ExTra: Transfer-guided Exploration Supplementary Material

Paper #1073

1 EXTRA VARIANTS OF TRADITIONAL EXPLORATION STRATEGIES

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

step = 0
while step < MAXSTEPS **do**
  with probability ϵ
    with probability ϵ_{bisim}
      $a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)$
    with probability $1 - \epsilon_{bisim}$
      $a_2 \sim uniform(A_2)$
  with probability $1 - \epsilon$
    $a_2 \leftarrow arg\max_{a_2'} Q_2(s_2, a_2')$
  $r = take\_step(a_2)$
  $update\_Q(Q_2(s_2, a_2), r)$
  step = step + 1
end while

Algorithm 2: ExTra + Softmax

step = 0
while step < MAXSTEPS **do**
  with probability ϵ
    $a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)$
  with probability $1 - \epsilon$
    $a_2 \sim softmax(Q_2(s_2, \cdot))$
  $r = take\_step(a_2)$
  $update\_Q(Q_2(s_2, a_2), r)$
  step = step + 1
end while

Algorithm 3: ExTra + Pursuit

step = 0
$\pi_{pursuit} = Uniform(A_2)$
while step < MAXSTEPS **do**
  with probability ϵ
    $a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)$
  with probability $1 - \epsilon$
    $a \leftarrow arg\max_{a_2'} Q_2(s_2, a_2')$
    $update\_\pi_{pursuit}(a)$
    $a_2 \leftarrow Sample(\pi_{pursuit})$
  $r = take\_step(a_2)$
  $update\_Q(Q_2(s_2, a_2), r)$
  step = step + 1
end while

Algorithm 4: ExTra + MBIE-EB

step = 0
while step < MAXSTEPS **do**
  with probability ϵ
    $a_2 \sim \pi_{ExTra}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)$
  with probability $1 - \epsilon$
    $a_2 \leftarrow arg\max_{b'} Q_2(s_2, a_2')$
  $r = take\_step(a_2) + \frac{\beta}{\sqrt{n(s_2, a_2)}}$
  $update\_Q(Q_2(s_2, a_2), r)$
  step = step + 1
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

https://doi.org/doi

2 HYPERPARAMETERS FOR OPTIMISTIC BISIMULATION TRANSFER

Transfer SixLarge	Parameters Rooms	FourLarge NineLarge	Rooms 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
	ϵ	0.5
Softmax	Q Learning Rate	0.2
	τ	8.1
MBIE-EB	Q Learning Rate	0.2
	cb-β	0.005
	ϵ	0.2
Pursuit	Q Learning Rate	0.2
	β	0.007
ExTra	Q Learning Rate	0.5
	ϵ	0.2
	α	1e-6