Monte Carlo Methods

estimate value functions

discover optimal policies

Monte Carlo : experience sample episodes

sample sequences of states, actions, and rewards from actual or simulated interaction with an environment

actual experience

simulated experience

"Average sample returns"

Experiences → episodes

DP : compute value functions

MC : learn value functions

"Simply average the returns observed after visits to that state"

↓

"As more returns are observed, the average should converge to the expected value."

A set of episodes

↑ visit

state s

first-visit MC method

estimate $V_\pi(s)$ as the average of the returns

every-visit MC method

average all the returns following first visit to s

$S_0, A_0, R_1, S_1, A_1, R_2, \cdots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$t = T-1, T-2, \cdots, 0$

$G \leftarrow \gamma G + R_{t+1}$

Unless S_t in $S_0, S_1, \cdots, S_{t-1}$

$$\begin{cases} \text{Append } G \text{ to } Return(S_t) \\ V(S_t) \leftarrow \text{average}(Returns(S_t)) \end{cases}$$

$n \uparrow \rightarrow$ error $\downarrow$ $\frac{1}{\sqrt{n}}$

MC : estimates for each state are independent.

DO NOT bootstrap

① learn from actual experience

② learn from simulated experience

③ independent.

Maintaining exploration (Sufficient exploration)

exploring starts: episodes begin with state-action pairs randomly selected to cover all possibilities

action-value function q

$$\pi(s) \doteq \arg\max_a q(s,a)$$

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a))$$

$$= \max_a q_{\pi_k}(s,a)$$

$$\geq q_{\pi_k}(s, \pi_k(s))$$

$$\geq v_{\pi_k}(s)$$

every pair has a non-zero probability of being selected as the start

MC Control with Exploring Start:

episode $t = T-1, T-2, \ldots, 0$

$$G \leftarrow \gamma G + R_{t+1}$$

if S_t, A_t in $S_0, A_0, S_1, A_1, \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$$Q(S_t, A_t) \leftarrow average(Returns(S_t, A_t))$$

$$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$$

MC Control without Exploring Starts

On-policy: evaluate or improve the policy that is used to make decisions

off-policy: evaluate or improve the policy different from that used to generate data (may be unrelated to the policy followed)

① On-policy Control:

policy "soft" $\pi(a|s) > 0$

ε-greedy policies

ε probability select an action at random

$$\begin{cases} \dfrac{\varepsilon}{|A(s)|} & \text{probability non-greedy action} \\ 1 - \varepsilon + \dfrac{\varepsilon}{|A(s)|} & \text{greedy action} \end{cases}$$

Episode: $t = T-1, T-2, \ldots, 0$

$$G \leftarrow \gamma G + R_{t+1}$$

if S_t, A_t in $S_0, A_0, S_1, A_1, \ldots, S_{t-1}, A_{t-1}$

Append G to $Returns(S_t, A_t)$

$$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$$

$$A^* \leftarrow \underset{a}{\text{argmax}}\ Q(S_t, a)$$

For all $a \in \mathcal{A}(S_t)$:

$$\pi(a \mid S_t) \leftarrow \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{|\mathcal{A}(S_t)|} & \text{if } a = A^* \\ \dfrac{\varepsilon}{|\mathcal{A}(S_t)|} & \text{if } a \neq A^* \end{cases}$$

Problem: sufficient exploration

(maintaining exploration)

with exploring start

without exploring start

- on-policy $\pi(a \mid s) > 0 \longrightarrow$ ~~deterministic~~ optimal policy
- off-policy

Monte Carlo Control without Exploring Starts

$$q_\pi(s, \pi'(s)) = \sum_a \pi'(a|s)\, q_\pi(s, a)$$

$$= \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \max_a q_\pi(s,a)$$

$$\geq \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \sum_a \frac{\pi(a|s) - \frac{\varepsilon}{|A(s)|}}{1-\varepsilon}\, q_\pi(s,a)$$

$$= \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) - \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + \sum_a \pi(a|s)\, q_\pi(s,a)$$

$$= V_\pi(s)$$

ε-soft : $\pi(a|s) \geq \frac{\varepsilon}{|A(s)|}$

On-policy method : use ε-greedy policy

Select an action at random ε probability

$$\begin{cases} \dfrac{\varepsilon}{|A(s)|} & \text{for non-greedy actions} \\[2ex] 1 - \varepsilon + \dfrac{\varepsilon}{|A(s)|} & \text{for greedy actions} \end{cases}$$

Off-policy Prediction

On-policy method : learn action values not for the optimal policy, but for a near-optimal policy that still explores

target policy : learned policy → optimal policy

behavior policy : generate policy → more exploratory

Estimate V_π, q_π, b $(b \neq \pi)$

b: behavior policy

π: target policy

Converage $\pi(a|s) > 0$, $b(a|s) > 0$

Importance Sampling (ratio)

weight returns according to the relative probability of their trajectories occurring under the target and behavior policies.

State-action trajectory:

$$A_t, S_{t+1}, A_{t+1}, \cdots, S_T$$

$$\Pr\{A_t, S_{t+1}, A_{t+1}, \cdots, S_T \mid S_t, A_{t:T-1} \sim \pi\}$$

$$= \pi(A_t \mid S_t)\, p(S_{t+1} \mid S_t, A_t)\, \pi(A_{t+1} \mid S_{t+1}) \cdots$$

$$p(S_T \mid S_{T-1}, A_{T-1})$$

$$= \prod_{k=t}^{T-1} \pi(A_k \mid S_k)\, p(S_{k+1} \mid S_k, A_k)$$

$$\rho_{t:T-1} \doteq \frac{\prod_{k=t}^{T-1} \pi(A_k \mid S_k)\, p(S_{k+1} \mid S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k \mid S_k)\, p(S_{k+1} \mid S_k, A_k)}$$

$$= \prod_{k=t}^{T-1} \frac{\pi(A_k \mid S_k)}{b(A_k, S_k)}$$

$$\mathbb{E}[\rho_{t:T-1} G_t \mid S_t = s] = v_\pi(s)$$

Scale the return by the ratios and average the returns:

$$V(s) \doteq \frac{\sum_{t \in \mathcal{J}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{J}(s)|}$$

ordinary importance sampling

$$V(s) \doteq \frac{\sum_{t \in \mathcal{J}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{J}(s)} \rho_{t:T(t)-1}}$$

weighted average

(preferred)

Variance of the importance sampling-scaled returns:

$$\mathrm{Var}[X] \doteq \mathbb{E}[(X - \bar{X})^2]$$

$$= \mathbb{E}[X^2 - 2X\bar{X} + \bar{X}^2]$$

$$= \mathbb{E}[X^2] - \bar{X}^2$$

Expected square of the improtance-sampling-scaled return:

$$\mathbb{E}_b\left[\left(\prod_{t=0}^{T-1}\frac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}G_0\right)^2\right]$$