Reinforcement Learning based Controller synthesis for Linear Temporal Logic Specifications Using Limit-Deterministic Generalized Büchi Automata

一般化 **Linit-Deterministic Büchi** オートマトンを用いた線形時相論理制約に対する強化学習に基づく制御器設計

学籍番号：09C18707　潮 研究室　大浦 稜平

1　緒論

近年，LTL 式に対する ω-オートマトンとして（一般化）limit-deterministic Büchi オートマトン（LD(G)BA) が注目されており，LDBA または LDGBA を用いた MDP に対する制御器の強化学習法が提案されている [1, 2]．本研究では，報酬関数のスパース性を緩和する拡張 LDGBA を提案し，割引率を十分 1 に近づけることで LTL 式を満たす最適方策が学習できることを示す．

2　拡張 LDGBA と制御器の強化学習

制御対象はラベル付き MDP でモデル化され，$M = (S, A, P, s_{init}, AP, L)$ で表現される．S は状態の有限集合，A は行動の有限集合，$P : S \times S \times A \to [0,1]$ は状態の遷移確率，$s_{init} \in S$ はシステムの初期状態，AP は原子命題の有限集合，$L : S \times A \times S \to 2^{AP}$ は各遷移に原子命題を割り当てるラベル関数である．状態 s で行動 a を起こした下で，状態 s' に遷移する確率が $P(s'|s,a)$ であり，P はマルコフ性を有するものとする．

LTL 式 φ に対する遷移ベースの LDGBA（tLDGBA）は $B_\varphi = (X,\ x_{init},\ \Sigma,\ \delta,\ \mathcal{F})$ で表現される．X は状態の有限集合，x_{init} は初期状態，Σ は文字の有限集合，δ は状態遷移の集合，$\mathcal{F} = \{F_j\}_{j=1}^n$ は受理条件であり，各 $j \in \{1,\dots,n\}$ に対して $F_j \subset \delta$ である．状態集合 X は 2 つの部分集合 $X_{initial}$,X_{final} に分割でき，$X_{initial}$ 内と X_{final} 内の遷移は決定的である．$X_{initial}$ から X_{final} への遷移は非決定的で，それらは ε-遷移で表現される．また，全ての F_i の要素は X_{final} 内の遷移である．X_{final} から $X_{initial}$ への遷移は存在しない．

V を n 次元 2 値ベクトルの集合とする．B_φ を拡張するため，以下に 3 つの関数 $visitf : \delta \to V$, $reset : V \to V$, $Max : V \times V \to V$ を導入する．任意の $e \in \delta$ に対して，$visitf(e) = (v_1, \dots, v_n)^T$，ここで $e \in F_i$ ならば $v_i = 1$，それ以外ならば $v_i = 0$ である．任意の $v \in V$ に対して，$v = \mathbf{1}$ ならば $reset(v) = \mathbf{0}$，それ以外ならば $reset(v) = v$ である．任意の $v, u \in V$ に対して，$Max(v,u) = (l_1, \dots, l_n)^T$, ここで各 $i \in \{1,\dots,n\}$ に対して $l_i = \max\{v_i, u_i\}$.

$B_\varphi = (X, x_{init}, \Sigma, \delta, \mathcal{F})$ に対する拡張オートマトンを次の tLDGBA $\bar{B}_\varphi = (\bar{X}, \bar{x}_{init}, \bar{\Sigma}, \bar{\delta}, \bar{\mathcal{F}})$ で定義する..

- $\bar{X} = X \times V$：状態の有限集合.
- $\bar{x}_{init} = (x_{init}, \mathbf{0})$：初期状態.
- $\bar{\Sigma} = \Sigma$：文字の有限集合.
- $\bar{\delta} = \{((x,v), \bar{\sigma}, (x', v')) \in \bar{X} \times \bar{\Sigma} \times \bar{X}\ ;\ (x, \bar{\sigma}, x') \in \delta,\ v' = reset(Max(v, visitf((x, \bar{\sigma}, x'))))\}$: 状態遷移の集合.
- $\bar{\mathcal{F}} = \{\bar{F}_1, \dots, \bar{F}_n\}$：受理条件．各 $j \in \{1,\dots,n\}$ に対して $\bar{F}_j = \{((x,v), \bar{\sigma}, (x', v')) \in \bar{\delta}\ ;\ (x, \bar{\sigma}, x') \in F_j,\ v_j = 0,\ visitf((x, \bar{\sigma}, x'))_j = 1\}$ と定義される．ここで，$visitf((x, \bar{\sigma}, x'))_j$ は $visitf((x, \bar{\sigma}, x'))$ の j 番目の要素を表す.

MDP M と拡張 tLDGBA $\bar{B}_\varphi$ による合成 MDP $M^\otimes = M \otimes \bar{B}_\varphi$ を次の $(S^\otimes, A^\otimes, s^\otimes_{init}, P^\otimes, \delta^\otimes, \mathcal{F}^\otimes)$ で定義する．

- $S^\otimes = S \times \bar{X}$：状態の有限集合.
- $A^\otimes = A \cup \{\varepsilon_{x'}; \exists x' \in X \text{ s.t. } (x, \varepsilon_{x'}, x') \in \delta\}$：行動の有限集合．ここで，$\varepsilon_{x'}$ は ε-遷移 $(x, \varepsilon_{x'}, x') \in \delta$ に対する行動である．
- $s^\otimes_{init} = (s_init, \bar{x}_{init})$：合成 MDP の初期状態.
- $P^\otimes : S^\otimes \times S^\otimes \times A^\otimes \to [0,1]$：状態の遷移確率．以下のように定義される．

$$P^\otimes(s^{\otimes\prime}|s^\otimes, a) = \begin{cases} P(s'|s,a) & \text{if } (\bar{x}, L((s,a,s')), \bar{x}') \in \bar{\delta}, a \in \mathcal{A}(s), \\ 1 & \text{if } s{=}s', v{=}v', (x, \varepsilon_{x'}, x') {\in} \bar{\delta}, a = \varepsilon_{x'}, \\ 0 & \text{otherwise,} \end{cases}$$

- $\delta^\otimes = \{(s^\otimes, a, s^{\otimes\prime}) \in S^\otimes \times A^\otimes \times S^\otimes; P^\otimes(s^{\otimes\prime}|s^\otimes, a) > 0\}$：状態遷移の集合.
- $\mathcal{F}^\otimes = \{\bar{F}_i^\otimes\}_{i=1}^n$：受理条件．各 $j \in \{1,\dots,n\}$ に対して，$\bar{F}_i^\otimes = \{((s,\bar{x}), a, (s', \bar{x}')) \in \delta^\otimes\ ;\ (\bar{x}, L(s,a,s'), \bar{x}') \in \bar{F}_i\}$ と定義される．

報酬関数 $\mathcal{R} : S^\otimes \times A^\otimes \times S^\otimes \to \mathbb{R}$ を次のように定義する．

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

ここで r_p は正の実数である．

3　定理

定理 1. *ある MDP M と与えられた LTL 式 φ に対応する拡張 tLDGBA $\bar{B}_\varphi$ の合成 MDP $M^\otimes$ 及び式 (1) で与えられる報酬関数に対して，$M^\otimes$ 上に φ を確率非 0 で満たす決定論的定常方策が存在すれば，ある割引率 γ^* が存在し，$\gamma > \gamma^*$ の下で状態価値関数を最大化するアルゴリズムはそのような方策の一つを見つける．*

4　結論

本研究では報酬関数のスパース性を緩和するオートマトンを提案し，それを用いた学習により LTL 式を確率非 0 で満たす方策が得られることを示した．LTL の充足確率を最大化するアルゴリズムを示すことが主要な課題の一つである．

参考文献

[1] M. Hasanbeig, A. Abate, and D. Kroening, "Logically-constrained reinforcement learning," *arXiv:1801.08099v8*, Feb. 2019.

[2] 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.