$$\begin{aligned} Q^{\pi}\left(s_t, a_t\right) =& x \\ & y \end{aligned} \tag{1}$$

$$\nabla \tag{2}$$

$$\begin{aligned} \nabla_{\theta^{\mu}} J &\approx \mathbb{E}_{s_t \sim S}\left[\nabla_{\theta^{\mu}} Q\left(s, a \mid \theta^{Q}\right) \mid_{s=s_t, a=\pi\left(s_t \mid \theta_t^{\mu}\right)}\right] \\ &= \mathbb{E}_{s_t \sim S}[\nabla_a Q\left(s, a \mid \theta^{Q}\right) \mid_{s=s_t, a=\pi\left(s_t \mid \theta_t^{\mu}\right)} \\ & \nabla_{\theta^{\mu}} \pi\left(s \mid \theta^{\mu}\right) \mid_{s=s_t}] \end{aligned} \tag{3}$$

$$L\left(\theta^{E}\right) = \mathbb{E}\left[\left(E\left(o_t \mid \theta_t^{E}\right) - a_t\right)^2\right] \tag{4}$$

1 introduction

aaaaaaaaaaaaaaaaaaaasssssssssssssssssssssssssssddddddddddddddddddaaaaaaaaaaaaaaaaaaaaaasssssssssssssssss
sssssssssddddddddddddddddddaaaaaaaaaaaaaaaaaaaaaasssssssssssssssssssssssssssssddddddddddddddddddaaaaaaaaaaa
aaaaaaasssssssssssssssssssssssssssssddddddddddddddddddaaaaaaaaaaaaaaaaaaaaaasssssssssssssssssssssssssssddddddddd
ddddaaaaaaaaaaaaaaaaaaaaaasssssssssssssssssssssssssssddddddddddddddddddaaaaaaaaaaaaaaaaaaaaaasssssssssssssssss
ssddddddddddddddddddaaaaaaaaaaaaaaaaaaaaaasssssssssssssssssssssssssssddddddddddddddddddaaaaaaaaaaaaaaaaa
sssssssssssssddddddddddddddddddaaaaaaaaaaaaaaaaaaaaaasssssssssssssssssssssssddddddddddddddddd

Algorithm 1 PACEE

(The encoder layers are included in actor network)

Randomly initialize critic network and actor network with weights θ_Q, $\theta_1^\mu, \theta_2^\mu, \cdots, \theta_n^\mu$

Randomly initialize experience network with weightsθ^E

Initialize target network $\theta^{Q'} \leftarrow \theta^Q, \theta_1^{\mu'} \leftarrow \theta_1^\mu, \theta_2^{\mu'} \leftarrow \theta_2^\mu, \cdots, \theta_n^{\mu'} \leftarrow \theta_n^\mu$

Randomly initialize experience network with weights θ^E

Initialize replay buffer $D_1, D_2, \cdots, D_n, D'$

For $episode = 1, 2, \cdots, M$ do:

  Initialize episode return $G = 0$

  Initialize an empty episode track Tr

  Receive initial state s_1

  For $t = 1, 2, \cdots, T$ do:

    Judge the stage of the time step n

    Select action and get observation:$a_t, o_t = \pi\left(s_t, \mid \theta_{t,n}^\mu\right)$

    Add positive guidance:$a_t = a_t + \left(E\left(o_t \mid \theta_{t,n}^E\right) - a_t\right)\xi$

    Execute action a_t and observe reward r_{t+1} and new state s_{t+1}

    Accumulate return $G = G + r_{t+1}$

    Store $(s_t, a_t, r_{t+1}, s_{t+1})$ and $(o_t, a_t, r_{t+1}, s_{t+1})$ in D_n and Tr respectively

    Sample a random minibatch of N transitions $(s_i, a_i, r_{i+1}, s_{i+1})$ from D_n

    Calculate:$q_{i+1} = Q\left(s_{i+1}, \pi(s_{i+1} \mid \theta_{t,n}^{\mu'}) \mid \theta_t^{Q'}\right)$

    Add positive guidance:$q_{i+1} = q_{i+1} + \left(Q\left(s_{i+1}, E\left(o_{i+1} \mid \theta_{t,n}^E\right)\theta_t^{Q'}\right) - q_{i+1}\right)\xi$

    Set:$y_i = r_{i+1} + \gamma q_{i+1}\phi$

    Calculate gradients wrt θ^Q and update critic network:

$$d\theta^Q \leftarrow \frac{1}{N}\nabla_{\theta^Q}\sum_i\left[\left(Q\left(s_i, a_i \mid \theta_t^Q\right) - y_i\right)^2\right]$$

    Calculate gradients wrt θ_n^μ and update actor network:

$$d\theta_n^\mu \leftarrow \frac{1}{N}\sum_i\left[\nabla_a Q\left(s, a \mid \theta_t^Q\right)\mid_{s=s_i, a=\pi\left(s_i \mid \theta_{t,n}^\mu\right)} \nabla_{\theta_n^\mu}\pi\left(s \mid \theta_{t,n}^\mu\right)\mid_{s=s_i}\right]$$

    Sample a random minibatch of N transitions $(o_k, a_k, r_{k+1}, s_{k+1})$ from D'

    Calculate gradients wrt θ^E and update experience network:

$$d\theta^E \leftarrow \frac{1}{N}\nabla_{\theta^E}\sum_k\left[\left(E\left(o_k \mid \theta_t^E\right) - a_k\right)^2\right]$$

    Update the target network:$\theta_{t+1}^{Q'} \leftarrow \tau\theta_t^Q + (1-\tau)\,\theta_t^{Q'}, \theta_{t+1,n}^{\mu'} \leftarrow \tau\theta_{t,n}^\mu + (1-\tau)\,\theta_{t,n}^{\mu'}$

  End for

  If $G \geq \bar{R}_K$,then store Tr intoD'

End for