Appendix

Proof of Theorem 1

Considering a stationary energy trading policy π of MG *i*, the state transition probability can be expressed as follows.

$$\varUpsilon_{oo'}^{a} = \varUpsilon(o_{i,t+1} = o' | o_{i,t} = o, a_{i,t} = a) \tag{1}$$

where o' is the next observation of o. The $V(o,\pi)$ and $Q(o,a)$ denote the state value and state-action value functions, respectively. At each training slot, following the strategy π and selecting action a under observation o, the expected cumulative discounted reward can be expressed as:

$$V(o,\pi) = \mathbb{E}_{\pi}\left[r(o,\pi(o)) + \gamma \cdot \sum_{o'} \varUpsilon_{oo'}^{a} \cdot V(o',\pi)\right] \tag{2}$$

where $r(o,\pi(o))$ represents the overall reward for selecting action $\pi(o)$ under observation o, γ is the discount factor. Then the maximum state value can be expressed as follows by decomposing Eq. (2) into the Bellman equation.

$$V^{*}(o,\pi) = \max_{a} \sum_{o'} \varUpsilon_{oo'}^{a} \cdot (r(o,\pi(o)) + \gamma \cdot V^{*}(o',\pi)) \tag{3}$$

Similar to the Eq. (3), the optimal value function of state-action pair can be denoted as:

$$Q^{*}(o,a) = \sum_{o'} \varUpsilon_{oo'}^{a} \cdot \left(r(o,a) + \gamma \cdot \max_{a'} Q^{*}(o',a')\right) \tag{4}$$

where o' denotes the next observation of o, a' denotes the trading action that performed under observation o'.

The state transition probability $\varUpsilon_{oo'}^{a}$ in Eq. (1) is stationary due to the state space and action space are limited. Considering the given (o_t, a_t, r_t, o_{t+1}), the updating rule of the target network is:

$$Q(o_t,a_t) = Q(o_t,a_t) + \alpha\left[r_t + \gamma \max_{a'} Q(o_{t+1},a') - Q(o_t,a_t)\right] \tag{5}$$

where o_{t+1} is the next observation of o_t, a' denotes the performing action under observation o_{t+1}, α is the learning rate and γ is the discount factor. Then subtract $Q^{*}(o_t,a_t)$ from both sides, which obtaining

$$\Delta(o_t,a_t) = Q(o_t,a_t) - Q^{*}(o_t,a_t) \tag{6}$$

thus yields

$$\Delta(o_t,a_t) = (1-\alpha)\Delta(o_t,a_t) + \alpha H(o_t,a_t) \tag{7}$$

$$H(o_t,a_t) = \left[r_t + \gamma \max_{a'} Q(o_{t+1},a') - Q^{*}(o_t,a_t)\right] \tag{8}$$

According to [1], given $0 \leqslant \alpha < 1$, then

$$\sum_{i=1}^{\infty} \alpha_{n^{i}(o,a)} = \infty, \ \sum_{i=1}^{\infty} [\alpha_{n^{i}(o,a)}]^{2} < \infty \tag{9}$$

where $n^i(o,a)$ denotes the index of the ith time that action a is performed under observation o. Therefore, the $\Delta(o_t,a_t) \to 0$ with probability 1 if :

(1) $\|\mathbb{E}[H(o_t,a_t)|H]\|_\infty \leqslant \gamma\|\Delta(o_t,a_t)\|_\infty$, with $\gamma < 1$.

(2) $\mathrm{var}[H(o_t,a_t)|H] \leqslant \xi(1+\|\Delta(o_t,a_t)\|_\infty^2)$, with $\xi > 0$.

First, we can derive the following equation.

$$\begin{aligned}\|\mathbb{E}[H(o_t,a_t)|H]\|_\infty &= \Upsilon_{oo'}^a H(o_t,a_t)\\ &\leqslant \gamma\|Q(o_t,a_t) - Q^*(o_t,a_t)\|_\infty \\ &= \gamma\|\Delta(o_t,a_t)\|_\infty\end{aligned} \tag{10}$$

Then, the following can be obtained.

$$\mathrm{var}[H(o_t,a_t)\,|H] = \mathrm{var}\left[r_t + \gamma\max_{a'} Q(o_{t+1},a')\,|H\right] \tag{11}$$

The following is true, due to the r_t is bounded.

$$\mathrm{var}[H(o_t,a_t)\,|H] \leqslant \xi\left(1+\|\Delta(o_t,a_t)\|_\infty^2\right) \tag{12}$$

where ξ is a positive constant. Thus, $\Delta(o_t,a_t) \to 0$ with probability 1, which means the target network of h-MADQN algorithm converges to $Q^*(o_t,a_t)$.

References

[1] C. J. Watkins and P. Dayan, “Q-learning,” Machine learning, vol. 8, no. 3-4, pp. 279–292, 1992.