

Reinforcement Learning

Borun D. Chowdhury

January 4, 2023

1 Bandits

1.1 Value Based Models

Suppose for every action there is a true value

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

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

We want to

- Find the true values for each action

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

- Take the action with the max value

$$A_t := \operatorname{argmax}_a Q_t(a) \quad (3)$$

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](#))

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

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

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

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

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

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

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

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

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

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

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

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

We then have

$$\begin{aligned} E[R_{t+1}|S_0] &= \sum_{S_T, S_t, R_{t+1}} R_{t+1} p(S_T, S_t) p(\textcolor{teal}{R}_{t+1}, \textcolor{teal}{S}_t) p(S_t, S_0) \\ &= \sum_{S_t, R_{t+1}} R_{t+1} p(\textcolor{teal}{R}_{t+1}, \textcolor{teal}{S}_t) p(S_t, S_0) \end{aligned} \quad (11)$$

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

$$(S_0, A_0, \widehat{R_0}), (S_1, A_1, R_1), \dots \quad (12)$$

Note that

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

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

Furthermore,

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

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

Thus we have

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

Thus the states and actions after t do not matter.

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

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

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

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

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

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

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

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

and likewise other expressions.

We then get

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

Similarly, we get

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

From these we see

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

Further we have

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

We get for the optimal policy

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

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

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

$$v_\pi(s), q_\pi(s, a) \tag{29}$$

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

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

and under a different policy we have

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

So clearly we have

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

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.

Here is an 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 $p(L) = p$ and the exploration policy has $p(L) = \tilde{p}$. We evaluate the value and variance of the state under the first visit method. Then we have

$$\begin{aligned}\mathbb{E}_\pi[G_0] &= pq \sum_{k=0}^{\infty} (p(1-q))^k \\ &= \frac{pq}{1 - p(1-q)}\end{aligned}\tag{33}$$

We also have

$$\begin{aligned}\mathbb{E}_\pi[G_0^2] &= pq \sum_{k=0}^{\infty} (p(1-q))^k \\ &= \frac{pq}{1-p(1-q)}\end{aligned}\tag{34}$$

Thus the variance is 0.

If we compute this with importance sampling we get

$$\begin{aligned}\mathbb{E}_b[G_0 \prod_{t=0}^{\infty} \frac{\pi(A_t|S_t)}{b(A_t|S_t)}] &= \tilde{p}q \sum_{k=0}^{\infty} (\tilde{p}(1-q))^k \left(\frac{p}{\tilde{p}}\right)^{k+1} \\ &= \frac{pq}{1-p(1-q)}\end{aligned}\tag{35}$$

which is the same as under the target policy.

We also get

$$\begin{aligned}\mathbb{E}_b\left[\left(G_0 \prod_{t=0}^{\infty} \frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right)^2\right] &= \tilde{p}q \sum_{k=0}^{\infty} (\tilde{p}(1-q))^k \left(\frac{p}{\tilde{p}}\right)^{2(k+1)} \\ &= \frac{qp^2}{\tilde{p}} \sum_{k=0}^{\infty} \left(\frac{(1-q)p^2}{\tilde{p}}\right)^k\end{aligned}\tag{36}$$

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

$$v_\pi(s) = \mathbb{E}_b[\rho_{t:T-1} G_t | S_t = s]\tag{37}$$

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

$$q_\pi(s, a) = \mathbb{E}_b[\rho_{t+1:T-1} G_t | S_t = s, A_t = a]\tag{38}$$

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

$$\begin{aligned} q(s, a) &= \frac{G_2 + G_0 \rho_{1:2}}{2} \\ &= \frac{R_3 + (R_1 + \gamma R_2 + \gamma^2 R_3) \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)}}{2} \end{aligned} \quad (39)$$

$$\begin{aligned} q(s', a') &= G_1 \rho_{2:2} \\ &= (R_2 + \gamma R_3) \frac{\pi(A_2|S_2)}{b(A_2|S_2)} \end{aligned} \quad (40)$$

whereas for weighted importance sampling we have

$$\begin{aligned} q(s, a) &= \frac{G_2 + G_0 \rho_{1:2}}{1 + \rho_{1:2}} \\ &= \frac{R_3 + (R_1 + \gamma R_2 + \gamma^2 R_3) \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)}}{1 + \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)}} \end{aligned} \quad (41)$$

$$\begin{aligned} q(s', a') &= G_1 \\ &= (R_2 + \gamma R_3) \end{aligned} \quad (42)$$

Now note that because of causality we have

$$\mathbb{E}_b[\rho_{t:T-1} R_{t+k}] = \mathbb{E}_b[\rho_{t:t+k-1} R_{t+k}] \quad (43)$$

so we could also simply take

$$q(s, a) = \frac{R_3 + (R_1 + \gamma R_2 \frac{\pi(A_1|S_1)}{b(A_1|S_1)} + \gamma^2 R_3 \frac{\pi(A_1|S_1)\pi(A_2|S_2)}{b(A_1|S_1)b(A_2|S_2)})}{2} \quad (44)$$

$$q(s', a') = (R_2 + \gamma R_3 \frac{\pi(A_2|S_2)}{b(A_2|S_2)}) \quad (45)$$