Supplementary materials

Knowledge-assistant Deep Reinforcement Learning for Multi-agent Region Protection

Since the rewards are newly designed, the convergence of the critical neural network needs to be proven. Once the error of the critical neural network converges, the Bellman Optimality Equation can be applied, ensuring the feasibility and convergence of the solution to Problem 1.
Specifically, to approach the cumulative reward R_i^t, the goal of the critic network Q_i^π is to minimize the sum of square loss $L(\varepsilon_i)$, between the critic network output $\widehat{Q}_i$ and R_i^t, i.e., $L(\varepsilon_i) = \min_{\varepsilon_i} \mathbb{E}[(\widehat{Q}_i - R_i)^2]$. Thus, the gradient of critic networks $\nabla_{\varepsilon_i} L(\varepsilon_i)$ is calculated by

$$\nabla_{\varepsilon_i} L(\varepsilon_i) = \mathbb{E}[(Q_i^\pi(s, a_{1,\dots,Ni}) - y) \cdot \nabla_{\varepsilon_i} Q_i^\pi(s, a_{1,\dots,Ni})], \tag{1}$$

where $y = r_i^t + \gamma Q_i^\pi(s^{t+1}, a_{1,\dots,N}^{t+1}|_{a_i^{t+1}=\pi_i(o_i)})$, which is an approximate value calculated by the temporal difference (TD).

Theorem : For an the learning process of critic networks by $\varepsilon_i^{k+1} = \varepsilon_i^k - \alpha_c \nabla_{\varepsilon_i} L(\varepsilon_i)$, where α_c is the critic networks' learning rate. Assuming that the time interval Δt is constant, the parameters ε_i^k of each defender asymptotically converges to the optimum value $\varepsilon_i^* = \underset{\epsilon_i}{\operatorname{argmin}} E[(\widehat{Q}_i - R_i)^2]$ with iteration index $k \to \infty$, if the learning rate α_c satisfies

$$0 < \alpha_c < 2. \tag{2}$$

In other words, the error of the critical neural network can converge.

Proof. For defender i, denote the critic network output $\widehat{Q}_i^k(x)$ by $\widehat{Q}_i^k(x) = (\varepsilon_i^k)^\top \phi(x)$, where x is the critic network input, including the states s and the actions a of all the defenders $\mathcal{V}_d$, and

$$\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3}$$

is the hyperbolic activation function of critic networks. For conciseness, we omit the symbol '(x)' in the following context. The error e_i^k between the reward prediction $\widehat{Q}_i^k$ and the real reward R_i writes $e_i^k = \widehat{Q}^k - R_i, L_i^k = \frac{1}{2}\left(e_i^k\right)^2$, where L_i^k is the square loss for obtaining gradients. To minimize L_i^k, ε_i^k is updated by

$$\varepsilon_i^{k+1} = \varepsilon_i^k - \alpha_c \partial L_i^k / \partial \varepsilon_i^k = \varepsilon_i^k - \alpha_c e_i^k, \tag{4}$$

where α_c is the learning rate of critic networks, and $\alpha_c \geq 0$. then the error of the parameters $\overline{\varepsilon}_i^k$ can be expressed as $\overline{\varepsilon}_i^k = \varepsilon_i^k - \varepsilon_i^*$. Thus, one has $\overline{\varepsilon}_i^{k+1} = \overline{\varepsilon}_i^k - \alpha_c e_i^k \phi$. Now, design a discrete-time Lyapunov candidate as $V(k) = (\overline{\varepsilon}_i^k)^\top \overline{\varepsilon}_i^k$, whose difference writes:

$$\begin{aligned}
\Delta V(k) &= V(k+1) - V(k) \\
&= (\overline{\varepsilon}_i^{k+1})^\top \overline{\varepsilon}_i^{k+1} - (\overline{\varepsilon}_i^k)^\top \overline{\varepsilon}_i^k \\
&= [(\overline{\varepsilon}_i^k)^\top - \alpha_c \phi^\top (e_i^k)^\top](\overline{\varepsilon}_i^k - \alpha_c e_i^k \phi) - (\overline{\varepsilon}_i^k)^\top (\overline{\varepsilon}_i^k) \\
&= -\alpha_c \phi^\top (e_i^k)^\top \overline{\varepsilon}_i^k - \alpha_c (\overline{\varepsilon}_i^k)^\top e_i^k \phi + \alpha_c^2 (e_i^k \phi)^\top e_i^k \phi \\
&= -\alpha_c \phi^\top (e_i^k)^\top (\varepsilon_i^k - \varepsilon_i^*) - \alpha_c (\varepsilon_i^k - \varepsilon_i^*)^\top e_i^k \phi \\
&\quad + \alpha_c^2 \|e_i^k\|^2 \|\phi\|^2 \\
&= \alpha_c \|e_i^k\|^2 \left(-2 + \alpha_c \|\phi\|^2\right).
\end{aligned} \tag{5}$$

According to (2) and (3), one has $-2 + \alpha_c \|\phi\|^2 < 0$, and hence $\Delta V < 0$, implying that ε_i^k asymptotically converges to ε_i^*. The error of the critical neural network can converge to 0, which completes the proof. ■