Appendices for Safe Explicable Planning

Paper ID: 86

The document includes three appendices. First, the complete proofs for Lemma. 1, Lemma. 2, Theorem 1, and Theorem 2 are presented in Appendix A. Second, the domain descriptions for the small cliff world (CS), large cliff world (CL), wumpus world (W), and the physical robot experiments are presented in Appendix B. Appendix B also includes the link to the demo of the physical robot experiments. Lastly, the link to the code of SEP is present in Appendix C, which will be made available if accepted.

1 Appendix A: Proofs

Lemma 1. *The set of policies after action pruning based on Eqn.* (4) *is a superset of the set of policies that satisfy the constraint in Eqn.* (2)*, i.e.,* $\widetilde{\Pi} \supseteq \Pi_\delta$.

Proof. To prove this we show that an action pruned by any state per Eqn. (4) is guaranteed to introduce policies that do not satisfy the constraint in Eqn. (2). Consider *any* state $s \in \mathcal{S}$ and *any* pruned action $a' \in \mathcal{A}(s) \backslash \widetilde{\mathcal{A}}(s)$ in that state. Consider choosing a' in s once and thereafter choosing actions as per the optimal policy. The expectation for this is given by

$$Q^{\pi^*}_{\mathcal{M}_R}(s, a') = \mathbb{E}^{\pi^*}_{\mathcal{T}_R}[r_R(s^1) + \gamma_R V^{\pi^*}_{\mathcal{M}_R}(s^1) | s^0 = s, a^0 = a'].$$

From Eqn. (4), w.k.t.

$$Q^{\pi^*}_{\mathcal{M}_R}(s, a') < \delta \mathbb{E}^{\pi^*}_{\mathcal{T}_R}[r_R(s^1) + \gamma_R V^{\pi^*}_{\mathcal{M}_R}(s^1) | s^0 = s, a^0 = \pi^*(s)]$$

Since the future states already choose the optimal actions, there is no room to improve the value of $Q^{\pi^*}_{\mathcal{M}_R}(s, a')$ and hence cannot satisfy δ. Thus, any policy that chooses a' for s cannot satisfy the constraint in that state. □

Lemma 2. *Let* π *and* π' *be two deterministic policies that differ by only a single action in some state i.e.,* $\exists s_i \in \mathcal{S}\ [\pi'(s_i) \neq \pi(s_i)] \wedge \forall s_j \in \mathcal{S} \setminus \{s_i\}\ [\pi'(s_j) = \pi(s_j)]$ *and satisfy* $Q^{\pi}_{\mathcal{M}_R}(s_i, \pi'(s_i)) \leq V^{\pi}_{\mathcal{M}_R}(s_i)$*. Then, policy* π' *is a descendant of* π *in PDT, i.e., policy* π' *is no better than* π*, or more formally,* $\forall s \in \mathcal{S}\ [V^{\pi'}_{\mathcal{M}_R}(s) \leq V^{\pi}_{\mathcal{M}_R}(s)]$.

Proof. This is an extension of the policy improvement theorem (Sutton and Barto 2018) but in the opposite direction (hence referred to as a policy descent). We know that $\forall s \in \mathcal{S}$,

$$Q^{\pi}_{\mathcal{M}_R}(s, \pi(s)) = V^{\pi}_{\mathcal{M}_R}(s)$$
$$\mathbb{E}^{\pi}_{\mathcal{T}_R}[r_R(s^1) + \gamma_R V^{\pi}_{\mathcal{M}_R}(s^1) | s^0 = s, a^0 = \pi(s)] = V^{\pi}_{\mathcal{M}_R}(s)$$

Consider a temporary non-stationary policy π'_1 that chooses the action as per π' once in the initial state s^0 and follows π thereafter. $\forall s \in \mathcal{S}$,

$$Q^{\pi'_1}_{\mathcal{M}_R}(s, \pi'_1(s)) \leq Q^{\pi}_{\mathcal{M}_R}(s, \pi(s))$$
$$Q^{\pi'_1}_{\mathcal{M}_R}(s, \pi'_1(s)) \leq V^{\pi}_{\mathcal{M}_R}(s)$$
$$\mathbb{E}^{\pi'_1}_{\mathcal{T}_R}[r_R(s^1) + \gamma_R V^{\pi'_1}_{\mathcal{M}_R}(s^1) | s^0{=}s, a^0{=}\pi'_1(s)] \leq V^{\pi}_{\mathcal{M}_R}(s)$$

This is because s^0 can either be $s_i \in \mathcal{S}$ or any $s_j \in \mathcal{S}\backslash\{s_i\}$.
If $s^0{=}s_i$ then, $Q^{\pi}_{\mathcal{M}_R}(s_i, \pi'(s_i)) \leq V^{\pi}_{\mathcal{M}_R}(s_i)$ (given).
If $s^0{=}s_j$ then, $Q^{\pi}_{\mathcal{M}_R}(s_j, \pi'(s_j)) = V^{\pi}_{\mathcal{M}_R}(s_j)$ as π'_1 differs from π only in one action in one state s_i and follows π in all future states.
Similarly, consider another temporary non-stationary policy π'_2 that chooses the actions as per π' once in s^0, again in s^1, and follows π thereafter. $\forall s \in \mathcal{S}$,

$$Q^{\pi'_2}_{\mathcal{M}_R}(s, \pi'_2(s)) \leq Q^{\pi'_1}_{\mathcal{M}_R}(s, \pi'_1(s))$$
$$Q^{\pi'_2}_{\mathcal{M}_R}(s, \pi'_2(s)) \leq Q^{\pi}_{\mathcal{M}_R}(s, \pi(s))$$
$$\mathbb{E}^{\pi'_2}_{\mathcal{T}_R}[r_R(s^1) + \gamma_R V^{\pi'_2}_{\mathcal{M}_R}(s^1) | s^0{=}s, a^0{=}\pi'_2(s)] \leq V^{\pi}_{\mathcal{M}_R}(s)$$

Consider repeating this until we always choose actions as per π'.

$$\mathbb{E}^{\pi'}_{\mathcal{T}_R}[r_R(s^1) + \gamma_R r_R(s^2) + \gamma_R^2 r_R(s^3) + ... | s^0{=}s, a^0{=}\pi'(s)] \leq V^{\pi}_{\mathcal{M}_R}(s)$$
$$V^{\pi'}_{\mathcal{M}_R}(s) \leq V^{\pi}_{\mathcal{M}_R}(s)$$

□

Theorem 1. *PDT+ returns all Pareto optimal policies in* $\Pi^*_{\mathcal{E}}$.

Proof. To prove this, we show that there exists a policy descent path from any optimal policy (denoted by π^*) in $\mathcal{M}_R$ (i.e., the root node in PDT) to any Pareto optimal policy (denoted by $\pi^*_{\mathcal{E}}$) in $\Pi^*_{\mathcal{E}}$ by induction. Let n denote the number of actions a policy differs from π^*.

When $n = 1$, i.e. $\pi^*_{\mathcal{E}}$ differs from π^* in a single action i.e. $\exists s_i \in \mathcal{S}\ [\pi^*_{\mathcal{E}}(s_i) \neq \pi^*(s_i)] \wedge \forall s_j \in \mathcal{S}\backslash\{s_i\}\ [\pi^*_{\mathcal{E}}(s_i) = \pi^*(s_i)]$ then, $\forall s \in \mathcal{S}\ Q^{\pi^*}_{\mathcal{M}_R}(s, \pi^*_{\mathcal{E}}(s)) \leq V^{\pi^*}_{\mathcal{M}_R}(s)$ as π^* is the optimal policy. This makes $\pi^*_{\mathcal{E}}$ one of the direct descendants of π^* (by Lem. 2). Hence, $\pi^*_{\mathcal{E}}$ is expanded by PDT.

When $n = k$, i.e. $\pi^*_{\mathcal{E}}$ differs from π^* in any k actions i.e. $\exists \mathcal{S}_k \subseteq \mathcal{S}\ \forall s_i \in \mathcal{S}_k\ [\pi^*_{\mathcal{E}}(s_i){\neq}\pi^*(s_i)] \wedge \forall s_j \in \mathcal{S}\backslash\mathcal{S}_k\ [\pi^*_{\mathcal{E}}(s_j){=}\pi^*(s_j)]$, assume $\pi^*_{\mathcal{E}}$ is expanded by PDT.

When $n = k + 1$, i.e. $\pi^*_{\mathcal{E}}$ differs from π^* in any $k + 1$ actions i.e. $\exists \mathcal{S}_{k+1} \subseteq \mathcal{S}\ \forall s_i \in \mathcal{S}_{k+1}$ $[\pi^*_{\mathcal{E}}(s_i){\neq}\pi^*(s_i)] \wedge \forall s_j \in \mathcal{S}\backslash\mathcal{S}_{k+1}\ [\pi^*_{\mathcal{E}}(s_j){=}\pi^*(s_j)]$, there must be at least one action out of the $k + 1$ actions aligning with π^* that improves, or is same as, the value of $\pi^*_{\mathcal{E}}$.
Assume this is false i.e. all the $k + 1$ actions aligning with π^* worsen $\pi^*_{\mathcal{E}}$. The policy introduced by aligning all $k+1$ actions with π^* is π^* itself (as all actions other than the $k+1$ actions in $\pi^*_{\mathcal{E}}$ are same as π^*). W.k.t. π^* is optimal and cannot be worse than $\pi^*_{\mathcal{E}}$, which is a contradiction. Thus, $\exists \pi\ \exists s_i \in \mathcal{S}_{k+1}\ [\pi^*_{\mathcal{E}}(s_i) \neq \pi(s_i) = \pi^*(s_i)] \wedge \forall s_j \in \mathcal{S}\backslash\{s_i\}\ [\pi(s_j) = \pi^*_{\mathcal{E}}(s_j)]$ that satisfies $Q^{\pi^*_{\mathcal{E}}}_{\mathcal{M}_R}(s_i, \pi(s_i)) \geq V^{\pi^*_{\mathcal{E}}}_{\mathcal{M}_R}(s_i)$ and by Lem. 2, π is no-worse than $\pi^*_{\mathcal{E}}$. W.k.t. π is

expanded (induction assumption). Consequently, $\pi^*_{\mathcal{E}}$ must be one of the direct descendants of π in PDT and hence $\pi^*_{\mathcal{E}}$ is expanded.

Therefore, the result holds for any n by the principle of induction. Finally, the same conclusion holds for PDT+ (by Lem. 1). □

Theorem 2. *PAG+ returns a policy in the Pareto set* $\Pi^*_{\mathcal{E}}$.

Proof. The PAG search process stops when it can no longer improve or find a policy that is equivalent in values to $\pi_{\mathcal{E}}$ under $\mathcal{M}^H_R$ while satisfying the safety constraint. This translates to that there does not exist a state-action update that implements a policy ascent step under the constraint i.e., $\neg\exists(s' \in \mathcal{S}, a' \in \mathcal{A})$
$[Q^{\pi_{\mathcal{E}}}_{\mathcal{M}^H_R}(s', a'{=}\pi'(s)) \geq Q^{\pi_{\mathcal{E}}}_{\mathcal{M}^H_R}(s', \pi_{\mathcal{E}}(s'))$, s.t. $\forall s' \in \mathcal{S}\ [V^{\pi'}_{\mathcal{M}_R}(s') \geq \delta V^{\pi^*}_{\mathcal{M}_R}(s')]]$. However, if $\pi_{\mathcal{E}} \notin \Pi^*_{\mathcal{E}}$, there must exist another policy $\pi \in \Pi^*_{\mathcal{E}}$ that dominates $\pi_{\mathcal{E}}$ i.e. $\exists(s \in \mathcal{S}, a \in \mathcal{A})$
$[Q^{\pi_{\mathcal{E}}}_{\mathcal{M}^H_R}(s, a{=}\pi(s)) > Q^{\pi_{\mathcal{E}}}_{\mathcal{M}^H_R}(s, \pi_{\mathcal{E}}(s))$, s.t. $\forall s \in \mathcal{S}\ [V^{\pi}_{\mathcal{M}_R}(s) \geq \delta V^{\pi^*}_{\mathcal{M}_R}(s)]]$. This contradicts with the fact that no policy ascent step exists.
Therefore, $\pi_{\mathcal{E}} \in \Pi^*_{\mathcal{E}}$.
Finally, the same conclusion holds for PAG+ (by Lem. 1).

□

2 Appendix B: Domain Descriptions

2.1 Simulations

1) Cliff Worlds (CS & CL): In the cliff worlds, the agent is required to navigate alongside the edge of a cliff to reach the goal. We created a small 4×5 grid, referred to as CS (shown in Fig. 5) to evaluate exact methods and a large 4×100 grid, referred to as CL (shown in Fig. 4) to evaluate approximate methods. To apply approximate methods to CL, the states were aggregated based on features such as distance to the cliff, and agent's position in the grid (e.g., along the edge or at the ends). For CL, we aggregated all non-terminal states into 10 clusters and retained the terminal states (cliffs and goal) as is.

The ground truth ($\mathcal{M}_R$) is that the agent can traverse alongside the edge without slipping off the cliff. Accordingly, $\mathcal{T}_R$ is designed such that the agent heads in the right direction with a probability of 0.9 and remains in the same state with a probability of 0.1. The human's belief ($\mathcal{M}^H_R$) is that the agent may slip off from the edge with some probability, and the terrain closer to the cliff is more uneven and hence more difficult for the agent to traverse. Accordingly, $\mathcal{T}^H_R$ is designed such that the agent heads in the right direction with a probability of 0.7, steers in either direction with a probability of 0.1 each, and remains in the same location with a probability of 0.1.

The reward functions $\mathcal{R}_R$ and $\mathcal{R}^H_R$ are shown in Figs. 4(a) and 4(b) respectively for CL. For CS, the rewards for non-terminal states remain the same but for the cliff states it is -100 instead of -1000, and for the goal state it is 100 instead of 1000.

2) Wumpus World (W): In the wumpus world, we created a 5×5 grid referred to as W (shown in Fig. 3). To apply approximate methods to W, the non-terminal states were aggregated into 15 clusters based on features such as the relative direction of the wumpus from agent and collection status of the gold coins. The agent is required to exit the 5×5

cave while collecting gold coins on its way out and avoiding encounters with the wumpus (i.e., staying in the same location). The cave has a moving wumpus, two gold coins, and an exit location.

The ground truth ($\mathcal{M}_R$) is that the agent's movement is deterministic and the wumpus's movement is stochastic. Accordingly, $\mathcal{T}_R$ is designed such that the wumpus always chooses to move toward the agent's current location with uniform probability. The human's belief $\mathcal{M}_R^H$ is that the actions of both agents are stochastic. Under such a belief, the human would consider it dangerous for the agent to move close to the wumpus. Accordingly, $\mathcal{T}_R^H$ is designed such that the dynamics of the agent's movement are the same as that in the cliff world and the dynamics of the wumpus's movement is to move close to the agent's current location with uniform probability.

In reward functions $\mathcal{R}_R$ and $\mathcal{R}_R^H$, collecting each gold coin gives a reward of $+30$. The game terminates if the agent encounters the wumpus with a reward of -100 or if it exits the cave with a reward of $+100$. The living reward is -0.1.

2.2 Physical Robot Experiment

Robot Assistant Domain: In this domain, a Kinova MOVO robot is assisting a human user with setting up the dining table (refer to Fig. 6). The state was modeled to include the location of an object (which was a napkin in this experiment), the location of an obstacle (which was a vase in this experiment), and the location of the robot. The possible locations for the napkin are viz., on the side table (shown in the top sub-figures), on the front table next to the vase which is away from the human, on the front table next to the plate which is near the human (shown in the bottom sub-figures), and in the robot's gripper (shown in the middle sub-figures). The possible locations for the vase are near the robot (shown in left sub-figures) and away from the robot (shown in right sub-figures), and fallen (if tipped over which is not shown). The possible locations for the robot are close to the side table (shown in the right sub-figures) and close to the human (shown in the middle-left and middle-right sub-figures).

The robot is required to fetch a napkin for the user from another table. In the robot's model ($\mathcal{M}_R$), movement of the arms is restricted by a vase on the table such that placing the napkin close to the user may tip over the vase containing water, resulting in a safety risk. Hence, the robot's optimal behavior is to pick the napkin from the side table and place it next to the vase, which is further away from the user as shown in Fig. 7. Accordingly, $\mathcal{T}_R$ is designed such that moving the napkin from the side table directly toward the human with the obstacle in the way (w.r.t. the robot position) will result in tipping the vase with probability 1.0. However, clearing the vase could successfully displace it with a probability of 0.9, tip it over with a probability of 0.05 and leave it as is with a probability of 0.05.

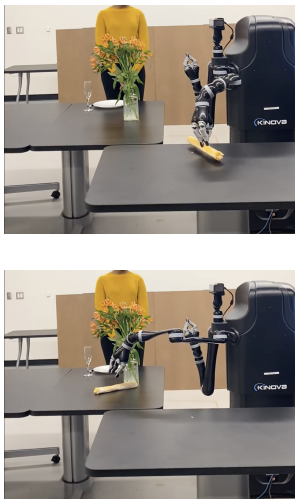

Figure 7: π^* in $\mathcal{M}_R$

The user does not fully understand the kinematic con-

straints of the robot arms and hence expects the robot to place the napkin next to the plate which is close to her. Accordingly, $\mathcal{T}_R^H$ is designed such that the robot can access any location on the table irrespective of its own location or the location of the obstacle. In both $\mathcal{T}_R$ and $\mathcal{T}_R^H$, the probability of executing an action successfully is 0.9 and the probability of failing is 0.1.

The environment terminates when the napkin is placed in any location on the table or if the flower vase is tipped over. In $\mathcal{R}_R$, there is an equal reward of 10 for placing the napkin anywhere on the table and -10 for hitting the vase. In $\mathcal{R}_R^H$, there is a reward of 10 for placing the napkin close to the human and 0 reward for placing the napkin anywhere else on the table.

Please refer to this demo video for further illustration.

3 Appendix C: Code

The code for SEP is available in this repository.