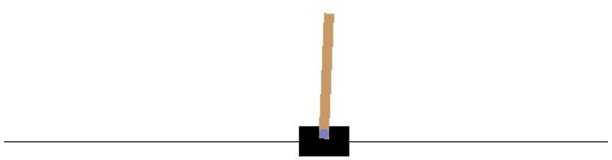
Temporal-Logic-Based Reward Shaping for Continuing Reinforcement Learning Tasks

***Yuqian Jiang*[1]**, Suda Bharadwaj[3], Bo Wu[3], Rishi Shah[1,4], Ufuk Topcu[3], Peter Stone[1,2]

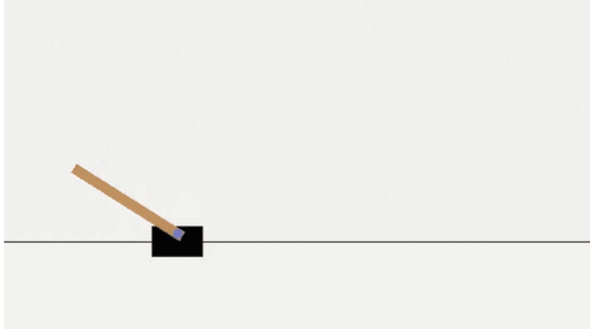
[1]Department of Computer Science, University of Texas at Austin
[2]Sony AI America
[3]Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin
[4]Amazon

Continuing RL tasks

- Continuing vs. episodic tasks
  - No termination, no reset of environment
  - Cart Pole



treated as an episodic task



treated as a continuing task

Average Reward Setting

- Total discounted reward: $\mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k, s_{k+1})\right]$

  The discount factor can lead to undesirable behaviors since the agent sacrifices long-term benefits for short-term gains

- Average reward: $\liminf_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\sum_{k=0}^{n-1} R(s_k, a_k, s_{k+1})\right]$

- Optimal differential Q-function satisfies the Bellman Equation:

$$Q^*_{\mathcal{M}}(s,a) = \mathbb{E}\left[R(s,a,s') + \max_{a'\in\mathcal{A}}(Q^*_{\mathcal{M}}(s',a'))\right] - \rho^*_{\mathcal{M}}$$

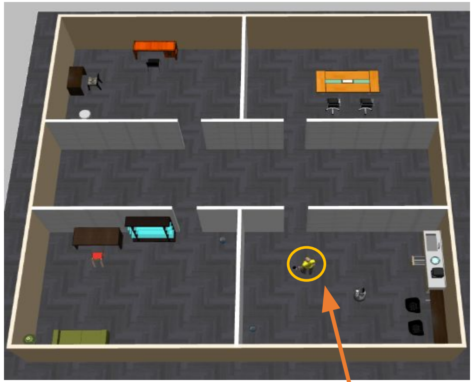
expected average reward of the optimal policy

Reward Shaping

- Reward shaping is a method to inject advice by providing **additional rewards**
- Manually constructing the shaping rewards can be non-trivial

Trash **only** appears in the **kitchen**

Trash appears in **all rooms** but appears in **kitchen with high probability**



Trash appears in the **corridor** and near **humans with high probability**

Encoding Advice

- Reward shaping without manually constructing rewards
  - Learning the shaping functions (Grze´s and Kudenko 2010, Marthi 2007)
  - Temporal logic specifications
    - Safe RL via shielding (Alshiekh et al. 2018)
  - What if the advice is not exactly correct
- Want to convert temporal logic specifications to a reward shaping function that **does not affect the optimal policy.**

Potential-Based Reward Shaping (PBRS)

- Given an MDP with reward function R, F is a **shaping function** such that the optimal policy **does not change** under the augmented reward R' = R + F
- Shaping rewards are expressed as the difference of a **potential function**

$$F = \gamma\Phi(s') - \Phi(s)$$

Discount factor

Potential function

Problem Statement

- How to adapt PBRS for the **average-reward** setting?
- How to construct the potential function Φ given advice as a temporal logic specification?

- Given
  - A Markov decision process $\mathcal{M} = (\mathcal{S}, s_I, \mathcal{A}, R, P)$
  - A linear temporal logic (LTL) formula
    - “Always human visible”

Design potential function Φ and shaping function F that encodes the LTL formula such that, in the augmented MDP $\mathcal{M}' = (\mathcal{S}, s_I, A, R', P)$ with $R' = R + F$, we can recover the optimal policy in $\mathcal{M}$

Reward Shaping for Average-Reward Setting

- Define the shaping function F as:

$$F(s,a,s') = \Phi\left(s', \arg\max_{a'}(Q^*_{\mathcal{M}}(s',a'))\right) - \Phi(s,a)$$

- We do not directly learn the optimal policy of $\mathcal{M}'$, but instead learn another value function $\hat{Q}_{\mathcal{M}'}$ that satisfies a different Bellman equation:

$$\hat{Q}_{\mathcal{M}'}(s,a) = \mathbb{E}[R'(s,a,s') + \hat{Q}_{\mathcal{M}'}(s',a^*)] - \rho^{\pi^*}_{\mathcal{M}'}$$

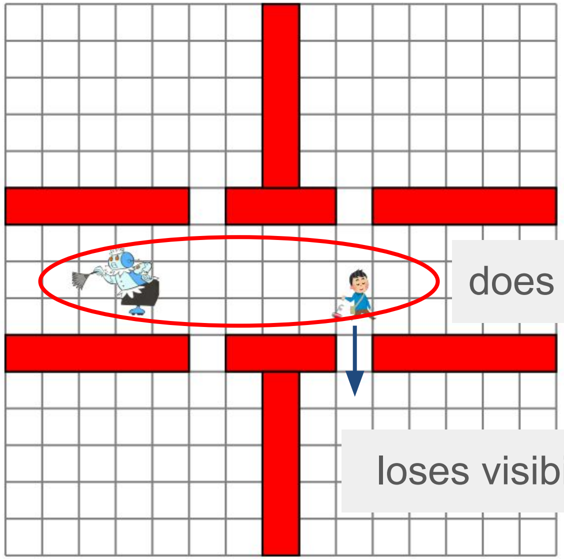
where $a^* = \arg\max_{a'}(\hat{Q}_{\mathcal{M}'}(s',a') + \Phi(s',a'))$

from which we recover the optimal policy $\pi^*(s)$ as:

$$\boxed{\pi^*(s) = \arg\max_{a\in\mathcal{A}}(\hat{Q}_{\mathcal{M}'}(s,a) + \Phi(s,a))}$$

Synthesis of Φ

- **Penalize** the agent for visiting states from which violation of the specification **can occur with a non-zero probability.**



- Must plan ahead
  - Almost-sure winning region: set of state-action pairs from which the probability of violation is 0

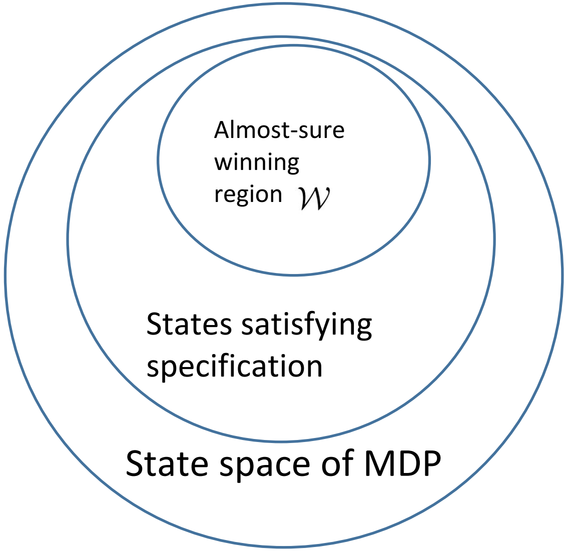
Synthesis of Φ

- Construct the almost-sure winning region $\mathcal{W}$ in the MDP $\mathcal{M}$ using graph-based methods

- After finding $\mathcal{W}$, we construct Φ as

$$\Phi(s,a) = \begin{cases} C & (s,a) \in \mathcal{W} \\ d(s,a) & (s,a) \notin \mathcal{W} \end{cases}$$

where C is an arbitrary constant and $d : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is any function such that $d(s,a) < C$



Experiments

- Compare our method with
  - Baseline deep differential Q-learning (Wan, Naik, and Sutton 2020)
  - **Shielding** – stops all actions that violate the given specification (Alshiekh et al. 2018)

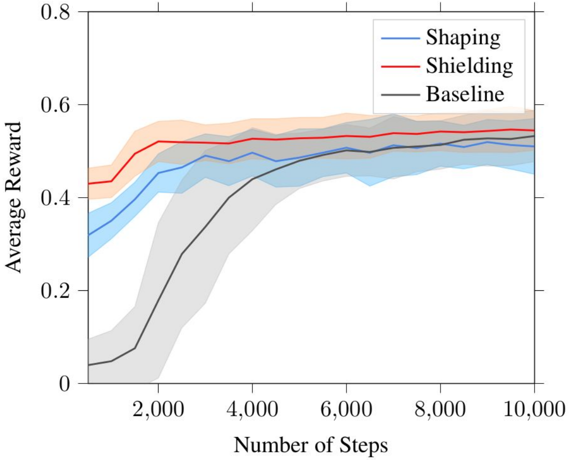
- Test in conditions where the advice is **not exactly correct** or a **conjunction** of specifications is given
  - Our method will still learn the optimal policy

Experiments

Always clean in the kitchen

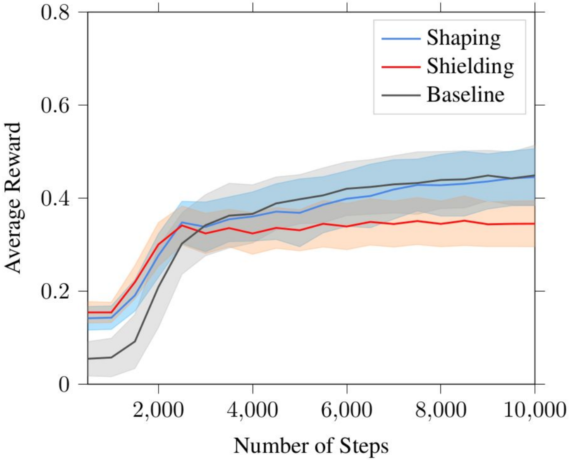
Exactly correct

Trash **only** appears in the **kitchen**



Not exactly correct

Trash appears in **all rooms** but appears in **kitchen with high probability**

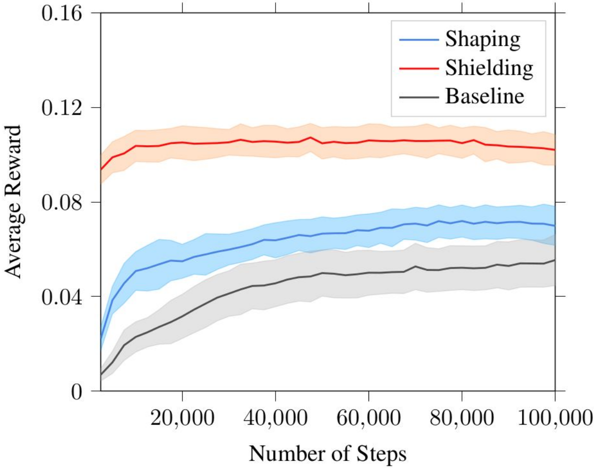


Experiments

Always keep human in view

Exactly correct

Trash **only** appears near humans

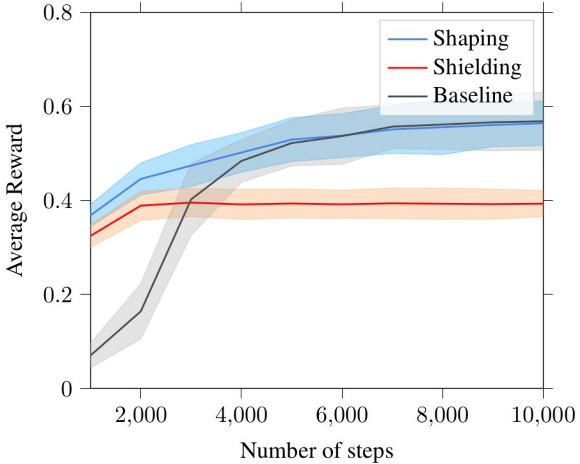


Conjunction of imperfect specifications:

Always keep human in view and always stay in the corridor

Exactly correct

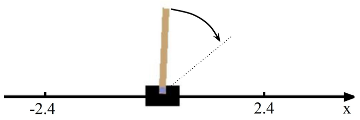
Trash appears near humans and appears in the corridor



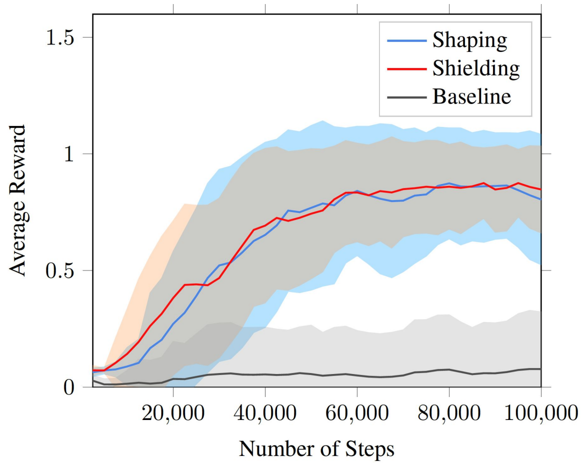
Experiments

Always keep cart in [-2.4, 2.4]

Exactly correct

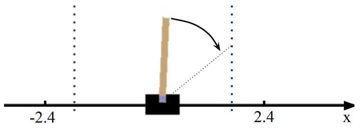


Correct range

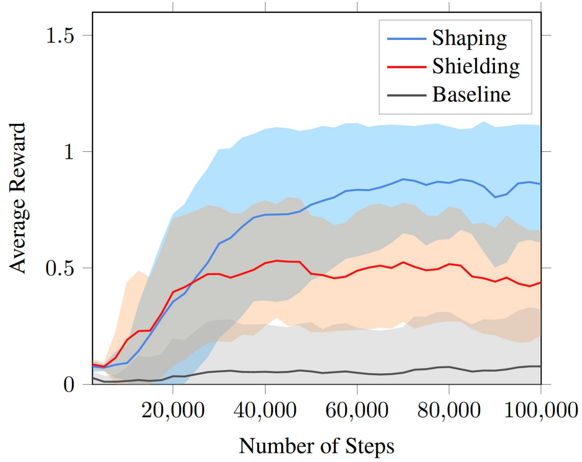


Always keep cart in [-2, 2]

Not Exactly correct



Incorrect range



Summary

- Potential-based **reward shaping** for **average-reward** reinforcement learning
- Construct potential functions from advice given in the form of **temporal logic specifications**
- Robust to **imperfect advice**, **conjunction** of specifications, and approximate dynamics
- Future work:
  - Unknown dynamics or models
  - Adversarial advice

Temporal-Logic-Based Reward Shaping for Continuing Reinforcement Learning Tasks

***Yuqian Jiang*[1]**, Suda Bharadwaj[3], Bo Wu[3], Rishi Shah[1,4], Ufuk Topcu[3], Peter Stone[1,2]

[1]Department of Computer Science, University of Texas at Austin
[2]Sony AI America
[3]Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin
[4]Amazon