ExTra: Transfer-guided Exploration Supplementary Material

Paper #1073

1 EXTRA VARIANTS OF TRADITIONAL EXPLORATION STRATEGIES

Algorithm 1: ExTra + vanilla ϵ -greedy Q-learning

```
2: while step < MAXSTEPS do
         with probability \epsilon
               with probability \epsilon_{bisim}
 5:
                  a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)
               with probability 1 - \epsilon_{bisim}
                  a_2 \sim uniform(A_2)
         with probability 1 - \epsilon
 8:
               a_2 \leftarrow arg \max_{a_2'} Q_2(s_2, a_2')
 9:
10:
         r = take\_step(a_2)
         update_Q(Q_2(s_2, a_2), r)
         step = step + 1
13: end while
```

Algorithm 2: ExTra + Softmax

```
1: step = 0

2: while step < MAXSTEPS do

3: with probability \epsilon

4: a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)

5: with probability 1 - \epsilon

6: a_2 \sim softmax(Q_2(s_2, \cdot))

7: r = take\_step(a_2)

8: update\_Q(Q_2(s_2, a_2), r)

9: step = step + 1

10: end while
```

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand

© 2020 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved. https://doi.org/doi

Algorithm 3: ExTra + Pursuit

```
1: step = 0

2: \pi_{pursuit} = Uniform(A_2)

3: while step < MAXSTEPS do

4: with probability \epsilon

5: a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)

6: with probability 1 - \epsilon

7: a \leftarrow arg \max_{a'_2} Q_2(s_2, a'_2)

8: update\_\pi_{pursuit}(a)

9: a_2 \leftarrow Sample(\pi_{pursuit})

10: r = take\_step(a_2)

11: update\_Q(Q_2(s_2, a_2), r)

12: step = step + 1

13: end while
```

Algorithm 4: ExTra + MBIE-EB

```
1: step = 0

2: while step < MAXSTEPS do

3: with probability \epsilon

4: a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)

5: with probability 1 - \epsilon

6: a_2 \leftarrow arg \max_{b'} Q_2(s_2, a_2')

7: r = take\_step(a_2) + \frac{\beta}{\sqrt{n(s_2, a_2)}}

8: update\_Q(Q_2(s_2, a_2), r)

9: step = step + 1

10: end while
```

2 HYPERPARAMETERS FOR OPTIMISTIC BISIMULATION TRANSFER

Transfer	Parameters	FourLarge	Rooms
SixLarge	Rooms	NineLarge	Rooms
c_R	0.1	0.2	0.1
c_T	0.9	0.9	0.9
Threshold	0.01	0.01	0.01

3 HYPERPARAMETERS FOR BASELINE EXPLORATION STRATEGIES

ϵ -greedy	Q Learning Rate	0.2
e-greedy	ϵ	0.5
Softmax	Q Learning Rate	0.2
Sommax	τ	8.1
MBIE-EB	Q Learning Rate	0.2
	cb-β	0.005
	ϵ	0.2
Pursuit	Q Learning Rate	0.2
1 ursuit	β	0.007
ExTra	Q Learning Rate	0.5
	ϵ	0.2
	α	1e-6