Reinforcement Learning with LTL and ω-Regular Objectives via Optimality-Preserving Translation to Average Rewards

Xuan-Bach Le **Dominik Wagner**

Leon Witzman Alexander Rabinovich Luke Ong

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

NEURAL INFORMATION PROCESSING SYSTEMS

Aim of Reinforcement Learning:

Agent learns to accomplish *task* in unknown environment

Aim of Reinforcement Learning:

Agent learns to accomplish *task* in unknown environment

How to specify tasks?

Reward Function

$\mathcal{R}(s_0) \coloneqq 0.2, \mathcal{R}(s_1) \coloneqq 5.3, \ldots$

✗ lack of interpretability

Linear Temporal Logic (LTL)

(more generally ω-regular languages)

✓ interpretable

Reward Function

$\mathcal{R}(s_0) := 0.2, \mathcal{R}(s_1) := 5.3, \ldots$

✗ lack of interpretability

Linear Temporal Logic (LTL)

(more generally ω-regular languages)

Objective: maximise probability to satisfy formula

e.g. visit target state exactly once

✓ interpretable

Reward Function

$\mathcal{R}(s_0) \coloneqq 0.2, \mathcal{R}(s_1) \coloneqq 5.3, \ldots$

✗ lack of interpretability

$\left.\begin{array}{l}\text{✓ rich theory} \\ \text{✓ practical algorithms}\end{array}\right\}$

Linear Temporal Logic (LTL)

(more generally ω-regular languages)

Objective: maximise probability to satisfy formula

e.g. visit target state exactly once

✓ interpretable

✗ much less explored

Reward Function

Linear Temporal Logic (LTL)

(more generally ω-regular languages)

$\longleftarrow$

Translation?

Reward Function **Linear Temporal Logic (LTL)**

(more generally ω-regular languages)

$\longleftarrow$

Translation?

Optimality-Preserving

Reward Function

Linear Temporal Logic (LTL)

(more generally ω-regular languages)

$\longleftarrow$

Translation?

Optimality-Preserving

- *Impossible** if rewards are aggregated by *discounting*

$$\sum_{i \in \mathbb{N}} \gamma^i \cdot R_i$$

---

*without prior knowledge of MDP

Reward Function **Linear Temporal Logic (LTL)**

(more generally ω-regular languages)

$\xleftarrow{\hspace{6em}}$

Translation?

Optimality-Preserving

- *Impossible** if rewards are aggregated by *discounting*

$$\sum_{i \in \mathbb{N}} \gamma^i \cdot R_i$$

- **In this paper:** study *limit-average* rewards instead

$$\liminf_{t \in \mathbb{N}} \mathbb{E}\left[\frac{1}{t}\sum_{i=0}^{t-1} R_i\right]$$

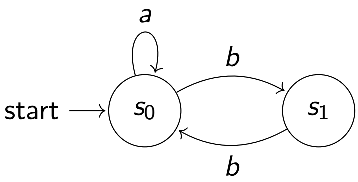
*without prior knowledge of MDP

Negative Result for Reward Functions

Proposition. There is *no optimality-preserving specification translation* from LTL objectives to limit-average rewards given by a *memoryless reward function* $\mathcal{R}$.

Negative Result for Reward Functions

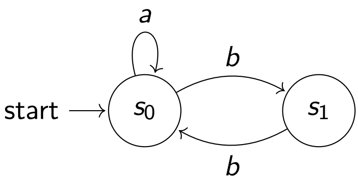
Proposition. There is *no optimality-preserving specification translation* from LTL objectives to limit-average rewards given by a *memoryless reward function* $\mathcal{R}$.



objective: visit s_1 exactly once

Negative Result for Reward Functions

Proposition. There is *no optimality-preserving specification translation* from LTL objectives to limit-average rewards given by a *memoryless reward function* $\mathcal{R}$.



objective: visit s_1 exactly once

Use reward machines for reward assignment

(rewards can depend on internal state)

Main Result

Theorem. There exists an optimality-preserving translation from LTL/ω-regular objectives to *reward machines* with *limit-average* aggregation.

Main Result

Open Problem* ✓

Theorem. There exists an optimality-preserving translation from LTL/ω-regular objectives to *reward machines* with *limit-average* aggregation.

*Alur et al.: A framework for transforming specifications in reinforcement learning, 2022.

Main Result

Open Problem* ✓

Theorem. There exists an optimality-preserving translation from LTL/ω-regular objectives to *reward machines* with *limit-average* aggregation.

Proof Sketch.

- Synchronise MDP with automaton expressing objective to obtain product MDP.

*Alur et al.: A framework for transforming specifications in reinforcement learning, 2022.

Main Result

Open Problem* ✓

Theorem. There exists an optimality-preserving translation from LTL/ω-regular objectives to *reward machines* with *limit-average* aggregation.

Proof Sketch.

- Synchronise MDP with automaton expressing objective to obtain product MDP.
- If transitions with positive probability are known, compute *minimal accepting end components* and give reward 1 to transitions *staying* in them.

*Alur et al.: A framework for transforming specifications in reinforcement learning, 2022.

Main Result

Open Problem* ✓

Theorem. There exists an optimality-preserving translation from LTL/ω-regular objectives to *reward machines* with *limit-average* aggregation.

Proof Sketch.

- Synchronise MDP with automaton expressing objective to obtain product MDP.
- If transitions with positive probability are known, compute *minimal accepting end components* and give reward 1 to transitions *staying* in them.
- Otherwise, keep track of previously taken transitions and give rewards based on the *current knowledge* of transitions with positive probability.

*Alur et al.: A framework for transforming specifications in reinforcement learning, 2022.

Machine

Reward ~~Function~~ (limit-average)

Linear Temporal Logic (LTL)

(more generally ω-regular languages)

$\longleftarrow$

optimality-preserving

Translation ✓

Machine

Reward ~~Function~~ (limit-average) **Linear Temporal Logic (LTL)**

(more generally ω-regular languages)

algorithms: R-learning, RVI Q-learning, Differential Q-learning, ...

$\longleftarrow$

optimality-preserving

Translation ✓

Reward ~~**Function**~~ **Machine** (limit-average)

Linear Temporal Logic (LTL)

(more generally ω-regular languages)

algorithms: R-learning, RVI Q-learning, Differential Q-learning, ...

$\underbrace{\hphantom{algorithms: R-learning, RVI Q-learning, Differential Q-learning}}_{\textit{Provable convergence?}}$

without assumptions on MDP?

$\longleftarrow$

optimality-preserving

Translation ✓

Convergence for Average Rewards and LTL

Theorem. Optimal policies for RL with limit average rewards can be learned in the limit (*without* assumptions on MDP).

Convergence for Average Rewards and LTL

Theorem. Optimal policies for RL with limit average rewards can be learned in the limit (*without* assumptions on MDP).

Algorithm sketch. Solve a sequence of problems with discount factor $\gamma \nearrow 1$ with (black-box) PAC-algorithm for discounted RL.

Convergence for Average Rewards and LTL

Theorem. Optimal policies for RL with limit average rewards can be learned in the limit (*without* assumptions on MDP).

Algorithm sketch. Solve a sequence of problems with discount factor $\gamma \nearrow 1$ with (black-box) PAC-algorithm for discounted RL.

Corollary. Optimal Policies for LTL can be learned in the limit. $\boxed{\text{Open Problem* ✓}}$

*Alur et al.: A framework for transforming specifications in reinforcement learning, 2022.

Summary

Machine
Reward ~~Function~~ (limit-average)

Linear Temporal Logic (LTL)
(more generally ω-regular languages)

$\longleftarrow$

optimality-preserving

Translation ✓

Summary

Machine
Reward ~~Function~~ (limit-average)

Linear Temporal Logic (LTL)
(more generally ω-regular languages)

algorithm with *provable convergence* ✓

$\longleftarrow$

optimality-preserving

Translation ✓