Supplementary Material

Paper #939

NETWORK STRUCTURE

The network structure is the same for all methods: the actor network has two fully-connected hidden layers both with 64 hidden units, the output layer is a fully-connected layer that outputs the action probabilities for all actions; the critic network contains two fully-connected hidden layers both with 64 hidden units and a fully-connected output layer with a single output: the state value; the option-value network contains two fully-connected hidden layers both with 32 units; two output layers, one outputs the option-values for all options, and the other outputs the termination probability of the selected option.

Grid world

The input consists of the following information: the coordinate of the agent and the environmental information (i.e., each of surrounding eight grids is a wall or not) which is encoded as a one-hot vector.

Pinball

The input contains the position of the ball (x and y) and the velocity of the ball in the $x - y$ plane.

Reacher

The input contains the positions of the finger (x and y), the relative distance to the target position, and the velocity of in the $x - y$ plane.

Parameter Settings

Table 1: CAPS Hyperparameters.

Hyperparameter	Value
Discount factor(γ)	0.99
Optimizer	Adam
Learning rate	$3e-4$
ϵ decrement	$1e-3$
ϵ-start	1.0
ϵ-end	0.05
Batch size	32
Number of episodes replacing the target network	1000

Table 2: A3C Hyperparameters.

Hyperparameter	Value
Number of processes	8
Discount factor(γ)	0.99
Optimizer	Adam
Learning rate	$3e-4$
Entropy term coefficient	$1e-4$

Table 3: PPO Hyperparameters.

Hyperparameter	Value
Discount factor(γ)	0.99
Optimizer	Adam
Learning rate	$3e-4$
Clip value	0.2
Entropy term coefficient	0.005

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

Table 4: PTF Hyperparameters.

Hyperparameter	Value
Discount factor(γ)	0.99
Optimizer	Adam
Learning rate for the policy network	$3e-4$
Learning rate for the option network	$1e-3$
$f(t)$	$\frac{1+\tanh(3-0.001t)}{2}$
ϵ decrement	$1e-3$
ϵ-start	1.0
ϵ-end	0.05
Batch size	32
Number of episodes replacing the target network	1000