Reinforcement Learning Equations

Aaron Hao Tan

1 Finite Markov Decision Processes

Components of MDP : Decision epochs, State, Action, Transition probability, Rewards

$$\{T, S, A_s, p_t(\cdot|s,a), r_t(s,a)\} \tag{1}$$

A state is *Markov* if and only if

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \ldots, S_t] \tag{2}$$

State-transition probabilities (p characterize the environment's dynamics). More often used than equations 4 and 5.

$$\begin{aligned} p(s'|s,a) &= Pr[S_t = s'|S_{t-1} = s, A_{t-1} = a] \\ &= \sum_{r\in\mathcal{R}} p(s', r|s,a) \end{aligned} \tag{3}$$

Expected rewards for state-action pairs and for state-action-next-state triples. The current reward depends on the previous state and action.

$$r(s,a) \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r|s,a) \tag{4}$$

$$r(s,a,s') \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r\in\mathcal{R}} r \frac{p(s', r|s,a)}{p(s'|s,a)} \tag{5}$$

Policy
- Probabilistic is inferior to deterministic

$$\pi(a|s) = Pr(A_t = a|S_t = s) \tag{6}$$

Returns (sum of rewards)

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \tag{7}$$

Discounted Returns (γ = discount rate). If $\gamma < 1$, the infinite sum in equation 8 would have a finite value. If $\gamma = 0$, the agent is concerned with only maximizing immediate rewards (myopic). If the reward is +1, the return is equivalent to $G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$.

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{8}$$

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots\right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{9}$$

State Value Function
- Value of terminal state (if any) is always zero.
- The expected return (expected cumulative future discounted reward) starting from that state

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s] \quad \forall s \in S \tag{10}$$

Bellman Equation for v_π : the value of the start state must equal the (discounted) value of the expected next state, plus the reward expected along the way. Assumes discrete and is invariant of time (time doesn't matter).

$$\begin{aligned} v_\pi(s) &= E_\pi[R_{t+1} + \gamma \cdot G_{t+1}|S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']] \\ &= \sum_a \pi(a|s) \sum_{s', r} p\left(s', r|s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right] \quad \forall s \in S \end{aligned} \tag{11}$$

Action Value Function

$$q_\pi(s, a) \doteq \mathbb{E}_\pi\left[G_t|S_t = s, A_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right] \tag{12}$$

Bellman Equation for $q_\pi(s, a)$

$$q_\pi(s, a) = \sum_{s', r} p\left(s', r|s, a\right)\left[r + \gamma \sum_{a'} \pi\left(a'|s'\right) q_\pi\left(s', a'\right)\right] \tag{13}$$

The following figure represents the state value and action state value functions graphically.

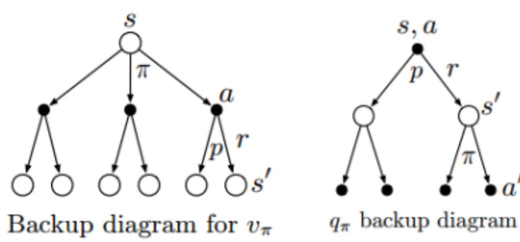

Figure 1: Value and State-Value Functions

Optimal State Value Function A policy π is better or equal to another policy π' if its expected return is greater than or equal to that of π' for all states ($\pi' \geq \pi'$ if and only if $v_\pi(s) \geq v_{\pi'}(s)$).

$$v_*(s) \doteq \max_\pi v_\pi(s) \tag{14}$$

Optimal Action Value Function and its relation to v_* . Equation 15 gives the expected return for taking action a in state s and thereafter following an optimal policy. Comparing with equation 11, $v_\pi(s)$ represents the value of a state as the summation of future returns. In equation 15, $v_*\left(S_{t+1}\right)$ gives the optimal value of all future states from the given state and action.

$$\begin{aligned} q_*(s, a) &\doteq \max_\pi q_\pi(s, a) \\ &= \mathbb{E}\left[R_{t+1} + \gamma v_*\left(S_{t+1}\right)|S_t = s, A_t = a\right] \end{aligned} \tag{15}$$

Relationship between q_* and v_*
- The value of a state under an optimal policy must equal the expected return for the best action from that state.
- Second last line of equation 16 shows that $G_t = R_{t+1} + \gamma v_*(S_{t+1})$ when following π_*.
- The last line is the same as equation 17. For finite MDPs, the last line of equation 16 has a unique solution.
- R_{t+1} has nothing to do with policy because the first action was not taken with policy
- max operator turns equation 16 in to a system of nonlinear equations

$$\begin{aligned} v_*(s) &= \max_{a \in A(s)} q_*(s, a) \\ &= \max_a q_*(s, a) \\ &= \max_a \mathbb{E}_{\pi_*}[G_t|S_t = s, A_t = a] \\ &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a] \\ &= \max_a \sum_{s', r} p(s', r|s, a)[r + \gamma v_*(s')] \end{aligned} \tag{16}$$

Bellman Optimality Equations

$$\begin{aligned} v_*(s) &= \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*\left(S_{t+1}\right)|S_t = s, A_t = a\right] \\ &= \max_a \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma v_*\left(s'\right)\right] \end{aligned} \tag{17}$$

$$\begin{aligned} q_*(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*\left(S_{t+1}, a'\right)|S_t = s, A_t = a\right] \\ &= \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma \max_{a'} q_*\left(s', a'\right)\right] \end{aligned} \tag{18}$$

In v_*, the next best action is selected based on the expected reward and value of future states. In q_*, the state and action is given as well as the reward. From the new state s', the best action is chosen.

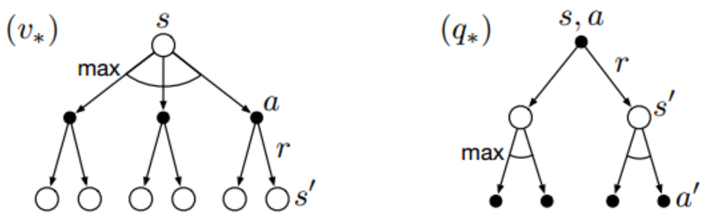

Figure 2: Bellman Optimality Equations

2 Dynamic Programming

Policy Evaluation
- turn Bellman equation from equation 11 in to an update rule
- maximum bootstrapping: $v_k(s')$ is using a previous estimate

$$\begin{aligned} v_{k+1}(s) &\doteq \mathbb{E}_\pi\left[R_{t+1} + \gamma v_k\left(S_{t+1}\right)|S_t = s\right] \\ &= \sum_a \pi(a|s) \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma v_k\left(s'\right)\right] \end{aligned} \tag{19}$$

Policy Evaluation Steps
- for state in states
—- for action in actions available in each state
——— calculate value of state using equation 19
——— (value = sum all values of each action together for each state (ie. 4 actions in a state = 1 total value))
—- replace old value of state with newly calculated value
—- calculate the change in value of each state with only the maximum difference remembered
- stop looping if the maximum change in state values falls below a threshold

Policy Improvement
- greedy policy = guaranteed to be optimal
- one-step greedy lookahead

$$
\begin{aligned}
\pi'(s) &\doteq \arg\max_a q_\pi(s,a) \\
&= \arg\max_a \mathbb{E}\left[R_{t+1} + \gamma v_\pi\left(S_{t+1}\right) | S_t = s, A_t = a\right] \\
&= \arg\max_a \sum_{s',r} p\left(s', r | s, a\right)\left[r + \gamma v_\pi\left(s'\right)\right]
\end{aligned}
\tag{20}
$$

Policy is represented $s \times a$ table where each row represents a state and each column represents an action. The value within each entry of the table represent the probability of taking that action, when in that state, using a probabilistic policy.

At this stage, there is a value calculated for every state using policy evaluation already.

Policy Improvement Steps
- for state in states
—- choose the best action with the current policy (argmax)
—- for action in actions available in each state
——— calculate action values for each action using equation 20 without the argmax
———(this creates a $1 \times a$ vector where each element is an action value q)
—- select the best action in the $1 \times a$ vector (equation 20 with argmax)
—- if action chosen with the current policy is not the calculated best action
——— policy is unstable
—- update the policy with the new best action

Policy Iteration
- built on the fundamentals of value iteration
- more efficient (CPU time) than value iteration

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
       $\Delta \leftarrow 0$
       Loop for each $s \in \mathcal{S}$:
           $v \leftarrow V(s)$
           $V(s) \leftarrow \sum_{s',r} p(s',r|s,\pi(s))\big[r + \gamma V(s')\big]$
           $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
       old-action $\leftarrow \pi(s)$
       $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s',r|s,a)\big[r + \gamma V(s')\big]$
       If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Figure 3: Policy Iteration

Value Evaluation

- Policy iteration is faster because policy converges quicker than value functions
- Bellman Optimality Equation (equation 17) turned in to an update rule
- Formally requires infinite number of iterations to converge exactly to v_* but we stop before then
- Only guarantees $\epsilon - \mathit{optimality}$ (theoretically)
- All algorithms converge to an optimal policy for discounted finite MDPs
- When γ is close to 1, iteration goes forever = making value iteration inefficient
- Policy iteration converges faster even for smaller γ - Main difference between Value Iteration and Policy Iteration

– In value iteration, instead of summing values from all actions per state, take only the best action value per state as the value of that state

- Value iteration update is identical to policy evaluation except that the max is taken over all actions

$$\begin{aligned} v_{k+1}(s) &\doteq \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k\left(S_{t+1}\right) | S_t = s, A_t = a\right] \\ &= \max_a \sum_{s',r} p\left(s', r | s, a\right)\left[r + \gamma v_k\left(s'\right)\right] \end{aligned} \tag{21}$$

Value Iteration Steps

- for state in states

—- for action in actions available in each state

——— calculate action values for each action using equation 21 without the max

———(this creates a $1 \times a$ vector where each element is an action value q)

—- value = the best action value in the $1 \times a$ vector (equation 21 with max)

—- calculate the change in value of each state with only the maximum difference remembered

- stop looping if the maximum change in state values falls below a threshold
- for state in states

—- for action in actions available in each state

——— calculate action values for each action using equation 21 without the max

———(this creates a $1 \times a$ vector where each element is an action value q)

—- select the best action in the $1 \times a$ vector (equation 21 with argmax)

—- overwrite policy table with best action

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s', r | s, a)\left[r + \gamma V(s')\right]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r | s, a)\left[r + \gamma V(s')\right]$

Figure 4: Value Iteration

3 Monte Carlo Methods

- Solving RL based on averaging sample returns
- Return after taking an action in one state depends on the actions taken in later states in the same episode
- No boot strapping (DP = max bootstrapping)
- can be considered as a way to go from π to v

First-visit MC Method
- average of the returns following first visit to s
- more popular, samples are independent (faster convergence)
- poor sample efficiency

Every-visit MC Method
- averages the returns following all visits to s
- more natural to function approximation/eligibility traces
- possibly statistically unstable

The following shows first-visit MC methods.

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated
Initialize:
    $V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$
    $Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):
    Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:
            Append G to $Returns(S_t)$
            $V(S_t) \leftarrow \text{average}(Returns(S_t))$

Figure 5: First-visit MC Prediction

Problem with prior, there will only be one action from each state if following a deterministic policy. To fix this:
- Look at state and action pairs as oppose to just states
- Give all state action pair a nonzero probability of being selected (line 6)
- Includes a greedy policy at the end to pick the best action in the next iteration (line 14)

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:
    $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
    $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
    $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
    Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0
    Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
            Append G to $Returns(S_t, A_t)$
            $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
            $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

Figure 6: Exploring Start MC Control

On-policy vs Off-policy Methods
On-policy methods attempt to evaluate/improve the policy that is used to make decisions, whereas off-policy methods evaluate/improve a policy different from that used to generate data. On-policy approach is actually a compromise - it learns action values not for the optimal policy, but for a near optimal policy that still explores.
- On-policy: unbiased, low bias and variance
- Off-policy: biased, high bias and high variance. Slower to converge.

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$
Initialize:
  $\pi \leftarrow$ an arbitrary ε-soft policy
  $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
  $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):
  Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
      Append G to $Returns(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
      $A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)
      For all $a \in \mathcal{A}(S_t)$:

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

Figure 7: On-policy first-visit MC Control

ϵ - greedy Exploration
- with probability ϵ, select an action at random
- all non-greedy actions are given the minimal probability of selection, $\frac{\varepsilon}{|\mathcal{A}(s)|}$
- remaining bulk of probability is given to the greedy action, $1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|}$
- the following equation either chooses the greedy action A^* or not based on the probabilities stated

$$\pi\left(a|S_t\right) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/\left|\mathcal{A}\left(S_t\right)\right| & \text{if } a = A^* \\ \varepsilon/\left|\mathcal{A}\left(S_t\right)\right| & \text{if } a \neq A^* \end{cases} \tag{22}$$

Off-policy Evaluation via Importance Sampling
- Target policy: policy being learned about
- Behavior policy: policy used to generate behavior

Importance Sampling Ratio
In order to use episodes from b to estimate values for π, we require that every action taken under π is also taken, at least occasionally, under b. That is, the following is assumed to be true (known as *coverage*).

$$\pi(a|s) > 0 \quad \text{implies} \quad b(a|s) > 0 \tag{23}$$

We apply importance sampling to off-policy learning by weighting returns according to the relative probability of their trajectories occurring under the target and behavior policies. For example, given a start state S_t, the probability of the subsequent state-action trajectory occuring under any policy π is the following.

$$\begin{aligned} &\Pr\left\{A_t, S_{t+1}, A_{t+1}, \ldots, S_T | S_t, A_{t:T-1} \sim \pi\right\} \\ &\quad = \pi\left(A_t|S_t\right) p\left(S_{t+1}|S_t, A_t\right) \pi\left(A_{t+1}|S_{t+1}\right) \cdots p\left(S_T|S_{T-1}, A_{T-1}\right) \\ &\quad = \textstyle\prod_{k=t}^{T-1} \pi\left(A_k|S_k\right) p\left(S_{k+1}|S_k, A_k\right) \end{aligned} \tag{24}$$

The relative probability of the trajectory under the target and behavior policies is shown below. Note, the ratio depends on the policies and not on the MDP.

$$\rho_{t:T-1} \doteq \frac{\prod_{k=t}^{T-1} \pi\left(A_k|S_k\right) p\left(S_{k+1}|S_k, A_k\right)}{\prod_{k=t}^{T-1} b\left(A_k|S_k\right) p\left(S_{k+1}|S_k, A_k\right)} = \prod_{k=t}^{T-1} \frac{\pi\left(A_k|S_k\right)}{b\left(A_k|S_k\right)} \tag{25}$$

Convert returns G_t due to the behavior policy to expected returns under the target policy.

$$\mathbb{E}\left[\rho_{t:T-1} G_t | S_t = s\right] = v_\pi(s) \tag{26}$$

Ordinary Importance Sampling
- $\mathcal{T}(s)$ represent the time steps in which state s is visited (number of samples that we have)
- To estimate $v_\pi(s)$, simply scale the returns by the ratios and average the results
- First visit: unbiased, high variance (unbounded)
- Every visit: biased

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|} \tag{27}$$

Weighted Importance Sampling
- Weighted average
- First visit: biased (though the bias converges asymptotically to zero), variance converges to zero
- In practice, this method has dramatically lower variance and is preferred.
- Every visit: biased

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}} \tag{28}$$

Comparison of Ordinary and Weighted
If you only have one sample, weighted importance would equal G_t as the ratio in the numerator and denominator cancel out. For ordinary, the denominator would equal to 1, and the result is equivalent to $ISR * G_t$.

Incremental Implementation: Ordinary Importance Sampling

$$V_n = \frac{\sum_{k=1}^{n-1} w_k G_t}{n-1} \tag{29}$$

$$V_{n+1} = V_n + \frac{1}{n}\left(w_n G_n - V_n\right) \tag{30}$$

Incremental Implementation: Weighted Importance Sampling

$$V_n = \frac{\sum_{k=1}^{n-1} w_k G_k}{\sum_{k=1}^{n-1} w_k} \tag{31}$$

$$V_{n+1} = V_n + \frac{w_n}{C_n}\left[G_n - V_n\right] \tag{32}$$

$$C_{n+1} = C_n + w_{n+1} \tag{33}$$

Off-policy MC Prediction

Off-policy MC prediction (policy evaluation) for estimating $Q \approx q_\pi$

Input: an arbitrary target policy π
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \in \mathbb{R}$ (arbitrarily)
  $C(s,a) \leftarrow 0$

Loop forever (for each episode):
  $b \leftarrow$ any policy with coverage of π
  Generate an episode following b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$, while $W \neq 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

Figure 8: Off-policy MC Prediction

Off-policy MC Control

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \in \mathbb{R}$ (arbitrarily)
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(s,a)$ (with ties broken consistently)

Loop forever (for each episode):
  $b \leftarrow$ any soft policy
  Generate an episode using b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

Figure 9: Off-policy MC Control