Teaching Multiple Tasks to a Reinforcement Learning Agent using Linear Temporal Logic (LTL)

Rodrigo Toro Icarte Toryn Q. Klassen
Richard Valenzano Sheila A. McIlraith

Computer Science
UNIVERSITY OF TORONTO

ELEMENT AI

VECTOR INSTITUTE | INSTITUT VECTEUR

Learning by Instruction Workshop,
NeurIPS 2018

Outline

1. Motivation
2. Related work
3. Approach:
    - Why (and how) we use LTL for specifying tasks in RL.
    - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

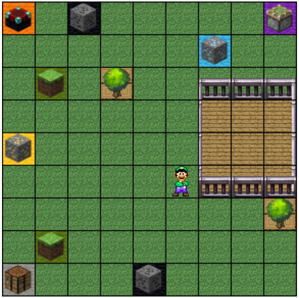
Outline

1. **Motivation**
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

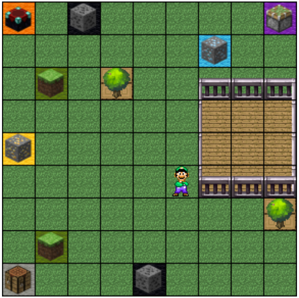
Motivation

How can we instruct RL agents?

Running example



Running example

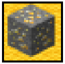


Luigi can collect raw materials:

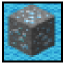
wood

grass

iron

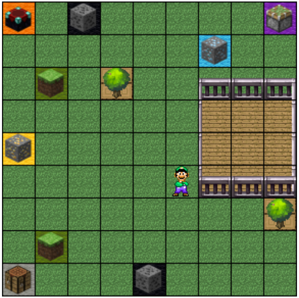


gold



gems

Running example

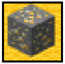


Luigi can collect raw materials:

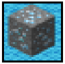
wood

grass

iron



gold



gems

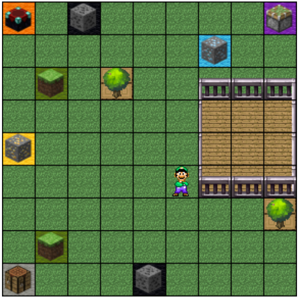
... and make new objects in:

factory

toolshed

workbench

Running example



Luigi can collect raw materials:

wood grass iron gold gems

... and make new objects in:

factory toolshed workbench

Make a bridge: get wood, iron, and use the factory

Running example

Task type	Example
Single goal	get wood
Sequence of goals	get wood and then use the factory
Disjunctive goals	get wood or iron
Conjunctive goals	get grass and iron
Safety constraints	do not leave the shelter at night

Running example

Task type	Example
Single goal	get wood
Sequence of goals	get wood and then use the factory
Disjunctive goals	get wood or iron
Conjunctive goals	get grass and iron
Safety constraints	do not leave the shelter at night

How can we instruct RL agents?

Motivation

Question: How can we instruct RL agents?
Proposal: Let's use language to describe tasks!

Motivation

Question: How can we instruct RL agents?
Proposal: Let's use language to describe tasks!

Desirable properties:

- It is expressive.
- RL agents understand it:
  - Task description → Reward function.
  - Task description → Learn faster.
- Humans understand it.

Outline

1. Motivation
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

Outline

1. Motivation
2. **Related work**
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

Related work

Hindsight Experience Replay by Andrychowicz et al. (NIPS-17)

Language: Single goal condition
Advantage: Learn to achieve goals in parallel (off-policy RL)

Task	HER
get wood	✓
get wood and then use the factory	
get wood or iron	
get grass and iron	
do not leave the shelter at night	
Off-policy learning	✓

Related work

Modular Multitask RL with Policy Sketches by Andreas et al. (ICML-17)

Language: Sequence of sub-goals (called sketch)
Advantage: Decompose the problem using the sketch.

Task	HER	Sketches
get wood	✓	✓
get wood and then use the factory		✓
get wood or iron		
get grass and iron		
do not leave the shelter at night		
Off-policy learning	✓	
Task decomposition		✓

Related work

Teaching Multiple Tasks to an RL Agent using LTL

Language: Linear Temporal Logic (LTL)

Advantage: Decompose and use off-policy RL to learn subtasks.

Task	HER	Sketches	LTL
get wood	✓	✓	✓
get wood and then use the factory		✓	✓
get wood or iron			✓
get grass and iron			✓
do not leave the shelter at night			✓
Off-policy learning	✓		✓
Task decomposition		✓	✓

Outline

1. Motivation
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

Outline

1. Motivation
2. Related work
3. Approach:
   - **Why (and how) we use LTL for specifying tasks in RL.**
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

Defining tasks using LTL

Idea: Let's give the RL agent a set of high-level event detectors.

Defining tasks using LTL

Idea: Let's give the RL agent a set of high-level event detectors.

Example

$$\mathcal{P} = \{\texttt{got\_wood}, \texttt{got\_iron}, \texttt{got\_grass}, \texttt{used\_workbench}, \texttt{used\_factory}, \texttt{is\_night}, \texttt{at\_shelter}, \ldots\}$$

Defining tasks using LTL

Idea: Let's give the RL agent a set of high-level event detectors.

Example

$$\mathcal{P} = \{\texttt{got\_wood}, \texttt{got\_iron}, \texttt{got\_grass}, \texttt{used\_workbench}, \texttt{used\_factory}, \texttt{is\_night}, \texttt{at\_shelter}, \ldots\}$$

Use LTL to define tasks by composing occurrences of events in $\mathcal{P}$

Defining tasks using LTL

Linear Temporal Logic (syntax)

LTL augments propositional logic with temporal operators $\bigcirc$ (*next*), $\Diamond$ (*eventually*), and U (*until*):

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \bigcirc \varphi \mid \Diamond\varphi \mid \varphi_1 \,\mathsf{U}\, \varphi_2 \text{ with } p \in \mathcal{P}$$

Defining tasks using LTL

Linear Temporal Logic (syntax)

LTL augments propositional logic with temporal operators $\bigcirc$ (*next*), $\Diamond$ (*eventually*), and U (*until*):

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \bigcirc \varphi \mid \Diamond\varphi \mid \varphi_1 \,\mathsf{U}\, \varphi_2 \text{ with } p \in \mathcal{P}$$

Examples

$\Diamond\texttt{got\_wood}$

$\Diamond(\texttt{got\_grass} \wedge \Diamond\texttt{used\_factory})$

$\Diamond\texttt{got\_wood} \vee \Diamond\texttt{got\_iron}$

$\Diamond\texttt{got\_grass} \wedge \Diamond\texttt{got\_iron}$

$(\texttt{is\_night} \rightarrow \texttt{at\_shelter}) \,\mathsf{U}\, \texttt{got\_wood}$

Defining tasks using LTL

Linear Temporal Logic (syntax)

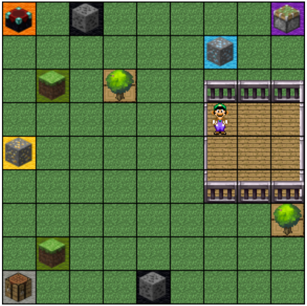
LTL augments propositional logic with temporal operators $\bigcirc$ (*next*), $\Diamond$ (*eventually*), and U (*until*):

$$\varphi ::= p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \bigcirc\varphi \mid \Diamond\varphi \mid \varphi_1 \,\mathsf{U}\, \varphi_2 \text{ with } p \in \mathcal{P}$$

Examples

eventually `got_wood`
eventually (`got_grass` **and eventually** `used_factory`)
eventually `got_wood` **or eventually** `got_iron`
eventually `got_grass` **and eventually** `got_iron`
(`is_night→at_shelter`) **until** `got_wood`

Example

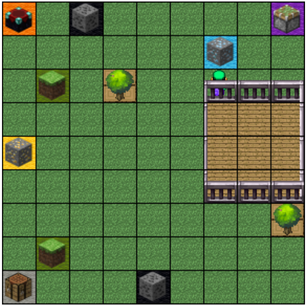


Detected events

`at_shelter`

$\varphi =$ **eventually**(`got_iron` **and eventually** `used_factory`)
and eventually `got_gold`

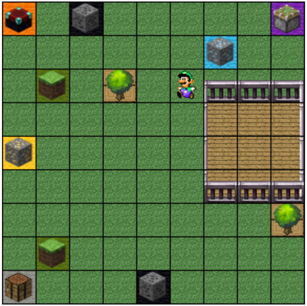
Example



Detected events
none

$$\varphi = \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory})$$
$$\textbf{and eventually}\ \texttt{got\_gold}$$

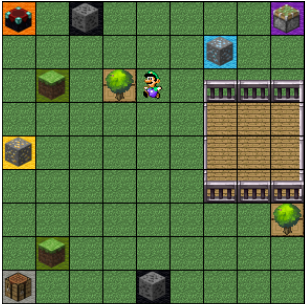
Example



Detected events
none

$$\varphi = \textbf{eventually}(\texttt{got\_iron} \textbf{ and eventually } \texttt{used\_factory}) \\ \textbf{and eventually } \texttt{got\_gold}$$

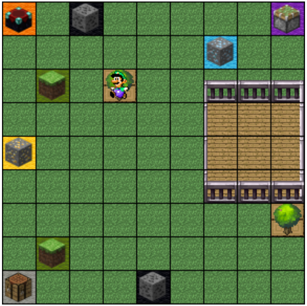
Example



Detected events
none

$$\varphi = \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory})$$
$$\textbf{and eventually}\ \texttt{got\_gold}$$

Example

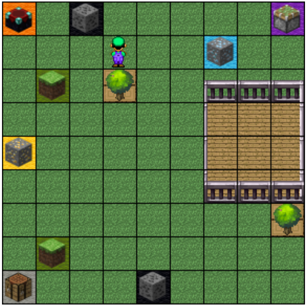


Detected events

`got_wood`

$\varphi =$ **eventually**(`got_iron` **and eventually** `used_factory`)
and eventually `got_gold`

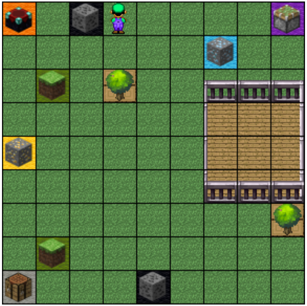
Example



Detected events
`none`

$\varphi =$ **eventually**(`got_iron` **and eventually** `used_factory`)
and eventually `got_gold`

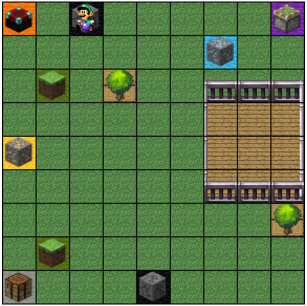
Example



Detected events
`none`

$\varphi =$ **eventually**(`got_iron` **and eventually** `used_factory`)
and eventually `got_gold`

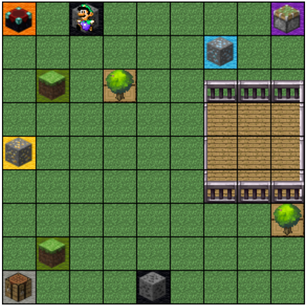
Example



Detected events
got_iron

$$\varphi = \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory})$$
$$\textbf{and eventually}\ \texttt{got\_gold}$$

Example

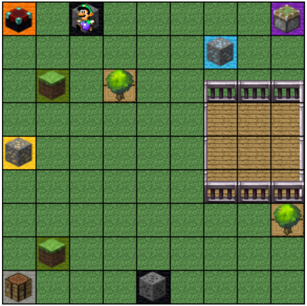


Detected events

`got_iron`

$\varphi =$ **eventually**(~~`got_iron`~~ **and eventually** `used_factory`)
and eventually `got_gold`

Example



Detected events
got_iron

$\varphi =$ **eventually** used_factory **and eventually** got_gold

LTL progression

LTL progression

Given an LTL formula φ and state s, we can *progress* φ using s:

- $\text{prog}(s, p) = \text{true}$ if $p \in L(s)$, where $p \in \mathcal{P}$
- $\text{prog}(s, p) = \text{false}$ if $p \notin L(s)$, where $p \in \mathcal{P}$
- $\text{prog}(s, \neg\varphi) = \neg\,\text{prog}(s, \varphi)$
- $\text{prog}(s, \varphi_1 \wedge \varphi_2) = \text{prog}(s, \varphi_1) \wedge \text{prog}(s, \varphi_2)$
- $\text{prog}(s, \bigcirc\varphi) = \varphi$
- $\text{prog}(s, \Diamond\varphi) = \text{prog}(s, \varphi) \vee \Diamond\varphi$
- $\text{prog}(s, \varphi_1 \,\mathsf{U}\, \varphi_2) = \text{prog}(s, \varphi_2) \vee (\text{prog}(s, \varphi_1) \wedge \varphi_1 \,\mathsf{U}\, \varphi_2)$

LTL progression

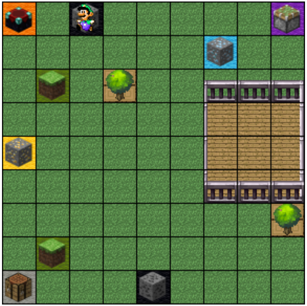
LTL progression

Given an LTL formula φ and state s, we can *progress* φ using s:

- $\text{prog}(s, p) = \text{true}$ if $p \in L(s)$, where $p \in \mathcal{P}$
- $\text{prog}(s, p) = \text{false}$ if $p \notin L(s)$, where $p \in \mathcal{P}$
- $\text{prog}(s, \neg\varphi) = \neg\,\text{prog}(s, \varphi)$
- $\text{prog}(s, \varphi_1 \wedge \varphi_2) = \text{prog}(s, \varphi_1) \wedge \text{prog}(s, \varphi_2)$
- $\text{prog}(s, \bigcirc\varphi) = \varphi$
- $\text{prog}(s, \Diamond\varphi) = \text{prog}(s, \varphi) \vee \Diamond\varphi$
- $\text{prog}(s, \varphi_1 \,\mathsf{U}\, \varphi_2) = \text{prog}(s, \varphi_2) \vee (\text{prog}(s, \varphi_1) \wedge \varphi_1 \,\mathsf{U}\, \varphi_2)$

This is a **correct** and **well-defined** procedure!

Example

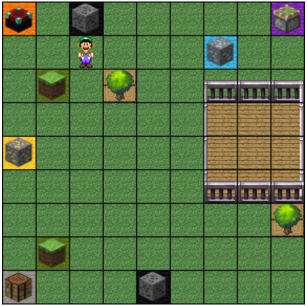


Detected events

got_iron

$\varphi =$ **eventually** `used_factory` **and eventually** `got_gold`

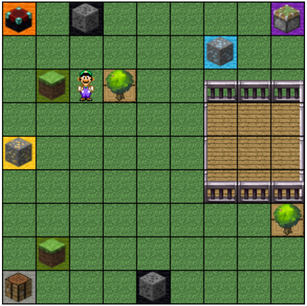
Example



Detected events
`none`

$\varphi =$ **eventually** `used_factory` **and eventually** `got_gold`

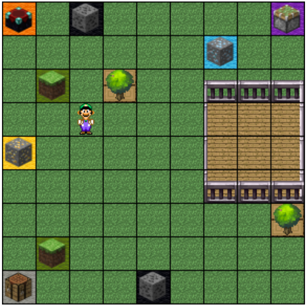
Example



Detected events
none

$\varphi =$ **eventually** `used_factory` **and eventually** `got_gold`

Example

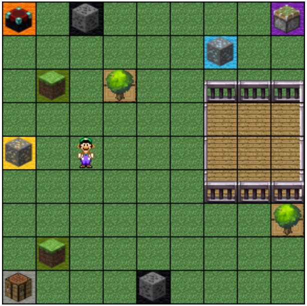


Detected events

`none`

$\varphi =$ **eventually** `used_factory` **and eventually** `got_gold`

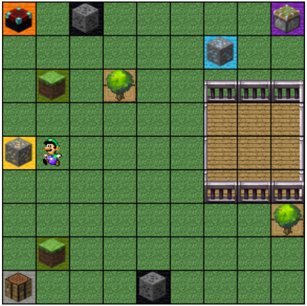
Example



Detected events
none

$\varphi =$ **eventually** `used_factory` **and eventually** `got_gold`

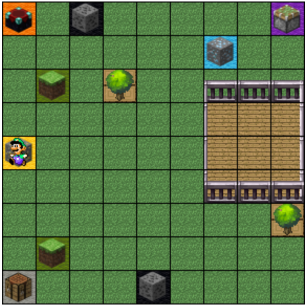
Example



Detected events
`none`

$\varphi =$ **eventually** `used_factory` **and eventually** `got_gold`

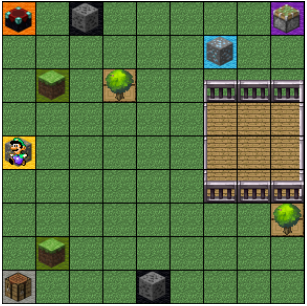
Example



Detected events
got_gold

$\varphi =$ **eventually** `used_factory` **and eventually** `got_gold`

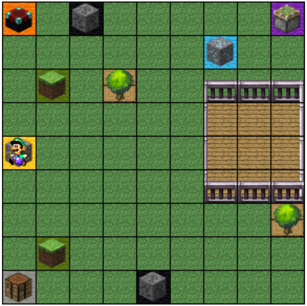
Example



Detected events
got_gold

$\varphi =$ **eventually** used_factory **and eventually** ~~got_gold~~

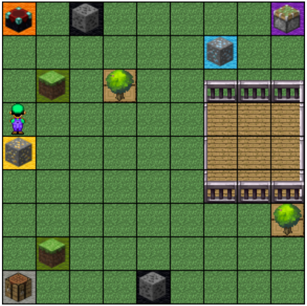
Example



Detected events
got_gold

$$\varphi = \textbf{eventually}\ \texttt{used\_factory}$$

Example

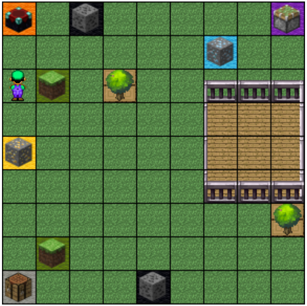


Detected events

`none`

$\varphi =$ **eventually** `used_factory`

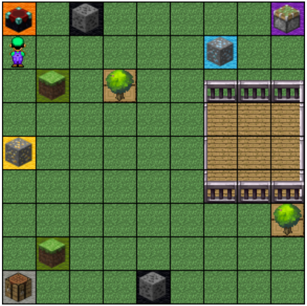
Example



Detected events
`none`

$\varphi =$ **eventually** `used_factory`

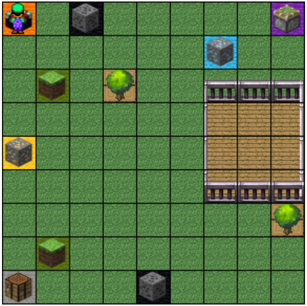
Example



Detected events
none

$\varphi =$ **eventually** `used_factory`

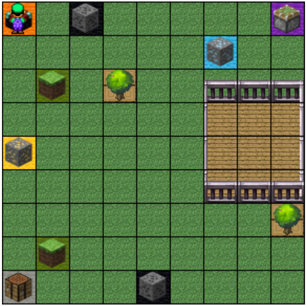
Example



Detected events
`used_factory`

$\varphi =$ **eventually** `used_factory`

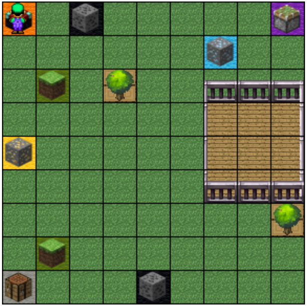
Example



Detected events
`used_factory`

$\varphi =$ **eventually** ~~used_factory~~

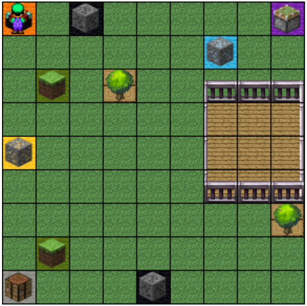
Example



Detected events
`used_factory`

$\varphi =$ true (+1 reward)

Example

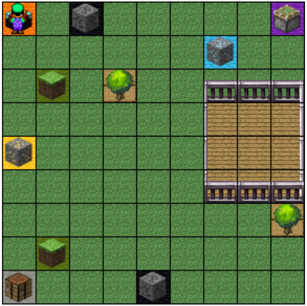


Detected events

`used_factory`

LTL formulas → Rewards

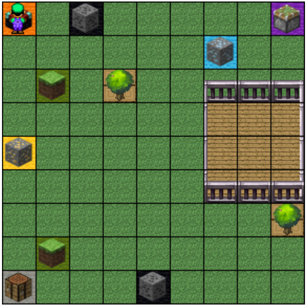
Example



Detected events
`used_factory`

(φ can be learned using standard RL)

Example



Detected events
`used_factory`

We can do better than this!

Outline

1. Motivation
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

Outline

1. Motivation
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - **Why (and how) LPOPL speeds up learning of multiple tasks.**
4. Results
5. Concluding remarks

LPOPL: An example

Suppose Luigi has to learn two tasks:

LPOPL: An example

Suppose Luigi has to learn two tasks:

$$\varphi_1 = \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory})\ \textbf{and eventually}\ \texttt{got\_gold}$$

$$\varphi_2 = \textbf{eventually}([\texttt{got\_grass}\ \textbf{or}\ \texttt{got\_wood}]\ \textbf{and eventually}\ \texttt{used\_factory})$$

LPOPL: An example

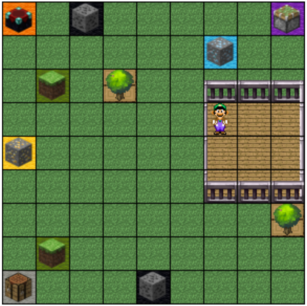
Suppose Luigi has to learn two tasks:

$$\varphi_1 = \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory})\ \textbf{and}\ \textbf{eventually}\ \texttt{got\_gold}$$

$$\varphi_2 = \textbf{eventually}([\texttt{got\_grass}\ \textbf{or}\ \texttt{got\_wood}]\ \textbf{and eventually}\ \texttt{used\_factory})$$

... and begins by trying to solve φ_1

LPOPL: An example



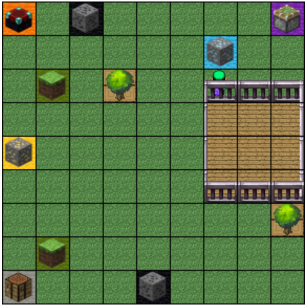
Detected events

`at_shelter`

$\varphi_1 =$ **eventually**(`got_iron` **and eventually** `used_factory`)
and eventually `got_gold`

$\varphi_2 =$ **eventually**([`got_grass` **or** `got_wood`] **and eventually** `used_factory`)

LPOPL: An example

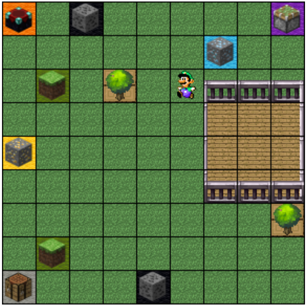


Detected events
`none`

$\varphi_1 =$ **eventually**(`got_iron` **and eventually** `used_factory`) **and eventually** `got_gold`

$\varphi_2 =$ **eventually**([`got_grass` **or** `got_wood`] **and eventually** `used_factory`)

LPOPL: An example

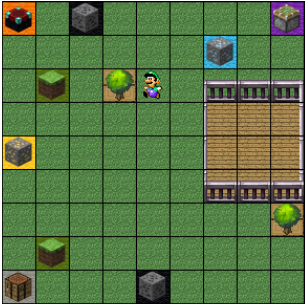


Detected events
`none`

$\varphi_1 =$ **eventually**(`got_iron` **and eventually** `used_factory`) **and eventually** `got_gold`

$\varphi_2 =$ **eventually**([`got_grass` **or** `got_wood`] **and eventually** `used_factory`)

LPOPL: An example

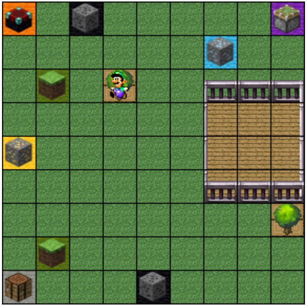


Detected events
none

$$\varphi_1 = \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory})\ \textbf{and eventually}\ \texttt{got\_gold}$$

$$\varphi_2 = \textbf{eventually}([\texttt{got\_grass}\ \textbf{or}\ \texttt{got\_wood}]\ \textbf{and eventually}\ \texttt{used\_factory})$$

LPOPL: An example

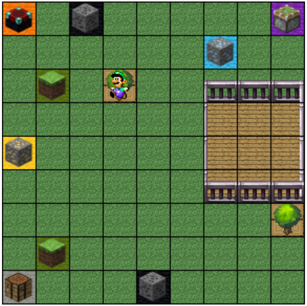


Detected events
got_wood

$$\varphi_1 = \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory})\ \textbf{and eventually}\ \texttt{got\_gold}$$

$$\varphi_2 = \textbf{eventually}([\texttt{got\_grass}\ \textbf{or}\ \texttt{got\_wood}]\ \textbf{and eventually}\ \texttt{used\_factory})$$

LPOPL: An example

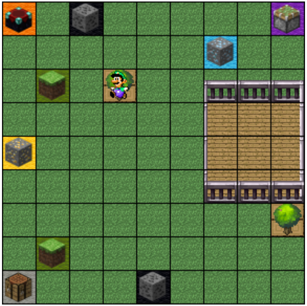


Detected events
got_wood

$\varphi_1 =$ **eventually**(got_iron **and eventually** used_factory) **and eventually** got_gold

$\varphi_2 =$ **eventually**([got_grass **or** got_wood] **and eventually** used_factory)

LPOPL: An example



Detected events

got_wood

$\varphi_1 =$ **eventually**(got_iron **and eventually** used_factory) **and eventually** got_gold

$\varphi_2 =$ **eventually**([got_grass **or** got_wood] **and eventually** used_factory)

LPOPL learns all the tasks in parallel!

LPOPL overview

Step 1: Decompose the tasks into subtasks using LTL progression.

LPOPL overview

Step 1: Decompose the tasks into subtasks using LTL progression.

$\varphi_1 =$ **eventually**(`got_iron` **and eventually** `used_factory`)
and eventually `got_gold`

$\varphi_2 =$ **eventually**([`got_grass` **or** `got_wood`]
and eventually `used_factory`)

LPOPL overview

Step 1: Decompose the tasks into subtasks using LTL progression.

$$\begin{aligned}
\varphi_1 &= \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory}) \\
&\quad \textbf{and eventually}\ \texttt{got\_gold} \\
\varphi_2 &= \textbf{eventually}([\texttt{got\_grass}\ \textbf{or}\ \texttt{got\_wood}] \\
&\quad \textbf{and eventually}\ \texttt{used\_factory}) \\
\varphi_3 &= \textbf{eventually}(\texttt{got\_iron}\ \textbf{and eventually}\ \texttt{used\_factory}) \\
\varphi_4 &= \textbf{eventually}\ \texttt{used\_factory}\ \textbf{and eventually}\ \texttt{got\_gold} \\
\varphi_5 &= \textbf{eventually}\ \texttt{used\_factory} \\
\varphi_6 &= \textbf{eventually}\ \texttt{got\_gold} \\
\varphi_7 &= \text{true}
\end{aligned}$$

LPOPL overview

Step 1: Decompose the tasks into subtasks using LTL progression.

$$\begin{aligned}
\varphi_1 &= \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \\
&\quad \wedge\Diamond\texttt{got\_gold} \\
\varphi_2 &= \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \\
&\quad \wedge\Diamond\texttt{used\_factory}) \\
\varphi_3 &= \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \\
\varphi_4 &= \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold} \\
\varphi_5 &= \Diamond\texttt{used\_factory} \\
\varphi_6 &= \Diamond\texttt{got\_gold} \\
\varphi_7 &= \text{true}
\end{aligned}$$

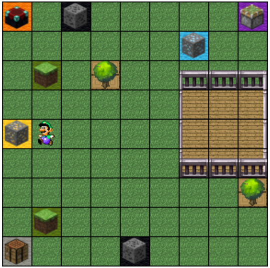
LPOPL overview

Step 1: Decompose the tasks into subtasks using LTL progression.

$$\begin{aligned}
\varphi_1 &= \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \\
&\quad \wedge \Diamond\texttt{got\_gold} \\
\varphi_2 &= \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \\
&\quad \wedge \Diamond\texttt{used\_factory}) \\
\varphi_3 &= \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \\
\varphi_4 &= \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold} \\
\varphi_5 &= \Diamond\texttt{used\_factory} \\
\varphi_6 &= \Diamond\texttt{got\_gold} \\
\varphi_7 &= \text{true}
\end{aligned}$$

Step 2: Learn one policy per subtask using off-policy learning.

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

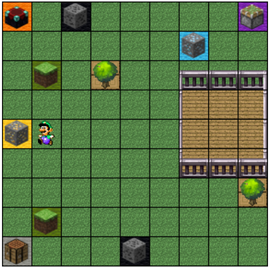
$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `none`

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `none`

Q-updates

$Q_{\varphi_1}(s,a)$

$Q_{\varphi_2}(s,a)$

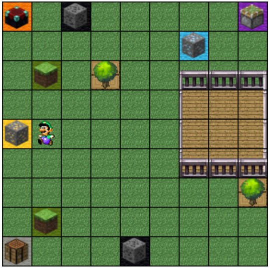
$Q_{\varphi_3}(s,a)$

$Q_{\varphi_4}(s,a)$

$Q_{\varphi_5}(s,a)$

$Q_{\varphi_6}(s,a)$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

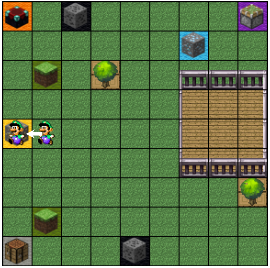
$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `none`

Q-updates

$Q_{\varphi_1}(s,a)$	$Q_{\varphi_4}(s,a)$
$Q_{\varphi_2}(s,a)$	$Q_{\varphi_5}(s,a)$
$Q_{\varphi_3}(s,a)$	$Q_{\varphi_6}(s,a)$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

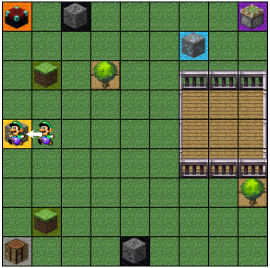
$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `got_gold`

Q-updates

$Q_{\varphi_1}(s,a)$	$Q_{\varphi_4}(s,a)$
$Q_{\varphi_2}(s,a)$	$Q_{\varphi_5}(s,a)$
$Q_{\varphi_3}(s,a)$	$Q_{\varphi_6}(s,a)$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

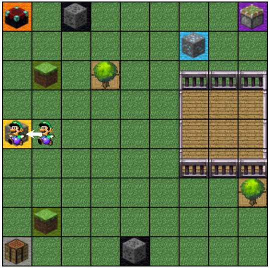
$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `got_gold`

Q-updates

$Q_{\varphi_1}(s,a)$	$Q_{\varphi_4}(s,a)$
$Q_{\varphi_2}(s,a)$	$Q_{\varphi_5}(s,a)$
$Q_{\varphi_3}(s,a)$	$Q_{\varphi_6}(s,a)$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

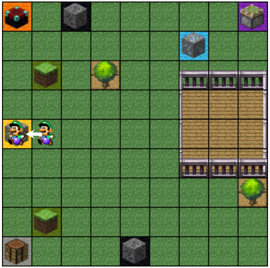
$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `got_gold`

Q-updates

$Q_{\varphi_1}(s,a)$	$Q_{\varphi_4}(s,a)$
$Q_{\varphi_2}(s,a)$	$Q_{\varphi_5}(s,a)$
$Q_{\varphi_3}(s,a)$	$Q_{\varphi_6}(s,a)$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

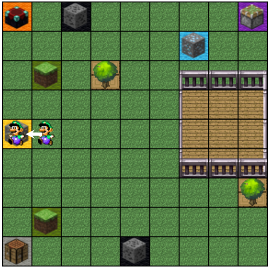
$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `got_gold`

Q-updates

$Q_{\varphi_1}(s, a)$	$Q_{\varphi_4}(s, a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_5}(s', a')$
$Q_{\varphi_2}(s, a)$	$Q_{\varphi_5}(s, a)$
$Q_{\varphi_3}(s, a)$	$Q_{\varphi_6}(s, a)$

Example



Subtasks

$$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$$

$$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$$

$$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$$

$$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$$

$$\varphi_5 = \Diamond\texttt{used\_factory}$$

$$\varphi_6 = \Diamond\texttt{got\_gold}$$

Detected events: got_gold

Q-updates

$$Q_{\varphi_1}(s, a)$$

$$Q_{\varphi_2}(s, a)$$

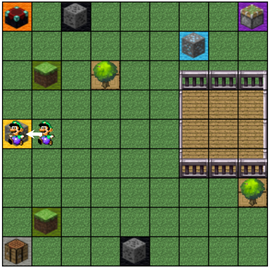
$$Q_{\varphi_3}(s, a)$$

$$Q_{\varphi_4}(s, a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_5}(s', a')$$

$$Q_{\varphi_5}(s, a)$$

$$Q_{\varphi_6}(s, a)$$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

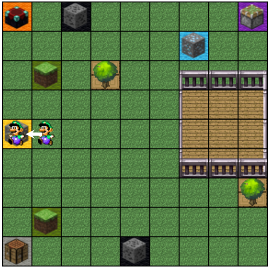
$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `got_gold`

Q-updates

$Q_{\varphi_1}(s,a)$	$Q_{\varphi_4}(s,a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_5}(s',a')$
$Q_{\varphi_2}(s,a)$	$Q_{\varphi_5}(s,a)$
$Q_{\varphi_3}(s,a)$	$Q_{\varphi_6}(s,a) \xleftarrow{\alpha} 1$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `got_gold`

Q-updates

$Q_{\varphi_1}(s,a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_3}(s',a')$

$Q_{\varphi_2}(s,a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_2}(s',a')$

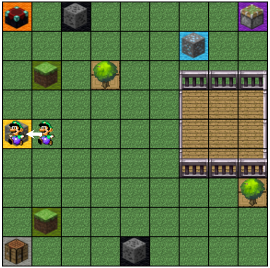
$Q_{\varphi_3}(s,a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_3}(s',a')$

$Q_{\varphi_4}(s,a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_5}(s',a')$

$Q_{\varphi_5}(s,a) \xleftarrow{\alpha} \gamma \max_{a'} Q_{\varphi_5}(s',a')$

$Q_{\varphi_6}(s,a) \xleftarrow{\alpha} 1$

Example



Subtasks

$\varphi_1 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory}) \wedge \Diamond\texttt{got\_gold}$

$\varphi_2 = \Diamond([\texttt{got\_grass} \vee \texttt{got\_wood}] \wedge \Diamond\texttt{used\_factory})$

$\varphi_3 = \Diamond(\texttt{got\_iron} \wedge \Diamond\texttt{used\_factory})$

$\varphi_4 = \Diamond\texttt{used\_factory} \wedge \Diamond\texttt{got\_gold}$

$\varphi_5 = \Diamond\texttt{used\_factory}$

$\varphi_6 = \Diamond\texttt{got\_gold}$

Detected events: `got_gold`

Theorem

LPOPL using tabular q-learning converges to an optimal policy.

Outline

1. Motivation
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

Outline

1. Motivation
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. **Results**
5. Concluding remarks

Experiments

Goal

- Study LPOPL + DQN
- Compare with standard RL
- Compare with alternative decomposition methods for RL

Experiments

Goal

- Study LPOPL + DQN
- Compare with standard RL
- Compare with alternative decomposition methods for RL

Baselines

- DQN-L: Standard DQN[1] (no decomposition).
- HRL-E: Hierarchical Deep RL proposed by Kulkarni et al.[2]
- HRL-L: HRL-E but exploiting LTL to prune *useless* options.

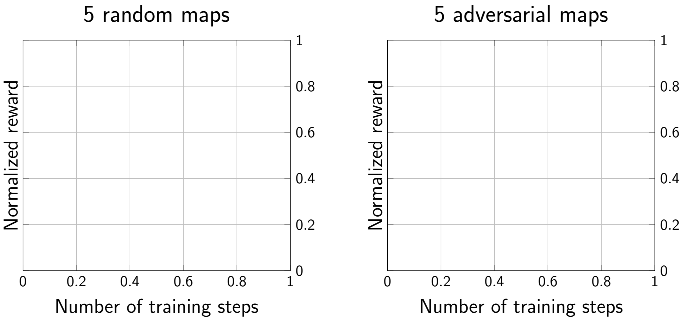
[1] Human-level control through deep reinforcement learning (Nature-15)

[2] HDRL: Integrating Temporal Abstraction and Intrinsic Motivation (NIPS-16)

Experiment 1: Interleaving subtasks

Description

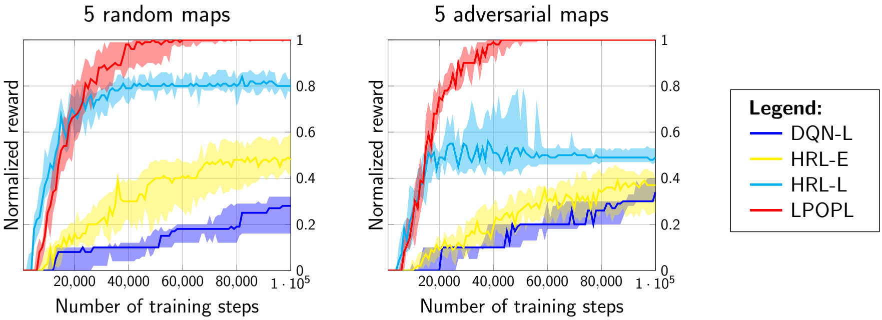
Tasks from (Andreas et al., 2017) w/o unneeded order constraints.
e.g. (get iron and get wood), then use factory.



Experiment 1: Interleaving subtasks

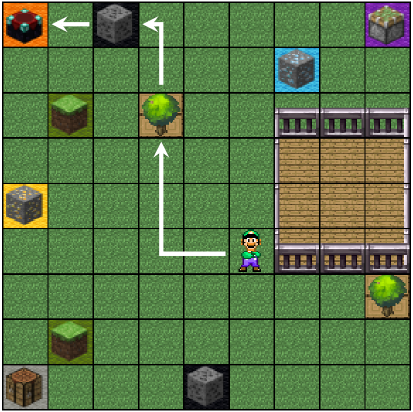
Description

Tasks from (Andreas et al., 2017) w/o unneeded order constraints. e.g. (get iron and get wood), then use factory.

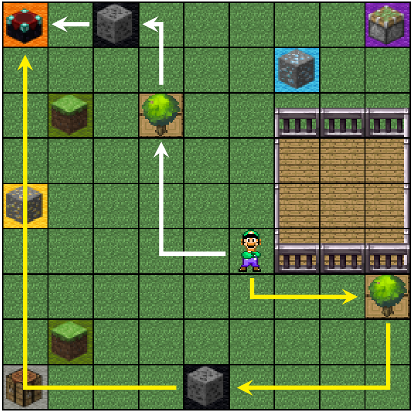

5 random maps

5 adversarial maps

Hierarchical RL might converge to suboptimal policies



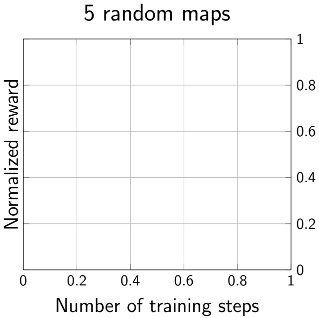
Hierarchical RL might converge to suboptimal policies

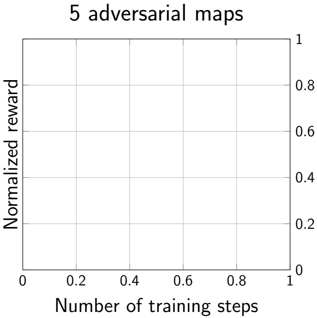


Experiment 2: Safety constraints

Description

Same set of 10 tasks but including the safety constraint of being at the shelter during the night.

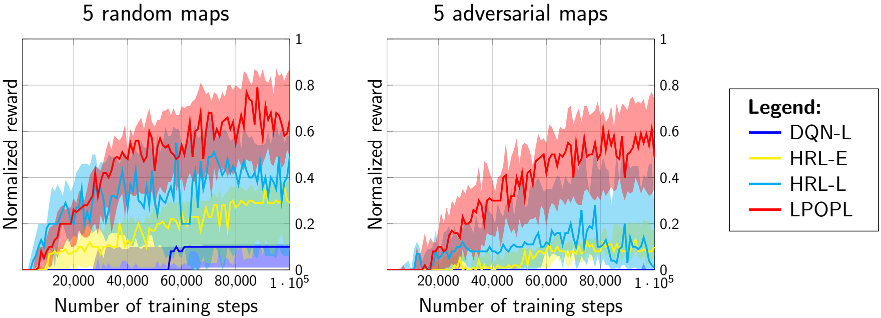




Experiment 2: Safety constraints

Description

Same set of 10 tasks but including the safety constraint of being at the shelter during the night.


5 random maps

5 adversarial maps

Outline

1. Motivation
2. Related work
3. Approach:
    - Why (and how) we use LTL for specifying tasks in RL.
    - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. Concluding remarks

Outline

1. Motivation
2. Related work
3. Approach:
   - Why (and how) we use LTL for specifying tasks in RL.
   - Why (and how) LPOPL speeds up learning of multiple tasks.
4. Results
5. **Concluding remarks**

Concluding remarks

Question: How can we instruct RL agents?

Concluding remarks

Question: How can we instruct RL agents?
Proposal: Define tasks using Linear Temporal Logic.

Concluding remarks

Question: How can we instruct RL agents?
Proposal: Define tasks using Linear Temporal Logic.

Desirable properties:

- It is expressive.

Concluding remarks

Question: How can we instruct RL agents?
Proposal: Define tasks using Linear Temporal Logic.

Desirable properties:

- It is expressive.
- RL agents understand it:
  - Task description → Reward function.
  - Task description → Learn faster.

Concluding remarks

Question: How can we instruct RL agents?
Proposal: Define tasks using Linear Temporal Logic.

Desirable properties:

- It is expressive.
- RL agents understand it:
  - Task description → Reward function.
  - Task description → Learn faster.
- Humans understand it.

Concluding remarks

Question: How can we instruct RL agents?
Proposal: Define tasks using Linear Temporal Logic.

Desirable properties:

- It is expressive.
- RL agents understand it:
  - Task description → Reward function.
  - Task description → Learn faster.
- Humans understand it.
  - NL → LTL by Dzifcak et al. (ICRA-09).

Concluding remarks

Question: How can we instruct RL agents?
Proposal: Define tasks using Linear Temporal Logic.

Desirable properties:

- It is expressive.
- RL agents understand it:
  - Task description → Reward function.
  - Task description → Learn faster.
- Humans understand it.
  - NL → LTL by Dzifcak et al. (ICRA-09).

Code: `https://bitbucket.org/RToroIcarte/lpopl`