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

1. $\delta^{\otimes}_{\pi,i} \cap \bar{F}^{\otimes}_{j} \neq \emptyset$, $\forall j \in \{1, \dots, n\}$,
2. $\delta^{\otimes}_{\pi,i} \cap \bar{F}^{\otimes}_{j} = \emptyset$, $\forall j \in \{1, \dots, n\}$.

Let $\mathcal{SV}^*$ be the set of optimal supervisors. Let $D^{\otimes}_{SV^*}$ denote the product DES $D^{\otimes}$ controlled by the optimal supervisor SV^*.

For a Markov chain $MC^{\otimes}_{SV}$ induced by a product MDP $D^{\otimes}$ with a supervisor SV, let $S^{\otimes}_{SV} = T^{\otimes}_{SV} \sqcup R^{\otimes 1}_{SV} \sqcup \ldots \sqcup R^{\otimes h}_{SV}$ be the set of states in $MC^{\otimes}_{SV}$, where $T^{\otimes}_{SV}$ is the set of transient states and $R^{\otimes i}_{SV}$ is the recurrent class for each $i \in \{1, \dots, h\}$, and let $R(MC^{\otimes}_{SV})$ be the union of all recurrent classes in $MC^{\otimes}_{SV}$. Let $\delta^{\otimes}_{SV,i}$ be the set of transtions in a recurrent class $R^{\otimes i}_{SV}$, namely $\delta^{\otimes}_{SV,i} = \{(s^{\otimes}, e, s^{\otimes\prime}) \in \delta^{\otimes}; s^{\otimes} \in R^{\otimes i}_{SV},\ P^{\otimes}_{T}(s^{\otimes\prime}|s^{\otimes}, e) > 0, P^{\otimes}_{E}(e|s^{\otimes}, SV(s^{\otimes})) > 0\}$, and let $P^{\otimes}_{SV} : S^{\otimes}_{SV} \times S^{\otimes}_{SV} \to [0,1]$ such that $P^{\otimes}_{SV} = \sum_{e \in SV(s^{\otimes})} P^{\otimes}_{T}(s^{\otimes\prime}|s^{\otimes}, e) P^{\otimes}_{E}(e|s^{\otimes}, SV(s^{\otimes}))$ be the transition probability under SV.

Theorem 1 *Let $D^{\otimes}$ be the product DES corresponding to a DES D and an LTL formuula φ. If there exists a supervisor SV satisfying φ, then there exist a discount factor γ^* and a positive reward r_p such that any algorithm that maximizes the expected discounted reward with $\gamma > \gamma^*$ and $r_p > ||\mathcal{R}_1||_{\infty}$ will find a supervisor satisfying φ.*

Proof 1 *Suppose that SV^* be an optimal supervisor but does not satisfy the LTL formula φ. Then, for any recurrent class $R^{\otimes i}_{SV^*}$ in the Markov chain $MC^{\otimes}_{SV^*}$ and any accepting set $\bar{F}^{\otimes}_{j}$ of the product DES $D^{\otimes}$, $\delta^{\otimes}_{SV^*,i} \cap \bar{F}^{\otimes}_{j} = \emptyset$ holds by Lemma 1. Thus, the agent under the policy π^* 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 $Post(T^{\otimes}_{\pi^*})$ be the set of states in recurrent classes that can be transitioned from states in $T^{\otimes}_{\pi^*}$ by one event occurrence. For the initial state $s^{\otimes}_{init}$ in the set of transient states, it holds that*

$$\begin{aligned} V^{SV^*}(s^{\otimes}_{init}) = & \sum_{k=0}^{\infty} \sum_{s^{\otimes} \in T^{\otimes}_{\pi^*}} \gamma^k p^k(s^{\otimes}_{init}, s^{\otimes}) \\ & \sum_{s^{\otimes\prime} \in T^{\otimes}_{\pi^*} \cup Post(T^{\otimes}_{\pi^*})} \sum_{e \in SV(s^{\otimes})} P^{\otimes}_{T}(s^{\otimes\prime}|s^{\otimes}, e) P^{\otimes}_{E}(e|s^{\otimes}, SV(s^{\otimes})) \mathcal{R}(s^{\otimes}, SV(s^{\otimes}), e, s^{\otimes\prime}) \\ \leq & \ r_p \sum_{k=0}^{\infty} \sum_{s^{\otimes} \in T^{\otimes}_{\pi^*}} \gamma^k p^k(s^{\otimes}_{init}, s^{\otimes}). \end{aligned}$$

By the property of the transient states, for any state $s^{\otimes}$ in $T^{\otimes}_{\pi^*}$, there exists a bounded positive value m such that $\sum_{k=0}^{\infty} \gamma^k p^k(s^{\otimes}_{init}, s^{\otimes}) \leq \sum_{k=0}^{\infty} p^k(s^{\otimes}_{init}, s^{\otimes}) <$

m [1]. Therefore, there exists a bounded positive value $\bar{m}$ such that $V^{\pi^*}(s^{\otimes}_{init}) < \bar{m}$. Let $\bar{S}V$ be a supervisor satisfying φ. We consider the following two cases.

1. *Assume that the initial state $s^{\otimes}_{init}$ is in a recurrent class $R^{\otimes i}_{\bar{\pi}}$ for some $i \in \{1, \ldots, h\}$. For any accepting set $\bar{F}^{\otimes}_j$, $\delta^{\otimes}_{\bar{\pi},i} \cap \bar{F}^{\otimes}_j \neq \emptyset$ holds by the definition of $\bar{\pi}$. The expected discounted reward for $s^{\otimes}_{init}$ is given by*

$$
\begin{aligned}
V^{\bar{\pi}}(s^{\otimes}_{init}) = &\sum_{k=0}^{\infty} \sum_{s^{\otimes} \in R^{\otimes i}_{\bar{S}V}} \gamma^k p^k(s^{\otimes}_{init}, s^{\otimes}) \\
&\sum_{s^{\otimes\prime} \in R^{\otimes i}_{\bar{S}V}} \sum_{e \in SV(s^{\otimes})} P^{\otimes}_T(s^{\otimes\prime} | s^{\otimes}, e) P^{\otimes}_E(e | s^{\otimes}, SV(s^{\otimes})) \mathcal{R}(s^{\otimes}, \bar{S}V(s^{\otimes}), e, s^{\otimes\prime}).
\end{aligned}
$$

Since $s^{\otimes}_{init}$ is in $R^{\otimes i}_{\bar{\pi}}$, there exists a set of positive numbers $K = \{k \ ; \ k \geq n, p^k(s^{\otimes}_{init}, s^{\otimes}_{init}) > 0\}$ [1]. We consider the worst scenario in this case. For the stopping time of first returning the initial state, it holds that

$$
\begin{aligned}
V^{\bar{\pi}}(s^{\otimes}_{init}) >& \mathbb{E}[\gamma^k r_p - (1 + \ldots + \gamma^k) ||\mathcal{R}_1||_{\infty} + \gamma^k V^{\bar{\pi}}(s^{\otimes}_{init})] \\
\geq& \gamma^{\mathbb{E}[k]} r_p - (1 + \ldots + \gamma^{\mathbb{E}[k]}) ||\mathcal{R}_1||_{\infty} + \gamma^{\mathbb{E}[k]} V^{\bar{\pi}}(s^{\otimes}_{init}) \\
=& \frac{\gamma^{\mathbb{E}[k]} r_p - (1 + \ldots + \gamma^{\mathbb{E}[k]}) ||\mathcal{R}_1||_{\infty}}{1 - \gamma^{\mathbb{E}[k]}},
\end{aligned}
$$

References

[1] R. Durrett, *Essentials of Stochastic Processes*, 2nd Edition. ser. Springer texts in statistics. New York; London; Springer, 2012.

[2] L. Breuer, "Introduction to Stochastic Processes," [Online]. Available: https://www.kent.ac.uk/smsas/personal/lb209/files/sp07.pdf

[3] S.M. Ross, *Stochastic Processes*, 2nd Edition. University of California, Wiley, 1995.

[4] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesv́ari, "Convergence results for single-step on-policy reinforcement learning algorithms" *Machine Learning*, vol. 38, no. 3, pp, 287–308, 1998.

[5] J. Kretínský, T. Meggendorfer, S. Sickert, "Owl: A library for ω-words, automata, and LTL," in *Proc. 16th International Symposium on Automated Technology for Verification and Analysis*, 2018, pp. 543–550.