# Reinforcement Learning

Borun D. Chowdhury

January 4, 2023

## 1 Bandits

### 1.1 Value Based Models

Suppose for every action there is a true value

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

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 $\epsilon$-greedy approach where we exploit with probability $1 - \epsilon$ and explore with probability $\epsilon$,
- 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 $\alpha$ 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 Finite Markov Decision Process

The notation we follow is

$$(S_0, A_0, \cancel{R_0}), (S_1, A_1, R_1), \dots \tag{7}$$

That is we start in state $S_0$, do and action $A_0$ to end up in state $S_1$ and get reward $R_1$. Then the agent does an action $A_1$ and so on.

Thus naturally we have the probability of getting to the next state and getting a reward $p', r$ respectively given we are in state $s$ and take action $a$.

$$p(s', r|s, a) := P(S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a) \tag{8}$$

A useful notation is

$$r(s, a) := \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] \tag{9}$$

$$r(s, a, s') := \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] \tag{10}$$

We define a discounted (stochastic) reward as

$$\begin{aligned} G_t &:= \sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{11}$$

Under a policy $\pi$ the value function of a state is *defined* as the reward one gets from

*starting* in the state and following the policy.

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t|S_t = s;\pi] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s;\pi] \\
&= \mathbb{E}[R_{t+1}|S_t = s;\pi] + \gamma\mathbb{E}[G_{t+1}|S_t = s;\pi] \\
&= \mathbb{E}[R_{t+1}|S_t = s;\pi] + \gamma\mathbb{E}[\mathbb{E}[G_{t+1}|S_{t+1} = s',\pi]|S_t = s;\pi] \\
&= \mathbb{E}[R_{t+1}|S_t = s;\pi] + \gamma\mathbb{E}[v_\pi(s')|S_t = s;\pi] \\
&= \sum_a \left[\mathbb{E}[R_{t+1}|S_t = s, a] + \gamma\mathbb{E}[v_\pi(s')|S_t = s, a]\right]\pi(a|s) \\
&= \sum_a \left[\sum_r rp(r|s,a) + \gamma\sum_{s'} v_\pi(s')p(s'|s,a)\right]\pi(a|s) \qquad (12)
\end{aligned}$$

Similarly, the action value function of a state action pair is *defined* as the reward one gets from *starting in that state and taking that particular action*

$$\begin{aligned}
q_\pi(s,a) &= \mathbb{E}[G_t|S_t = s, A_t = a;\pi] \\
&= \mathbb{E}[R_{t+1}|S_t = s, A_t = a;\pi] + \gamma\mathbb{E}[G_{t+1}|S_t = s, A_t = a] \\
&= \mathbb{E}[R_{t+1}|S_t = s, A_t = a;\pi] + \gamma\mathbb{E}[\mathbb{E}[G_{t+1}|S_{t+1} = s', A_{t+1} = a';\pi]|S_t = s, A_t = a;\pi] \\
&= \mathbb{E}[R_{t+1}|S_t = s, A_t = a;\pi] + \gamma\mathbb{E}[q_\pi(s',a')||S_t = s, A_t = a;\pi] \\
&= \sum_r rp(r|s,a) + \gamma\sum_{s',a'} q_\pi(s',a')p(s',a'|s,a) \\
&= \sum_r rp(r|s,a) + \gamma\sum_{s',a'} q_\pi(s',a')\pi(a'|s')p(s'|s,a) \qquad (13)
\end{aligned}$$

From these we see

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t|S_t = s;\pi] \\
&= \mathbb{E}[\mathbb{E}[G_t|S_t = s, A_t = a;\pi]] \\
&= \mathbb{E}[q_\pi(s,a)] \\
&= \sum_a q_\pi(s,a)\pi(a|s) \qquad (14)
\end{aligned}$$

Further we have

$$q_\pi(s,a) = \sum_r rp(r|s,a) + \gamma\sum_{s'} v_\pi(s')p(s'|s,a) \qquad (15)$$

We get for the optimal policy

$$v_\star(s) = \max_a \left[ \sum_r rp(r|s,a) + \gamma \sum_{s'} v_\pi(s')p(s'|s,a) \right] \tag{16}$$

$$q_\star(s,a) = \sum_r rp(r|s,a) + \gamma \sum_{s'} \left( \max_{a'} q_\pi(s',a') \right) p(s'|sa) \tag{17}$$

## 2.1 Writing the value function as a path integral

It is instructive for later (Monte Carlo, Importance Sampling and Policy Gradients) to write the value function as well as state action functions as path integrals. First note that we have

$$p(S_T|S_0) = \sum_{S_t,S_0} p(S_T|S_t)p(S_t|S_0) \tag{18}$$

$$\sum_{S_T} p(S_T|S_t) = 1 \tag{19}$$

Furthermore,

$$p(S_T|S_0) = \sum_{S_t,A_t} p(S_T|S_t,A_t)\pi(A_t|S_t)p(S_t|S_0) \tag{20}$$

$$\sum_{S_T} p(S_T|S_t,A_t) = 1 \tag{21}$$

Thus we have

$$\begin{aligned} \mathbb{E}[R_{t+1}] &= \sum_{S_0,S_T,S_t,A_t} R_{t+1}p(S_T|S_t,A_t)p(R_{t+1}|S_t,A_t)\pi(A_t|S_t)p(S_t|S_0)p(S_0) \\ &= \sum_{S_0,S_t,A_t} R_{t+1}p(R_{t+1}|S_t,A_t)\pi(A_t|S_t)p(S_t|S_0)p(S_0) \end{aligned} \tag{22}$$

Thus the states and actions after $t$ do not matter.

To this end note that

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t|S_t=s] \\ &= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}[R_{t+1+k}|S_t=s] \end{aligned} \tag{23}$$

and

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}[G_t|S_t=s,A_t=a] \\ &= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}[R_{t+1+k}|S_t=s,A_t=a] \end{aligned} \tag{24}$$

Now we can write

$$\begin{aligned}
\mathbb{E}[R_{t+1+k}] &= \sum_{R_{t+1+k}} R_{t+1+k} p(R_{t+1+k}) \\
&= \sum_{R_{t+1+k}, S_t} R_{t+1+k} p(R_{t+1+k}|S_t) p(S_t) \\
&= \sum_{R_{t+1+k}, S_t, A_t} R_{t+1+k} p(R_{t+1+k}|S_t, A_t) \pi(A_t|S_t) p(S_t) \\
&= \sum_{R_{t+1+k}, S_t, A_t, \ldots S_{t+k}, A_{t+k}} R_{t+1+k} \Big[ \\
&\qquad \pi(A_t|S_t) p(S_t) \\
&\qquad \times \pi(A_{t+1}|S_{t+1}) p(S_{t+1}|S_t, A_t) \\
&\quad \ldots \\
&\qquad \times \pi(A_{t+n}|S_{t+n}) p(S_{t+n}|S_{t+n-1}, A_{t+n-1}) \\
&\quad \ldots \\
&\qquad \times \pi(A_{t+k}|S_{t+k}) p(S_{t+k}|S_{t+k-1}, A_{t+k-1}) \\
&\quad \Big] \\
&\quad \times p(R_{t+1+k}|S_{t+k}, A_{t+k}) \qquad (25)
\end{aligned}$$

From this we can get the value function and state-action terms by taking the blue and blue and red terms to 1 in the above expression respectively.

Thus if an episode lasts till time $T$ then starting from $t$ we have the sequence $(S_t, A_t), (S_{t+1}, A_{t+1}, R_{t+1}) \ldots (S_{T-1}, A_{T-1}, R_{T-1}), (S_T, R_T)$ and the probability of a path

$$p(S_T|S_{T-1}, A_{T-1}) \left( \prod_{n=0}^{T-(t+1)-1} \pi(A_{t+n+1}|S_{t+n+1}) p(S_{t+n+1}|S_{t+n}, A_{t+n}) \right) \pi(A_t|S_t) p(S_t) \qquad (25)$$

Note also that

$$\sum_{S,A} p(S_T|S_{T-1}, A_{T-1}) \left( \prod_{n=0}^{T-(t+1)-1} \pi(A_{t+n+1}|S_{t+n+1}) p(S_{t+n+1}|S_{t+n}, A_{t+n}) \right) \pi(A_t|S_t) p(S_t) = 1 \qquad (26)$$

$$\sum_{k=0}^{T-(t+1)} \gamma^k \mathbb{E}[R_{t+1+k}] = \sum_{R,S,A} \left( \prod_{n=0}^{T-(t+1)-1} \pi(A_{t+n+1}|S_{t+n+1}) p(S_{t+n+1}|S_{t+n}, A_{t+n}) \right) \pi(A_t|S_t) p(S_t)$$
$$\times \sum_{k=0}^{T-(t+1)} \gamma^k R_{t+1+k} p(R_{t+1+k}|S_{t+k}, A_{t+k}) \tag{26}$$

Another useful way of writing this is

$$\begin{aligned}
\mathbb{E}[R_{t+k}] &= \sum_{S_T, S_{t+k-1}, S_t, R_{t+k}} R_{t+k} p(R_{t+k}|S_{t+k-1}) p(S_T|S_{t+k-1}) p(S_{t+k-1}|S_t) p(S_t) \\
&= \sum_{S_T, S_{t+k-1}, A_{t+k-1}, S_t, R_{t+k}} R_{t+k} \pi(A_{t+k-1}|S_{t+k-1}) p(R_{t+k}|S_{t+k-1}, A_{t+k-1}) \\
&\quad \times p(S_T|S_{t+k-1}) p(S_{t+k-1}|S_t) p(S_t) \\
&= \sum_{S_{t+k-1}, A_{t+k-1}, S_t, R_{t+k}} R_{t+k} \pi(A_{t+k-1}|S_{t+k-1}) p(R_{t+k}|S_{t+k-1}, A_{t+k-1}) p(S_{t+k-1}|S_t) p(S_t)
\end{aligned} \tag{23}$$

# 3 Monte Carlo

## 3.1 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

$$v_\pi(s), q_\pi(s,a) \tag{23}$$

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.

## 3.2 Importance Sampling

From section 2.1 we see that if we write the expectations under two policies - the target $\pi$ and the one used for sampling (when we do MC) $b$ we have

$$\begin{aligned}
\mathbb{E}_\pi[R_{t+1+k}] &= \sum_{R_{t+1+k}, S_t, A_t, \ldots S_{t+k}, A_{t+k}} R_{t+1+k}\ p(R_{t+1+k}|S_{t+k}, A_{t+k}) \\
&\quad \times \left( \prod_{n=0}^{k-1} \pi(A_{t+n+1}|S_{t+n+1}) p(S_{t+n+}|S_{t+n}, A_{t+n}) \right) \\
&\quad \times \pi(A_t|S_t) p(S_t)
\end{aligned} \tag{22}$$

Similarly, under a different policy we have

$$\begin{aligned}\mathbb{E}_b[R_{t+1+k}] &= \sum_{R_{t+1+k},S_t,A_t,\ldots S_{t+k},A_{t+k}} R_{t+1+k}\; p(R_{t+1+k}|S_{t+k},A_{t+k}) \\ &\times\left(\prod_{n=0}^{k-1} b(A_{t+n+1}|S_{t+n+1})p(S_{t+n+1}|S_{t+n},A_{t+n})\right) \\ &\times b(A_t|S_t)p(S_t) \end{aligned} \tag{21}$$

So clearly we have

$$\mathbb{E}_\pi[R_{t+1+k}] = \mathbb{E}_b[\left(\prod_{n=0}^{k-1}\frac{\pi(A_{t+n+1}|S_{t+n+1})}{b(A_{t+n+1}|S_{t+n+1})}\right)\frac{\pi(A_t|S_t)}{b(A_t|S_t)}R_{t+1+k}] \tag{22}$$

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

$$\begin{aligned}\mathbb{E}_\pi[G_0] &= pq\sum_{k=0}^{\infty}(p(1-q))^k \\ &= \frac{pq}{1-p(1-q)}\end{aligned} \tag{22}$$

We also have

$$\begin{aligned}\mathbb{E}_\pi[G_0^2] &= pq\sum_{k=0}^{\infty}(p(1-q))^k \\ &= \frac{pq}{1-p(1-q)}\end{aligned} \tag{22}$$

Thus the variance is 0.

If we compute this with importance sampling we get

$$\begin{aligned}\mathbb{E}_b[G_0\prod_{t=0}^{\infty}\frac{\pi(A_t|S_t)}{b(A_t|S_t}] &= \tilde{p}q\sum_{k=0}^{\infty}(\tilde{p}(1-q))^k\left(\frac{p}{\tilde{p}}\right)^{k+1} \\ &= \frac{pq}{1-p(1-q)}\end{aligned} \tag{22}$$

which is the same as under the target policy.

We also get

$$\begin{aligned}\mathbb{E}_b[\left(G_0\prod_{t=0}^{\infty}\frac{\pi(A_t|S_t)}{b(A_t|S_t}\right)^2] &= \tilde{p}q\sum_{k=0}^{\infty}(\tilde{p}(1-q))^k\left(\frac{p}{\tilde{p}}\right)^{2(k+1)} \\ &= \frac{qp^2}{\tilde{p}}\sum_{k=0}^{\infty}\left(\frac{(1-q)p^2}{\tilde{p}}\right)^k \qquad (22)\end{aligned}$$

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