

Definition 1. The two reward functions $\mathcal{R}_1 : S^\otimes \times 2^{E^\otimes} \rightarrow \mathbb{R}$ and $\mathcal{R}_2 : S^\otimes \times E^\otimes \times S^\otimes \rightarrow \mathbb{R}$ are defined as follows.

$$\mathcal{R}_1(s^\otimes, \pi) = \begin{cases} r_n |\pi| & \text{if } \llbracket s^\otimes \rrbracket_q \notin \text{SinkSet}, \\ 0 & \text{if } \llbracket s^\otimes \rrbracket_q \in \text{SinkSet}, \end{cases} \quad (1)$$

where $|E|$ means number of elements in the set E and r_n is a positive value.

$$\mathcal{R}_2(s^\otimes, e, s^{\otimes'}) = \begin{cases} r_p & \text{if } \exists i \in \{1, \dots, n\}, (s^\otimes, e, s^{\otimes'}) \in \bar{F}_i^\otimes, \\ r_{\text{sink}} & \text{if } \llbracket s^{\otimes'} \rrbracket_q \in \text{SinkSet}, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where r_p and r_{sink} are the positive and negative value, respectively.

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

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

For a Markov chain MC_{SV}^\otimes induced by a product MDP D^\otimes with a supervisor SV , let $S_{SV}^\otimes = T_{SV}^\otimes \sqcup R_{SV}^{\otimes 1} \sqcup \dots \sqcup R_{SV}^{\otimes h}$ be the set of states in MC_{SV}^\otimes , where T_{SV}^\otimes is the set of transient states and $R_{SV}^{\otimes i}$ is the recurrent class for each $i \in \{1, \dots, h\}$, and let $R(MC_{SV}^\otimes)$ be the union of all recurrent classes in MC_{SV}^\otimes . Let $\delta_{SV, i}^\otimes$ be the set of transitions in a recurrent class $R_{SV}^{\otimes i}$, namely $\delta_{SV, i}^\otimes = \{(s^\otimes, e, s^{\otimes'}) \in \delta^\otimes; s^\otimes \in R_{SV}^{\otimes i}, P_T^\otimes(s^{\otimes'}|s^\otimes, e) > 0, P_E^\otimes(e|s^\otimes, SV(s^\otimes)) > 0\}$, and let $P_{SV}^\otimes : S_{SV}^\otimes \times S_{SV}^\otimes \rightarrow [0, 1]$ such that $P_{SV}^\otimes(s^{\otimes'}|s^\otimes) = \sum_{e \in SV(s^\otimes)} P_T^\otimes(s^{\otimes'}|s^\otimes, e) P_E^\otimes(e|s^\otimes, SV(s^\otimes))$ be the transition probability under SV .

Definition 2. An accepting recurrent class is defined as the recurrent class whose at least one accepting transition in each accepting set \bar{F}_j^\otimes with $j \in \{1, \dots, n\}$.

Theorem 1. Let M^\otimes be the product DES corresponding to a DES M and an LTL formula φ . Let \mathcal{R}_1 be a reward function for control patterns. If there exists a supervisor SV satisfying φ and it satisfies that there is no state $s^\otimes \in S_{SV}^\otimes$ reachable from initial state s_{init}^\otimes such that $\llbracket s^\otimes \rrbracket_q \in \text{SinkSet}$, then there exist a discount factor γ^* , a positive reward $r_p^*(\mathcal{R}_1)$ that is a function of \mathcal{R}_1 and satisfies $r_p^*(\mathcal{R}_1) > \|\mathcal{R}_1\|_\infty$, and a negative reward r_{sink}^* that is a function of r_p and \mathcal{R}_1 and satisfies $r_{\text{sink}}^*(r_p, \mathcal{R}_1) < -(r_p + \|\mathcal{R}_1\|_\infty)$ such that any algorithm that maximizes the expected discounted reward with $\gamma > \gamma^*$, $r_p > r_p^*(\mathcal{R}_1)$, and $r_{\text{sink}} < r_{\text{sink}}^*(r_p, \mathcal{R}_1)$ will find, with probability one, a supervisor satisfying φ and it satisfies that there is no state $s^\otimes \in S_{SV}^\otimes$ reachable from initial state s_{init}^\otimes such that $\llbracket s^\otimes \rrbracket_q \in \text{SinkSet}$.

Proof. Suppose that SV^* be an optimal supervisor but does not satisfy the LTL formula φ or there is a state s_{sink}^\otimes reachable from the initial state such that $\llbracket s_{sink}^\otimes \rrbracket_q \in SinkSet$ under the supervisor SV^* . Then, for any recurrent class $R_{SV^*}^{\otimes i}$ in the Markov chain $MC_{SV^*}^\otimes$ and any accepting set \bar{F}_j^\otimes of the product DES M^\otimes , $\delta_{SV^*,i}^\otimes \cap \bar{F}_j^\otimes = \emptyset$ holds for the first case by Lemma 1 and there is a recurrent class $R_{SV^*}^{\otimes i}$ such that $s_{sink}^\otimes \in R_{SV^*}^{\otimes i}$ for the second case. We consider the two cases separately.

1. Assume that SV^* does not the LTL formula φ . By the assumption, the system under the supervisor SV^* 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 $Post(T_{SV^*}^\otimes)$ be the set of states in recurrent classes that can be transitioned from states in $T_{SV^*}^\otimes$ by one event occurrence. For the initial state s_{init}^\otimes in the set of transient states, it holds that

$$\begin{aligned} V^{SV^*}(s_{init}^\otimes) &= \sum_{k=0}^{\infty} \sum_{s^\otimes \in T_{SV^*}^\otimes} \gamma^k p^k(s_{init}^\otimes, s^\otimes) \\ &\quad \sum_{s^{\otimes'} \in T_{SV^*}^\otimes \cup Post(T_{SV^*}^\otimes)} \sum_{e \in SV(s^\otimes)} P_T^\otimes(s^{\otimes'} | s^\otimes, e) P_E^\otimes(e | s^\otimes, SV(s^\otimes)) \mathcal{R}(s^\otimes, SV(s^\otimes), e, s^{\otimes'}) \\ &\leq r_p \sum_{k=0}^{\infty} \sum_{s^\otimes \in T_{SV^*}^\otimes} \gamma^k p^k(s_{init}^\otimes, s^\otimes) + \sum_{k=0}^{\infty} \gamma^k \|\mathcal{R}_1\|_\infty. \end{aligned}$$

By the property of the transient states, for any state s^\otimes in $T_{SV^*}^\otimes$, there exists a bounded positive value m such that $\sum_{k=0}^{\infty} \gamma^k p^k(s_{init}^\otimes, s^\otimes) \leq \sum_{k=0}^{\infty} p^k(s_{init}^\otimes, s^\otimes) < m$ [1]. Therefore, there exists a bounded positive value \bar{m} such that $V^{SV^*}(s_{init}^\otimes) < \bar{m} + \frac{1}{1-\gamma} \|\mathcal{R}_1\|_\infty$.

2. Assume that there is a state s_{sink}^\otimes reachable from the initial state such that $\llbracket s_{sink}^\otimes \rrbracket_q \in SinkSet$ under SV^* . By the assumption, there is a recurrent class $R_{SV^*}^{\otimes i}$ reachable from the initial state such that $s_{sink}^\otimes \in R_{SV^*}^{\otimes i}$. We consider the best scenario in the assumption. We assume that all of the recurrent classes except for $R_{SV^*}^{\otimes i}$ are the accepting recurrent classes and there exist a number $l > 0$, a state s_{sink}^\otimes in $Post(T_{SV^*}^\otimes) \cap R_{SV^*}^{\otimes i}$, and a subset of transient states $\{s_1^\otimes, \dots, s_{l-1}^\otimes\} \subset T_{SV^*}^\otimes$ such that $p(s_{init}^\otimes, s_1^\otimes) > 0$, $p(s_i^\otimes, s_{i+1}^\otimes) > 0$ for $i \in \{1, \dots, l-2\}$, and $p(s_{l-1}^\otimes, s_{sink}^\otimes) > 0$ by the property of transient states. We have

$$V^{SV^*}(s_{init}^\otimes) < Pr_{SV^*}^{M^\otimes}(s_{init}^\otimes \models \varphi) \sum_{k=0}^{\infty} \gamma^k (r_p + \|\mathcal{R}_1\|_\infty) + \gamma^l p^l(s_{init}^\otimes, s_{sink}^\otimes) \sum_{k=0}^{\infty} \gamma^k r_{sink}$$

$$\begin{aligned}
& +Pr_{SV^*}^{M^\otimes}(s_{init}^\otimes \models \varphi)(r_p + \|\mathcal{R}_1\|_\infty) \sum_{k=0}^{\infty} \sum_{s^\otimes \in T_{\pi^*}^\otimes} \gamma^k p^k(s_{init}^\otimes, s^\otimes) \\
& < \frac{1}{1-\gamma} \{Pr_{SV^*}^{M^\otimes}(s_{init}^\otimes \models \varphi)(r_p + \|\mathcal{R}_1\|_\infty) + \gamma^l p^l(s_{init}^\otimes, s_{sink}^\otimes) r_{sink}\} + \bar{m}',
\end{aligned}$$

where \bar{m}' is a constant such that $\bar{m}' > Pr_{SV^*}^{M^\otimes}(s_{init}^\otimes \models \varphi)(r_p + \|\mathcal{R}_1\|_\infty) \sum_{k=0}^{\infty} \sum_{s^\otimes \in T_{\pi^*}^\otimes} \gamma^k p^k(s_{init}^\otimes, s^\otimes)$.

Therefore, if it holds that $r_{sink} \leq -\frac{Pr_{SV^*}^{M^\otimes}(s_{init}^\otimes \models \varphi)}{\gamma^l p^l(s_{init}^\otimes, s_{sink}^\otimes)}(r_p + \|\mathcal{R}_1\|_\infty)$, we then have $V^{SV^*}(s_{init}^\otimes) < \bar{m}'$ for any $\gamma \in [0, 1)$.

Let \bar{SV} be a supervisor satisfying φ and it satisfies that there is no state $s^\otimes \in S_{\bar{SV}}^\otimes$ reachable from initial state s_{init}^\otimes such that $\llbracket s^\otimes \rrbracket_q \in SinkSet$. We consider the following two cases.

1. Assume that the initial state s_{init}^\otimes 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 , $\delta_{\bar{\pi}, i}^\otimes \cap \bar{F}_j^\otimes \neq \emptyset$ holds by the definition of $\bar{\pi}$. The expected discounted reward for s_{init}^\otimes is given by

$$V^{\bar{SV}}(s_{init}^\otimes) = \mathbb{E}^{\bar{SV}} \left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k, \pi_k, e_k, s_{k+1}) \mid s_0 = s_{init}^\otimes \right] \quad (3)$$

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

$$\begin{aligned}
V^{\bar{\pi}}(s_{init}^\otimes) & > \mathbb{E}^{\bar{SV}}[\gamma^{k-1} r_p + \gamma^{k-1} V^{\bar{\pi}}(s_{init}^\otimes) \mid s_0 = s_{init}^\otimes] \\
& \geq \gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]} r_p + \gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]} V^{\bar{\pi}}(s_{init}^\otimes) \\
& = \frac{\gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]} r_p}{1 - \gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]}},
\end{aligned}$$

where second inequality holds since it holds that $\mathbb{E}^{\bar{SV}}[\gamma^k \mid s_0 = s_{init}^\otimes] \geq \gamma^{\mathbb{E}^{\bar{SV}}[k \mid s_0 = s_{init}^\otimes]}$ by Jensen's inequality. If it holds that $\gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]} r_p^* > (1 + \dots + \gamma^{\mathbb{E}^{\bar{SV}}[k \mid s_0 = s_{init}^\otimes]}) \|\mathcal{R}_1\|_\infty$, we then have $\frac{\gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]} r_p}{1 - \gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]}} r_p > \frac{1}{1-\gamma} \|\mathcal{R}_1\|_\infty$. Therefore, for any positive value r_{sink} for the reward function \mathcal{R}_1 , there exist $\gamma^* < 1$ and a positive reward r_p^* that satisfies $\gamma^{\mathbb{E}^{\bar{SV}}[k-1 \mid s_0 = s_{init}^\otimes]} r_p^* > (1 + \dots + \gamma^{\mathbb{E}^{\bar{SV}}[k \mid s_0 = s_{init}^\otimes]}) \|\mathcal{R}_1\|_\infty$ such that $\gamma > \gamma^*$ and $r_p > r_p^*$ imply $V^{\bar{SV}}(s_{init}^\otimes) > V^{SV^*}(s_{init}^\otimes)$.

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

$$\begin{aligned}
& V^{SV}(s_{init}^\otimes) \\
&= \mathbb{E}^{SV} \left[\sum_{m=0}^{\infty} \gamma^m \mathcal{R}(s_m, \bar{SV}(s_m), e_m, s_{m+1}) \mid s_0 = s_{init}^\otimes \right] \\
&\geq \mathbb{E}^{SV} \left[\gamma^l \sum_{m=0}^{\infty} \gamma^m \mathcal{R}(s_{m+l}, \bar{SV}(s_{m+l}), e_{m+l}, s_{m+l+1}) \mid s_0 = s_{init}^\otimes \right] \\
&\geq \gamma^l p^l(s_{init}^\otimes, \hat{s}^\otimes) \mathbb{E}^{SV} [\gamma^{k-1} r_p + \gamma^{k-1} V^{SV}(\hat{s}^\otimes) \mid s_l = \hat{s}^\otimes] \\
&\geq \gamma^l p^l(s_{init}^\otimes, \hat{s}^\otimes) \{ \gamma^{\mathbb{E}^{SV}[k-1 \mid s_l = \hat{s}^\otimes]} r_p + \gamma^{\mathbb{E}^{SV}[k-1 \mid s_l = \hat{s}^\otimes]} V^{SV}(\hat{s}^\otimes) \} \\
&= \gamma^l p^l(s_{init}^\otimes, \hat{s}^\otimes) \frac{\gamma^{\mathbb{E}^{SV}[k-1 \mid s_l = \hat{s}^\otimes]} r_p}{1 - \gamma^{\mathbb{E}^{SV}[k-1 \mid s_l = \hat{s}^\otimes]}},
\end{aligned}$$

where $\bar{l} = \mathbb{E}^{SV}[l \mid p^l(s_{init}^\otimes, \hat{s}^\otimes) > 0]$. If it holds that $\gamma^l p^l(s_{init}^\otimes, \hat{s}^\otimes) \gamma^{\mathbb{E}^{SV}[k-1 \mid s_0 = s_{init}^\otimes]} r_p > (1 + \dots + \gamma^{\mathbb{E}^{SV}[k \mid s_0 = s_{init}^\otimes]}) \|\mathcal{R}_1\|_\infty$, we then have $\gamma^l p^l(s_{init}^\otimes, \hat{s}^\otimes) \frac{\gamma^{\mathbb{E}^{SV}[k-1 \mid s_l = \hat{s}^\otimes]} r_p}{1 - \gamma^{\mathbb{E}^{SV}[k-1 \mid s_l = \hat{s}^\otimes]}} r_p > \frac{1}{1-\gamma} \|\mathcal{R}_1\|_\infty$ for any $\gamma \in [0, 1)$. Therefore, for any positive value r_{sink} for the reward function \mathcal{R}_1 , there exist $\gamma^* < 1$ and a positive reward r_p^* that satisfies $\gamma^l p^l(s_{init}^\otimes, \hat{s}^\otimes) \gamma^{\mathbb{E}^{SV}[k \mid s_l = \hat{s}^\otimes]} r_p^* > (1 + \dots + \gamma^{\mathbb{E}^{SV}[k \mid s_l = \hat{s}^\otimes]}) \|\mathcal{R}_1\|_\infty$ such that $\gamma > \gamma^*$ and $r_p > r_p^*$ imply $V^{SV}(s_{init}^\otimes) > V^{SV^*}(s_{init}^\otimes)$.

The results contradict the optimality assumption of SV^* \square

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. Szepesvári, “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.