Reinforcement Learning Formulation and Training for the Advanced Network Environment

April 2025

Contents

1 MDP Formulation

We model the problem as a Markov Decision Process (MDP) defined by the tuple

$$(\mathcal{S}, \mathcal{A}, T, R, \gamma),$$

where:

$\mathcal{S}$: The state space,

$\mathcal{A}$: The action space,

T: The state transition dynamics,

R: The reward function,

γ: The discount factor.

2 Environment and State Space

2.1 Environment Details

Common ports:

$$\mathcal{P}_{\text{common}} = \{80, 443, 8080, 22, 53\}.$$

Suspicious ports:

$$\mathcal{P}_{\text{suspicious}} = \{4444, 31337, 6667\}.$$

The agent's suspicion level is capped (e.g., at 100) and an episode terminates either upon detection or after a maximum of $T_{\max}$ steps.

2.2 State Space $\mathcal{S}$

At time step t, the state $s_t \in \mathcal{S}$ is represented as a 16-dimensional vector:

$$s_t = \begin{bmatrix} \dfrac{\text{Suspicion Level}_t}{\text{Max Suspicion}} \\ \dfrac{\text{Current Port}_t}{65535} \\ \text{Packet Mean}_t \quad (\text{normalized by } 1500) \\ \text{Packet Max}_t \quad (\text{normalized by } 1500) \\ \text{Fraction of Large Packets}_t \\ \text{Normalized Unique Port Count}_t \; = \; \frac{|\text{Unique Ports in History}|}{\text{History Length}} \\ \text{Time Normalization}_t \; = \; \min\left(\frac{t}{T_{\max}}, 1\right) \\ \text{One-hot action}_{t,1} \\ \text{One-hot action}_{t,2} \\ \text{One-hot action}_{t,3} \\ \text{One-hot action}_{t,4} \\ \text{One-hot action}_{t,5} \\ \dfrac{p_{t-3}}{65535} \\ \dfrac{p_{t-2}}{65535} \\ \dfrac{p_{t-1}}{65535} \\ \dfrac{p_t}{65535} \end{bmatrix} \in \mathbb{R}^{16}.$$

Component Descriptions:

- Suspicion Level_t quantifies the proximity to detection.
- Current Port_t denotes the port currently used by the agent.
- Packet Mean_t and Packet Max_t are normalized by a maximum packet size (e.g., 1500).
- Fraction of Large Packets_t is the ratio of packets in the recent history that exceed a size threshold (e.g., 1200 bytes).
- Normalized Unique Port Count_t represents the diversity of port usage.
- Time Normalization_t indicates the relative progress within the episode.
- The one-hot encoded actions represent the last executed action (from 5 possible actions).
- p_{t-k} for $k = 0, 1, 2, 3$ are the most recent port values (normalized by 65535).

2.3 Action Space $\mathcal{A}$

The discrete action space is defined as:

$$\mathcal{A} = \{0, 1, 2, 3, 4\},$$

with the following interpretations:

$$
\begin{aligned}
&0: \text{ Send small packet (size} = 200), \\
&1: \text{ Send large packet (size} = 1500), \\
&2: \text{ Delay (record packet size as 0)}, \\
&3: \text{ Change port (choose new port from } \mathcal{P}_{\text{common}} \setminus \{\text{Current Port}\}), \\
&4: \text{ Stealth combo (a sequence comprising delay, small packet, and port change).}
\end{aligned}
$$

3 Transition Dynamics and Reward Function

3.1 Transition Dynamics T

The transition function $T(s_{t+1} \mid s_t, a_t)$ is defined by the following updates:

1. **Packet History Update:**

$$\text{packet\_history}_{t+1} = \text{Append}\Big(\text{packet\_history}_t,\ f_{\text{pkt}}(a_t)\Big),$$

   where

$$f_{\text{pkt}}(a_t) = \begin{cases} 200, & a_t = 0, \\ 1500, & a_t = 1, \\ 0, & a_t = 2, \\ (\text{a combination for } a_t = 4), & \end{cases}$$

2. **Port History Update:** When $a_t \in \{3, 4\}$,

$$\text{port\_history}_{t+1} = \text{Append}\Big(\text{port\_history}_t,\ p_{\text{new}}\Big),$$

   where

$$p_{\text{new}} \in \{p \in \mathcal{P}_{\text{common}} : p \neq \text{Current Port}_t\}$$

   is chosen uniformly at random.

3. **Action History Update:** The current action a_t is appended to a fixed-length action history.

4. **Detection Mechanism:** The environment evaluates detection by checking:

   a. **Packet Size Check:** If in the last N_p packets (e.g., $N_p = 10$), at least k_p (e.g., $k_p = 4$) packets exceed a threshold $S_{\text{th}} = 1200$:

$$\sum_{i=1}^{N_p} \mathbb{1}_{\{\text{packet}_i > 1200\}} \geq k_p,$$

   then detection is triggered.

   b. **Port Scan Check:** If the number of unique ports in the last N_{port} entries (e.g., $N_{\text{port}} = 5$) exceeds

$$\text{port\_scan\_threshold} + 2,$$

   detection is triggered.

c. **Suspicious Port Check:** If Current $\text{Port}_t \in \mathcal{P}_{\text{suspicious}}$, then with probability

$$\delta_t = \min\Bigl\{\text{base\_detection\_probability} + 0.1\Bigl(\frac{\text{Episode Count}}{10}\Bigr), 1\Bigr\},$$

detection is triggered.

Define the detection indicator d_t as:

$$d_t = \begin{cases} 1, & \text{if any detection condition is met,} \\ 0, & \text{otherwise.} \end{cases}$$

5. **Termination:** The episode terminates if $d_t = 1$ or when $t = T_{\max}$.

3.2 Reward Function R

The immediate reward is defined as:

$$R(s_t, a_t) = \begin{cases} -100 - 0.5\,(T_{\max} - t), & \text{if } d_t = 1, \\ r_{\text{survival}}(s_t, a_t), & \text{if } d_t = 0, \end{cases}$$

where the survival reward is given by

$$r_{\text{survival}}(s_t, a_t) = 0.2 + 0.5\,\mathbb{1}_{\{\text{Current Port} \in \mathcal{P}_{\text{common}}\}} - 0.5\,\mathbb{1}_{\{a_t = 3\}} + 0.2\,\mathbb{1}_{\{\text{diverse action history}\}},$$

with an additional bonus of 10 if the agent reaches $t = T_{\max}$ without detection:

$$\text{if } t = T_{\max} \text{ and } d_t = 0, \quad r(s_t, a_t) \leftarrow r(s_t, a_t) + 10.$$

4 Reinforcement Learning Objective and PPO Formulation

4.1 Policy and Value Function

The agent learns a stochastic policy

$$\pi(a \mid s) : \mathcal{S} \to \Delta(\mathcal{A}),$$

parameterized by a neural network. In addition, a value function $V(s)$ approximates the expected return from state s.

4.2 Objective

The learning objective is to maximize the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{T_{\max}} \gamma^t\, R(s_t, a_t)\right],$$

with $\gamma \in [0, 1)$ denoting the discount factor.

4.3 Advantage Estimation

Using Generalized Advantage Estimation (GAE) [?], the advantage function is computed as:

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l},$$

where

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t),$$

and $\lambda \in [0, 1]$ is the GAE parameter.

4.4 PPO Surrogate Objective

The Proximal Policy Optimization (PPO) algorithm optimizes the following clipped surrogate objective:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)\hat{A}_t,\ \text{clip}\big(r_t(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\hat{A}_t\Big)\Big],$$

where

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

is the probability ratio, and ϵ is a hyperparameter (e.g., 0.2).

4.5 Total Loss Function

The complete loss combines the policy loss, value function loss, and an entropy bonus:

$$L(\theta) = L^{\text{CLIP}}(\theta) - c_1\, L^{\text{VF}}(\theta) + c_2\, S[\pi_\theta],$$

with:

- $L^{\text{VF}}(\theta) = \mathbb{E}_t\Big[\big(V_\theta(s_t) - R_t^{\text{target}}\big)^2\Big]$, the value function loss,
- $S[\pi_\theta]$ is the entropy bonus,
- c_1 and c_2 are coefficients balancing the losses (e.g., $c_1 = 0.7$, $c_2 = 0.02$).

5 PPO-based Training Algorithm

The training process follows the PPO framework as detailed in the pseudo-code below.

Algorithm 1 PPO Training for the Advanced Network Environment

Input: Total timesteps T_{total}, update frequency K, clip parameter ϵ
Initialize policy parameters θ and value function parameters
Initialize environment and corresponding histories
for $t = 0, 1, \ldots, T_{\text{total}} - 1$ **do**
  Observe current state s_t
  Sample action $a_t \sim \pi_\theta(\cdot \mid s_t)$
  Execute action a_t in the environment
  Observe reward $r_t = R(s_t, a_t)$, next state s_{t+1}, detection flag d_t
  Store transition $(s_t, a_t, r_t, s_{t+1}, d_t)$
  if episode terminates (i.e., $d_t = 1$ or $t = T_{\max}$) **then**
    Compute returns and advantages $\{\hat{A}_t\}$ using GAE
    for epoch = 1 to K **do**
    Update policy using the PPO surrogate loss:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)\hat{A}_t, \text{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_t\Big)\Big]$$

    Update value function by minimizing:

$$L^{\text{VF}}(\theta) = \mathbb{E}_t\Big[\big(V_\theta(s_t) - R_t^{\text{target}}\big)^2\Big]$$

    Incorporate an entropy bonus $S[\pi_\theta]$ to encourage exploration.

    Reset episode-specific histories and update environment (curriculum adjustments, etc.)
  end if
end for
Save the trained model parameters θ

6 Summary of Learning Dynamics

At each iteration:

1. The agent observes s_t and selects an action a_t according to the policy $\pi_\theta(a_t \mid s_t)$.
2. The environment applies the action, updates histories (packet, port, and action histories), and transitions to a new state s_{t+1}.
3. The immediate reward r_t is computed, and the detection mechanism evaluates the current state.
4. When an episode terminates, GAE computes the advantages, and PPO performs multiple epochs of updates on the policy and value function using the clipped surrogate objective.
5. The procedure continues until a predefined total timestep limit is reached.