

# Reinforcement Learning of Control Policy for Linear Temporal Logic Specifications Using Limit-Deterministic Generalized Büchi Automata

Ryohei Oura, Ami Sakakibara, and Toshimitsu Ushio

**Abstract**—This letter proposes a novel reinforcement learning method for the synthesis of a control policy satisfying a control specification described by a linear temporal logic formula. We assume that the controlled system is modeled by a Markov decision process (MDP). We transform the specification to a limit-deterministic generalized Büchi automaton (LDGBA) with several accepting sets that accepts all infinite sequences satisfying the formula. The LDGBA is augmented so that it explicitly records the previous visits to accepting sets. We take a product of the augmented LDBA and the MDP, based on which we define a reward function. The agent gets rewards whenever state transitions are in an accepting set that has not been visited for a certain number of steps. Consequently, sparsity of rewards is relaxed and optimal circulations among the accepting sets are learned. We show that the proposed method can learn an optimal policy when the discount factor is sufficiently close to one.

**Index Terms**—Reinforcement Learning, Linear Temporal Logic, Limit-Deterministic Büchi Automata.

## I. INTRODUCTION

Reinforcement learning (RL) is a useful approach to learning an optimal policy from sample behaviors of the controlled system [6]. In RL, we use a reward function that assigns a reward to each transition in the behaviors and evaluate a control policy by the return that is an expected (discounted) sum of the rewards along the behaviors. Thus, to apply RL to the synthesis of a control policy for the LTL specification, it is an important issue how to introduce the reward function. Previously, a reward function based on the acceptance condition of a Rabin automaton was proposed in [5]. They converted a given LTL specification to a Rabin automaton and introduced the reward function that assigns a positive and negative value to observations corresponding to an acceptance pair of the Rabin automaton. It was applied to a control problem where the controller optimizes a given control cost under the LTL constraint [7]. It is known that the LTL specification is converted into an  $\omega$ -automaton such as a nondeterministic Büchi automaton and a deterministic Rabin automaton [1], [2]. A Büchi automaton (BA) and a generalized one (GBA) are also an  $\omega$ -automaton. When constructing a BA from a GBA, however, the order of visits to accepting sets of the GBA is fixed. The construction causes the sparsity of the reward based on the acceptance

condition of a BA. It is critical against an RL-based controller synthesis.

Recently, a limit-deterministic Büchi automaton (LDBA) or a generalized one (LDGBA) is paid much attention to as an  $\omega$ -automaton corresponding to the LTL specification [8]. The RL-based approaches to the synthesis of a control policy using LDBAs or LDGBAs have been proposed in [9]–[12]. To deal with the acceptance condition of an LDGBA that accepts behaviors visiting all accepting sets infinitely often, the accepting frontier function was introduced in [9], [11]. The reward function is defined based on the function. However, the function is memoryless, that is, it does not provide the information of accepting sets that have been visited, which is important to improve learning performance. In this letter, we propose a novel method to augment an LDBA converted from a given LTL formula. Then, we define a reward function based on the acceptance condition of the product MDP of the augmented LDBA and the controlled system. As a result, we can improve the sparsity of rewards and expand the class of policies that satisfy the LTL specification compared to [9].

The rest of the letter is organized as follows. Section II reviews an MDP, LTL, and automata. Section III proposed a novel RL-based method for the synthesis of a control policy. Section IV presents a numerical example for which the previous method cannot learn a control policy but the proposed one can.

**Notations:** For sets  $A$  and  $B$ ,  $AB$  denotes their concatenation.  $A^\omega$  denotes the infinite concatenation of the set  $A$  and  $A^*$  denotes the finite one.  $\mathbb{N}_0$  is the nonnegative integers.  $\mathbb{R}_0$  is nonnegative real numbers.

## II. PRELIMINARIES

### A. Markov Decision Process

**Definition 1:** A (labeled) Markov decision process (MDP) is a tuple  $M = (S, A, P, s_{init}, AP, L)$ , where  $S$  is a finite set of states,  $A$  is a finite set of actions,  $P : S \times S \times A \rightarrow [0, 1]$  is a transition probability,  $s_{init} \in S$  is the initial state,  $AP$  is a finite set of atomic propositions, and  $L : S \times A \times S \rightarrow 2^{AP}$  is a labeling function that assigns a set of atomic propositions to each transition. Let  $\mathcal{A}(s) = \{a \in A; \text{the action } a \text{ is feasible at the state } s\}$ . Note that  $\sum_{s' \in S} P(s'|s, a) = 1$  holds for any state  $s \in S$  and action  $a \in \mathcal{A}(s)$ .

In the MDP  $M$ , an infinite path starting from a state  $s_0 \in S$  is defined as a sequence  $\rho = s_0 a_0 s_1 \dots \in S(AS)^\omega$  such that  $P(s_{i+1}|s_i, a_i) > 0$  for any  $i \in \mathbb{N}_0$ . A finite path is a finite sequence in  $S(AS)^*$ . In addition, we sometimes represent  $\rho$  as  $\rho_{init}$  to emphasize that  $\rho$  starts from  $s_0 = s_{init}$ . For

This work was partially supported by JST-ERATO HASUO Project Grant Number JPMJER1603, Japan, JST-Mirai Program Grant Number JPMJMI18B4, Japan, and JSPS KAKENHI Grant Number JP19J13487, Japan.

The authors are with the Graduate School of Engineering Science, Osaka University, Toyonaka 560-8531, Japan (e-mail: r-oura, sakakibara@hopf.sys.es.osaka-u.ac.jp; ushio@sys.es.osaka-u.ac.jp).

a path  $\rho = s_0 a_0 s_1 \dots$ , we define the corresponding labeled path  $L(\rho) = L(s_0, a_0, s_1) L(s_1, a_1, s_2) \dots \in (2^{AP})^\omega$ .  $InfPath^M$  (resp.,  $FinPath^M$ ) is defined as the set of infinite (resp., finite) paths starting from  $s_0 = s_{init}$  in the MDP  $M$ . For each finite path  $\rho$ ,  $last(\rho)$  denotes its last state.

**Definition 2:** A policy on an MDP  $M$  is defined as a mapping  $\pi : FinPath^M \times \mathcal{A}(last(\rho)) \rightarrow [0, 1]$ . A policy  $\pi$  is a *positional* policy if for any  $\rho \in FinPath^M$  and any  $a \in \mathcal{A}(last(\rho))$ , it holds that  $\pi(\rho, a) = \pi(last(\rho), a)$  and there exists  $a' \in \mathcal{A}(last(\rho))$  such that  $\pi(\rho, a) = 1$  if  $a = a'$ , otherwise  $\pi(\rho, a) = 0$ .

Let  $InfPath_\pi^M$  (resp.,  $FinPath_\pi^M$ ) be the set of infinite (resp., finite) paths starting from  $s_0 = s_{init}$  in the MDP  $M$  under a policy  $\pi$ . The behavior of an MDP  $M$  under a policy  $\pi$  is defined on a probability space  $(InfPath_\pi^M, \mathcal{F}_{InfPath_\pi^M}, Pr_\pi^M)$ .

A Markov chain induced by an MDP  $M$  with a positional policy  $\pi$  is a tuple  $MC_\pi = (S_\pi, P_\pi, s_{init}, AP, L)$ , where  $S_\pi = S$ ,  $P_\pi(s'|s) = P(s'|s, a)$  for  $s, s' \in S$  and  $a \in \mathcal{A}(s)$  such that  $\pi(s, a) = 1$ . The state set  $S_\pi$  of  $MC_\pi$  can be represented as a disjoint union of a set of transient states  $T_\pi$  and closed irreducible sets of recurrent states  $R_\pi^j$  with  $j \in \{1, \dots, h\}$ , i.e.,  $S_\pi = T_\pi \cup R_\pi^1 \cup \dots \cup R_\pi^h$  [13]. In the following, we say a “recurrent class” instead of a “closed irreducible set of recurrent states” for simplicity.

In an MDP  $M$ , we define a reward function  $R : S \times A \times S \rightarrow \mathbb{R}_0$ . The function denotes the immediate reward received after the agent performs an action  $a$  at a state  $s$  and reaches a next state  $s'$  as a result.

**Definition 3:** For a policy  $\pi$  on an MDP  $M$ , any state  $s \in S$ , and a reward function  $R$ , we define the expected discounted reward as

$$V^\pi(s) = \mathbb{E}^\pi \left[ \sum_{n=0}^{\infty} \gamma^n R(S_n, A_n, S_{n+1}) | S_0 = s \right],$$

where  $\mathbb{E}^\pi$  denotes the expected value given that the agent follows the policy  $\pi$  from the state  $s$  and  $\gamma \in [0, 1)$  is a discount factor. The function  $V^\pi(s)$  is often referred to as a state-value function under the policy  $\pi$ . For any state-action pair  $(s, a) \in S \times A$ , we define an action-value function  $Q^\pi(s, a)$  under the policy  $\pi$  as follows.

$$Q^\pi(s, a) = \mathbb{E}^\pi \left[ \sum_{n=0}^{\infty} \gamma^n R(S_n, A_n, S_{n+1}) | S_0 = s, A_0 = a \right].$$

**Definition 4:** For any state  $s \in S$ , a policy  $\pi^*$  is optimal if

$$\pi^* \in \arg \max_{\pi \in \Pi^{pos}} V^\pi(s),$$

where  $\Pi^{pos}$  is the set of positional policies over the state set  $S$ .

## B. Linear Temporal Logic and Automata

In our proposed method, we use linear temporal logic (LTL) formulas to describe various constraints or properties and to systematically assign corresponding rewards. LTL formulas are constructed from a set of atomic propositions, Boolean operators, and temporal operators. We use the

standard notations for the Boolean operators:  $\top$  (true),  $\neg$  (negation), and  $\wedge$  (conjunction). LTL formulas over a set of atomic propositions  $AP$  are defined as

$$\varphi ::= \top \mid \alpha \in AP \mid \varphi_1 \wedge \varphi_2 \mid \neg \varphi \mid \mathbf{X}\varphi \mid \varphi_1 \mathbf{U} \varphi_2,$$

where  $\varphi$ ,  $\varphi_1$ , and  $\varphi_2$  are LTL formulas. Additional Boolean operators are defined as  $\perp := \neg \top$ ,  $\varphi_1 \vee \varphi_2 := \neg(\neg \varphi_1 \wedge \neg \varphi_2)$ , and  $\varphi_1 \Rightarrow \varphi_2 := \neg \varphi_1 \vee \varphi_2$ . The operators  $\mathbf{X}$  and  $\mathbf{U}$  are called “next” and “until”, respectively. Using the operator  $\mathbf{U}$ , we define two temporal operators: (1) *eventually*,  $\mathbf{F}\varphi := \top \mathbf{U} \varphi$  and (2) *always*,  $\mathbf{G}\varphi := \neg \mathbf{F} \neg \varphi$ .

Let  $M$  be an MDP. For an infinite path  $\rho = s_0 a_0 s_1 \dots$  of  $M$  with  $s_0 \in S$ , let  $\rho[i]$  be the  $i$ -th state of  $\rho$  i.e.,  $\rho[i] = s_i$  and let  $\rho[i:]$  be the  $i$ -th suffix  $\rho[i:] = s_i a_i s_{i+1} \dots$ .

**Definition 5:** For an LTL formula  $\varphi$ , an MDP  $M$ , and an infinite path  $\rho = s_0 a_0 s_1 \dots$  of  $M$  with  $s_0 \in S$ , the satisfaction relation  $M, \rho \models \varphi$  is recursively defined as follows.  $M, \rho \models \top$ ;  $M, \rho \models \alpha \in AP$  iff  $\alpha \in L(s_0, a_0, s_1)$ ;  $M, \rho \models \varphi_1 \wedge \varphi_2$  iff  $M, \rho \models \varphi_1$  and  $M, \rho \models \varphi_2$ ;  $M, \rho \models \neg \varphi$  iff  $M, \rho \not\models \varphi$ ;  $M, \rho \models \mathbf{X}\varphi$  iff  $M, \rho[1:] \models \varphi$ ; and  $M, \rho \models \varphi_1 \mathbf{U} \varphi_2$  iff  $\exists j \geq 0$ ,  $M, \rho[j:] \models \varphi_2$  and  $\forall i, 0 \leq i < j$ ,  $M, \rho[i:] \models \varphi_1$ . The next operator  $\mathbf{X}$  requires that  $\varphi$  is satisfied by the next state suffix of  $\rho$ . The until operator  $\mathbf{U}$  requires that  $\varphi_1$  holds true until  $\varphi_2$  becomes true over the path  $\rho$ . In the following, we write  $\rho \models \varphi$  for simplicity without referring to MDP  $M$ .

For any policy  $\pi$ , we denote the probability of all paths starting from  $s_{init}$  on the MDP  $M$  that satisfy an LTL formula  $\varphi$  under the policy  $\pi$  as

$$Pr_\pi^M(s_{init} \models \varphi) := Pr_\pi^M(\{\rho_{init} \in InfPath_\pi^M; \rho_{init} \models \varphi\}).$$

We say that an LTL formula  $\varphi$  is satisfied by a positional policy  $\pi$  if

$$Pr_\pi^M(s_{init} \models \varphi) > 0.$$

Any LTL formula  $\varphi$  can be converted into various automata, namely finite state machines that recognize all words satisfying  $\varphi$ . We define a generalized Büchi automaton at the beginning, and then introduce a limit-deterministic Büchi automaton.

**Definition 6:** A transition-based generalized Büchi automaton (tGBA) is a tuple  $B = (X, x_{init}, \Sigma, \delta, \mathcal{F})$ , where  $X$  is a finite set of states,  $x_{init} \in X$  is the initial state,  $\Sigma$  is an input alphabet,  $\delta \subset X \times \Sigma \times X$  is a set of transitions, and  $\mathcal{F} = \{F_1, \dots, F_n\}$  is an acceptance condition, where for each  $j \in \{1, \dots, n\}$ ,  $F_j \subset \delta$  is a set of accepting transitions and called an accepting set. We refer to a tGBA with one accepting set as a tBA.

Let  $\Sigma^\omega$  be the set of all infinite words over  $\Sigma$  and let an infinite run be an infinite sequence  $r = x_0 \sigma_0 x_1 \dots \in X(\Sigma X)^\omega$  where  $(x_i, \sigma_i, x_{i+1}) \in \delta$  for any  $i \in \mathbb{N}_0$ . An infinite word  $w = \sigma_0 \sigma_1 \dots \in \Sigma^\omega$  is accepted by  $B_\varphi$  if and only if there exists an infinite run  $r = x_0 \sigma_0 x_1 \dots$  starting from  $x_0 = x_{init}$  such that  $inf(r) \cap F_j \neq \emptyset$  for each  $F_j \in \mathcal{F}$ , where  $inf(r)$  is the set of transitions that occur infinitely often in the run  $r$ .

**Definition 7:** A transition-based limit-deterministic generalized Büchi automaton (tLDGBA) is a tGBA  $B = (X, x_{init}, \Sigma, \delta, \mathcal{F})$  such that  $X$  is partitioned into two disjoint sets  $X_{initial}$  and  $X_{final}$  such that

- $F_j \subset X_{final} \times \Sigma \times X_{final}, \forall j \in \{1, \dots, n\}$ ,
- $|\{(x, \sigma, x') \in \delta; x' \in X_{final}\}| \leq 1, \forall x \in X_{final}, \forall \sigma \in \Sigma$ ,
- $|\{(x, \sigma, x') \in \delta; x' \in X_{initial}\}| = 0, \forall x \in X_{final}, \forall \sigma \in \Sigma$ .

$X_{initial}$  and  $X_{final}$  are connected by a single “guess”. An  $\varepsilon$ -transition enables the tLDGBA to change its state with no input. Then,  $\varepsilon$ -transitions reflect the single “guess” from  $X_{initial}$  to  $X_{final}$ . The final part has all accepting sets. Note that by the construction in [8], the transitions in each part have determinism except for  $\varepsilon$ -transitions into  $X_{final}$  from  $X_{initial}$ . It is known that, for any LTL formula  $\varphi$ , there exists a tLDGBA that accepts all words satisfying  $\varphi$  [8]. We refer to a tLDGBA with one accepting set as a tLDBA. In particular, we represent a tLDGBA recognizing an LTL formula  $\varphi$  as  $B_\varphi$ , whose input alphabet is given by  $\Sigma = 2^{AP}$ .

### III. REINFORCEMENT-LEARNING-BASED SYNTHESIS OF CONTROL POLICY

We introduce an automaton augmented with binary-valued vectors. The automaton can explicitly represent whether transitions in each accepting set occur at least once, and ensure transitions in each accepting set occur infinitely often.

Let  $V = \{(v_1, \dots, v_n)^T; v_i \in \{0, 1\}, i \in \{1, \dots, n\}\}$  be a set of binary-valued vectors, and let  $\mathbf{1}$  and  $\mathbf{0}$  be the  $n$ -dimensional vectors with all elements 1 and 0, respectively. In order to augment a tLDGBA  $B_\varphi$ , we introduce three functions  $visitf: \delta \rightarrow V$ ,  $reset: V \rightarrow V$ , and  $Max: V \times V \rightarrow V$  as follows. For any  $e \in \delta$ ,  $visitf(e) = (v_1, \dots, v_n)^T$ , where  $v_i = 1$  if  $e \in F_i$  and  $v_i = 0$  otherwise. For any  $v \in V$ ,  $reset(v) = \mathbf{0}$  if  $v = \mathbf{1}$  and  $reset(v) = v$  otherwise. For any  $v, u \in V$ ,  $Max(v, u) = (l_1, \dots, l_n)^T$ , where  $l_i = \max\{v_i, u_i\}$  for any  $i \in \{1, \dots, n\}$ .

Intuitively, each vector  $v$  represents which accepting sets have been visited. The function  $visitf$  returns a binary vector whose  $i$ -th element is 1 if and only if a transition in the accepting set  $F_i$  occurs. The function  $reset$  returns the zero vector  $\mathbf{0}$  if at least one transition in each accepting set has occurred after the latest reset. Otherwise, it returns the input vector without change.

**Definition 8:** For a tLDGBA  $B_\varphi = (X, x_{init}, \Sigma, \delta, \mathcal{F})$ , its augmented automaton is a tLDGBA  $\bar{B}_\varphi = (\bar{X}, \bar{x}_{init}, \bar{\Sigma}, \bar{\delta}, \bar{\mathcal{F}})$ , where  $\bar{X} = X \times V$ ,  $\bar{x}_{init} = (x_{init}, \mathbf{0})$ ,  $\bar{\Sigma} = \Sigma$ ,  $\bar{\delta}$  is defined as  $\bar{\delta} = \{((x, v), \bar{\sigma}, (x', v')) \in \bar{X} \times \bar{\Sigma} \times \bar{X}; (x, \bar{\sigma}, x') \in \delta, v' = reset(Max(v, visitf((x, \bar{\sigma}, x'))))\}$ , and  $\bar{\mathcal{F}} = \{\bar{F}_1, \dots, \bar{F}_n\}$  is defined as  $\bar{F}_j = \{((x, v), \bar{\sigma}, (x', v')) \in \bar{\delta}; (x, \bar{\sigma}, x') \in F_j, v_j = 0, visitf((x, \bar{\sigma}, x'))_j = 1\}$  for each  $j \in \{1, \dots, n\}$ , where  $visitf((x, \bar{\sigma}, x'))_j$  is the  $j$ -th element of  $visitf((x, \bar{\sigma}, x'))$ .

The augmented tLDGBA  $\bar{B}_\varphi$  keeps track of previous visits to the accepting sets of  $B_\varphi$  as a state. For example, We show an augmented tLDGBAs with the corresponding tLDGBA associated with the LTL formula  $\mathbf{GF}a \wedge \mathbf{GF}b \wedge \mathbf{G}\neg c$  in Figs. 1 and 2.

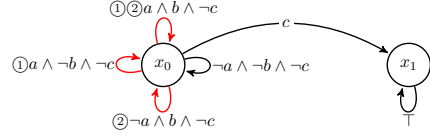


Fig. 1. The tLDGBA recognizing the LTL formula  $\mathbf{GF}a \wedge \mathbf{GF}b \wedge \mathbf{G}\neg c$ , where the initial state is  $x_0$ . Red arcs are accepting transitions that are numbered in accordance with the accepting sets they belong to.

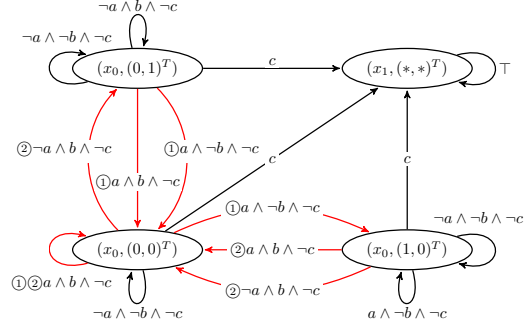


Fig. 2. The augmented automaton for the tLDGBA in Fig. 1 recognizing the LTL formula  $\mathbf{GF}a \wedge \mathbf{GF}b \wedge \mathbf{G}\neg c$ , where the initial state is  $(x_0, (0, 0)^T$ . Red arcs are accepting transitions that are numbered in accordance with the accepting sets they belong to. All states corresponding to  $x_1$  are merged into  $(x_1, (*, *)^T)$ .

**Definition 9:** Given an augmented tLDGBA  $\bar{B}_\varphi$  and an MDP  $M$ , a tuple  $M \otimes \bar{B}_\varphi = M^\otimes = (S^\otimes, A^\otimes, s_{init}^\otimes, P^\otimes, \delta^\otimes, \mathcal{F}^\otimes)$  is a product MDP, where  $S^\otimes = S \times \bar{X}$  is the finite set of states,  $A^\otimes$  is the finite set of actions such that  $A^\otimes = A \cup \{\varepsilon_{x'}; \exists x' \in X \text{ s.t. } (x, \varepsilon_{x'}, x') \in \delta\}$  where  $\varepsilon_{x'}$  is the action for the  $\varepsilon$ -transition to the state  $x' \in X$ ,  $s_{init}^\otimes = (s_{init}, \bar{x}_{init})$  is the initial states,  $P^\otimes: S^\otimes \times S^\otimes \times A^\otimes \rightarrow [0, 1]$  is the transition probability defined as  $P^\otimes(s^\otimes | s^\otimes, a) = P(s' | s, a)$  if  $(\bar{x}, L((s, a, s')), \bar{x}') \in \bar{\delta}$  and  $a \in \mathcal{A}(s)$ ,  $P^\otimes(s^\otimes | s^\otimes, a) = 1$  if  $s = s', v = v', (x, \varepsilon_{x'}, x') \in \bar{\delta}$ , and  $a = \varepsilon_{x'}$ , otherwise  $P^\otimes(s^\otimes | s^\otimes, a) = 0$  where  $s^\otimes = (s, (x, v))$  and  $s'^\otimes = (s', (x', v'))$ ,  $\delta^\otimes = \{(s^\otimes, a, s'^\otimes) \in S^\otimes \times A^\otimes \times S^\otimes; P^\otimes(s'^\otimes | s^\otimes, a) > 0\}$  is the set of transitions, and  $\mathcal{F}^\otimes = \{\bar{F}_1^\otimes, \dots, \bar{F}_n^\otimes\}$  is the acceptance condition, where  $\bar{F}_i^\otimes = \{((s, \bar{x}), a, (s', \bar{x}')) \in \delta^\otimes; (\bar{x}, L(s, a, s'), \bar{x}') \in \bar{F}_i\}$  for each  $i \in \{1, \dots, n\}$ .

**Definition 10:** The reward function  $\mathcal{R}: S^\otimes \times A^\otimes \times S^\otimes \rightarrow \mathbb{R}_0$  is defined as

$$\mathcal{R}(s^\otimes, a, s'^\otimes) = \begin{cases} r_p & \text{if } \exists i \in \{1, \dots, n\}, (s^\otimes, a, s'^\otimes) \in \bar{F}_i^\otimes, \\ 0 & \text{otherwise,} \end{cases}$$

where  $r_p$  is a positive value.

The reward can be obtained if the first visits to the accepting sets is occurred regardless of its orders after a reset to  $v = \mathbf{0}$ .

**Remark 1:** When constructing a tBA from a tGBA, the order of visits to accepting sets of the tGBA is fixed. Consequently, the rewards based on the acceptance condition of the tBA tends to be sparse and the sparsity is critical against RL-based controller synthesis problems. The augmentation

of tGBA relaxes the sparsity since the augmented tGBA has all of the order of visits to all accepting sets of the original tGBA. For the acceptance condition  $\mathcal{F}$  of a tGBA, the size of the state space of the augmented tGBA is about  $\frac{2^{|\mathcal{F}|-1}}{|\mathcal{F}|}$  times larger than the tBA, however, the ratio of the number of accepting transitions to the number of all transitions of the augmented tGBA is much greater than the tBA. Therefore, our proposed method is expected more sample efficient than using tLDBA.

Under the product MDP  $M^\otimes$  and the reward function  $\mathcal{R}$ , which is based on the acceptance condition of  $M^\otimes$ , we show that if there exists a positional policy  $\pi$  satisfying the LTL specification  $\varphi$  on  $M^\otimes$ , maximizing the expected discounted reward with large enough  $\gamma$  produces a positional policy satisfying  $\varphi$  on  $M^\otimes$ .

For a Markov chain  $MC_\pi^\otimes$  induced by a product MDP  $M^\otimes$  with a positional policy  $\pi$ , let  $S_\pi^\otimes = T_\pi^\otimes \cup R_\pi^{\otimes 1} \cup \dots \cup R_\pi^{\otimes h}$  be the set of states in  $MC_\pi^\otimes$ , where  $T_\pi^\otimes$  is the set of transient states and  $R_\pi^{\otimes i}$  is the recurrent class for each  $i \in \{1, \dots, h\}$ , and let  $R(MC_\pi^\otimes)$  be the set of all recurrent states in  $MC_\pi^\otimes$ . Let  $\delta_\pi^{\otimes i}$  be the set of transitions in a recurrent class  $R_\pi^{\otimes i}$ , namely  $\delta_\pi^{\otimes i} = \{(s^\otimes, a, s'^\otimes) \in \delta^\otimes; s^\otimes \in R_\pi^{\otimes i}, P^\otimes(s'^\otimes | s^\otimes, a) > 0\}$ , and let  $P_\pi^\otimes : S_\pi^\otimes \times S_\pi^\otimes \rightarrow [0, 1]$  be the transition probability under  $\pi$ . In the following, for simplicity, we omit the superscript  $\otimes$  of product states, but remain some superscripts and subscripts to emphasize that we consider a product MDP  $M^\otimes$  with a positional policy  $\pi$ .

**Lemma 1:** For any policy  $\pi$  and any recurrent class  $R_\pi^{\otimes i}$  in the Markov chain  $MC_\pi^\otimes$ ,  $MC_\pi^\otimes$  satisfies one of the following conditions.

- 1)  $\delta_\pi^{\otimes i} \cap \bar{F}_j^\otimes \neq \emptyset, \forall j \in \{1, \dots, n\}$ ,
- 2)  $\delta_\pi^{\otimes i} \cap \bar{F}_j^\otimes = \emptyset, \forall j \in \{1, \dots, n\}$ .

*Proof:* Suppose that  $MC_\pi^\otimes$  satisfies neither conditions 1 nor 2. Then, there exists a policy  $\pi, i \in \{1, \dots, h\}$ , and  $j_1, j_2 \in \{1, \dots, n\}$  such that  $\delta_\pi^{\otimes i} \cap \bar{F}_{j_1}^\otimes = \emptyset$  and  $\delta_\pi^{\otimes i} \cap \bar{F}_{j_2}^\otimes \neq \emptyset$ . In other words, there exists a nonempty and proper subset  $J \in 2^{\{1, \dots, n\}} \setminus \{\{1, \dots, n\}, \emptyset\}$  such that  $\delta_\pi^{\otimes i} \cap \bar{F}_j^\otimes \neq \emptyset$  for any  $j \in J$ . For any transition  $(s, a, s') \in \delta_\pi^{\otimes i} \cap \bar{F}_j^\otimes$ , the following equation holds by the properties of the recurrent states in  $MC_\pi^\otimes$  [13].

$$\sum_{k=0}^{\infty} p^k((s, a, s'), (s, a, s')) = \infty, \quad (1)$$

where  $p^k((s, a, s'), (s, a, s'))$  is the probability that the transition  $(s, a, s')$  occurs again after the occurrence of itself in  $k$  time steps. Eq. (1) means that the agent obtains a reward infinitely often. This contradicts the definition of the acceptance condition of the product MDP  $M^\otimes$ . ■

Lemma 1 implies that for an LTL formula  $\varphi$  if a path  $\rho$  under a policy  $\pi$  does not satisfy  $\varphi$ , then the agent obtains no reward in recurrent classes; otherwise there exists at least one recurrent class where the agent obtains rewards infinitely often.

**Theorem 1:** For a product MDP  $M^\otimes$  of an MDP  $M$  and an augmented tLDBA  $\bar{B}_\varphi$  corresponding to a given LTL

formula  $\varphi$  and the reward function given by Def. 10, if there exists a positional policy satisfying  $\varphi$  on  $M^\otimes$ , then there exists a discount factor  $\gamma^*$  such that any algorithm that maximizes the expected discounted reward with  $\gamma > \gamma^*$  will find a positional policy satisfying  $\varphi$  on  $M^\otimes$ .

*Proof:* Suppose that  $\pi^*$  is an optimal policy but does not satisfy the LTL formula  $\varphi$ . Then, for any recurrent class  $R_{\pi^*}^{\otimes i}$  in the Markov chain  $MC_{\pi^*}^\otimes$  and any accepting set  $\bar{F}_j^\otimes$  of the product MDP  $M^\otimes$ ,  $\delta_{\pi^*}^{\otimes i} \cap \bar{F}_j^\otimes = \emptyset$  holds by Lemma 1. Thus, the agent under the policy  $\pi^*$  can obtain rewards only in the set of transient states. We consider the best scenario in the assumption. Let  $p^k(s, s')$  be the probability of going to a state  $s'$  in  $k$  time steps after leaving the state  $s$ , and let  $Post(T_{\pi^*}^\otimes)$  be the set of states in recurrent classes that can be transitioned from states in  $T_{\pi^*}^\otimes$  by one action. For the initial state  $s_{init}^\otimes$  in the set of transient states, it holds that

$$\begin{aligned} V^{\pi^*}(s_{init}) &= \sum_{k=0}^{\infty} \sum_{s \in T_{\pi^*}^\otimes} \gamma^k p^k(s_{init}, s) \\ &\quad \sum_{s' \in T_{\pi^*}^\otimes \cup Post(T_{\pi^*}^\otimes)} P_{\pi^*}^\otimes(s' | s) \mathcal{R}(s, a, s') \\ &\leq r_p \sum_{k=0}^{\infty} \sum_{s \in T_{\pi^*}^\otimes} \gamma^k p^k(s_{init}, s), \end{aligned}$$

where the action  $a$  is selected by  $\pi^*$ . By the property of the transient states, for any state  $s^\otimes$  in  $T_{\pi^*}^\otimes$ , there exists a bounded positive value  $m$  such that  $\sum_{k=0}^{\infty} \gamma^k p^k(s_{init}, s) \leq \sum_{k=0}^{\infty} p^k(s_{init}, s) < m$  [13]. Therefore, there exists a bounded positive value  $\bar{m}$  such that  $V^{\pi^*}(s_{init}^\otimes) < \bar{m}$ . Let  $\bar{\pi}$  be a positional policy satisfying  $\varphi$ . We consider the following two cases.

- 1) Assume that the initial state  $s_{init}$  is in a recurrent class  $R_{\bar{\pi}}^{\otimes i}$  for some  $i \in \{1, \dots, h\}$ . For any accepting set  $\bar{F}_j^\otimes$ ,  $\delta_{\bar{\pi}}^{\otimes i} \cap \bar{F}_j^\otimes \neq \emptyset$  holds by the definition of  $\bar{\pi}$ . The expected discounted reward for  $s_{init}$  is given by

$$\begin{aligned} V^{\bar{\pi}}(s_{init}) &= \sum_{k=0}^{\infty} \sum_{s \in R_{\bar{\pi}}^{\otimes i}} \gamma^k p^k(s_{init}, s) \\ &\quad \sum_{s' \in R_{\bar{\pi}}^{\otimes i}} P_{\bar{\pi}}^\otimes(s' | s) \mathcal{R}(s, a, s'), \end{aligned}$$

where the action  $a$  is selected by  $\bar{\pi}$ . Since  $s_{init}$  is in  $R_{\bar{\pi}}^{\otimes i}$ , there exists a positive number  $\bar{k} = \min\{k; k \geq n, p^k(s_{init}, s_{init}) > 0\}$  [13]. We consider the worst scenario in this case. It holds that

$$\begin{aligned} V^{\bar{\pi}}(s_{init}^\otimes) &\geq \sum_{k=n}^{\infty} p^k(s_{init}, s_{init}) \sum_{i=1}^n \gamma^{k-i} r_p \\ &\geq \sum_{k=1}^{\infty} p^{k\bar{k}}(s_{init}, s_{init}) \sum_{i=0}^{n-1} \gamma^{k\bar{k}-i} r_p \\ &> r_p \sum_{k=1}^{\infty} \gamma^{k\bar{k}} p^{k\bar{k}}(s_{init}, s_{init}), \end{aligned}$$

whereas all states in  $R(MC_{\bar{\pi}}^{\otimes})$  are positive recurrent because  $|S^{\otimes}| < \infty$  [14]. Obviously,  $p^{kk}(s_{init}, s_{init}) \geq (p^k(s_{init}, s_{init}))^k > 0$  holds for any  $k \in (0, \infty)$  by the Chapman-Kolmogorov equation [13]. Furthermore, we have  $\lim_{k \rightarrow \infty} p^{kk}(s_{init}, s_{init}) > 0$  by the property of irreducibility and positive recurrence [15]. Hence, there exists  $\bar{p}$  such that  $0 < \bar{p} < p^{kk}(s_{init}, s_{init})$  for any  $k \in (0, \infty]$  and we have

$$V^{\bar{\pi}}(s_{init}) > r_p \bar{p} \frac{\gamma^{\bar{k}}}{1 - \gamma^{\bar{k}}}.$$

Therefore, for any  $\bar{m} \in (V^{\pi^*}(s_{init}), \infty)$  and any  $r_p < \infty$ , there exists  $\gamma^* < 1$  such that  $\gamma > \gamma^*$  implies  $V^{\bar{\pi}}(s_{init}) > r_p \bar{p} \frac{\gamma^{\bar{k}}}{1 - \gamma^{\bar{k}}} > \bar{m}$ .

- 2) Assume that the initial state  $s_{init}$  is in the set of transient states  $T_{\bar{\pi}}^{\otimes}$ .  $P_{\bar{\pi}}^{M^{\otimes}}(s_{init} \models \varphi) > 0$  holds by the definition of  $\bar{\pi}$ . For a recurrent class  $R_{\bar{\pi}}^{\otimes i}$  such that  $\delta_{\bar{\pi}}^{\otimes i} \cap F_j^{\otimes} \neq \emptyset$  for each accepting set  $F_j^{\otimes}$ , there exist a number  $\bar{l} > 0$ , a state  $\hat{s}$  in  $Post(T_{\bar{\pi}}^{\otimes}) \cap R_{\bar{\pi}}^{\otimes i}$ , and a subset of transient states  $\{s_1, \dots, s_{\bar{l}-1}\} \subset T_{\bar{\pi}}^{\otimes}$  such that  $p(s_{init}, s_1) > 0$ ,  $p(s_i, s_{i+1}) > 0$  for  $i \in \{1, \dots, \bar{l}-2\}$ , and  $p(s_{\bar{l}-1}, \hat{s}) > 0$  by the property of transient states. Hence, it holds that  $p^{\bar{l}}(s_{init}, \hat{s}) > 0$  for the state  $\hat{s}^{\otimes}$ . Thus, by ignoring rewards in  $T_{\bar{\pi}}^{\otimes}$ , we have

$$\begin{aligned} V^{\bar{\pi}}(s_{init}) &\geq \gamma^{\bar{l}} p^{\bar{l}}(s_{init}, \hat{s}) \sum_{k=0}^{\infty} \sum_{s' \in R_{\bar{\pi}}^{\otimes i}} \gamma^k p^k(\hat{s}, s') \\ &\quad \sum_{s'' \in R_{\bar{\pi}}^{\otimes i}} P_{\bar{\pi}}^{\otimes}(s''|s') \mathcal{R}(s', a, s'') \\ &> \gamma^{\bar{l}} p^{\bar{l}}(s_{init}, \hat{s}) r_p \bar{p} \frac{\gamma^{\bar{k}'}}{1 - \gamma^{\bar{k}'}} \end{aligned}$$

where  $\bar{k}' \geq n$  is a constant and  $0 < \bar{p} < p^{k\bar{k}'}(\hat{s}, \hat{s})$  for any  $k \in (0, \infty]$ . Therefore, for any  $r_p < \infty$  and any  $\bar{m} \in (V^{\pi^*}(s_{init}), \infty)$ , there exists  $\gamma^* < 1$  such that  $\gamma > \gamma^*$  implies  $V^{\bar{\pi}}(s_{init}) > \gamma^{\bar{l}} p^{\bar{l}}(s_{init}, \hat{s}) \frac{r_p \bar{p} \gamma^{\bar{k}'}}{1 - \gamma^{\bar{k}'}} > \bar{m}$ .

The results contradict the optimality assumption of  $\pi^*$ . ■

Theorem 1 implies that for the product MDP  $M^{\otimes}$  of an MDP  $M$  and an augmented tLDBA corresponding to the given LTL formula  $\varphi$ , we can obtain a feasible positional policy satisfying  $\varphi$  on  $M^{\otimes}$  by an algorithm maximizing the expected discounted reward with large enough discount factor if there exists a positional policy on  $M^{\otimes}$  satisfying  $\varphi$  with non-zero probability.

#### IV. EXAMPLE

In this section, we compare our proposed method with the method in [9], [11] and the method using tLDBA. In the method using tLDBA, the reward function given by like Def. 10, namely it is based on the one accepting set of the corresponding product MDP. We consider a path planning problem of a robot in an environment consisting of eight rooms and one corridor as shown in Fig. 3. The state  $s_7$  is the initial state and the action space is

specified with  $\mathcal{A}(s) = \{Right, Left, Up, Down\}$  for any state  $s \neq s_4$  and  $\mathcal{A}(s_4) = \{to_{s_0}, to_{s_1}, to_{s_2}, to_{s_3}, to_{s_5}, to_{s_6}, to_{s_7}, to_{s_8}\}$ , where  $to_{s_i}$  means attempting to go to the state  $s_i$  for  $i \in \{0, 1, 2, 3, 5, 6, 7, 8\}$ . The robot moves in the intended direction with probability 0.9 and it stays in the same state with probability 0.1 if it is in the state  $s_4$ . In the states other than  $s_4$ , it moves in the intended direction with probability 0.9 and it moves in the opposite direction to that it intended to go with probability 0.1. If the robot tries to go to outside the environment, it stays in the same state. The labeling function is as follows.

$$L((s, a, s')) = \begin{cases} \{c\} & \text{if } s' = s_i, i \in \{2, 3, 5, 6\}, \\ \{a\} & \text{if } (s, a, s') = (s_4, to_{s_0}, s_0), \\ \{b\} & \text{if } (s, a, s') = (s_4, to_{s_8}, s_8), \\ \emptyset & \text{otherwise.} \end{cases}$$

In the example, the robot tries to take two transitions that we want to occur infinitely often, represented by arcs labeled by  $\{a\}$  and  $\{b\}$ , while avoiding unsafe transitions represented by the arcs labeled by  $\{c\}$ . This is formally specified by the following LTL formula.

$$\varphi = \mathbf{GF}a \wedge \mathbf{GF}b \wedge \mathbf{G}\neg c.$$

The above LTL formula requires the robot to keep on entering the two rooms  $s_0$  and  $s_8$  from the corridor  $s_4$  regardless of the order of entries, while avoiding entering the four rooms  $s_2, s_3, s_5$ , and  $s_6$ . They have two accepting sets.

The tLDBA  $B_{\varphi} = (X, x_{init}, \Sigma, \delta, \mathcal{F})$  and its augmented automaton  $\bar{B}_{\varphi} = (\bar{X}, \bar{x}_{init}, \bar{\Sigma}, \bar{\delta}, \bar{\mathcal{F}})$  are shown in Figs. 1 and 2, respectively. Specifically, the acceptance condition  $\mathcal{F}$  of the tLDBA is given by  $\mathcal{F} = \{F_1, F_2\}$ , where  $F_1 = \{(x_0, \{a\}, x_0), (x_0, \{a, b\}, x_0)\}$  and  $F_2 = \{(x_0, \{b\}, x_0), (x_0, \{a, b\}, x_0)\}$ .

For the three methods, we use Q-learning<sup>1</sup> with  $\varepsilon$ -greedy policy and gradually reduce  $\varepsilon$  to 0 to learn an optimal policy asymptotically. We set the positive reward  $r_p = 2$ , the epsilon greedy parameter  $\varepsilon = \frac{0.95}{n_t(s^{\otimes})}$ , where  $n_t(s^{\otimes})$  is the number of visits to state  $s^{\otimes}$  within  $t$  time steps [16], and the discount factor  $\gamma = 0.95$ . The learning rate  $\alpha$  varies in accordance

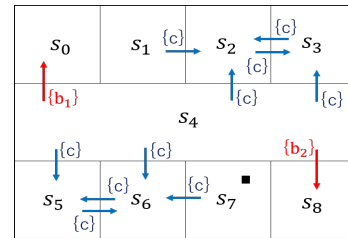


Fig. 3. The environment consisting of eight rooms and one corridor. Red arcs are the transitions that we want to occur infinitely often, while blue arcs are the transitions that we never want to occur.  $s_7$  is the initial state.

<sup>1</sup>We employ Q-learning here but any algorithm that maximizes the discounted expected reward can be applied to our proposed method.

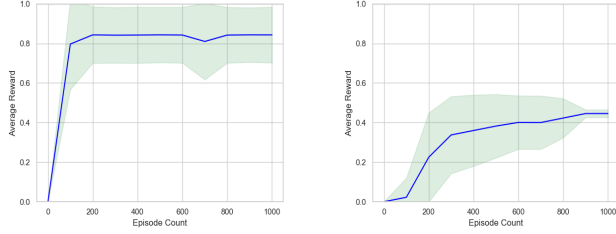


Fig. 4. The arithmetic mean of average reward in each episode for 20 learning sessions obtained from our proposed method (left) and the method using tLDBA (right). They are plotted per 100 episodes and the green areas represent the range of standard deviations. Each theoretical optimal convergence value are about

with the *Robbins-Monro condition*. We train the agent in 10000 iterations and 1000 episodes for 20 learning sessions.

In [9], [11], the accepting frontier function is proposed, which is  $Acc : X \times 2^X \rightarrow 2^X$  where  $X$  is the set of states of the state-based LDBA. Under initializing a set of states  $\mathbb{F}$  with the union of the all accepting sets of the state-based LDBA, the function receives the state  $x$  after each transition and the set  $\mathbb{F}$ . If  $x$  is in  $\mathbb{F}$ , then  $Acc$  removes the accepting sets containing  $x$  from  $\mathbb{F}$ . For the product MDP of an MDP  $M$  and an state-based LDBA  $B_\varphi$ , the reward function depends on the removed states of  $B_\varphi$ .

Fig. 4 shows the average rewards obtained by our proposed method and the method using tLDBA, respectively. Both of methods eventually acquire an optimal policy satisfying  $\varphi$ . As can be seen from Fig. 4, however, our proposed method converges faster. It suggest that our proposed method is expected to be more sample efficient than using tLDBA. It is expected that such learning delays are remarkable as the number of accepting sets increases.

Fig. 5 shows the obtained optimal policies by our proposed method and the method in [9], [11], respectively.

The result for the method in [9], [11] is due to that it is impossible the transitions labeled by  $\{a\}$  and  $\{b\}$  occur from  $s_4$  infinitely often by any positional policy with the tLDGBA shown in Fig. 1. In detail, the state of the tLDGBA is always  $x_0$  while the agent does not move to states  $s_2$ ,  $s_3$ ,  $s_5$ , and  $s_6$ . Thus, the state of the product MDP is always  $(s_4, x_0)$  while the agent stays in  $s_4$ . While our proposed method can recognize the previous visits as a state. Thus, our proposed method can synthesize a positional policy satisfying  $\varphi$  on the product MDP, while the method in [9], [11] cannot depending on a tLDGBA associated with  $\varphi$ . That is, in order to obtain a positional policy satisfying a given LTL formula  $\varphi$  with the method in [9], [11], we may have to make an LDGBA associated with  $\varphi$  redundant. However, the redundancy can cause a sparsity of rewards or large search space or something bad.

## V. CONCLUSIONS

The letter proposed a novel RL-based method for the synthesis of a control policy for an LTL specification using a limit-deterministic generalized Büchi automaton. The proposed method improved the learning performance compared

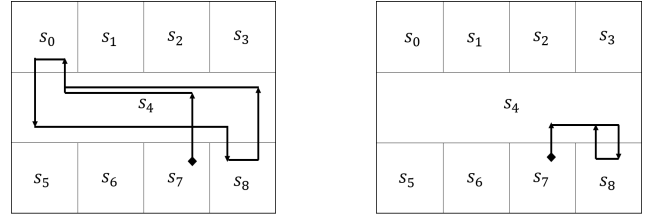


Fig. 5. The optimal policy obtained from our proposed method (left) and the method in [9], [11] (right).

to existing methods. It is future work to investigate a method that maximizes the satisfaction probability.

## REFERENCES

- [1] C. Baier and J.-P. Katoen, *Principles of Model Checking*. MIT Press, 2008.
- [2] C. Belta, B. Yordanov, and E. A. Gol, *Formal Methods for Discrete-Time Dynamical Systems*. Springer, 2017.
- [3] M. L. Puterman, *Markov Decision Processes, Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.
- [4] E. M. Wolff, U. Topcu, and R. M. Murray, "Robust control of uncertain Markov decision processes with temporal logic specifications," in *Proc. 51st IEEE Conference on Decision and Control*, 2012, pp. 3372–3379.
- [5] D. Sadigh, E. S. Kim, A. Coogan, S. S. Sastry, and S. Seshia, "A learning based approach to control synthesis of Markov decision processes for linear temporal logic specifications," in *Proc. 53rd IEEE Conference on Decision and Control*, pp. 1091–1096, 2014.
- [6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd Edition. MIT Press, 2018.
- [7] M. Hiromoto and T. Ushio, "Learning an optimal control policy for a Markov decision process under linear temporal logic specifications," in *Proc. 2015 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 2015, pp. 548–555.
- [8] S. Sickert, J. Esparaza, S. Jaax, and J. Křetínský, "Limit-deterministic Büchi automata for linear temporal logic," in *International Conference on Computer Aided Verification*, 2016, pp. 312–332.
- [9] M. Hasanbeig, A. Abate, and D. Kroening, "Logically-constrained reinforcement learning," *arXiv:1801.08099v8*, Feb. 2019.
- [10] E. M. Hahn, M. Perez, S. Schewe, F. Somenzi, A. Triverdi, and D. Wojtczak, "Omega-regular objective in model-free reinforcement learning," *Lecture Notes in Computer Science*, no. 11427, pp. 395–412, 2019.
- [11] M. Hasanbeig, Y. Kantaros, A. Abate, D. Kroening, G. J. Pappas, and I. Lee, "Reinforcement learning for temporal logic control synthesis with probabilistic satisfaction guarantee," *arXiv:1909.05304v1*, 2019.
- [12] A. K. Bozkurt, Y. Wang, M. Zavlanos, and M. Pajic, "Control synthesis from linear temporal logic specifications using model-free reinforcement learning," *arXiv:1909.07299*, 2019.
- [13] R. Durrett, *Essentials of Stochastic Processes*, 2nd Edition. ser. Springer texts in statistics. New York; London; Springer, 2012.
- [14] L. Breuer, "Introduction to Stochastic Processes," [Online]. Available: <https://www.kent.ac.uk/smsas/personal/lb209/files/sp07.pdf>
- [15] S.M. Ross, *Stochastic Processes*, 2nd Edition. University of California, Wiley, 1995.
- [16] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement learning algorithms" *Machine Learning*, vol. 38, no. 3, pp. 287–308, 1998.
- [17] J. Křetínský, T. Meggendorfer, S. Sickert, and C. Ziegler, "Rabinizer 4: From ltl to your favourite deterministic automaton," *In CAV*, 2018. To appear.