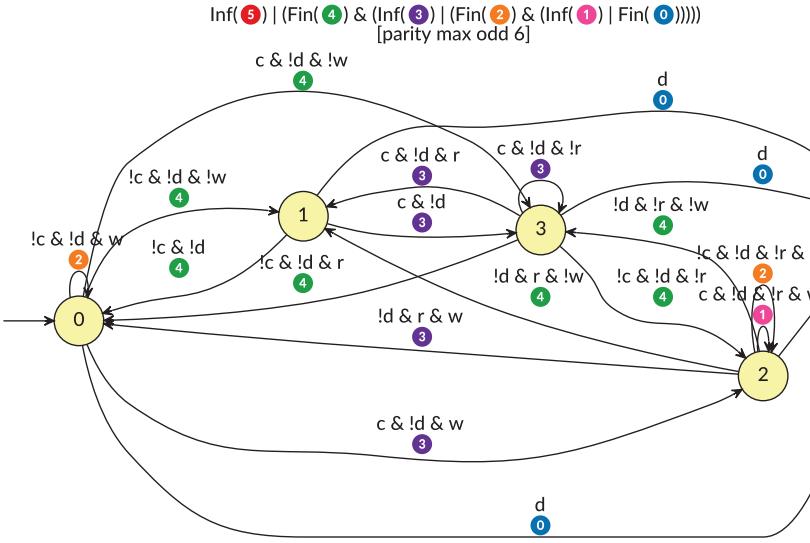


Fig. 6. DPA constructed from the LTL task specification  $\varphi_1$  from (36) using *Spot* [57].

probability. PG could not converge until  $10^8$  steps due to the large discount factors. APG did not converge even with  $10^9$  steps as the probability of getting a positive reward is  $10^{-8}$ .

In the third environment, learning to perform (III) is considerably harder as two cells need to be repeatedly visited without reaching the danger zone in the presence of adversarial actions. The controller must take the action *left* in the cells (0, 3), (1, 3),

and (2, 3) because otherwise, the adversary can eventually pull the robot to the danger zone. MPG significantly outperforms the other methods as shown in Fig. 2(c), and PG is slowly converging. TMPG failed to converge as this task requires two colors, 3 and 1 to be focused, corresponding to (II), and the reaching and staying in the cell (2, 3). If the controller focuses on 3, the adversary can pull the robot to the cell (2, 3), which

requires focusing on 1, and if the controller focuses on 1, the adversary can pull the robot out of (2, 3). APG, in this case as well, failed to converge.

Overall, our MPG method outperforms our methods PG and TMPG, as well as the existing method APG for a general LTL task. TMPG performs better than MPG if the task can be performed by focusing on only one color; however, it could perform poorly otherwise. Finally, the existing APG method is significantly outperformed by our methods due to its prohibitively sparse rewards.

**2) Continuous Environments:** In this case study, the environments are the continuous version of the environments in the first case study. The state space is defined as the continuous (y,x)-position of the robot where the origin is at the top left corner and the positive directions of y and x are downwards and rightwards, respectively. Similar to the first case study, the robot can take four actions to move in four directions. The robot moves 1 unit in the intended direction perturbed by bivariate Gaussian noise with a variance of  $\sigma^2 I$  where  $\sigma$  is 0.05 and  $I$  is a  $2 \times 2$  identity matrix. The noise is truncated to fit in a circle with a radius of  $4\sigma$ .

The adversary is located within the obstacle at (2.5, 2.5), and can take the actions *push* and *pull* to move the robot towards or away from the obstacle. Specifically, let  $(y, x)$  denote the position of the robot; then the adversary can perturb the x-position of the robot if  $y$  is in  $[2, 3]$  and can perturb the y-position of the robot if  $x$  is in  $[2, 3]$ . The amount of perturbation is drawn from a uniform distribution over the interval  $[0, 2]$ . Similar to the previous case study, the robot cannot leave a trap region, cannot go inside the obstacle, and is constrained by the borders of the environment; if the robot attempts, the robot moves and remains on the borders or the edges of the obstacle. A position  $(y, x)$  in the continuous versions of the environments has the characteristics of the corresponding grid cell, e.g., same label, being a trap or an obstacle. The robot needs to learn how to map these continuous positions to actions to perform the LTL task  $\varphi_1$  from (36) against any adversary strategy.

We adopted deep RL techniques to learn the controller and adversary strategies. We modified deep Q-Network (DQN) [49], [50] to include adversarial actions and minimax-Q updates. We used the default parameters with two hidden layers of size 64. We set  $\varepsilon = 0.25$  to increase discounting and thereby stability as performing  $\varphi_1$  in these environments does not require long-horizon planning.

In this case study, the optimal adversary strategies and the corresponding minimum satisfaction probabilities cannot be derived using the model checking tools such as PRISM [48] due to the continuous state-space and the nonlinear dynamics. Instead, in our evaluation, we used a manually crafted optimal adversary strategy that is to push the robot if the robot is on the right-hand side of the obstacle, and to pull otherwise.

In addition, since we cannot directly compute the minimum satisfaction probabilities for the learned controller strategies (as that would require reasoning on infinite traces), we estimated them as follows. We considered an episode satisfying if the largest color in the second half of the episode is an odd number and occurs at least 20% of the time (a proxy for satisfaction of repeated reachability condition)—for the environments shown in Fig. 2, it is reasonable to assume that, on average, one of five transitions will be colored with the largest odd number visited.

The results, however, are not sensitive to this ratio. We calculated the estimated satisfaction probability as the ratio of the number of satisfying paths to the number of all evaluation episodes, and used 1000 evaluation episodes of length  $\mathcal{T} = 100$ .

The learning curves of the methods for each environment against the optimal adversary are shown in the third row of Fig. 2; the estimated satisfaction probabilities are averaged over 4 simulations. All of our methods converged to an optimal controller strategy around  $10^6$  steps in the first and the second environments while APG struggled to converge about until  $10^7$  steps and exhibited high variance. In the last environment, MPG outperformed the other methods. PG was able to quickly learn a near-optimal controller strategy; however, it failed to converge to an optimal one. Similar to the first case study, TMPG performed worse than the rest in this environment.

**3) Long-Horizon Planning:** This case study considers a larger grid to evaluate the performance of the methods for long-horizon planning. Here, the robot can deterministically move to an adjacent cell using the same four actions, and the adversary can manipulate the position of the garbage area ( $g$ ) and the danger zone ( $d$ ) within the designated regions. Fig. 3 shows the environment and the designated regions considered in this case study. The initial position of the garbage area is (2, 6), and the adversary can change it to one of (2, 5), (2, 6), and (2, 7) at any time step w.p. 0.1. Similarly, the initial position of the danger zone is (5, 1), and the adversary can change it to one of (5, 0), (5, 1), and (5, 2) at any step w.p. 0.1.

The robot needs to either pick up the garbage ( $g$ ) and trash it out ( $t$ ) in at most two time steps, or repeatedly monitor the assigned region ( $r$ ) and go to the charging station ( $c$ ). The robot can be formally defined as the following LTL formula:

$$\varphi_2 := \left( \square (\Diamond g \wedge [g \rightarrow (\bigcirc t \vee \bigcirc \bigcirc t)]) \right. \\ \left. \vee (\Diamond \square w \wedge \square \Diamond c \wedge \square \Diamond r) \right) \wedge \square \neg d. \quad (37)$$

The DPA obtained from  $\varphi_2$  has 9 states, and 4 colors [see Fig. 6(b) in our corresponding technical report [51]].

There are two possible ways to perform this task: (i) the robot can repeatedly pick up the garbage at (2, 6) and trash out at (0, 6), or (ii) potentially go through the danger zone and repeatedly visit the assigned region at (6, 9) and charging station at (9, 9). Against an optimal policy, (i) cannot be satisfied because the adversary will eventually move the garbage area to either (2, 5) or (2, 7) from which the robot needs to take at least three time steps to trash it out. However, the robot can satisfy (ii) if the robot safely passes the designated region for the danger zone. The optimal controller strategy is, therefore, to try to pass the region while avoiding the danger zone and satisfy (ii); and the optimal adversary strategy is to move the danger zone to the cell the robot uses to pass the fifth row. As a result, the task can be almost surely satisfied in the lower part of the grid, and can be satisfied w.p. 0.9 in the upper part.

We used minimax-Q to learn the optimal controller strategies using the same parameters from the first case study. We conducted three experiments where the robot starts in cells: (0, 0), (9, 0), and (9, 9) and navigates for  $\mathcal{T} = 1000$  environment steps. Fig. 4 shows the derived learning curves; the satisfaction probabilities are averaged over ten simulations.

MPG learned a near-optimal controller strategy for the case where the robot starts in  $(0, 0)$ , which is far from the assigned region and the charging station. PG was able to learn a controller strategy satisfying the task w.p. around 0.5; however, it could not converge. The other methods failed to learn any reasonable strategy. We note that the maximum satisfaction probability against the optimal adversary is 0.9 in this case. For the case where the robot starts in  $(9, 0)$ ; i.e., the robot does not need to pass through the danger zone, MPG converged to an optimal strategy faster than the other methods. PG converged slower than MPG, and TMPG learned a reasonable strategy only after  $10^8$  steps. APG could not learn any strategy that yields a positive satisfaction probability.

For the last case where the robot starts in the charging station at  $(9, 9)$ , PG learned faster than the other methods although it struggled to stably converge to an optimal strategy. MPG converged to an optimal almost as fast as PG. TMPG learned an optimal strategy only after about  $10^8$  steps. APG, like in the other cases, could not learn any meaningful strategy. Overall, MPG significantly outperformed the other methods and APG could not learn any controller strategy that satisfies the task with a positive probability.

### B. Robotic Arm

We consider two motion planning tasks where a robotic arm should repeatedly reach some target ball positions where the adversary can manipulate the position of one of the balls. We implemented a simulation environment based on the Fetch environments [52] in OpenAI Gym [53]. In this environment, the state space consists of the positions of the gripper and the balls. The action space of the controller is continuous and consists of three dimensions that represent the position that the arm intends to move the gripper to. In the beginning of each episode, the balls are placed in random positions that are at least 10 cm apart and within the reach of the arm, including the air. The red ball is placed outside the boundaries of the smallest cuboid containing the blue and green balls; however, the adversary can move the red ball to any point in the state space w.p. 0.1. The environment is illustrated in Fig. 5(a).

The first task is repeatedly visiting two given random positions, represented by a green and a blue ball while staying within the boundaries of the cuboid of these balls. The robotic arm does not need to be within the cuboid if it can repeatedly reach both of the balls at the same time. The robotic arm should stay away at all costs from the red ball, which is controlled by the adversary.

The task can be represented as the following LTL formula:

$$\varphi_3 = \left( (\square \Diamond \text{green\_ball} \wedge \square \Diamond \text{blue\_ball} \wedge \Diamond \square \text{boundary}) \right. \\ \left. \vee (\square \Diamond (\text{green\_ball} \wedge \text{blue\_ball})) \wedge \square \neg \text{red\_ball} \right) \quad (38)$$

which is translated to a DPA with 3 states and 4 colors (see Fig. 6(c) in our technical report [51]). The optimal adversary strategy for this task is to try to put the red ball in the position of either the green ball or the blue ball. In this way, the adversary can prevent the controller performing this task w.p. 0.1.

The second task includes reaching and staying in the position of the red ball outside the cuboid formed by the other balls. If the red ball is in the same position as the green or the blue ball, then reaching the red ball is sufficient; the robot does not need to stay in its position or to be outside the cuboid. Alternatively, the robot can repeatedly reach a position where both the green and the blue balls are in, but the red ball is not. This task can be formally defined as the LTL formula

$$\varphi_4 = \left( \Diamond \square (\text{red\_ball} \wedge \neg \text{boundary}) \right. \\ \left. \vee \Diamond (\text{red\_ball} \wedge \text{green\_ball}) \vee \Diamond (\text{red\_ball} \wedge \text{blue\_ball}) \right. \\ \left. \vee \square \Diamond (\text{green\_ball} \wedge \text{blue\_ball} \wedge \neg \text{red\_ball}) \right) \quad (39)$$

which is translated to a DPA with 2 states and 4 colors (see Fig. 6(d) in our report [51]). The optimal adversary strategy, in this task, is to try to move the red ball to a position that is away from the green and the blue balls but inside the boundaries of the cuboid of these balls. Against such an adversary strategy, the maximum probability that the robot successfully performs this task is 0.9.

We integrated our framework into Truncated Quantile Critics (TQC) [50], [54], an off-policy RL algorithm for continuous actions. We used a neural network with two hidden layers of size 64 for both the actor and the critic. Similar to the case in Section VI-A2, we used  $\varepsilon = 0.25$  to increase discounting and stability since the robotic arm can quickly reach any position in this environment. We used two separate TQC for the controller and the adversary. The adversary gets the cumulative discounted rewards the controller obtained throughout an episode as the reward signal. During the learning, each episode starts in a random state, and the adversary tries to move the red ball. The controller observes the new position of the red ball and takes actions for  $\mathcal{T} = 100$  environment steps to learn to perform the task. We trained the controller and the adversary in turns consisting of  $10^4$  exploration steps.

Similar to the case study in Section VI-A2, we manually crafted the optimal adversary strategies and estimated the minimum satisfaction probabilities via simulation as deriving them via model checking tools is not feasible. We evaluated the performance by following the deterministic version of the controller strategy for 1000 evaluation episodes of length 100 against the optimal adversary strategy. To estimate the satisfaction probabilities for the first task, we calculated the ratio of the evaluation episodes that visited both the green and the blue balls inside the boundaries of their cuboid more than 20% of the time in the second half of the episode without visiting the red ball—since the robotic arm can quickly reach positions, 20% is reasonable; however, the results are not sensitive to this number. Similarly, for the second task, we estimated the satisfaction probability using the ratio of the evaluation episodes that visited the red ball either with the green ball or the blue ball, or outside the cuboid of these balls. We did not consider the case where the green and blue balls are jointly visited, as it is not possible in this environment.

Fig. 5(b) and (c) shows the results averaged over four simulations for each task. MPG quickly learned a near-optimal controller strategy and successfully converged to an optimal one

for the first task. TMPG converged to an optimal strategy slower than MPG. APG learned a near-optimal strategy only after  $10^6$  steps. PG could not learn any reasonable strategy. For the second task, MPG immediately learned a near-optimal strategy and then converged to an optimal one. PG converged slower than MPG but faster than TMPG, and TMPG converged faster than APG. Overall, MPG significantly outperformed the other methods.

## VII. CONCLUSION

In this work, we presented a model-free RL approach to synthesize optimal controller strategies for *any* LTL task in SGs. We provided an approach to craft rewards and discount factors from the parity condition of the DPAs translated from the given LTL tasks. We showed that any controller strategy maximizing the sum of discounted rewards in the worst case also maximizes the minimum probability of satisfying the LTL specification for some sufficiently small parameter. We then introduced our scalable lazy color generation method providing distinct rewards and discount factors only when necessary and thereby improving the learning scalability. In addition, we provided an approximate method that is highly efficient when the controller can perform the task by focusing on a single color. Finally, we demonstrated the applicability of our methods in several case studies and showed that our methods outperform the existing methods for learning from LTL tasks in SGs. Generalization of our approach to partially observable environments and multiobjectives that include prioritized safety/task constraints (in the form of LTL specifications) and secondary (control-related) cost minimization objectives, as done in [55] and [56] for MDPs, respectively, is left as a future work.

## REFERENCES

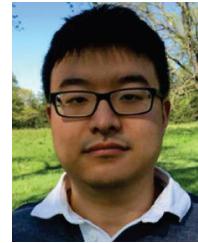
- C. Baier and J.-P. Katoen, *Principles of Model Checking*. Cambridge, MA, USA: MIT Press, 2008.
- M. Kloetzer and C. Belta, “Temporal logic planning and control of robotic swarms by hierarchical abstractions,” *IEEE Trans. Robot.*, vol. 23, no. 2, pp. 320–330, Apr. 2007.
- H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas, “Temporal-logic-based reactive mission and motion planning,” *IEEE Trans. Robot.*, vol. 25, no. 6, pp. 1370–1381, Dec. 2009.
- M. Kloetzer and C. Belta, “Automatic deployment of distributed teams of robots from temporal logic motion specifications,” *IEEE Trans. Robot.*, vol. 26, no. 1, pp. 48–61, Feb. 2010.
- Y. Chen, X. C. Ding, A. Stefanescu, and C. Belta, “Formal approach to the deployment of distributed robotic teams,” *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 158–171, Feb. 2012.
- M. Lahijanian, S. B. Andersson, and C. Belta, “Temporal logic motion planning and control with probabilistic satisfaction guarantees,” *IEEE Trans. Robot.*, vol. 28, no. 2, pp. 396–409, Apr. 2012.
- X. Ding, S. L. Smith, C. Belta, and D. Rus, “Optimal control of Markov decision processes with linear temporal logic constraints,” *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1244–1257, May 2014.
- G. E. Fainekos, A. Girard, H. Kress-Gazit, and G. J. Pappas, “Temporal logic motion planning for dynamic robots,” *Automatica*, vol. 45, no. 2, pp. 343–352, 2009.
- S. Jiang and R. Kumar, “Failure diagnosis of discrete-event systems with linear-time temporal logic specifications,” *IEEE Trans. Autom. Control*, vol. 49, no. 6, pp. 934–945, Jun. 2004.
- P. Tabuada and G. J. Pappas, “Linear time logic control of discrete-time linear systems,” *IEEE Trans. Autom. Control*, vol. 51, no. 12, pp. 1862–1877, Dec. 2006.
- M. Kloetzer and C. Belta, “A fully automated framework for control of linear systems from temporal logic specifications,” *IEEE Trans. Autom. Control*, vol. 53, no. 1, pp. 287–297, Feb. 2008.
- T. Wongpiromsarn, U. Topcu, and R. M. Murray, “Receding horizon temporal logic planning,” *IEEE Trans. Autom. Control*, vol. 57, no. 11, pp. 2817–2830, Nov. 2012.
- B. Yordanov, J. Tumova, I. Cerna, J. Barnat, and C. Belta, “Temporal logic control of discrete-time piecewise affine systems,” *IEEE Trans. Autom. Control*, vol. 57, no. 6, pp. 1491–1504, Jun. 2012.
- J. Liu, N. Ozay, U. Topcu, and R. M. Murray, “Synthesis of reactive switching protocols from temporal logic specifications,” *IEEE Trans. Autom. Control*, vol. 58, no. 7, pp. 1771–1785, Jul. 2013.
- M. Zamani, P. Mohajeran Esfahani, R. Majumdar, A. Abate, and J. Lygeros, “Symbolic control of stochastic systems via approximately bisimilar finite abstractions,” *IEEE Trans. Autom. Control*, vol. 59, no. 12, pp. 3135–3150, Dec. 2014.
- J. Fu and U. Topcu, “Probably approximately correct MDP learning and control with temporal logic constraints,” in *Proc. Robot.: Sci. Syst. Conf.*, 2014. [Online]. Available: <https://www.roboticsproceedings.org/rss10/p39.pdf>
- T. Brázdil et al., “Verification of Markov decision processes using learning algorithms,” in *Proc. Int. Symp. Autom. Technol. Verification Anal.*, 2014, pp. 98–114.
- Q. Gao, M. Pajic, and M. M. Zavlanos, “Deep imitative reinforcement learning for temporal logic robot motion planning with noisy semantic observations,” in *Proc. Int. Conf. Robot. Autom.*, 2020, pp. 8490–8496.
- E. M. Hahn, M. Perez, S. Schewe, F. Somenzi, A. Trivedi, and D. Wojtczak, “Omega-regular objectives in model-free reinforcement learning,” in *Proc. Int. Conf. Tools Algorithms Construction Anal. Syst.*, 2019, pp. 395–412.
- A. K. Bozkurt, Y. Wang, M. M. Zavlanos, and M. Pajic, “Control synthesis from linear temporal logic specifications using model-free reinforcement learning,” in *Proc. Int. Conf. Robot. Automat.*, 2020, pp. 10349–10355.
- E. M. Hahn, G. Li, S. Schewe, A. Turrini, and L. Zhang, “Lazy probabilistic model checking without determinisation,” in *Proc. Int. Conf. Concurrency Theory*, 2015, Art. no. 354.
- H. Fawzi, P. Tabuada, and S. Diggavi, “Secure estimation and control for cyber-physical systems under adversarial attacks,” *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, Jun. 2014.
- K. Chatterjee and T. A. Henzinger, “A survey of stochastic  $\omega$ -regular games,” *J. Comput. Syst. Sci.*, vol. 78, no. 2, pp. 394–413, 2012.
- J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. Berlin, Germany: Springer, 1997.
- A. Neyman and S. Sorin, *Stochastic Games and Applications*, vol. 570. Dordrecht, The Netherlands: Kluwer, 2003.
- M. Wen and U. Topcu, “Probably approximately correct learning in adversarial environments with temporal logic specifications,” *IEEE Trans. Autom. Control*, vol. 67, no. 10, pp. 5055–5070, Oct. 2022.
- L. Niu and A. Clark, “Optimal secure control with linear temporal logic constraints,” *IEEE Trans. Autom. Control*, vol. 65, no. 6, pp. 2434–2449, Jun. 2020.
- L. Niu, J. Fu, and A. Clark, “Optimal minimum violation control synthesis of cyber-physical systems under attacks,” *IEEE Trans. Autom. Control*, vol. 66, no. 3, pp. 995–1008, Mar. 2021.
- K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, “Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents,” *IEEE Trans. Autom. Control*, vol. 66, no. 12, pp. 5925–5940, Dec. 2021.
- J. Fu, H. G. Tanner, J. Heinz, and J. Chandee, “Adaptive symbolic control for finite-state transition systems with grammatical inference,” *IEEE Trans. Autom. Control*, vol. 59, no. 2, pp. 505–511, Feb. 2014.
- J. Fu and U. Topcu, “Synthesis of joint control and active sensing strategies under temporal logic constraints,” *IEEE Trans. Autom. Control*, vol. 61, no. 11, pp. 3464–3476, Nov. 2016.
- M. Svoreňová and M. Kwiatkowska, “Quantitative verification and strategy synthesis for stochastic games,” *Eur. J. Control*, vol. 30, pp. 15–30, 2016.
- M. Elfar, Y. Wang, and M. Pajic, “Security-aware synthesis using delayed-action games,” in *Proc. Comput. Aided Verification*, 2019, pp. 180–199.
- M. Elfar, H. Zhu, M. L. Cummings, and M. Pajic, “Security-aware synthesis of human-UAV protocols,” in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 8011–8017.
- A. K. Bozkurt, Y. Wang, and M. Pajic, “Secure planning against stealthy attacks via model-free reinforcement learning,” in *Proc. Int. Conf. Robot. Autom.*, 2021, pp. 10656–10662.
- M. Svoreňová, J. Křetínský, M. Chmelík, K. Chatterjee, I. Černá, and C. Belta, “Temporal logic control for stochastic linear systems using abstraction refinement of probabilistic games,” *Nonlinear Anal.: Hybrid Syst.*, vol. 23, pp. 230–253, 2017.

- [37] S. Sadraddini and C. Belta, "Formal methods for adaptive control of dynamical systems," in *Proc. IEEE 56th Annu. Conf. Decis. Control*, 2017, pp. 1782–1787.
- [38] P. Ashok, J. Křetínsky, and M. Weininger, "Pac statistical model checking for Markov decision processes and stochastic games," in *Proc. Int. Conf. Comput. Aided Verification*, 2019, pp. 497–519.
- [39] E. M. Hahn, M. Perez, S. Schewe, F. Somenzi, A. Trivedi, and D. Wojtczak, "Model-free reinforcement learning for stochastic parity games," in *Proc. Int. Conf. Concurrency Theory*, 2020, vol. 171, Art. no. 21.
- [40] A. K. Bozkurt, Y. Wang, M. M. Zavlanos, and M. Pajic, "Model-free reinforcement learning for stochastic games with linear temporal logic objectives," in *Proc. Int. Conf. Robot. Autom.*, 2021, pp. 10649–10655.
- [41] J. Esparza, J. Křetínsky, J.-F. Raskin, and S. Sickert, "From LTL and limit-deterministic Büchi automata to deterministic parity automata," in *Proc. Int. Conf. Tools Algorithms Construction Anal. Syst.*, 2017, pp. 426–442.
- [42] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. Mach. Learn.*, 1994, pp. 157–163.
- [43] A. Hordijk and A. A. Yushkevich, "Blackwell optimality," in *Handbook of Markov Decision Processes*, vol. 40. Boston, MA, USA: Springer, 2002, pp. 231–267.
- [44] M. L. Littman and C. Szepesvári, "A generalized reinforcement-learning model: Convergence and applications," in *Proc. Int. Conf. Mach. Learn.*, 1996, vol. 96, pp. 310–318.
- [45] Z. Xuan, A. K. Bozkurt, M. Pajic, and Y. Wang, "On the uniqueness of solution for the Bellman equation of LTL objectives," 2024, *arXiv:2404.05074*.
- [46] C. Yang, M. L. Littman, and M. Carbin, "On the (in) tractability of reinforcement learning for LTL objectives," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 3650–3658.
- [47] J. Křetínský, T. Meggendorfer, and S. Sickert, "OWL: A library for omega-words, automata, and LTL," in *Proc. Int. Symp. Autom. Technol. Verification Anal.*, 2018, vol. 11138, pp. 543–550.
- [48] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of probabilistic real-time systems," in *Proc. Int. Conf. Comput. Aided Verification*, 2011, pp. 585–591.
- [49] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [50] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021.
- [51] A. K. Bozkurt, Y. Wang, M. M. Zavlanos, and M. Pajic, "Learning optimal strategies for temporal tasks in stochastic games," 2023, *arXiv:2102.04307*.
- [52] M. Plappert et al., "Multi-goal reinforcement learning: Challenging robotics environments and request for research," 2018, *arXiv:1802.09464*.
- [53] G. Brockman et al., "Openai gym," 2016. [Online]. Available: <https://github.com/openai/gym?tab=readme-ov-file>
- [54] A. Kuznetsov, P. Shvechikov, A. Grishin, and D. Vetrov, "Controlling overestimation bias with truncated mixture of continuous distributional quantile critics," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5556–5566.
- [55] Y. Wang, A. K. Bozkurt, and M. Pajic, "Reinforcement learning with temporal logic constraints for partially-observable Markov decision processes," 2021, *arXiv:2104.01612*.
- [56] A. K. Bozkurt, Y. Wang, and M. Pajic, "Model-free learning of safe yet effective controllers," in *Proc. Conf. Decis. Control*, 2021, pp. 6560–6565.
- [57] A. Duret-Lutz, A. Lewkowicz, A. Fauchille, T. Michaud, É. Renault, and L. Xu, "Spot 2.0—A framework for LTL and-automata manipulation," in *Proc. Int. Symp. Autom. Technol. Verification Anal.*, 2016, pp. 122–129.



**Alper Kamil Bozkurt** received the B.S. and M.S. degrees in computer engineering from Bogazici University, Istanbul, Turkey, in 2015 and 2018, respectively. He is currently working toward the Ph.D. degree in computer science with the Department of Computer Science, Duke University, Durham, NC, USA.

His research interests include the intersection of machine learning, control theory, and formal methods. In particular, he focuses on developing learning-based algorithms that synthesize provably safe and reliable controllers for cyber-physical systems.



**Yu Wang** (Senior Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2018.

He is currently an Assistant Professor with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA. He was a Postdoctoral Associate with the Department of Electrical and Computer Engineering, Duke University. His research interests include assured autonomy, cyber-physical systems, machine learning, and formal methods.



**Michael M. Zavlanos** (Senior Member, IEEE) received the Diploma in mechanical engineering from the National Technical University of Athens, Athens, Greece, in 2002, and the M.S.E. and Ph.D. degrees in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2005 and 2008, respectively.

He is currently an Associate Professor with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA. His research interests include control theory, optimization, and learning and, in particular, autonomous systems and robotics, networked and distributed control systems, and cyber-physical systems.

Dr. Zavlanos was the recipient of various awards including the 2014 ONR YIP Award and the 2011 NSF CAREER Award.



**Miroslav Pajic** received the Dipl.Ing. and M.S. degrees in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 2003 and 2007, respectively, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2010 and 2012, respectively.

He is currently the Dickinson Family Associate Professor with Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. His research interests include the design and analysis of high-assurance cyber-physical systems with varying levels of autonomy and human interaction, at the intersection of (more traditional) areas of embedded systems, AI, learning and controls, formal methods, and robotics.

Dr. Pajic was the recipient of various awards including the ACM SIGBED Early-Career Award, IEEE TCCPS Early-Career Award, NSF CAREER Award, ONR Young Investigator Award, ACM SIGBED Frank Anger Memorial Award, Joseph and Rosaline Wolf Best Dissertation Award from Penn Engineering, IBM Faculty Award, as well as eight Best Paper and Runner-up Awards.