Reinforcement Learning

Borun D. Chowdhury

February 13, 2023

1 Bandits

1.1 Value Based Models

Suppose there is a process where one can choose among a set of actions $\{a\}$ that give rewards R then it makes sense to take actions that maximize the reward

$$q_*(a) = \mathbb{E}[R|A = a] \tag{1}$$

which is the expectation of the reward given the action chose is a.

We want to

- Find the true values for each action

$$Q_t(a) := \frac{\sum_{i=1}^{t-1} R_i \delta_{A_i,a}}{\sum_{i=1}^{t-1} \delta_{A_i,1}} \tag{2}$$

- Take the action with the max value

$$A_t := \underset{a}{\text{argmax}}\ Q_t(a) \tag{3}$$

However, the tricky part is to do them together. For this we can

- Take an ϵ-greedy approach where we exploit with probability $1 - \epsilon$ and explore with probability ϵ,
- Use the upper-confidence-bound method (TODO: find the math behind this)

$$A_t := \underset{a}{\text{argmax}}\ \left[Q_t(a) + c\sqrt{\frac{\ln(t)}{N_t(a)}} \right] \tag{4}$$

The computation of the running mean for rewards eqn 2 per action naively requires keeping track of all rewards but we can instead do

$$Q_{n+1} = Q_n + \frac{1}{n}[R_n - Q_n] \tag{5}$$

Furthermore if the problem is non-stationary we can instead have a fixed parameter α to exponential weight the prior reward

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha[R_n - Q_n] \\ &= (1-\alpha)Q_n + \alpha R_n \\ &= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \dots \\ &= (1-\alpha)^n Q_1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{n-i} R_i \end{aligned} \tag{6}$$

2 Markov Reward Process

A Markov Reward Process is a process where one starts off in a state and at each time makes a transition to another state (possibly the same) while earning some rewards

$$(S_0, \cancel{R_0}), (S_1, R_1), \dots \tag{7}$$

The probability to get to state S_T at time step T is

$$p(S_T) = \sum_{S_0} p(S_T|S_0)p(S_0) \tag{8}$$

Usually the initial state is sharply peaked $p(S_0) = \delta_{S_0,s_0}$ for some s_0. We have

$$\sum_{S_T} p(S_T|S_0) = 1 \tag{9}$$

$$p(S_T|S_0) = \sum_{S_t} p(S_T|S_t)p(S_t, S_0) \tag{10}$$

We then have

$$\begin{aligned} E[R_{t+1}|S_0] &= \sum_{S_T, S_t, R_{t+1}} R_{t+1} p(S_T, S_t) p(R_{t+1}, S_t) p(S_t, S_0) \\ &= \sum_{S_t, R_{t+1}} R_{t+1} p(R_{t+1}, S_t) p(S_t, S_0) \end{aligned} \tag{11}$$

The states after time t do not matter for the expectation value of R_T due to causality.

3 Markov Decision Process

Where Markov Decision Process also involves and agent choosing actions at each step

$$(S_0, A_0, \cancel{R_0}), (S_1, A_1, R_1), \dots \tag{12}$$

Note that

$$p(S_T|S_0) = \sum_{S_t, S_0} p(S_T|S_t)p(S_t|S_0) \tag{13}$$

$$\sum_{S_T} p(S_T|S_t) = 1 \tag{14}$$

Furthermore,

$$p(S_T|S_0) = \sum_{S_t, A_t} p(S_T|S_t, A_t)\pi(A_t|S_t)p(S_t|S_0) \tag{15}$$

$$\sum_{S_T} p(S_T|S_t, A_t) = 1 \tag{16}$$

Thus we have

$$\begin{aligned}\mathbb{E}[R_{t+1}] &= \sum_{S_0, S_T, S_t, A_t} R_{t+1}p(S_T|S_t, A_t)p(R_{t+1}|S_t, A_t)\pi(A_t|S_t)p(S_t|S_0)p(S_0) \\ &= \sum_{S_0, S_t, A_t} R_{t+1}p(R_{t+1}|S_t, A_t)\pi(A_t|S_t)p(S_t|S_0)p(S_0)\end{aligned} \tag{17}$$

Thus the states and actions after t do not matter.

We can break the "propagator" into individual time steps so

$$
\begin{aligned}
\mathbb{E}[R_{t+1}] &= \sum_{R_{t+1},S_0,S_1,\ldots S_T,A_0,A_1,A_{T-1}} R_{t+1} \Big[\\
&\quad \pi(A_0|S_0)p(S_0) \\
&\quad \times\pi(A_1|S_1)p(S_1|S_0,A_0) \\
&\quad\quad \ldots \\
&\quad \times\pi(A_t|S_t)p(S_t|S_{t-1},A_{t-1}) \\
&\quad \times\pi(A_{t+1}|S_{t+1})p(S_{t+1}|S_t,A_t) \\
&\quad\quad \ldots \\
&\quad \times\pi(A_{T-1}|S_{T-1})p(S_{T-1}|S_{T-2},A_{T-2}) \\
&\quad \times p(S_T) \\
&\Big] \\
&\quad \times p(R_{t+1}|S_t,A_t) \\
&= \sum_{R_{t+1},S_0,S_1,\ldots S_t,A_0,A_1,A_t} R_{t+1} \Big[\\
&\quad \pi(A_0|S_0)p(S_0) \\
&\quad \times\pi(A_1|S_1)p(S_1|S_0,A_0) \\
&\quad\quad \ldots \\
&\quad \times\pi(A_t|S_t)p(S_t|S_{t-1},A_{t-1}) \\
&\Big] \\
&\quad \times p(R_{t+1}|S_t,A_t)
\end{aligned} \tag{18}
$$

Here the olive colored terms are redundant in that the sum on them can be done directly and will give one.

We define a discounted (stochastic) reward as

$$
\begin{aligned}
G_t &:= \sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \\
&= R_{t+1} + \gamma G_{t+1}
\end{aligned} \tag{19}
$$

The state value is defined as

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_0|S_0 = s] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\pi[R_{t+1}|S_0 = s] \end{aligned} \tag{20}$$

Thus, we see why we had the term $p(S_0)$ in blue as to compute state values we take this probability to be sharply peaked while summing over all subsequent paths.

Similarly, the action-value is defined as

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi[G_0|S_0 = s, A_0 = a] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\pi[R_{t+1}|S_0 = s, A_0 = a] \end{aligned} \tag{21}$$

and we see why we had the term $\pi(A_0|S_0)$ in red as in this case this term is also sharply peaked and we only sum over subsequent steps.

Since the process is Markovian the transition probabilities are independent of the time step. We define

$$p(s'|s,a) = p(S_{t+1} = s'|S_t = s, A_t = a) \tag{22}$$

and likewise other expressions.

We then get

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t|S_t = s;\pi] \\ &= \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s;\pi] \\ &= \mathbb{E}[R_{t+1}|S_t = s;\pi] + \gamma \mathbb{E}[G_{t+1}|S_t = s;\pi] \\ &= \mathbb{E}[R_{t+1}|S_t = s;\pi] + \gamma \mathbb{E}[\mathbb{E}[G_{t+1}|S_{t+1} = s',\pi]|S_t = s;\pi] \\ &= \mathbb{E}[R_{t+1}|S_t = s;\pi] + \gamma \mathbb{E}[v_\pi(s')|S_t = s;\pi] \\ &= \sum_a \left[\mathbb{E}[R_{t+1}|S_t = s,a] + \gamma \mathbb{E}[v_\pi(s')|S_t = s,a]\right]\pi(a|s) \\ &= \sum_a \left[\sum_r r p(r|s,a) + \gamma \sum_{s'} v_\pi(s') p(s'|s,a)\right]\pi(a|s) \end{aligned} \tag{23}$$

Similarly, we get

$$\begin{aligned}
q_\pi(s,a) &= \mathbb{E}[G_t|S_t=s,A_t=a;\pi] \\
&= \mathbb{E}[R_{t+1}|S_t=s,A_t=a;\pi]+\gamma\mathbb{E}[G_{t+1}|S_t=s,A_t=a] \\
&= \mathbb{E}[R_{t+1}|S_t=s,A_t=a;\pi]+\gamma\mathbb{E}[\mathbb{E}[G_{t+1}|S_{t+1}=s',A_{t+1}=a';\pi]|S_t=s,A_t=a;\pi] \\
&= \mathbb{E}[R_{t+1}|S_t=s,A_t=a;\pi]+\gamma\mathbb{E}[q_\pi(s',a')||S_t=s,A_t=a;\pi] \\
&= \sum_r rp(r|s,a)+\gamma\sum_{s',a'}q_\pi(s',a')p(s',a'|s,a) \\
&= \sum_r rp(r|s,a)+\gamma\sum_{s',a'}q_\pi(s',a')\pi(a'|s')p(s'|s,a) \qquad (24)
\end{aligned}$$

From these we see

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t|S_t=s;\pi] \\
&= \mathbb{E}[\mathbb{E}[G_t|S_t=s,A_t=a;\pi]] \\
&= \mathbb{E}[q_\pi(s,a)] \\
&= \sum_a q_\pi(s,a)\pi(a|s) \qquad (25)
\end{aligned}$$

Further we have

$$q_\pi(s,a)=\sum_r rp(r|s,a)+\gamma\sum_{s'}v_\pi(s')p(s'|s,a) \qquad (26)$$

We get for the optimal policy

$$v_\star(s) = \max_a\left[\sum_r rp(r|s,a)+\gamma\sum_{s'}v_\pi(s')p(s'|s,a)\right] \qquad (27)$$

$$q_\star(s,a) = \sum_r rp(r|s,a)+\gamma\sum_{s'}\left(\max_{a'}q_\pi(s',a')\right)p(s'|sa) \qquad (28)$$

4 Monte Carlo

4.1 The idea

When we do not know the dynamics of the system (i.e. $p(s'|s,a)$ and/or $p(r|s,a)$ then we can sample whole episodes several times to get

$$v_\pi(s), q_\pi(s,a) \qquad (29)$$

by starting off in state s or in state s and take action a and then follow through till the end of the episode. In doing so we can take the first visit approach where we update the averages for a state or state action pair only for the first visit or we can do this for all the visits. Both converse but proving theorems is easier for the former.

4.2 Importance Sampling

We have the expectation of a reward under policy π

$$\begin{aligned}\mathbb{E}_\pi[R_{t+1+k}] &= \sum_{R_{t+1+k},S_t,A_t,\ldots S_{t+k},A_{t+k}} R_{t+1+k}\; p(R_{t+1+k}|S_{t+k},A_{t+k}) \\ &\times\left(\prod_{n=0}^{k-1}\pi(A_{t+n+1}|S_{t+n+1})p(S_{t+n+}|S_{t+n},A_{t+n})\right) \\ &\times\pi(A_t|S_t)p(S_t) \end{aligned} \tag{30}$$

and under a different policy we have

$$\begin{aligned}\mathbb{E}_b[R_{t+1+k}] &= \sum_{R_{t+1+k},S_t,A_t,\ldots S_{t+k},A_{t+k}} R_{t+1+k}\; p(R_{t+1+k}|S_{t+k},A_{t+k}) \\ &\times\left(\prod_{n=0}^{k-1}b(A_{t+n+1}|S_{t+n+1})p(S_{t+n+1}|S_{t+n},A_{t+n})\right) \\ &\times b(A_t|S_t)p(S_t) \end{aligned} \tag{31}$$

So clearly we have

$$\begin{aligned}\mathbb{E}_\pi[R_{t+1+k}] &= \mathbb{E}_b[\left(\prod_{n=0}^{k-1}\frac{\pi(A_{t+n+1}|S_{t+n+1})}{b(A_{t+n+1}|S_{t+n+1})}\right)\frac{\pi(A_t|S_t)}{b(A_t|S_t)}R_{t+1+k}] \\ &= \mathbb{E}_b[\left(\prod_{n=k}^{T-1}\frac{\pi(A_{t+n+1}|S_{t+n+1})}{b(A_{t+n+1}|S_{t+n+1})}\right)\left(\prod_{n=0}^{k-1}\frac{\pi(A_{t+n+1}|S_{t+n+1})}{b(A_{t+n+1}|S_{t+n+1})}\right)\frac{\pi(A_t|S_t)}{b(A_t|S_t)}R_{t+1+k}] \end{aligned} \tag{32}$$

Where in the second line we have written the importance sampling contribution from the full path but the part in olive does not matter as it is after the reward.

4.3 Taking Averages

4.3.1 Ordinary Importance Sampling

For Monte Carlo we can talk about two kinds of averages across episodes. The first is *ordinary importance sampling*

$$V(s) = \frac{\sum_{t\in\mathcal{T}(s)}\rho_{t:T(t)-1}G_t}{|\mathcal{T}(s)|} \tag{33}$$

The way to interpret this is that $\mathcal{T}(s)$ is the union of all time steps at which the state was s. If we are discussing first visit MC then this would be restricted to first visit. T(t) is the

first time of termination following t. G_t are returns pertaining to time step t and $\rho_{t:T(t)-1}$ are the importance sampling ratios.

To show how this works we consider an example. We show how the values are updated for first visit and every visit MC for the first two episodes.

- Episode 1: $\{(s,a),(s',a',R_1),(s,a,R_2),(s_t,R_3\}$
  - First Visit

$$q(s,a) = (R_1+\gamma R_2+\gamma^2 R_3)\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)} \tag{34}$$

$$q(s',a') = (R_2+\gamma R_3)\frac{\pi(a|s)}{b(a|s)} \tag{35}$$

  - Every Visit

$$q(s,a) = \frac{(R_1+\gamma R_2+\gamma^2 R_3)\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)}+R_3}{2} \tag{36}$$

$$q(s',a') = (R_2+\gamma R_3)\frac{\pi(a|s)}{b(a|s)} \tag{37}$$

- Epsiode 2: $\{(s',a'),(s',a',R_4),(s,\bar{a},R_5),(s_t,R_6)\}$
  - First Visit

$$q(s,a) = (R_1+\gamma R_2+\gamma^2 R_3)\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)} \tag{38}$$

$$q(s',a') = \frac{(R_2+\gamma R_3)\frac{\pi(a|s)}{b(a|s)}+(R_4+\gamma R_5+\gamma^2 R_6)\frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)}}{2} \tag{39}$$

$$q(s,\bar{a}) = R_6 \tag{40}$$

  - Every Visit

$$q(s,a) = \frac{(R_1+\gamma R_2+\gamma^2 R_3)\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)}+R_3}{2} \tag{41}$$

$$q(s',a') = \frac{(R_2+\gamma R_3)\frac{\pi(a|s)}{b(a|s)}+(R_4+\gamma R_5+\gamma^2 R_6)\frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)}+(R_5+\gamma R_6)\frac{\pi(\bar{a}|s)}{b(\bar{a}|s)}}{3} \tag{42}$$

$$q(s,\bar{a}) = R_6 \tag{43}$$

4.3.2 Ordinary Importance Sampling: Per-importance importance sampling

At this point it makes sense to *re-write the above after dropping the importance sampling ratios after the reward* as they do not make an effect on expectation but increase variance.

- Episode 1: $\{(s,a),(s',a',R_1),(s,a,R_2),(s_t,R_3\}$
  - First Visit

$$q(s,a) = R_1 + \gamma R_2 \frac{\pi(a'|s')}{b(a'|s')} + \gamma^2 R_3 \frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)} \tag{44}$$

$$q(s',a') = R_2 + \gamma R_3 \frac{\pi(a|s)}{b(a|s)} \tag{45}$$

  - Every Visit

$$q(s,a) = \frac{R_1 + \gamma R_2 \frac{\pi(a'|s')}{b(a'|s')} + \gamma^2 R_3 \frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)} + R_3}{2} \tag{46}$$

$$q(s',a') = R_2 + \gamma R_3 \frac{\pi(a|s)}{b(a|s)} \tag{47}$$

- Episode 2: $\{(s',a'),(s',a',R_4),(s,\bar{a},R_5),(s_t,R_6)\}$
  - First Visit

$$q(s,a) = R_1 + \gamma R_2 \frac{\pi(a'|s')}{b(a'|s')} + \gamma^2 R_3 \frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)} \tag{48}$$

$$q(s',a') = \frac{R_2 + \gamma R_3 \frac{\pi(a|s)}{b(a|s)} + R_4 + \gamma R_5 \frac{\pi(a'|s')}{b(a'|s')} + \gamma^2 R_6 \frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)}}{2} \tag{49}$$

$$q(s,\bar{a}) = R_6 \tag{50}$$

  - Every Visit

$$q(s,a) = \frac{R_1 + \gamma R_2 \frac{\pi(a'|s')}{b(a'|s')} + \gamma^2 R_3 \frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)} + R_3}{2} \tag{51}$$

$$q(s',a') = \frac{R_2 + \gamma R_3 \frac{\pi(a|s)}{b(a|s)} + R_4 + \gamma R_5 \frac{\pi(a'|s')}{b(a'|s')} + \gamma^2 R_6 \frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)} + R_5 + \gamma R_6 \frac{\pi(\bar{a}|s)}{b(\bar{a}|s)}}{3} \tag{52}$$

$$q(s,\bar{a}) = R_6 \tag{53}$$

4.3.3 Weighted Importance Sampling

$$V(s) = \frac{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1}} \tag{54}$$

The way to interpret this is that $\mathcal{T}(s)$ is the union of all time steps at which the state was s. If we are discussing first visit MC then this would be restricted to first visit. T(t) is the

first time of termination following t. G_t are returns pertaining to time step t and $\rho_{t:T(t)-1}$ are the importance sampling ratios.

To show how this works we consider an example. We show how the values are updated for first visit and every visit MC for the first two episodes.

- Episode 1: $\{(s,a),(s',a',R_1),(s,a,R_2),(s_t,R_3\}$
  - First Visit

$$q(s,a) = (R_1+\gamma R_2+\gamma^2 R_3) \quad (55)$$
$$q(s',a') = (R_2+\gamma R_3) \quad (56)$$

  - Every Visit

$$q(s,a) = \frac{(R_1+\gamma R_2+\gamma^2 R_3)\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)}+R_3}{\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)}+1} \quad (57)$$
$$q(s',a') = (R_2+\gamma R_3) \quad (58)$$

- Epsiode 2: $\{(s',a'),(s',a',R_4),(s,\bar{a},R_5),(s_t,R_6)\}$
  - First Visit

$$q(s,a) = (R_1+\gamma R_2+\gamma^2 R_3) \quad (59)$$
$$q(s',a') = \frac{(R_2+\gamma R_3)\frac{\pi(a|s)}{b(a|s)}+(R_4+\gamma R_5+\gamma^2 R_6)\frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)}}{\frac{\pi(a|s)}{b(a|s)}+\frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)}} \quad (60)$$
$$q(s,\bar{a}) = R_6 \quad (61)$$

  - Every Visit

$$q(s,a) = \frac{(R_1+\gamma R_2+\gamma^2 R_3)\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)}+R_3}{\frac{\pi(a'|s')\pi(a|s)}{b(a'|s')b(a|s)}+1} \quad (62)$$
$$q(s',a') = \frac{(R_2+\gamma R_3)\frac{\pi(a|s)}{b(a|s)}+(R_4+\gamma R_5+\gamma^2 R_6)\frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)}+(R_5+\gamma R_6)\frac{\pi(\bar{a}|s)}{b(\bar{a}|s)}}{\frac{\pi(a|s)}{b(a|s)}+\frac{\pi(a'|s')\pi(\bar{a}|s)}{b(a'|s')b(\bar{a}|s)}+\frac{\pi(\bar{a}|s)}{b(\bar{a}|s)}} \quad (63)$$
$$q(s,\bar{a}) = R_6 \quad (64)$$

4.4 Algorithms

4.4.1 Every Visit Algorithms

Here is the algorithm for on-policy MC prediction.

On-policy prediction

Initialize $Q(s,a) = 0$ and $C(s,a) = 0 \; \forall \; s, a$
Loop over episodes:
    Using policy π generate the sequence $S_0, A_0, R_1, S_1, A_1 \ldots S_{T-1}, A_{T-1}, R_T, S_T$
    $G = 0$
    For t in $(T-1, T-2, \ldots 0)$:
        $G = \gamma G + R_{t+1}$
        $Q(S_t, A_t) = \frac{C_(S_t,A_t)Q_(S_t,A_t)+G}{C(S_t,A_t)+1}$
        $C(S_t, A_t) += 1$

The algorithm for off-policy oridinary importance sampling MC prediction is

Off-policy ordinary importance sampling prediction

Initialize $Q(s,a) = 0$ and $C(s,a) = 0 \; \forall \; s, a$
Loop over episodes:
    Using policy b generate the sequence $S_0, A_0, R_1, S_1, A_1 \ldots S_{T-1}, A_{T-1}, R_T, S_T$
    $G = 0$
    $W = 1$
    For t in $(T-1, T-2, \ldots 0)$ while $W \neq 0$:
        $G = \gamma G + R_{t+1}$
        $Q(S_t, A_t) = \frac{C_(S_t,A_t)Q_(S_t,A_t)+WG}{C(S_t,A_t)+1}$
        $C(S_t, A_t) += 1$
        $W* = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

The algorithm for off-policy weighted importance sampling MC prediction is

Off-policy weighted importance sampling prediction

Initialize $Q(s,a) = 0$ and $C(s,a) = 0 \; \forall \; s, a$
Loop over episodes:
    Using policy b generate the sequence $S_0, A_0, R_1, S_1, A_1 \ldots S_{T-1}, A_{T-1}, R_T, S_T$
    $G, W = 0$
    $W = 1$
    For t in $(T-1, T-2, \ldots 0)$ while $W \neq 0$:
        $G = \gamma G + R_{t+1}$
        $Q(S_t, A_t) = \frac{C_(S_t,A_t)Q_(S_t,A_t)+WG}{C(S_t,A_t)+W}$
        $C(S_t, A_t) += W$
        $W* = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

The algorithm for off-policy per decision oridinary importance sampling MC prediction is

Off-policy ordinary importance sampling prediction

Initialize $Q(s,a) = 0$ and $C(s,a) = 0 \; \forall \; s, a$
Loop over episodes:
    Using policy b generate the sequence $S_0, A_0, R_1, S_1, A_1 \ldots S_{T-1}, A_{T-1}, R_T, S_T$
    $G = 0$
    For t in $(T-1, T-2, \ldots 0)$:
        $G = \gamma G + R_{t+1}$
        $Q(S_t, A_t) = \frac{C_(S_t,A_t)Q_(S_t,A_t)+G}{C(S_t,A_t)+1}$
        $C(S_t, A_t) += 1$
        $G* = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

4.5 Example

Suppose we have a state s and one terminal state t. There are only two actions left L and right R. The left takes the agent to the terminal state with probability q and return 1. It takes the agent to the state s with probability $1-q$. Suppose further the target policy has $\pi(L|s) = p$ and the exploration policy has $b(L|s) = \tilde{p}$. We evaluate the value and variance of the state under the first visit method. Then we have

$$\begin{aligned} \mathbb{E}_\pi[G_0] &= pq \sum_{k=0}^{\infty} (p(1-q))^k \\ &= \frac{pq}{1-p(1-q)} \end{aligned} \tag{65}$$

We also have

$$\begin{aligned} \mathbb{E}_\pi[G_0^2] &= pq \sum_{k=0}^{\infty} (p(1-q))^k \\ &= \frac{pq}{1-p(1-q)} \end{aligned} \tag{66}$$

If we compute this with importance sampling we get

$$\begin{aligned} \mathbb{E}_b[G_0 \prod_{t=0}^{\infty} \frac{\pi(A_t|S_t)}{b(A_t|S_t}] &= \tilde{p}q \sum_{k=0}^{\infty} (\tilde{p}(1-q))^k \left(\frac{p}{\tilde{p}}\right)^{k+1} \\ &= \frac{pq}{1-p(1-q)} \end{aligned} \tag{67}$$

which is the same as under the target policy.

We also get

$$\begin{aligned}\mathbb{E}_b[\left(G_0 \prod_{t=0}^{\infty} \frac{\pi(A_t|S_t)}{b(A_t|S_t}\right)^2] &= \tilde{p}q \sum_{k=0}^{\infty} (\tilde{p}(1-q))^k \left(\frac{p}{\tilde{p}}\right)^{2(k+1)} \\ &= \frac{qp^2}{\tilde{p}} \sum_{k=0}^{\infty} \left(\frac{(1-q)p^2}{\tilde{p}}\right)^k \end{aligned} \tag{68}$$

This variance is divergent if $(1-q)p^2 \geq \tilde{p}$. In Barto and Sutton, they take $q = 0.1, p = 1$ and $\tilde{p} = 0.5$.

A Importance Sampling Example

Note that

$$v_\pi(s) = \mathbb{E}_b[\rho_{t:T-1} G_t | S_t = s] \tag{69}$$

where $\rho_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$ and

$$q_\pi(s,a) = \mathbb{E}_b[\rho_{t+1:T-1} G_t | S_t = s, A_t = a] \tag{70}$$

Suppose we have a sequence $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3$ where $S_0 = S_2 = s$ and $A_0 = A_2 = a$ and $S_1 = s', A_1 = a'$ then for ordinary importance sampling starting with $Q(S,A) = 0$ we get

$$\begin{aligned} q(s,a) &= \frac{G_2 + G_0 \rho_{1:2}}{2} \\ &= \frac{R_3 + (R_1 + \gamma R_2 + \gamma^2 R_3) \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)}}{2} \end{aligned} \tag{71}$$

$$\begin{aligned} q(s',a') &= G_1 \rho_{2:2} \\ &= (R_2 + \gamma R_3) \frac{\pi(A_2|S_2)}{b(A_2|S_2)} \end{aligned} \tag{72}$$

whereas for weighted importance sampling we have

$$\begin{aligned} q(s,a) &= \frac{G_2 + G_0 \rho_{1:2}}{1 + \rho_{1:2}} \\ &= \frac{R_3 + (R_1 + \gamma R_2 + \gamma^2 R_3) \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)}}{1 + \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)}} \end{aligned} \tag{73}$$

$$\begin{aligned} q(s',a') &= G_1 \\ &= (R_2 + \gamma R_3) \end{aligned} \tag{74}$$

Now note that because of causality we have

$$\mathbb{E}_b[\rho_{t:T-1}R_{t+k}] = \mathbb{E}_b[\rho_{t:t+k-1}R_{t+k}] \tag{75}$$

so we could also simply take

$$q(s,a) = \frac{R_3 + (R_1 + \gamma R_2 \frac{\pi(A_1|S_1)}{b(A_1|S_1)} + \gamma^2 R_3 \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)})}{2} \tag{76}$$

$$q(s',a') = (R_2 + \gamma R_3 \frac{\pi(A_2|S_2)}{b(A_2|S_2)}) \tag{77}$$