Chapter 5

- Monte Carlo methods can be used to estimate state values through experience by taking the average of rewards that follow each visit to a particular state.

- Monte Carlo methods can be used to compute the value function in cases that would be difficult to apply dynamic programming methods.

- Monte Carlo methods do not bootstrap, meaning the estimate for one state does not build upon the estimate of any other state.

- Estimate action values instead of state values when no model is available.

- Approximate optimal policies can be computed using policy iteration:

$$\pi_0 \xrightarrow{E} q_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} q_{\pi_1} \xrightarrow{I} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} q_{\pi_*}$$

evaluate improve

- Greedy policy for action-value function:

(5.1) $$\pi(s) = \underset{a}{\operatorname{argmax}}\; q(s,a)$$

- Policy improvement

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \underset{a}{\operatorname{argmax}}\; q_{\pi_k}(s,a))$$
$$= \max_a q_{\pi_k}(s,a)$$
$$\geq q_{\pi_k}(s, \pi_k(s))$$
$$= v_{\pi_k}(s)$$

$$\Downarrow$$

$$\pi'(s) = \underset{a}{\operatorname{argmax}}\; q_\pi(s,a)$$

- Exploring starts - Episodes start in a randomly chosen state-action pair. Each pair has a nonzero probability of being selected as the start. This guarantees that all state-action pairs are visited as the number of episodes approaches infinity.

- Without exploring starts, there are two approaches to ensure that all actions are explored:

  - On-policy methods
  - off-policy methods

- On-policy methods are generally "soft", meaning:

$$\pi(a|s) > 0 \quad \forall \quad s \in S, \; a \in A(s)$$

- An ε-greedy policy is an ε-soft policy such that:

$$\pi(a|s) \geq \frac{\varepsilon}{|A(s)|} \quad \forall \quad s \in S, \; a \in A(s)$$

$\frac{\varepsilon}{|A(s)|}$ is the minimum probability

$1 - \varepsilon + \frac{\varepsilon}{|A(s)|}$ is the remaining probability given to the greedy action

↳ See solution to Exercise 4.7

- ε-greedy policy improvement

When π is greedy: $\pi'(s) = \underset{a}{\operatorname{argmax}}\ q_\pi(s,a)$

When π is ε-greedy:

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

$$= \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + \left[\left(1-\varepsilon+\frac{\varepsilon}{|A(s)|}\right) - \frac{\varepsilon}{|A(s)|}\right] q_\pi(s,\pi'(s))$$

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

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

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

$$\geq q_\pi(s, \pi(s))$$

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

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

(cont.)

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

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

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

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

$$= v_\pi(s)$$

Note that $\sum_a \frac{\pi(a|s) - \frac{\varepsilon}{|A(s)|}}{(1-\varepsilon)}$ works out to be 1 for the greedy part and 0 for the nongreedy part.

Greedy part:

$$\frac{\left(1-\varepsilon+\frac{\varepsilon}{|A(s)|}\right) - \frac{\varepsilon}{|A(s)|}}{(1-\varepsilon)} = \frac{1-\varepsilon}{1-\varepsilon} = 1$$

Nongreedy part:

$$\frac{\frac{\varepsilon}{|A(s)|} - \frac{\varepsilon}{|A(s)|}}{(1-\varepsilon)} = \frac{0}{1-\varepsilon} = 0$$

(cont.)

Thus: $q_\pi(s, \pi'(s)) \geq v_\pi(s)$

- Let $\tilde{v}_*$ and $\tilde{q}_*$ denote optimal value functions for an environment that uses ε-soft policies.

$$\tilde{v}_*(s) = (1-\varepsilon)\max_a \tilde{q}_*(s,a) + \frac{\varepsilon}{|A(s)|}\sum_a \tilde{q}_*(s,a)$$

$$= (1-\varepsilon)\max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma \tilde{v}_*(s')\right] + \frac{\varepsilon}{|A(s)|}\sum_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma \tilde{v}_*(s')\right]$$

$\tilde{v}_*(s) = v_\pi(s)$ when π is no longer improved

- Policy iteration works for ε-soft policies.

- Off-policy methods estimate the target policy based on episodes generated from a different policy.

π is the target policy
μ is the behavior policy

- The assumption of coverage:

$\pi(a|s) > 0$ implies $\mu(a|s) > 0$

- Importance sampling ratio:

(5.3) $$\rho_t^T = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

↳ Weighting returns according to relative probability under target and behavior policies.

- Use policy μ to estimate $v_\pi(s)$:

$\mathcal{T}(s)$ is the set of all time steps in which state s is visited

$T(t)$ is the first time of termination following time t.

G_t is the return following t up to $T(t)$

(cont.)

$\{G_t\}_{t\in\mathcal{T}(s)}$ is the set of returns that pertain to state s.

$\{\rho_t^{T(t)}\}_{t\in\mathcal{T}(s)}$ is the set of corresponding importance sampling ratios that pertain to state s.

- Ordinary importance sampling:

$$V(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)} G_t}{|\mathcal{T}(s)|}$$

↳ estimation of $v_\pi(s)$ by scaling the returns by the ratios and averaging the results.

- Weighted importance sampling:

$$V(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)} G_t}{\sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)}}, \quad V(s) = 0 \text{ if } \sum_{t\in\mathcal{T}(s)} \rho_t^{T(t)} = 0$$

↳ estimation of $v_\pi(s)$ using a weighted average.

- Incremental implementation

weight $W_i = \rho_t^{T(t)}$

(5.6) $$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k} \qquad \text{where } n \geq 2$$

$$C_n = \sum_{k=1}^{n-1} W_k + W_n$$

$$V_{n+1} = \frac{V_n(C_n - W_n) + W_n G_n}{C_n}$$

$$= V_n + \frac{W_n G_n - W_n V_n}{C_n}$$

(5.7) $$= V_n + \frac{W_n (G_n - V_n)}{C_n} \qquad n \geq 1$$

$$C_{n+1} = C_n + W_{n+1} \qquad C_0 = 0$$

↳ off-policy weighted importance sampling

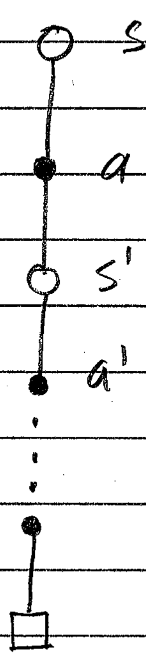
Exercise 5.1

When the player stands on a hand of 20 or 21, there is a very high probability of winning.

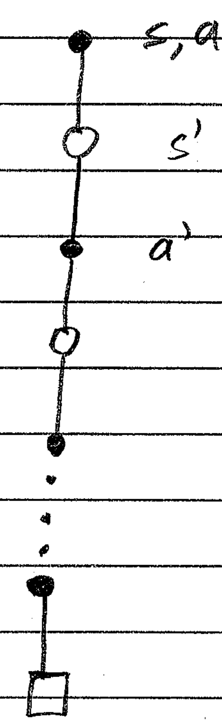
The chance of winning is not as high if the dealer has an ace.

If the player has a soft hand (i.e. a usable ace), then the player has a higher chance of winning.

Exercise 5.2



V_π



q_π

Chapter 5

Exercise 5.3 / 5.4

$V(s)$ is the weighted average of the returns for all episodes starting in state s and following policy μ.

$Q(s,a)$ is the weighted average of the returns for all episodes starting in state s, taking action a, and following policy μ thereafter.

$$V(s) = \frac{\sum_{t \in \tau(s)} \rho_t^{T(t)} G_t}{\sum_{t \in \tau(s)} \rho_t^{T(t)}}$$

$$\rho_t^{T(t)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

For $Q(s,a)$, the first ratio is always $\frac{\pi(a|s)}{\mu(a|s)}$

$$\therefore \rho_t^{T(t)} = \frac{\pi(a|s)}{\mu(a|s)} \prod_{k=t+1}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} = \frac{\pi(a|s)}{\mu(a|s)} \rho_{t+1}^{T(t)}$$

$$Q(s,a) = \frac{\frac{\pi(a|s)}{\mu(a|s)} \sum_{t \in \tau(s)} \rho_{t+1}^{T(t)} G_t}{\frac{\pi(a|s)}{\mu(a|s)} \sum_{t \in \tau(s)} \rho_{t+1}^{T(t)}} = \frac{\sum_{t \in \tau(s)} \rho_{t+1}^{T(t)} G_t}{\sum_{t \in \tau(s)} \rho_{t+1}^{T(t)}}$$

Exercise 5.5

The returns are not counted for all episodes. In episodes where the stochastic policy $\mu(a|s)$ chooses a different action than would have been chosen by the deterministic policy $\pi(s)$, the ratio $\rho_t^{T(t)}$ is zero because $\pi(a|s) = 0$ in one of the states visited in the episode. It might take several episodes on average before the returns start to count.

If $\rho_t^{T(t)} = 0$ on the first episode, the error is going to be $(0 - (-0.27726))^2 = 0.07687$.

In this example, only about 14% of the episodes are counted; $\rho_t^{T(t)} = 0$ about 86% of the time. It might take 7 or 8 episodes before most runs have counted at least one return.

The error increases at first because the first few returns that are counted do not represent a very broad sample of the returns. As more episodes are counted, a broader sample of returns is included in the average and the error decreases.

Chapter 5

Exercise 5.6

There is no reward until the final transition to the end state.

There is only one non-terminal state. The first visit to state s is no different than any intermediate visit to state s within the episode.

Exercise 5.7

$V_1 =$ initial estimate, $V_2 = G_1$, $V_3 = \frac{G_1 + G_2}{2}$

$$V_{n+1} = \frac{1}{n} \sum_{i=1}^{n} G_i = V_n + \frac{1}{n}\left(G_n - V_n\right)$$

Initialize:

$\pi \leftarrow$ policy to be evaluated
$V(s) \leftarrow 0$ or arbitrary initial estimate
$N(s) \leftarrow 0$

Repeat forever:

Generate an episode using π
For each state s appearing in the episode:
$G \leftarrow$ return following first occurence of s
$N(s) \leftarrow N(s) + 1$
$V(s) \leftarrow V(s) + \frac{1}{N(s)}\left(G - V(s)\right)$