AI Planning Annotation in Reinforcement Learning: Options and Beyond

Junkyu Lee, Michael Katz, Don Joven Agravante, Miao Liu,
Tim Klinger, Murray Campbell, Shirin Sohrabi, Gerald Tesauro

IBM Research AI

PRL workshop at ICAPS 2021

Summary

Planning Annotation in RL
Derive hierarchical RL architecture from AI planning task
Generate option specifications from planning operators

Solving Planning Annotated RL Task
Utilize AI planning and RL algorithms and improve sample efficiency

Future Work
Online approach interleaves option selection and intra-option learning
Learning AI planning task

Background – RL and Options Framework

Markov Decision Process $\mathcal{M} = \langle S, A, P, R, \gamma \rangle$

stationary stochastic policy $\pi(a|s) : S \times A \rightarrow [0,1]$

$$\mathrm{MEU} = \max_\pi \lim_{k\rightarrow\infty} \mathbb{E}_\pi[\sum_{t=0}^{t=k} \gamma^t r^t]$$

$$V^\pi(s) = \sum_a \pi(a|s)[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^\pi(s)]$$

Options Framework $\langle \mathcal{M}, O \rangle$ [Sutton, Precup, and Singh 1999]

$o = \langle I_o, \pi_o, \beta(o) \rangle$
- I_o : option Initiation set
- π_o : intra option policy function
- $\beta(o)$: option termination set

option level policy $\mu(o'|s,o) : S \times O \times O \rightarrow [0,1]$

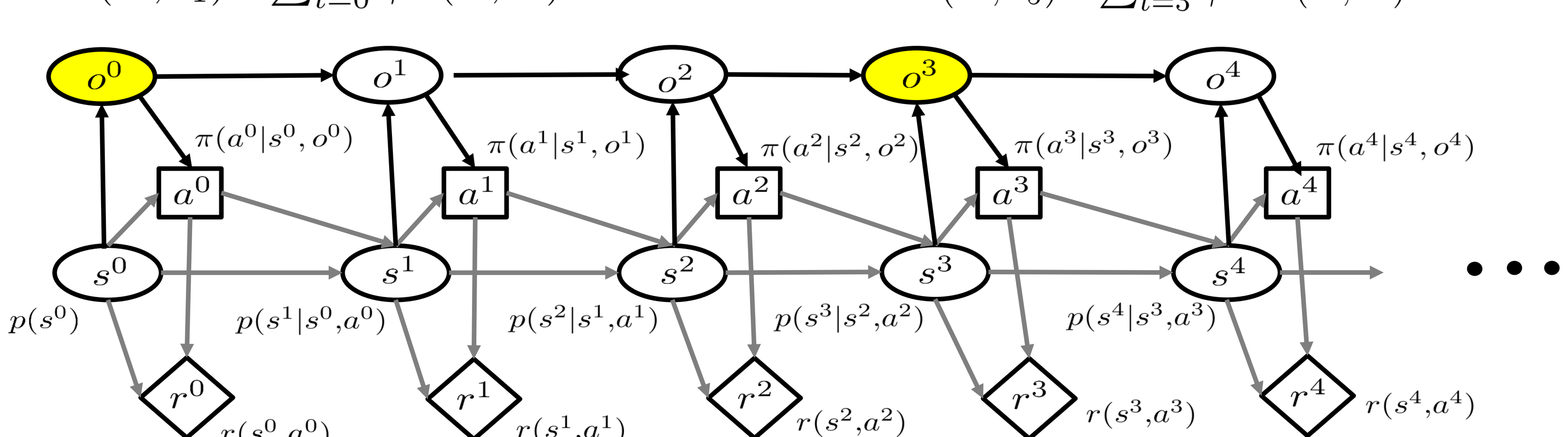
intra option policy $\{\pi_o(a|s,o) | o \in O\}$

Semi-MDP over Options

$o_1 = \arg\max_{o\in O} \bar{\mu}(o|s^0)$ $\quad$ $o_5 = \arg\max_{o \in O} \bar{\mu}(o|o_1, s^3)$

$s^0 \in I_{o_1}$ → $s^3 \in \beta_{o_1}$ $\quad$ $s^3 \in I_{o_5}$ → $s^4 \in \beta_{o_5}$

$\bar{r}(s^0, o_1) = \sum_{t=0}^{t=2} \gamma^t r(s^t, a^t)$ $\quad$ $\bar{r}(s^3, o_5) = \sum_{t=3}^{t=4} \gamma^{t-3} r(s^t, a^t)$



value function for SMDP over options:

$$\bar{V}^{\bar{\mu}}(s) = \sum_{o\in O} \bar{\mu}(o|s)[\bar{r}(s,o) + \gamma \sum_{s' \in S} \bar{p}(s'|s,o) \bar{V}^{\bar{\mu}}(s')]$$

$$\bar{p}(s'|s,o) = \sum_{j=0}^{\infty} \gamma^j p(s' = s^{t+j} | s = s^t)$$

Background – AI Planning Task

AI Planning Task $\mathcal{T} = \langle V', O', S'_G \rangle$

- variables $V' : \{V_0, V_1, \dots, V_{|V'|}\}$
- operators $O' : \{O_1, O_2, \dots, O_{|O'|}\}$
- goal states $S'_G : S'_G \subseteq S'$
- planning states $S' : \{(V_0 = v_0, V_1 = v_1, \dots, V_{|V'|} = v_{|V'|}) | V_i \in V'\}$

Related Works

Hierarchical RL [Kulkarni, et. al 2016]
Define master/slave architecture and master policy generates subgoals for each slave

Option Critic [Bacon and Precup 2017]
End-to-End approach for training intra option and option level policy functions

PEORL/SDRL [Yang, et. al 2018][Lyu, et. al 2019]
Derive a Planning task from BC action language

Taskable RL [Illanes, et. al 2020]
Derive a planning task from subtasks in RL problem

Planning Annotated RL Task

Planning Annotated RL Task (PaRL) $\langle \mathcal{M}, \mathcal{T}, L \rangle$

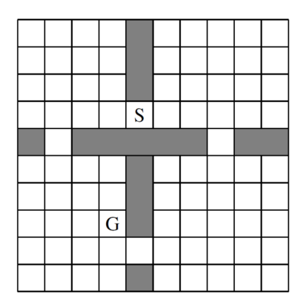
$\mathcal{M}$: MDP $\quad$ $\mathcal{T}$: Planning Task $\quad$ L : State mapping function

Options from AI Planning Task

$$I_{Op} = \{s \in S | \mathrm{precondition}(Op) \subseteq L(s)\}$$

$$\beta_{Op} = \begin{cases} T & \text{if prevail } (Op) \cup \text{ effect } (Op) \subseteq L(s) \\ F & \text{o.w.} \end{cases}$$

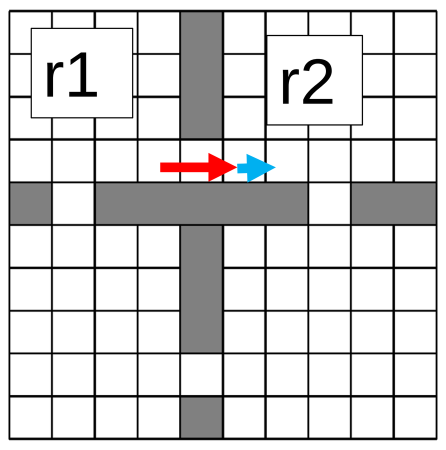
```
(:action move-in-room
    :parameters (?from - location ?to - location ?r - room)
    :precondition (and
        (IN ?from ?r)
        (IN ?to ?r)
        (CONNECTED ?from ?to)
        (in-room ?r)
        (at ?from)
    )
    :effect (and
        (not (at ?from))
        (at ?to)
    )
)

(:action move-out-room
    :parameters (?from - location ?to - location ?r - room ?s - room)
    :precondition (and
        (IN ?from ?r)
        (IN ?to ?s)
        (CONNECTED ?from ?to)
        (CONNECTED-ROOMS ?r ?s)
        (at ?from)
        (in-room ?r)
        (not (at ?to))
        (not (in-room ?s))
    )
    :effect (and
        (not (at ?from))
        (at ?to)
        (not (in-room ?r) )
        (in-room ?s)
    )
)
```



Grounded Operator
(Move from r5 to c-r5-r3)
(precondition): in-room(r5)
(effect): in-room(c-r5-r3)

Derived Option
$I_O = \{s \in \mathcal{S} | \textit{in-room(r5:room)} \subseteq L(s)\}$
$\beta_O = \{s \in \mathcal{S} | \textit{in-room(c-r5-r3:room)} \subseteq L(s)\}$

Solving PaRL



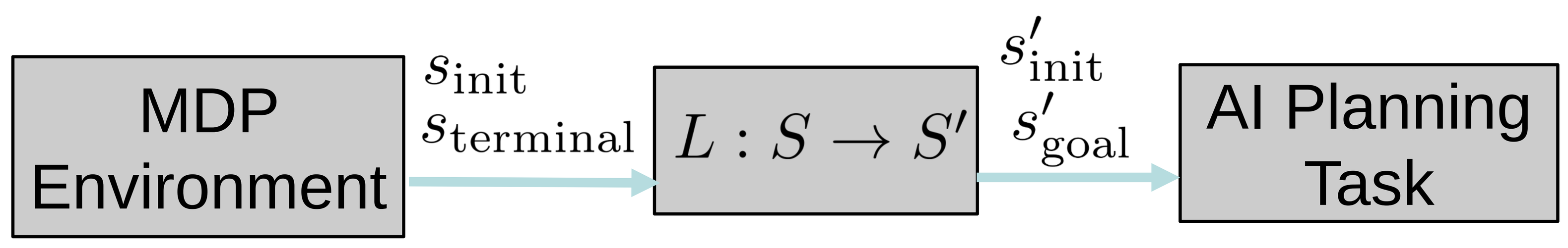
$o_1 :=$ (Move from r1 to c-r1-r2) $\quad$ $o_5 :=$ (Move from c-r1-r2 to r2)

$s^0 \in I_{o_1}$ → $s^3 \in \beta_{o_1}$ $\quad$ $s^3 \in I_{o_5}$ → $s^4 \in \beta_{o_5}$

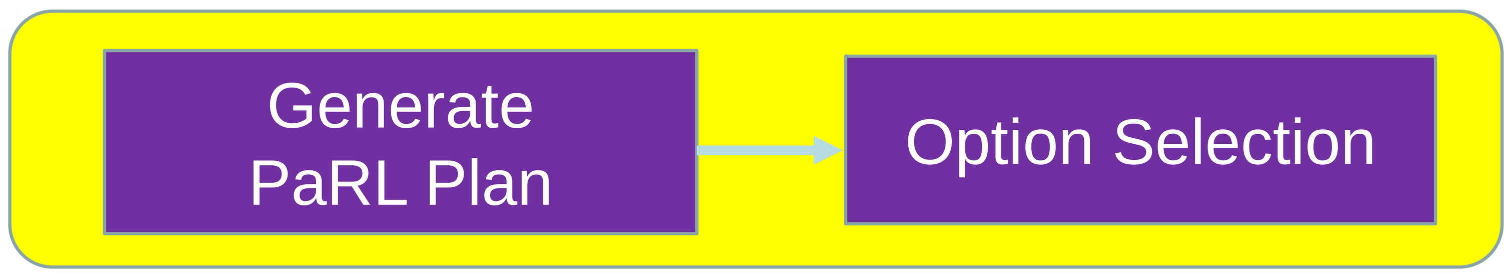
$\bar{r}(s^0, o_1) = \sum_{t=0}^{t=3} \gamma^t r(s^t, a^t)$ $\quad$ $\bar{r}(s^3, o_5) = \sum_{t=3}^{t=4} \gamma^{t-3} r(s^t, a^t)$

Offline Options Training with SMDP learning

Select options for a problem given a fixed initial/ terminal state.



Option Selection by Offline planning



SMDP Learning + PPO using pretrained options [Sutton, Precup, and Singh 1999] [Schulman, et. al 2017]