Continuous Motion Planning with Temporal Logic Specifications using Deep Neural Networks

Chuanzheng Wang, Yinan Li, Stephen L. Smith, Jun Liu

Abstract—In this paper, we propose a model-free reinforcement learning method to synthesize control policies for motion planning problems for continuous states and actions. The robot is modelled as a labeled Markov decision process (MDP) with continuous state and action spaces. Linear temporal logics (LTL) are used to specify high-level tasks. We then train deep neural networks to approximate the value function and policy using an actor-critic reinforcement learning method. The LTL specification is converted into an annotated limit-deterministic Büchi automaton (LDBA) for continuously shaping the reward so that dense reward is available during training. A naive way of solving a motion planning problem with LTL specifications using reinforcement learning is to sample a trajectory and, if the trajectory satisfies the entire LTL formula then we assign a high reward for training. However, the sampling complexity needed to find such a trajectory is too high when we have a complex LTL formula for continuous state and action spaces. As a result, it is very unlikely that we get enough reward for training if all sample trajectories start from the initial state in the automata. In this paper, we propose a method that samples not only an initial state from the state space, but also an arbitrary state in the automata at the beginning of each training episode. We test our algorithm in simulation using a car-like robot and find out that our method can learn policies for different working configurations and LTL specifications successfully.

I. INTRODUCTION

Traditionally, motion planning problems consider generating a trajectory for reaching a specific target while avoiding obstacles [10]. However, real-world applications often require more complex tasks than simply reaching a target. As a result, recent motion planning problems consider a class of high-level complex specifications that can be used to describe a richer class of tasks. A branch of planning approaches has been proposed recently that describes high-level tasks like reaching a sequence of goals or ordering a set of events using formal languages such as linear temporal logic (LTL) [15]. As a simple example, the task of reaching region A and then reaching region B can be easily expressed as an LTL formula. To deal with LTL specifications, an approach for dealing with point-mass robot model has been proposed in [5]. A control synthesis technique with receding horizon control has been proposed in [28] to handle a linear robot model. The approach in [2] takes advantage of the latest development in sampling-based methods to deal with nonlinear dynamic robot models. However, this method suffers from the curse of dimensionality when we have a high-dimensional system as in most of the sampling-based approaches. As a result, a method with more computational efficiency is needed to solve motion planning problems with LTL specifications for complex dynamical systems.

Reinforcement learning has achieved great success in the past decades with both theoretical results [26] and applications [22] [24]. It is a way of learning the best actions for a Markov decision process (MDP) by interacting with the environment [25]. It is efficient in solving problems of high-dimensional systems with or without knowing a model. Early works were mainly based on Q-learning [27] and policy gradient methods [26]. The actor-critic algorithm [19] is also widely used with two components, namely an actor and a critic. The actor is used as the policy, which tells the system what action should be taken at each state, and the critic is used to approximate the state-action value function. Modern reinforcement learning methods take advantage of deep neural networks to solve problems with large state and action spaces. A deep Q-network (DQN) [16] uses a deep neural network to approximate state-action values and learns an implicit control policy by improving this Q-network. In [23], a deterministic policy gradient method is proposed with better time efficiency and consequently the deep deterministic policy gradient method (DDPG) [14] leverages this idea of a deterministic policy and uses two deep neural networks, an actor network and a critic network, to solve problems of continuous state and action spaces.

Due to the recent developments, reinforcement learning algorithms have been applied to solve model-free robotic control problems with temporal logic specifications. In [20], a Q-learning method is used to solve an MDP problem with LTL specifications. The temporal logical formula is transformed into a deterministic Rabin automaton (DRA) and a real-valued reward function is designed in order to satisfy complex requirements. In [6], the LTL specification is also converted into a DRA and a reduced variance deep Q-learning method is used to approximate the state-action values of the product MDP with the help of deep neural networks. Another branch of methods convert the LTL formula into a limit-deterministic Büchi automaton (LDBA) and a synchronous reward function is designed based on the acceptance condition of the LDBA as in [9] and [7]. Also, the authors in [7] claim that one may fail to find optimal strategies by converting the LTL formula into a DRA. In [18], the limit-deterministic generalized Büchi automaton (LDGBA) is used to convert the LTL formula. Moreover, a continuous state space is considered in [8].

Almost all the previous papers on reinforcement learn-
ing using temporal logic specifications focused on discrete models without continuous actions, with the exception that in [12], a policy described using a time-varying linear Gaussian process is updated by maximizing the robustness function in each step. A good reward is obtained to improve the policy if a training episode produces a trajectory that successfully reaches an accepting condition in the DRA or LDBA. However, training deep networks for continuous controls is more challenging due to significantly increased sample complexity and most approaches achieve poor performance on hierarchical tasks, even with extensive hyperparameter tuning [4]. This is because that the reward function is so sparse if we only have a terminal reward when the accepting conditions are satisfied. As a result, in this paper, we demonstrate that using an annotated LDBA converted from the LTL specification and a simple idea that randomly samples from the automaton states without initializing it to a fixed initial state (as given by the translated automaton), we can effectively train deep networks to solve continuous control problems with temporal logic goals. We show in simulations that our method achieves a good performance for a nonlinear robot model with complex LTL specifications.

II. PRELIMINARY AND PROBLEM DEFINITION

A. Linear Temporal Logic

Linear temporal logic (LTL) formulas are composed over a set of atomic propositions \( AP \) by the following syntax:

\[
\varphi ::= \text{true} \mid p \mid \neg \varphi \mid \varphi \land \varphi \mid \varphi \lor \varphi \mid \varphi \ U \varphi \mid \varphi \ O \varphi,
\]

where \( p \in AP \) is an atomic proposition, \( \text{true}, \neg \) are propositional logic operators and \( \land, \lor, \ U, \ O \) are temporal operators.

Other propositional logic operators such as \( \text{false}, \lor, \land, \neg \) are temporal operators always (\( [] \)), eventually (\( [] \)), implication (\( \rightarrow \)); temporal operators we denote by \( w \models \varphi \) if \( w \) satisfies the LTL formula \( \varphi \).

Details on syntax and semantics of LTL can be found in [1].

For probabilistic systems such as Markov decision processes, it is sufficient to use a limit-deterministic Büchi automaton (LDBA) over the set of symbols \( \Sigma \), which are deterministic in the limit, to guide the verification or control synthesis with respect to an LTL formula. For any LTL formula \( \varphi \), there exists an equivalent LDBA that accepts exactly the words described by \( \varphi \) [21].

In the current paper, we use transitions-based LDBA, since it is often of smaller size than its state-based version. We begin by defining a transition-based Büchi automaton and then give a formal definition of an LDBA.

Definition 1: A transition-based generalized Büchi automaton (TGBA) is a tuple \( A = (Q, \Sigma, \delta, q_0, F) \), where \( Q \) is a set of states, \( \Sigma \) is a set of finite alphabet, \( \delta : Q \times \Sigma \rightarrow 2^Q \) is the state transition function, \( q_0 \in Q \) is the initial state, and \( F = \{F_1, \ldots, F_k\} \) with \( F_i \subseteq Q \times Q \times Q \) (i \( \in \{1, \ldots, k\}, k \geq 1 \)) is a set of accepting conditions.

A run of a TGBA \( A \) under an input word \( w = \sigma_0\sigma_1\cdots \) is an infinite sequence of transitions in \( Q \times \Sigma \times Q \), denoted by \( \xi = (v_0, \sigma_0, v_1)(v_1, \sigma_1, v_2)\cdots \), that satisfies \( v_{i+1} \in \delta(v_i, \sigma_i) \) and \( v_i \in Q \) for all \( i \in \mathbb{N} \). Let \( \xi[i] = (v_i, \sigma_i, v_{i+1}) \). OutProps(q) = \{\sigma \in \Sigma | \exists q' \in Q \text{ s.t. } q' \in \delta(q, \sigma) \} \text{ and OutEdges(q) = } \{(q, \sigma, q') \mid \exists q' \in Q, \sigma \in \Sigma \text{ s.t. } q' \in \delta(q, \sigma)\}. \text{ Denote by } (q, \sigma, q') \text{ the transition between } q, q' \in Q \text{ under the input } \sigma \in \Sigma. \text{ A word } w \text{ is accepted by } A \text{ if there exists a run } \xi \text{ such that } \text{Inf}(\xi) \cap F_j \neq \emptyset \text{ for all } j \in \{1, \ldots, k\}, \text{ where } \text{Inf}(\xi) = \{(v, \sigma, v') \in Q \times Q \times Q \mid \forall i, i+1 > j, i, \text{ s.t. } (v, \sigma, v') = \xi[j] \}. \text{ Let } Q_N \text{ be the set of transitions that occur infinitely often during the run } \xi.

Definition 2: A TGBA \( A = (Q, \Sigma, \{\delta, q_0, F\}) \) is a limit-deterministic Büchi automaton (LDBA) if \( Q = Q_N \cup Q_D \), \( Q_N \cap Q_D = \emptyset \), and

- \( \delta(q, \sigma) \subseteq Q_D \text{ and } |\delta(q, \sigma)| = 1 \) for all \( q \in Q_D \) and \( \sigma \in \Sigma \),
- \( \delta(q, \sigma) \cap Q_N = 1 \) for all \( q \in Q_N \) and \( \sigma \in \Sigma \),
- \( F \subseteq Q_D \times Q \times Q_D \) for all \( F \in F \).

The transitions from \( Q_N \) to \( Q_D \) are called \( \varepsilon \)-transitions defined in the last condition of Definition 2 which are taken without reading any input propositions in \( \Sigma \). The correctness of taking any of the \( \varepsilon \)-transitions can be checked by the transitions in \( Q_D \) [21]. For an LDBA \( A \), we define OutProps(q) = \{\sigma \in \Sigma | \exists q' \in Q \text{ s.t. } q' \in \delta(q, \sigma)\}.

B. Labeled Markov Decision Process

To capture the robot motion and working properties, we use a continuous labeled Markov decision process to describe the dynamics of the robot and its interaction with the environment.

Definition 3: A continuous labeled Markov Decision Process (MDP) is a tuple \( M = (S, A, P, R, \gamma, AP, L) \), where \( S \subseteq \mathbb{R}^n \) is a continuous state space, \( A \subseteq \mathbb{R}^m \) is a continuous action space, \( P : S \times A \rightarrow \mathcal{X}(S) \) is a transition probability kernel with \( \gamma(\cdot|s, a) \) defining the next-state distribution of taking action \( a \in A \) at state \( s \in S \), the function \( R : S \times A \times S \rightarrow \mathbb{R} \) specifies the reward, \( \gamma \) is a discount factor, \( AP \) is the set of atomic propositions, and \( L : S \rightarrow 2^{AP} \) is the labeling function that returns propositions that are satisfied at a state \( s \in S \). Here \( \mathcal{X}(S) \) denotes the set of all probability measures over \( S \).

The labeling function \( L \) is used to assign labels from a set \( AP \) of atomic propositions to each state in the state space \( S \). Given a sequence of states \( s = s_0s_1\cdots \), a sequence of symbols \( tr(s) = L(s_0)L(s_1)\cdots \), called the trace of \( s \), can be generated to verify if it meets a LTL specification \( \varphi \). If \( tr(s) \models \varphi \), we also write \( s \models \varphi \).

Definition 4: A deterministic policy \( \pi \) of a labeled MDP is a function \( \pi : S \rightarrow A \) that maps an action \( a \in A \) to a state \( s \in S \).

Given a labeled MDP, we can defined the accumulated reward starting from time step \( t \) as

\[
G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1},
\]
where \( R_t \) is the reward at time step \( t \).

C. Product MDP

Considering a robot operating in a working space to accomplish a high-level complex task described by an LTL-equivalent LDBA \( \mathcal{A} \), the mobility of the robot is captured by a labeled MDP \( M \) defined as above. We can combine the labeled MDP and the LDBA to obtain a product MDP.

**Definition 5:** A product MDP between a labeled MDP \( M = \{ S, A, P, R, \gamma, AP, L \} \) and an LDBA \( \mathcal{A} = (Q, \Sigma, \delta, q_0, F) \) is a tuple

\[
M_p = M \times \mathcal{A} := \{ S_p, A_p, P_p, R_p, AP, L, F_p, \gamma \},
\]

where

- \( S_p = S \times Q \) is the set of states,
- \( A_p = A \cup \{ \varepsilon \} \) is the set of actions,
- \( P_p : (S \times Q) \times A \to \varkappa(S, Q) \) is the transition probability kernel defined as

\[
P_p(s'_p|s_p, a_p) = \begin{cases} P(s'|s, a), & q' = \delta(q, L(s)), a_p \in A, \\ 1, & s = s', a_p = \varepsilon, \\ 0, & \text{otherwise}, \end{cases}
\]

for all \( s_p = (s, q), s'_p = (s', q') \in S_p \).
- \( R_p : S_p \times A_p \times S_p \to \mathbb{R} \) is the reward function, and
- \( F_p = \{ F^1_p, \ldots, F^k_p \} \) \((k \geq 1)\), where \( F^i_p = \{ ((s, q), a, (s', q')) \in S_p \times A_p \times S_p : (q, a, q') \in F_i \} \)

for all \( i \in \{1, \ldots, k\} \) is a set of accepting conditions.

Likewise, a run of a product MDP \( M_p = M \times \mathcal{A} \) is an infinite sequence of transitions of the form

\[
\xi = ((s_0, q_0), (s_1, q_1), (s_2, q_2), \ldots),
\]

where \((s_i, q_i), (s_{i+1}, q_{i+1}) \in S_p \times A_p \times S_p \). We say that \( \xi \) is an accepted run, denoted by \( \xi \models \mathcal{A} \), if \( \text{Inf}(\xi) \cap F^i_p \neq \emptyset \) for all \( i \in \{1, \ldots, k\} \), where \( \text{Inf}(\xi) \) is the set of transitions that occur infinitely often in \( \xi \).

D. Problem Formulation

We consider the problem in which a robot and its environment are modelled as an MDP \( M \), and the robot task is specified as an LTL formula \( \varphi \). Given an initial state \( s_0 \), we define the probability of an MDP \( M \) satisfying \( \varphi \) from \( s_0 \) as

\[
\mathbb{P}^M(\xi = \varphi) := \mathbb{P}(s \in \mathbb{P}^M | s \models \varphi, s = s_0s_1 \ldots),
\]

where \( \mathbb{P}^M \) is the set of all infinite sequences of states of the MDP that are induced from the policy \( \pi \). Then the problem we address in this paper is as follows.

**Problem 1:** Given a continuous labeled MDP \( M = \{ S, A, P, R, \gamma, AP, L \} \) and an LTL specification \( \varphi \), find a policy \( \pi^* \) such that \( \mathbb{P}^M(\xi = \varphi) \) is maximized for each \( s_0 \in S \).

As we have seen in Section II-A, an LTL formula \( \varphi \) can be translated into an LDBA \( \mathcal{A}_\varphi \). Therefore, solving Problem [1] is equivalent to solving the following control problem for the corresponding product MDP \( M_p \) of the given MDP \( M \) and \( \mathcal{A}_\varphi \) [3].
and the deterministic policy can be updated by
\[ \vartheta_{k+1} = \vartheta_k + \alpha E_{s \sim \rho^{\vartheta_k}}[\nabla \vartheta_k Q^\pi(s, \vartheta_k(s))], \]
where \( \alpha \in [0, 1] \) is a learning rate. By applying the chain rule,
\[ \vartheta_{k+1} = \vartheta_k + \alpha E_{s \sim \rho^{\vartheta_k}}[\nabla \vartheta_k \pi^{\vartheta_k}(s) \nabla_a Q^\pi(s, a)_{a=\pi^{\vartheta_k}(s)}]. \]

(2)

The DDPG method moves the parameter vector \( \vartheta \) greedily in the direction of the gradient of \( Q \) and is more efficient in solving MDP problems with continuous state and action spaces. As a result, we propose a learning method to solve motion planning problems with LTL specifications based on DDPG.

IV. REINFORCEMENT LEARNING WITH LDBA-GUIDED REWARD SHAPING

In this section, we introduce our method of solving a continuous state and action MDP with LTL specifications using deep reinforcement learning. The LTL specification is transformed into an annotated LDBA and a reward function is defined on the annotated LDBA for reward shaping in order to training the networks with dense reward.

A. Reward Shaping

Our definition of the reward function for the product MDP \( M_p \) depends on an annotated LDBA defined as follows.

**Definition 6:** An annotated LDBA \( (A, B) \) is an LDBA \( A = (Q, \Sigma, \delta, q_0, F) \) augmented by \( B = \{b_1, \ldots, b_k\} (k \geq 1) \), where \( F = \{F_1, \ldots, F_k\} \) and \( b_i : Q \times \Sigma \times Q \rightarrow \{0, 1\} (i \in \{1, \ldots, k\}) \) is a function assigning 0 or 1 to all the edges of \( A \) according to the following rules:

\[ b_i(q, \sigma, q') = \begin{cases} 1 & (q, \sigma, q') \in F_i, \\ 0 & \text{otherwise.} \end{cases} \]

For any \( i \in \{1, \ldots, k\} \), the map \( b_i \), which corresponds to \( F_i \), assigns 1 to all the accepting transitions in \( F_i \) and 0 to all others. The set \( B \) defined above, however, only marks the accepting transitions but not the other transitions that can be taken so that the accepting transitions can happen in some future steps. In order to also identify such transitions to guide the design of the reward function of the product MDP, we provide the following Algorithm 1 to pre-process the set \( B \) of boolean maps.

The function \( g : Q \rightarrow \{0, 1\} \) in line 1 of Algorithm 1 is defined to gradually mark every state in \( Q \) that has outgoing transitions annotated by 1. For each set \( F \in \mathcal{F} \), the function \( g \) is initialized (in line 3 and 4) to 1 for any state \( q \in Q \) that has at least one accepting outgoing transition and 0 for any other states. By using \( g \), the loop from line 5 to 12 in Algorithm 1 marks backwardly the state \( q \in Q \) with no outgoing transition marked 1 (i.e., \( g(q) = 0 \)), through which the accepting transitions can be taken. The loop terminates in a finite number of steps since the set \( Q \) of states is finite and \( g \) can only be marked to 1 not 0. After running Algorithm 1 for each \( i \in \{1, \ldots, k\} \), the map \( b_i \in \mathcal{B} \) marks 1 to the transitions that either are accepting in \( F_i \) or can lead to the occurrence of accepting transitions in \( F_i \). A state \( q \in Q \) with \( g(q) = 0 \) after the end of the loop (for all \( i \in \{1, \ldots, k\} \)) is marked as a trap, because accepting transitions do not occur in any run that passes through \( q \).

Since any accepting run of an LDBA \( A \) should contain infinitely many transitions from each \( F \in \mathcal{F} \), the status of whether there is at least one transition in any \( F \in \mathcal{F} \) is taken should be tracked. For this purpose, we let \( V \) be a Boolean vector of size \( k \times 1 \) and \( V[i] \) be the \( i \)-th element in \( V \), where \( k \) is the number of subsets in the accepting condition \( F \) and \( i \in \{1, \ldots, k\} \). The vector \( V \) is initialized to all ones and is updated according to the following rules:

- If a transition in set \( F_i \) is taken, then \( V[i] = 0 \).
- If all elements in \( V \) are 0, reset \( V \) to all ones.

Now we define a function \( b : (Q \times \Sigma \times Q) \rightarrow \{0, 1\} \) that is updated by vector \( V \) as follows:

\[ b(q, \sigma, q') = \bigvee_{i=1}^k (b_i(q, \sigma, q') \land V[i]). \]

(3)

For a transition \((q, \sigma, q')\), \( b(q, \sigma, q') = 1 \) if and only if there exists an \( F_i \) that has not been visited (i.e., \( V[i] = 1 \)) and \( b_i(q, \sigma, q') = 1 \).

Based on the above definitions, the reward function of the product MDP \( M_p \) is defined as:

\[ R_p(s_p, a, s'_p) = \begin{cases} r_n d(s, E)(1 - b(e)) + r_g b(e) & \exists \sigma, \text{s.t. } b(q, \sigma, q') = 1, \\ r_d & \text{otherwise,} \end{cases} \]

(4)

where \( s_p = (s, q), s'_p = (s', q'), e = (q, L(s), q') \), the numbers \( r_n \) and \( r_g \) satisfy \( r_d < r_n < 0 < r_g, |r_n| \ll
\( |r_q| \ll |r_g| \), the function \( b \) is given in \( 3 \). The set \( E \) is given by
\[
E = \bigcup_{\sigma \in \text{OutProps}(q)} L^{-1}(\sigma).
\] (5)
The term \( d(s, E) = \inf_{s' \in E} \{ d(s, s') \} \) measures the distance from the MDP state \( s \) to the set \( E \), where \( d(s, s') \) denotes the distance between the states \( s \) and \( s' \).

The large positive number \( r_q \) is used to reward taking an accepting transition or a transition that can lead to an accepting one, the small negative number \( r_n \) is used to guide the transitions in the state space of the MDP to encourage the occurrence of the desired transitions between LDBA states, and the negative reward \( r_d \) will be collected if the corresponding run in \( A \) hit a trap.

B. The Proposed Algorithm

The authors of [6] propose a method that initializes each episode with the initial Rabin state for a discrete product MDP model. The approach in [20] also resets the Rabin state with the initial state \( q_0 \) periodically. The main drawback of doing this is that we can only have a good reward if a training episode produces a trajectory that successfully reaches an accepting state in the DFA. However, for a product MDP with continuous state and action spaces, the sampling complexity of getting such a satisfactory trajectory is too high when we have a complex LTL formula and consequently, we cannot obtain enough reward to train the neural networks for a good performance. As a result, at the beginning of each episode, we sample a \( q_{\text{init}} \) instead of using the initial state \( q_0 \) as given by the translated automaton. Then the initial state of the product MDP is constructed by using this sampled \( q_{\text{init}} \).

We use DDPG [14] to train the neural networks. As most reinforcement learning algorithms in which data has to be independently and identically distributed, a buffer is used here for storing only the last \( N \) steps of transition data [16]. At each time step, the tuple \( \{ s_p, a, R_p, s'_p \} \) is stored into the buffer and a batch of data is uniformly sampled from the buffer for training the networks. As is shown in Algorithm 2 in line 16 and 17, the critic is updated with minimizing the loss function of the neural network and the actor is updated such that the average value is used to approximate the expectation as in Eq. [2]. It was discussed in [17] that directly implementing deep Q-learning with neural networks will be unstable because the \( Q \) value is also used for policy network training. As a result, a small change in the \( Q \) value may significantly change the policy and therefore change the data distribution. The authors proposed a way of solving this issue by cloning the \( Q \)-network to obtain a target network after each fixed number of updates. This modification makes the algorithm more stable compared with the standard deep Q-learning. We use two target networks \( Q^\omega(s_p, a) \) and \( \pi^{\theta'}(s_p) \) as in [14]. The target networks are copied from the actor and critic networks in the beginning and the weights of both networks are updated after every several steps by using \( \omega' \leftarrow \tau \omega + (1-\tau)\omega' \) and \( \theta' \leftarrow \tau \theta + (1-\tau)\theta' \) with \( \tau \ll 1 \).

The proposed method to solve a continuous MDP with LTL specifications is summarized in Algorithm 2

Algorithm 2 Actor-Critic Algorithm for Continuous Product MDP \( M_p \)

\textbf{Require:} labeled MDP \( M \), LDBA \( A \), product MDP \( M_p \)

\textbf{Ensure:} Policy \( \pi \)

1: Initialize the critic network \( Q^\omega \), the actor network \( \pi^{\theta} \) with arbitrary weights \( \omega \) and \( \theta \)
2: Copy target network \( Q^{\tau \omega} \) and \( \pi^{\tau \theta} \) with weights \( \omega' \leftarrow \omega \) and \( \theta' \leftarrow \theta \)
3: Initialize buffer \( B \)
4: for each episode do
5: Sample a state \( s_0 \) from \( S \) in \( M \)
6: Sample a \( q_{\text{init}} \) from \( Q \) in \( A \)
7: Construct initial state for \( M_p \) with \( s_{\text{init}} = (s_0, q_{\text{init}}) \)
8: while \( s_{\text{pt}}=(s_t, a_t) \) is not a terminal state do
9: Get an action \( a_t \) from \( \pi^{\theta} \)
10: Simulate from \( s_t \) to \( s'_t \)
11: Get LDBA state \( q_t' \leftarrow \delta(q_t, L(s_t)) \)
12: Get the next state \( s'_{pt} \leftarrow (s'_t, q_t') \)
13: Calculated process reward \( r(s_{pt}, a_t, s'_{pt}) \)
14: Store tuple \( \{ s_{pt}, a_t, r, s'_{pt} \} \) in buffer \( B \)
15: Sample \( N \) batches from the buffer and calculate target values for \( i \in N \) with
\[
y_i = r_i + \gamma Q^{\omega'}(s'_{pt}, \pi^{\theta'}(s'_{pt}))
\] (6)
16: Update the critic network by minimizing the loss function:
\[
L = \frac{1}{N} \sum_{i \in N} (y_i - Q^{\omega}(s_{pt}, \pi^{\theta}(s_{pt})))^2
\] (7)
17: Update the actor network according to:
\[
\theta_{k+1} \leftarrow \theta_k + \alpha \frac{1}{N} \sum_{i \in N} \nabla_{\theta} Q^{\omega}(s_{pt}, a_t) \nabla_{\theta} \pi^{\theta}(s_{pt})
\] (8)
18: Update state \( s_{pt} \leftarrow s'_{pt} \)
19: Update the target networks:
\[
\theta' \leftarrow \tau \theta + (1-\tau)\theta'
\]
\[
\omega' \leftarrow \tau \omega + (1-\tau)\omega'
\] (9)
20: end while
21: end for

C. Analysis of the Algorithm

While DDPG does not offer any convergence guarantees for approximating a general nonlinear value function, we prove in this section that, if the MDP is finite (e.g. obtained as a finite approximation of the underlying continuous-state MDP), the reward function defined by \( 4 \) does characterize
Problem correctly in the sense that maximizing the reward (value function) implies maximizing the satisfaction probability.

**Theorem 1:** Let $\varphi$ be an LTL formula and $M_p$ be the product MDP formed from the MDP $M$ and an LDBA translation $A_{\varphi}$ encoding $\varphi$. Then the optimal policy on $M_p$ that maximizes the expected accumulated reward using (4) also maximizes the probability of satisfying $A_{\varphi}$.

**Proof:** (Sketch of proof) Suppose that $\pi_i$ ($i = 1, 2$) are two policies such that $\pi_i$ has probability $p_i$ of satisfying $A_{\varphi}$ (i.e., producing accepting runs on $M_p$) from an initial state $s_0 \in S_p$. Let $p_1 > p_2$. We show that $v_{\pi_1}(s_0) > v_{\pi_2}(s_0)$. Note that

$$v_{\pi_i}(s_0) = E[G_t|s_t = s_0, \xi_t = A_{\varphi}]P_{\pi_i}(s_0) = A_{\varphi} + E[G_t|s_t = s_0, \xi_t \not= A_{\varphi}]P_{\pi_i}(s_0) \not= A_{\varphi},$$

where $\xi_t$ denotes a run of the product MDP under $\pi_i$ starting from $s_t$.

By carefully estimating the accumulated reward, we can get an upper bound for $V_{\pi_2}(s_0)$ and an lower bound for $V_{\pi_1}(s_0)$ as follows:

$$v_{\pi_2}(s_0) \leq p_2 r_g \frac{1}{1 - \gamma} + (1 - p_2) r_g C,$$

$$v_{\pi_1}(s_0) \geq p_1 \left( r_g \frac{\gamma^k}{1 - \gamma^k} - \frac{M}{1 - \gamma} \right) - (1 - p_1) \frac{M}{1 - \gamma},$$

where $M = \max(|r|d_{\max}, |r|d)$, $d_{\max}$ is the maximum value that can be taken by $d(\cdot, \cdot)$ in (4), and $C > 0$, $k > 0$ are constants (depending on the product MDP). Since $p_1 > p_2$, there exists $\gamma^* \in (0, 1)$ and a choice of a sufficiently large $r_g$ (depending on $\gamma^*$ and other constants) such that $v_{\pi_1}(s_0) > v_{\pi_2}(s_0)$ for all $\gamma \in (\gamma^*, 1)$. Since the MDP is finite, we can make such choices independent of $s_0$. Hence, a policy that maximizes $v_{\pi}(s)$ for all $s$ also maximizes the satisfaction probability for all $s$.

**V. SIMULATION RESULTS**

In this section, we test the proposed method with different LTL specifications using a car-like robot with the following dynamics [13]:

$$\dot{x} = \frac{v \cos (\gamma + \vartheta)}{\cos \gamma},$$
$$\dot{y} = \frac{v \sin (\gamma + \vartheta)}{\cos \gamma},$$
$$\dot{\vartheta} = v \tan \varphi,$$

where $(x, y)$ is the planar position of center of the vehicle, $\vartheta$ is its orientation, the control variables $v$ and $\varphi$ are the velocity and steering angle, respectively, and $\gamma = \arctan (\tan (\varphi))/2$. The state space is $X' = [-5, 5] \times [-5, 5]$ and the control space is $U = [-1, 1] \times [-1, 1]$.

**A. Example 1**

In the first example, we test our algorithm with a simple LTL specification

$$\varphi_1 = \Diamond (a \land \Diamond b),$$

where $a = [-3.5, -2] \times [-3.5, -2]$ and $b = [2.3, 3.5] \times [2.3, 3.5]$ are two regions in working space. This LTL formula specifies that the robot must reach $a$ first and then reach $b$. We compare our algorithm that samples a random $q$ with the standard idea that resets $q$ at the beginning of each episode. The neural networks are trained for 1 million steps with 200 steps in each episode. The simulation step is $\Delta t = 0.1s$. We use $r_g = 50$ and $r_n = -0.1$ for the reward function as in Eq. (4). The simulation result of example 1 is presented in Fig. 1. The areas marked as blue are the regions $a$ and $b$. We show the trajectories from an initial point at $(0, -2.5, 0)$ in Fig. 1(a). The black curve is the trajectory generated using the idea of fixing $q_0$ at the beginning of each episode and the red one is the trajectory from our method. It is shown that for this simple LTL specification, both ideas provide a successful trajectory. Fig. 1(b) shows the normalized reward during training for both ideas. The blue one is the normalized reward for the standard idea and the red curve is for our method. Our method collects a normalized reward of $-0.1$ for 500k steps of training and 0.2 for 1M steps of training while the standard method obtains a normalized reward of $-0.3$ and 0.1 for 500k steps and 1M steps, respectively.

**B. Example 2**

In the second example, we test our algorithm following LTL specification:

$$\varphi_2 = \Diamond (a \land \Diamond (b \land \Diamond (c \land \Diamond d))),$$

where $a = [-3, -1.5] \times [-3, -1.5]$, $b = [-3, -1.5] \times [1.5, 3]$, $c = [2, 3.5] \times [1.5, 3]$ and $d = [2, 3.5] \times [-3, -1.5]$ are four areas in the working space. In other words, we want the robot to visit $a$, $b$, $c$, $d$ sequentially. The LDBA obtained from the LTL specification is shown in Fig. 2.

We train the neural networks for 1 million steps. The system is also simulated using a time step $\Delta t = 0.1s$. The reward function is the same as in the first example, where
then it has to reach either a or b first. If it reaches a first, then it must next reach d without any other restrictions. If it reaches b first, then it has to reach d without entering c. We consider this specification with two different layouts of the regions a, b, c and d.

1) Case 1: In case 1, \(a = [-4, -3] \times [-3, -2], b = [-4, -3] \times [1, 2] \text{ and } d = [3, 4.5] \times [1.5, 3]\) are three goals marked as blue and \(c = [-1, 1.5] \times [-1, 3.5]\) is a restricted area in the working space marked as yellow as in Fig. 5(a) and Fig. 5(b). The LDBA obtained from the LTL specification is shown in Fig. 4.

The reward function is of \(r_n = -0.1\) and \(r_g = 100\). Since we have a constraint of not entering region c if it reaches b, we assign \(r_d = -10\) if this happens. We also train the networks for 1 million steps with 200 steps in each episode. The simulation time step is \(\Delta t = 0.1s\). The trajectories generated from two initial states \((-2, 1.5, 0)\) and \((-2, -4, 0)\) are shown in Fig. 5(a) and Fig. 5(b). For initial point at \((-2, 1.5, 0)\), it is closer to region b so that the trajectory reaches b first. According to the LTL specification, it has to avoid c before reaching d if it reaches b first. For initial point at \((-2, -4, 0)\), the trajectory first reaches region a and then, the trajectory can reach d without avoiding c. The simulation results show that our method can successfully generate a policy that satisfies the LTL specification for different initial points. The algorithm learns that the trajectory should choose the target that is closer to the initial point between a and b and then reach d according to the specification.

2) Case 2: In Case 2, a, b, d are the same regions as in Case 1. Region c is the area of \(C = [-4.5, -2.5] \times [0, 3]\). As in Fig. 5(c) a, b, d are marked as blue and c is the yellow area plus the area of region b. As is shown in the figure, b is enclosed by c, which means that if a trajectory enters b, it will also be in c. This implies that the automaton will be trapped in the deadlock between \(q = 0\) and \(q = 3\) and will never reach accepting condition \(q = 2\) as shown in Fig. 4. The learning algorithm is able to figure out that even if the initial point is closer to b, it still need to reach a first. The result is shown in Fig. 5(c).

\[
\varphi_3 = \Diamond (a \land \Diamond d) \lor (b \land \neg c \land U d).
\] (13)

In plain words, the specification encodes that the robot must reach either a or b first. If it reaches a first, then it must next reach d without any other restrictions. If it reaches b first, then it has to reach d without entering c. We consider this...
VI. CONCLUSIONS

In this paper, we proposed an actor-critic reinforcement learning method for motion planning problems with LTL specifications. We consider continuous state and action spaces for the robot model. The LTL specification is converted into an annotated LDBA and the deep deterministic policy gradient method is used to train the resulting product MDP. The annotated LDBA is used to continuously shape the reward so that dense reward is available for training. At the beginning of each episode of training, we sample not only a state from the state space of the robot, but also a state from the annotated LDBA. This idea mitigates the sampling complexity of generating a trajectory to satisfy the full LTL specification. We use a car-like robot to test our algorithm with three LTL specifications from different working configurations and initial positions in our simulation. Simulation with three LTL specifications from different working configurations and initial positions in our simulation. Simulation results show that our method achieves successful trajectories for each of the specifications. For future work, we found out in our simulation that the algorithm sometimes fails to deal with complex working configurations such as non-convex obstacles. Complex configurations make it easier for the algorithm to stick into a local optimal point. We will focus on doing research about improving the algorithm to deal with more complex scenarios and LTL specifications.

REFERENCES

[1] Christel Baier and Joost-Pieter Katoen. Principles of Model Checking. MIT press, 2008.
[2] Amit Bhatia, Lydia E Kavraki, and Moshe Y Vardi. Sampling-based motion planning with temporal goals. In Proc. of ICRA, pages 2689–2696. IEEE, 2010.
[3] Costas Courcoubetis and Mihalis Yannakakis. The complexity of probabilistic verification. J. ACM, 42(4):857–907, jul 1995.
[4] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In Proc. of ICML, pages 1329–1338, 2016.
[5] Georgios E Fainekos, Antoine Girard, Hadas Kress-Gazit, and George J Pappas. Temporal logic motion planning for dynamic robots. Automatica, 45(2):343–352, 2009.
[6] Qitong Gao, Davood Hajernazad, Yan Zhang, Yiannis Kantaros, and Michael M Zavlanos. Reduced variance deep reinforcement learning with temporal logic specifications. In Proc. of ICCPS, pages 237–248. ACM, 2019.
[7] Ernst Moritz Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. Omega-regular objectives in model-free reinforcement learning. In Proc. of TACAS, pages 395–412. Springer, 2019.
[8] Mohammadhasnein Hasanbeig, Alessandro Abate, and Daniel Kroening. Certified reinforcement learning with logic guidance. arXiv preprint arXiv:1902.00778, 2019.
[9] Mohammadhasnein Hasanbeig, Yiannis Kantaros, Alessandro Abate, Daniel Kroening, George J Pappas, and Insup Lee. Reinforcement learning for temporal logic control synthesis with probabilistic satisfaction guarantees. arXiv preprint arXiv:1909.05304, 2019.
[10] Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. The International Journal of Robotics Research, 30(7):846–894, 2011.
[11] Jan Kretínský, Tobias Meggendorfer, and Salomon Sickert. Owl: A library for ω-words, automata, and LTL. In Shuvendu K Lahiri and Chao Wang, editors, Proc. of ATVA, volume 11138 of Lecture Notes in Computer Science, pages 543–550. Springer, 2018.
[12] Xiao Li, Yao Ma, and Calin Belta. A policy search method for temporal logic specified reinforcement learning tasks. In Proc. of ACC, pages 240–245. IEEE, 2018.
[13] Yinan Li and Jun Liu. Robustly complete synthesis of memoryless controllers for nonlinear systems with reach-and-stay specifications. arXiv preprint arXiv:1802.09082, 2018.
[14] Timothy P Lillicrap, Jonathan H Hunt, Alexander Pritzel, et al. Continuous control with deep reinforcement learning. In Proc. of ICML, pages 2499–2508. J. ACM, 2015.
[15] Savvas G Loizou and Kostas J Kyriakopoulos. Automatic synthesis of multi-agent motion tasks based on LTL specifications. In Proc. of CDC, volume 1, pages 153–158. IEEE, 2004.
[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.
[17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.
[18] Ryohei Omi, Ami Sakakibara, and Toshimitsu Ushio. Reinforcement learning of control policy for linear temporal logic specifications using limi-t-deterministic büchi automata. arXiv preprint arXiv:2001.04669, 2020.
[19] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Natural actor-critic. In European Conference on Machine Learning, pages 280–291. Springer, 2005.
[20] Dorsa Sadigh, Eric S Kim, Samuel Coogan, S Shankar Sastry, and Sanjit A Seshia. A learning based approach to control synthesis of markov decision processes for linear temporal logic specifications. In Proc. of CDC, pages 1091–1096. IEEE, 2014.
[21] Salomon Sickert, Javier Esparza, Stefan Jaax, and Jan Kretínsky. Limit-Deterministic Büchi Automata for Linear Temporal Logic. In Proc. of FAVS, pages 312–332. 2016.
[22] David Silver, Aja Huang, Chris J Maddison, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484, 2016.
[23] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In ICML, 2014.
[24] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. Nature, 550(7676):354, 2017.
[25] Richard S Sutton and Andrew G Barto. *Introduction to Reinforcement Learning*, volume 135. MIT press Cambridge, 1998.

[26] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proc. of NeurIPS*, pages 1057–1063, 2000.

[27] Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King’s College, Cambridge, 1989.

[28] Tichakorn Wongpiromsarn, Ufuk Topcu, and Richard M Murray. Receding horizon temporal logic planning for dynamical systems. In *Proc. of CDC*, pages 5997–6004. IEEE, 2009.