First-Visit Monte Carlo On-Policy Evaluation

Theoretical Explanation

The goal of Monte Carlo (MC) policy evaluation is to estimate the value function:

$$V^\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s],$$

where G_t is the return from time t onward, following policy π.

Algorithm (First-Visit MC)

1. Initialize, for all states s:
$$N(s) = 0, \quad G(s) = 0$$
where $N(s)$ is a counter of visits and $G(s)$ accumulates total returns.

2. For each episode i:

   (a) Generate an episode:
   $$s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \ldots, s_{i,T}, r_{i,T}$$

   (b) For each time step t, define the return:
   $$G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots + \gamma^{T-t} r_{i,T}.$$

   (c) For each state s_t in the episode:
   - If s_t is the **first visit** to s in the episode:
   $$N(s) \leftarrow N(s) + 1, \quad G(s) \leftarrow G(s) + G_{i,t}$$
   - Update the estimate:
   $$V^\pi(s) = \frac{G(s)}{N(s)}.$$

This ensures that $V^\pi(s)$ converges to the true value function by averaging returns across many first visits of s.

—

Numerical Example

Let $\gamma = 0.9$. Consider two episodes:

Episode 1

Trajectory:

$$s_1 \xrightarrow{r=2} s_2 \xrightarrow{r=-1} s_3 \xrightarrow{r=3} \text{terminal}$$

Returns:

$$G(s_1) = 2 + 0.9(-1) + 0.9^2 \cdot 3 = 3.53$$

$$G(s_2) = -1 + 0.9 \cdot 3 = 1.7$$

$$G(s_3) = 3$$

Accumulators after Episode 1:

$$G(s_1) = 3.53, \quad N(s_1) = 1$$

$$G(s_2) = 1.70, \quad N(s_2) = 1$$

$$G(s_3) = 3.00, \quad N(s_3) = 1$$

Episode 2

Trajectory:

$$s_2 \xrightarrow{r=0} s_1 \xrightarrow{r=4} s_2 \xrightarrow{r=-2} s_3 \xrightarrow{r=1} \text{terminal}$$

First visits only:

$$G(s_2) = 0 + 0.9 \cdot 4 + 0.9^2(-2) + 0.9^3 \cdot 1 = 2.709$$

$$G(s_1) = 4 + 0.9(-2) + 0.9^2 \cdot 1 = 3.01$$

$$G(s_3) = 1$$

Accumulators after Episode 2:

$$G(s_1) = 3.53 + 3.01 = 6.54, \quad N(s_1) = 2$$

$$G(s_2) = 1.70 + 2.709 = 4.409, \quad N(s_2) = 2$$

$$G(s_3) = 3.00 + 1.00 = 4.00, \quad N(s_3) = 2$$

Final Estimates

$$V^\pi(s_1) = \frac{6.54}{2} = 3.27, \quad V^\pi(s_2) = \frac{4.409}{2} \approx 2.20, \quad V^\pi(s_3) = \frac{4.00}{2} = 2.00$$

Thus, the First-Visit Monte Carlo method produces estimates of the state values by averaging the returns observed from their first occurrences across episodes.