Problem statement

When the target is in the first frame ($t = 1$), the initial target state is S_1. After doing action a_1, the reward r_1 is obtained and the target state changes to S_2. Repeated the above process until the last frame T of the sequences, the whole process can be expressed as $\tau = \{S_1, a_1, r_1, S_2, a_2, r_2, S_2, \cdots, S_T\}$. The final reward can be calculated as :

$$R(\tau) = \sum_{t=1}^{T} r_t \tag{1}$$

We use a network $\pi(\theta)$ as the actor to do action and generate new target state, and θ is the learnable parameters of the network. Let the probability of the possible τ be $P(\tau|\theta)$, the expected value of final reward $R(\tau)$ can be expressed as:

$$\overline{R_\theta} = \sum_{\tau} R(\tau)P(\tau|\theta) \tag{2}$$

For reinforcement learning, we want to maximize the final reward $\overline{R_\theta}$:

$$\theta^* = \arg\max_{\theta} \overline{R_\theta} \tag{3}$$

Gradient ascent

- Start with θ^0
- $\theta^1 \leftarrow \theta^0 + \eta\nabla\overline{R_{\theta^0}}$
- $\theta^2 \leftarrow \theta^1 + \eta\nabla\overline{R_{\theta^1}}$
- ......

Here, η is the learning rate.

Solving

Solving for $\nabla\overline{R_\theta}$:

$$\begin{aligned} \nabla\overline{R_\theta} &= \sum_{\tau} R(\tau)\nabla P(\tau|\theta) \\ &= \sum_{\tau} R(\tau)P(\tau|\theta)\nabla \log P(\tau|\theta) \end{aligned} \tag{4}$$

According to the Law of Large Numers, Eq. (4) can be expressed as:

$$\nabla\overline{R_\theta} \approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla \log P(\tau^n|\theta) \tag{5}$$

Here, n is the number of τ which are input to the network for training, and N is the total number of n. According to the Markov Assumption, $P(\tau|\theta)$ in can be expressed as:

$$\begin{aligned} P(\tau|\theta) &= P(S_1)P(a_1|S_1,\theta)P(r_1,S_2|S_1,a_1)\cdots \\ &= P(S_1)\prod_{t=1}^{T} P(a_t|S_t,\theta)P(r_t,S_{t+1}|S_t,a_t) \end{aligned} \tag{6}$$

Taken logarithm to Eq. (6), we can obtain:

$$\begin{aligned} \log P(\tau|\theta) &= \log P(S_1) + \sum_{t=1}^{T} \log P(a_t|S_t,\theta) + \log P(r_t,S_{t+1}|S_t,a_t) \\ \nabla \log P(\tau|\theta) &= \sum_{t=1}^{T} \nabla \log P(a_t|S_t,\theta) \end{aligned} \tag{7}$$

Putting Eq. (7) into Eq. (5), we can obtain $\nabla\overline{R_\theta}$ for each n:

$$\nabla\overline{R_\theta} \approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n)\nabla \log P(a_t^n|S_t^n,\theta) \tag{8}$$