1 Problem Formulation and overview

This section discusses the motivation of the paper. We will firstly formalize and analyze safety problems in apprenticeship learning, then we will setting up a new learning goal.

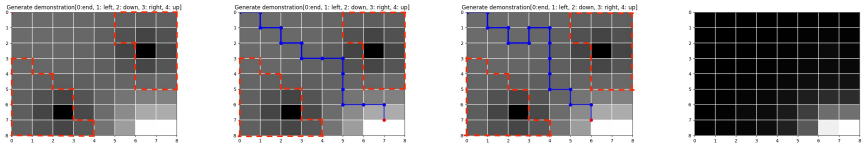
Given an $MDP\backslash R\ M = (S, A, T, \gamma, D)$ and a set of m trajectories $\{\tau_0, \tau_1, ..., \tau_m\}$. Now we define a set of atomic propositions $AP = \{'safe', 'unsafe', ...\}$ and use label function L to label each $s \in S$ 'safe' or 'unsafe'. A specification Φ defines requirements on learnt policies. Especially, safety specification Φ formalized with PCTL can be in such forms:

$$\Phi = [\forall s^{(0)} \in S \wedge D(s^{(0)}) \in (0, 1], s^{(0)} \models \phi] \tag{1}$$

$$or\ \Phi = [\sum_{s \in S} D(s^{(0)}) P_{=?|s^{(0)}}[\psi] \leq p^*] \tag{2}$$

(1) means that all initial states shall satisfy ϕ. (2) means that the probability of initializing from some state and traveling along a trajectory that satisfies ψ is smaller than p^*. Now that a policy π^* has been learnt via apprenticeship learning algorithm, the safety of the learnt policy π is defined as:

Definition 1. *Policy π for an MDP M is safe iff the DTMC D reduced from M and π satisfies Φ.*



(a) (b) (c) (d)

Fig. 1. (a) The 8 x 8 gridworld. Lighter grid cells indicates relative higher reward while darker ones indicates lower reward. It is regarded that the two black grid cells have the lowest rewards, while the two white ones, which constitute the goal area, have the the highest rewards. The grid cells surrounded by red dashed lines are labelled as unsafe. (b) The first trajectory that expert performed during demonstration. (c) The second trajectory that expert performed during demonstration.

In a 8 x 8 grid world navigation example in Fig. 1, agent needs to start from the upper-left corner and move from cell to cell until reaching the lower-right corner within a maximal step length $t < 64$. Meanwhile, there are several grids labelled as 'unsafe' surrounded by red dashed lines near upper-right and lower-left corners, where agent should try to avoid moving into. In each time step, agent can choose from 4 actions to move to the closest grid in each of the

four compass directions, but with 30% chance of moving in a random direction instead. Expert knows the goal area and the unsafe area, but only has an rough judgement on the distribution of the rewards, which can be seen in Fig. 1(a). To formalize the task, each grid cell is mapped to a state in a 8 x 8 state space S. For each state $s \in S$, the feature vector $\phi(s)$ consists of 4 feature functions $\phi_i(s),\ i = 0, 1, 2, 3$. All of them are radial basis functions which respectively depend on the squared Euclidean distances between s and 4 other states which are regarded by the expert as having highest or lowest rewards by expert. For some parameter vector $\omega, ||\omega||_2 \in [-1, 1]$, the reward of state s is $R(s) = \omega^T \phi(s)$.

This experiment considers 2 conditions. In the first one, to simulate abundant expert demonstrations, μ_E is directly generated from the optimal policy with respect to a predetermined ω_E that results in the reward mapping in Fig.1(a). The Apprenticeship Learning algorithm can successfully recover this policy. In the second, which is more realistic, expert only performs successful, absolutely safe, representative and limited number of demonstrations. As shown in Fig. 1(b) and Fig. 1(c), although only 2 trajectories are performed, both of them end up reaching the goal area without reaching any unsafe state.

Supposed that a safety specification Φ is given to enforce p^* to be the upper bound of the probability of reaching states labelled as 'unsafe' within $t = 64$ steps:

$$\mathbf{\Phi} = [\sum_{\mathbf{s} \in \mathbf{S}} \mathbf{D}(\mathbf{s^{(0)}}) \mathbf{P}_{=?|\mathbf{s^{(0)}}} [\mathbf{true\ U^{\leq t}\ 'unsafe'}] \leq \mathbf{p^*}] \tag{3}$$

Model checking result shows that the probability of reaching unsafe state for the policy learnt in the first condition, or expert policy π_E, is 0.11737245848. The reward function that induces the learnt policy in the second condition is shown in Fig. 1(d), where only goal states being white implies that agent only learns to reach goal area. Such lack of information of 'unsafe' states leads to a large probability 0.979873817876 of reaching 'unsafe' states.

Actually specification can be either inferred before the learning process or after a policy has already been learnt, in which regard, even the expert policy π_E would also be unsafe if an extremely strict safety specification, e.g. $p^* = 0.1$, is given. Therefore, our safety problem is not limited to be the condition when expert demonstrations are not enough. So we regard safety specification as one of the inputs and add safety into the goal of learning:

Problem 1. Given an $MDP \backslash R$, a set of m trajectories $\{\tau_0, \tau_1, ..., \tau_m\}$ demonstrated by expert, and a safe specification Φ, learn a policy π that satisfies Φ and is similar to expert policy π_E.