Integrated Planning and Reinforcement Learning for Compositional Domains

Harsha Kokel
Research Scientist
IBM Research
Integrated Planning and Reinforcement Learning for Compositional Domains — Harsha Kokel
Research Scientist
IBM Research
Reinforcement Learning

Environment

State, Reward

action

Relational MDP
Planning

A: Pre: P
   Add: Q
   Del: P

B: Pre: P, Q, R
   Add: T
   Del: Q, R

C: Pre: P, Q
   Add: U, V
   Del: P, Q

D: Pre: P, T
   Add: U
   Del: P

E: Pre: S, T
   Add: V, W
   Del: S, T

Initial state

Long and Fox (2002)

A state satisfying goals
Planning

• Search through space of states
• Relies on an explicit model of the environment

RL

• Search through space of policies
• Relies on trial & error by interaction
Integrating Planning and RL

knowledge

Planner

Environment

State, Reward

action

initial state & task

high-level plan
Integrating Planning and RL

knowledge → Planner → high-level plan → Environment

initial state & goal

State, Reward → action

OSU: AI Seminar
Decomposing GRMDP

Goal-directed RMDP:

\[ \langle S, A, P, R, \gamma, G \rangle \]

Multiple subgoal RMDP:

\[ \langle S, A, P_1, R_1, \gamma \rangle \]

\[ \langle S, A, P_2, R_2, \gamma \rangle \]

\[ \ldots \]

\[ \langle S, A, P_n, R_n, \gamma \rangle \]

\[
R_o(s, a, s') = \begin{cases} 
    t_R + R(s, a, s') & \text{if } s' \in \beta(o) \text{ and } s \notin \beta(o) \\
    0 & \text{if } s' \in \beta(o) \text{ and } s \in \beta(o) \\
    R(s, a, s') & \text{otherwise}
\end{cases}
\]

\[
P_o(s, a, s') = \begin{cases} 
    0 & \text{if } s \in \beta(o) \text{ and } s' \notin \beta(o) \\
    1 & \text{if } s \in \beta(o) \text{ and } s' \in \beta(o) \\
    P(s, a, s') & \text{otherwise}
\end{cases}
\]
Irrelevant variables

Factored MDP represented as Dynamic Bayesian Network (DBN)
Model-agnostic Abstraction

A model-agnostic abstraction $\phi(s)$ is such that for any action $a$ and an abstract state $\bar{s}$, $\phi(s_1) = \phi(s_2)$ if and only if

$$
\sum_{\{s'_1 | \phi(s'_1) = \bar{s}\}} R_0(s_1, a, s'_1) = \sum_{\{s'_2 | \phi(s'_2) = \bar{s}\}} R_0(s_2, a, s'_2)
$$

$$
\sum_{\{s'_1 | \phi(s'_1) = \bar{s}\}} P_0(s_1, a, s'_1) = \sum_{\{s'_2 | \phi(s'_2) = \bar{s}\}} P_0(s_2, a, s'_2)
$$

Dietterich NeurIPS 2000; Ravindran and Barto IJCAI 2003; Givan, Dean, and Greig AI 2003; Li, Walsh, and Littman ISAIM 2006
Graphical models

- Bayesian Network (BN)
- Dynamic BN
- Dynamic PLM
- Relational
- Time

Probabilistic Logic Models (PLM, PRM, BLP, LBN, DAPER)

Koller and Friedman 2009; Getoor and tasker 2007; Raedt et al. 2016
First Order Conditional Influence (FOCI) statements

\[ \text{if } \langle \text{condition}\rangle \quad \text{then } \langle \text{influents}\rangle \quad \text{QINF } \langle \text{resultant}\rangle \]

Dynamic FOCI statements

\[ [\langle \text{subgoal}\rangle]: \langle \text{influents}\rangle \quad \rightarrow_{[+1]} \quad \langle \text{resultant}\rangle \]

Natarajan, Tadepalli, Dietterich, and Fern 2008
D-FOCI

\{\text{action, taxi}_\text{at}(X)\} \xrightarrow{+1} \text{taxi}_\text{at}(X) \quad (3a)

\text{pick}(P) : \{\text{action, taxi}_\text{at}(X), \text{at}(P, Y), \}

\quad \text{in}_\text{taxi}(P) \xrightarrow{+1} \text{in}_\text{taxi}(P) \quad (3b)

\text{pick}(P) : \{\text{in}_\text{taxi}(P)\} \rightarrow \text{Reward} \quad (3c)

\text{drop}(P) : \{\text{at}_\text{dest}(P)\} \rightarrow \text{Reward} \quad (3d)

\text{drop}(P) : \{\text{at}(P, X), \text{dest}(P, D), \text{at}_\text{dest}(P)\}

\quad \rightarrow \text{at}_\text{dest}(P) \quad (3e)

\text{drop}(P) : \{\text{action, taxi}_\text{at}(X), \text{at}(P, Y), \}

\quad \text{in}_\text{taxi}(P) \xrightarrow{+1} \text{at}(P, K) \quad (3f)
Experiments

Domains
- Office World
- Minecraft World
- Relational Taxi
- Relational Box World
- Craft World
- Robotic Fetch domain

Baselines
- HRL (options framework)
- Tabular Q-learning
- Deep RL (DDQN, HDQN, SAC)
- Deep Relational RL
- Planning+RL (Taskable RL)
Experiments

Sample efficiency
Transfer across task

Office World

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>▲</td>
<td>Agent</td>
</tr>
<tr>
<td>⋊</td>
<td>Furniture</td>
</tr>
<tr>
<td>🚉</td>
<td>Coffee machine</td>
</tr>
<tr>
<td>📨</td>
<td>Mail room</td>
</tr>
<tr>
<td>🏢</td>
<td>Office</td>
</tr>
<tr>
<td>A, B, C, D</td>
<td>Marked locations</td>
</tr>
</tbody>
</table>

Kokel et al. ICAPS 2021
Deep Relational RL

1 Kokel et al. NCAA 2022a
2 Li et al. ICRA 2022
AI Planning + RL = Integrated Methods

- Domain independent heuristics
- Domain knowledge
- Symbolic
- Domain dependent reward
- Simulator and Data
- Neural
- Scalability
- Sample efficiency
- Neuro-Symbolic
Bridging the Gap Planning & RL
Planning

• Search through space of states
• Relies on an explicit model of the environment
• PDDL Task

RL

• Search through space of policies
• Relies on trial & error by interaction
• MDP
**PDDL Task**

\[ \langle L, O, I, G \rangle \]

**Lifted Action Models**\(O\)

(:action move
 :parameters (?curpos ?nextpos ?dir)
 :precondition (and (place ?curpos)
 (place ?nextpos) (at-robot ?curpos)
 :effect (and (at-robot ?nextpos)
 (not (at-robot ?curpos))))

move(c_1_1, c_1_2, east),
move(c_2_2, c_2_1, west), ...

**MDP**

\[ \langle S, A, T, R, \gamma \rangle \]

**Actions**

[east, west, north, south]
“Two actions that are applicable in the same state cannot have the same label”
Are all the parameters of LAM relevant?*

*Do they define different grounded actions that can be applied in a single state?

# of grounding = # of keys * # of rooms
Relevant parameters

Know,

(at key1 room1) ⊕ (at key1 room2)

So,

(pickup key1 room1) ⊕

(pickup key1 room2)

⊕: Mutually exclusive

(:action pickup
 :parameters (?k - key ?r - room)
 :precondition (and (at ?k ?r)
                  (at-agent ?r)
                  (empty-hand))
 :effect (and (not (at ?k ?r))
            (not (empty-hand))
            (carry ?k))
)

Applicable Action Mutex Group (AAMG)

\[\begin{align*}
\text{(pickup key1 room1),} \\
\text{(pickup key1 room2),} \\
\text{(pickup key1 room3),} \\
\vdots \\
\text{(pickup key1 \_\_)}
\end{align*}\]

\[\begin{align*}
\text{(pickup key2 room1),} \\
\text{(pickup key2 room2),} \\
\text{(pickup key2 room3),} \\
\vdots \\
\text{(pickup key2 \_\_)}
\end{align*}\]

\[\begin{align*}
\text{(pickup ?k - key ?r - room)} \\
\downarrow \\
\text{(pickup ?k - key \_\_)}
\end{align*}\]

Parameter Seed Set of pickup 
\{?k – key \}
Action Space Reduction

Figure 2: Comparison of label set sizes on (a) 14 IPC STRIPS domains and (b) 7 HTG domains.
Impact on learning RL policies

Figure 3: Learning curve in the (a) ferry, (b) gripper, (c) blocks, and (d) logistics; with and without action label reduction.
Integrating Planning and RL

- How to represent this knowledge?
- Can this knowledge be refined?
- How to obtain this knowledge?
- How to customize plan for agent’s skill?
- When/how to replan?
- How to represent the goal?

Initial state & task:

Environment:
- State, Reward
- Action

Planner:
- Knowledge
- High-level plan

How to represent this knowledge?

How to obtain this knowledge?

Can this knowledge be refined?

How to customize plan for agent’s skill?

When/how to replan?

How to represent the goal?
Reference

Harsha Kokel, Arjun Manoharan, Sriraam Natarajan, Balaraman Ravindran, Prasad Tadepalli, RePReL: Integrating Relational Planning and Reinforcement Learning for Effective Abstraction, In ICAPS 2021a.


Harsha Kokel, Sriraam Natarajan, Balaraman Ravindran, Prasad Tadepalli, RePReL: A Unified Framework for Integrating Relational Planning and Reinforcement Learning for Effective Abstraction in Discrete and Continuous Domains, In NCAA 2022a.

Harsha Kokel, Nikhilesh Prabhakar, Sriraam Natarajan, Balaraman Ravindran, Prasad Tadepalli, Hybrid Deep RePReL: Integrating Relational Planning and Reinforcement Learning for Information Fusion, In FUSION 2022b.

Harsha Kokel, Mayukh Das, Rakibul Islam, Julia Bonn, Jon Cai, Soham Dan, Anjali Narayan-Chen, Prashant Jayannavar, Janardhan Rao Doppa, Julia Hockenmaier, Sriraam Natarajan, Martha Palmer, Dan Roth, Lara -- Human-guided collaborative problem solver: Effective integration of learning, reasoning and communication, In ACS 2022c.

Harsha Kokel, Junkyu Lee, Michael Katz, Kavitha Srinivas, Shirin Sohrabi, Action Space Reduction for Planning Domains, IJCAI 2023
Questions?