Reinforcement Learning: Tutorial 4

Monte Carlo methods

Week 2
University of Amsterdam

Milena Kapralova
September 2024

Check-in

- How is it going?
- How is HW1?
- Are you ready to start HW2?
- If you have any feedback so far, please mail me at *m.kapralova@uva.nl*

Outline

1. Admin

2. Monte Carlo exercises

3. Ask anything about HW1 or 3.4 (HW2)

Admin

- Reminder that HW1 deadline is tomorrow @ 17:00
- Just to be clear about our lenience during grading - depends on the context: if the questions is asking to compare v_π for some policy π vs. the optimal value function $v*$, then omitting the indices makes it impossible for us to check if you understood. When the only v function in the whole question is $v*$, but you forget the $*$ once, or e.g. you put a subscript instead of a superscript, it is clear to us what you mean and while we put a short comment to remind we we wouldn't subtract points.
- Any questions?

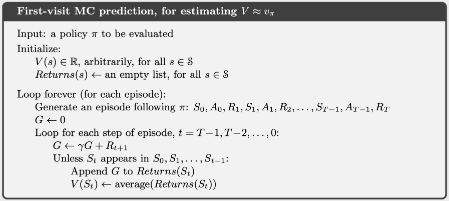
Tutorial 3 Overview

1. Monte Carlo exercises
2. Ask anything about HW1 or 3.4 (HW2)

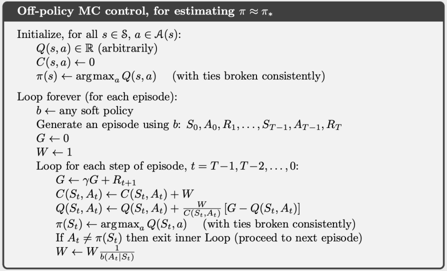
Tutorial 3 Overview

1. Monte Carlo exercises
    - Questions 3.1-3.3
2. Ask anything about HW1 or 3.4 (HW2)

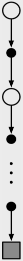
Theory Intermezzo: so many Monte Carlo's



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

Input: a policy $\pi$ 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 $\pi$: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    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 $Returns(S_t)$
            $V(S_t) \leftarrow \text{average}(Returns(S_t))$
```



```
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)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$    (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, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```



- We don't always know the transition dynamics, so we can't always use dynamic programming
- Value-based RL: value function → policy

Q 3.1 Monte Carlo

1. Consider an MDP with a single state s_0 that has a certain probability of transitioning back onto itself with a reward of 0, and will otherwise terminate with a reward of 3. Your agent has interacted with the environment and has gotten the following two sequences of rewards obtained: $[0, 0, 3], [0, 0, 0, 3]$. Use $\gamma = 0.8$.

Q 3.1 Monte Carlo

1. Consider an MDP with a single state s_0 that has a certain probability of transitioning back onto itself with a reward of 0, and will otherwise terminate with a reward of 3. Your agent has interacted with the environment and has gotten the following two sequences of rewards obtained: $[0, 0, 3], [0, 0, 0, 3]$. Use $\gamma = 0.8$.

   a Estimate the value of s_0 using first-visit MC.
   b Estimate the value of s_0 using every-visit MC.

Q 3.1 Monte Carlo

1. Consider an MDP with a single state s_0 that has a certain probability of transitioning back onto itself with a reward of 0, and will otherwise terminate with a reward of 3. Your agent has interacted with the environment and has gotten the following two sequences of rewards obtained: $[0, 0, 3], [0, 0, 0, 3]$. Use $\gamma = 0.8$.

   The returns are: $[1.92, 2.4, 3.0], [1.54, 1.92, 2.4, 3.0]$.

   a. Estimate the value of s_0 using first-visit MC.
   $\frac{1.92+1.54}{2} = 1.73$

   b. Estimate the value of s_0 using every-visit MC.
   $\frac{1.92+2.4+3.0+1.54+1.92+2.4+3.0}{7} \approx 2.31$

Q 3.2 Bias of v_π Monte Carlo estimators

1. Comment on the bias of *weighted importance sampling* compared to *ordinary importance sampling*. Why might we nevertheless use weighted importance sampling?

Q 3.2 Bias of v_π Monte Carlo estimators

1. Comment on the bias of *weighted importance sampling* compared to *ordinary importance sampling*. Why might we nevertheless use weighted importance sampling?

   While the ordinary importance sampling is unbiased, weighted importance sampling is not.
   However, the variance of ordinary importance sampling is unbounded:
   E.g. if $\pi(a|s) = 1, \mu(a|s) = 0.01$ then the ratio is 100, for one visit with return of 1 we would get a value estimate of 100.
   For this reason, it can be a good idea to consider weighted importance sampling.

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

2. $\sum_{i=1}^{T} \gamma^{i-1} R_i$.

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

2. $\sum_{i=1}^{T} \gamma^{i-1} R_i$.

This estimator is unbiased:

$$\mathbb{E}\left[\sum_{i=1}^{T} \gamma^{i-1} R_i\right] = \mathbb{E}\left[\sum_{k=0}^{T-1} \gamma^{k} R_{k+1}\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{T-1} \gamma^{k} R_{k+1} | S_0 = s\right] = v_\pi(s).$$

In the first equality, we relabel the indices of the summation to match the original definition in the lecture. While this is not completely necessary, it helps us check for correctness. The second equality holds since our episode is sampled from π and $S_0 = s$. Note that this is one Monte Carlo first-visit sample.

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \dots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

3. $\frac{1}{|J|} \sum_{j \in J} \sum_{i=1}^{T-j} \gamma^{i-1} R_{j+i}$, where J contains all indices j s.t. $S_j = s$.

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

3. $\frac{1}{|J|}\sum_{j\in J}\sum_{i=1}^{T-j}\gamma^{i-1}R_{j+i}$, where J contains all indices j s.t. $S_j = s$.

Rewriting this estimator in terms of returns:

$$\frac{1}{|J|}\sum_{j\in J}\sum_{i=1}^{T-j}\gamma^{i-1}R_{j+i} = \frac{1}{|J|}\sum_{j\in J}\sum_{k=0}^{T-1-j}\gamma^{k}R_{j+k+1} = \frac{1}{|J|}\sum_{j\in J}G_j.$$

Since $S_j = s$ for every $j \in J$, we have:

$$\mathbb{E}\left[\frac{1}{|J|}\sum_{j\in J}G_j\right] = \left[\frac{1}{|J|}\sum_{j\in J}\mathbb{E}\left[G_j\right]\right] = \left[\frac{1}{|J|}\sum_{j\in J}\mathbb{E}_\pi\left[G_j|S_j = s\right]\right].$$

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

3. $\frac{1}{|J|} \sum_{j \in J} \sum_{i=1}^{T-j} \gamma^{i-1} R_{j+i}$, where J contains all indices j s.t. $S_j = s$.

This is a Monte Carlo every-visit estimator. For any visit after the first, the corresponding G is a sample of the distribution of returns *conditioned on the fact that s has been visited* once, twice, etc before in the trajectory.
This means that the corresponding expected value of such G is not necessarily the same as the value function, which represents the expected value without any such condition, implying that this estimator is biased.

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

3. $\frac{1}{|J|} \sum_{j \in J} \sum_{i=1}^{T-j} \gamma^{i-1} R_{j+i}$, where J contains all indices j s.t. $S_j = s$.

Another way to look at this, is that for j' past the first visit, $G_{j'}$'s are systematically dropping the subsequence of rewards after the first visit up to the j'-th. This restriction on the sample space induces a bias if $v_\pi(s)$ is the target quantity. Therefore, not all of the G_j's are unbiased estimators, implying that this estimator is biased. See Theorem 7 of [1] for another proof.

[1] Reinforcement Learning with Replacing Eligibility Traces. Singh and Sutton 1995.

Q 3.2 Bias of v_π Monte Carlo estimators

Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

4. $\sum_{i=1}^{T-t_s} \gamma^{i-1} R_{t_s+i}$, where t_s is the latest time step such that $S_{t_s} = s$. How does this compare to the first-visit MC estimator?

Q 3.2 Bias of v_π Monte Carlo estimators

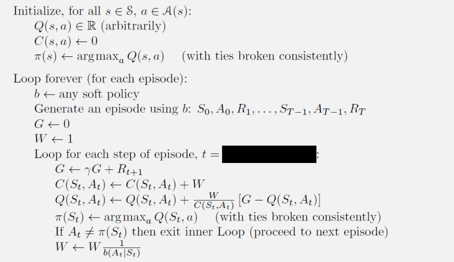
Consider one episode following π: $(S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T_1}, R_T)$, where $S_0 = s$. Determine and provide intuition on the biasedness of the estimators for $v_\pi(s)$:

4. $\sum_{i=1}^{T-t_s} \gamma^{i-1} R_{t_s+i}$, where t_s is the latest time step such that $S_{t_s} = s$. How does this compare to the first-visit MC estimator?

   Via a similar argument, we can derive the conditional expectation $\mathbb{E}_\pi \left[G_{t_s} | S_{t_s} = s\right]$. If t_s is not the first time index such that $S_j = s$, then this is not equal to the value function, which implies that the estimator is biased, via an argument similar to the above. Otherwise, this is a first-visit estimator, which is unbiased.

Q 3.3 Exam question: Monte Carlo for control

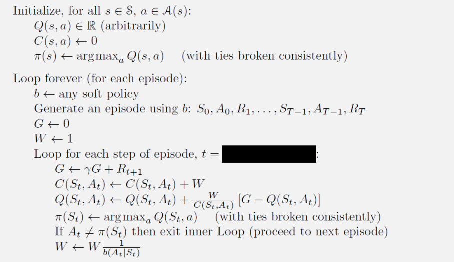
You are given the algorithm pseudocode:



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

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

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:

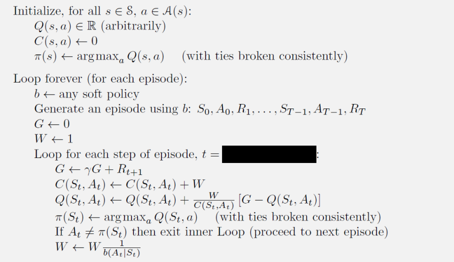


```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$    (with ties broken consistently)

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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

1. Part of the algorithm is covered by a black square. What is the missing information?

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:



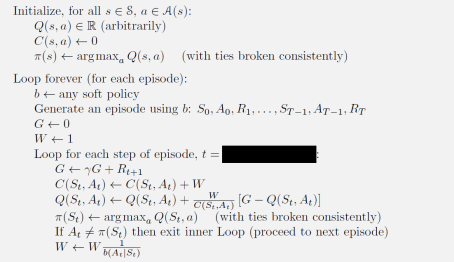
```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$    (with ties broken consistently)

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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

1. Part of the algorithm is covered by a black square. What is the missing information?

   T-1, T-2, ... 0 (decreasing numbers). Why is this?

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:

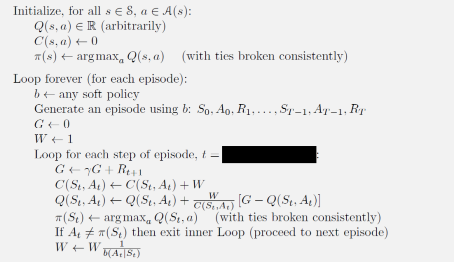


```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(s,a)$    (with ties broken consistently)

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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

2. Is this a Monte-Carlo algorithm or a TD-based algorithm? Explain your answer based on the given pseudo-code.

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:



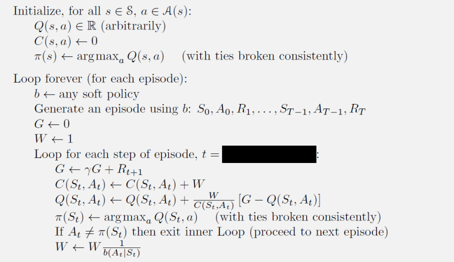
```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(s,a)$    (with ties broken consistently)

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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

2. Is this a Monte-Carlo algorithm or a TD-based algorithm? Explain your answer based on the given pseudo-code.

   Monte-Carlo, as the learning signal is the return without any bootstrapping. This can be seen from the target $(G - Q(S, A))$ and/or from the definition of G: $G \longleftarrow \gamma G + R_{t+1}$.

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:

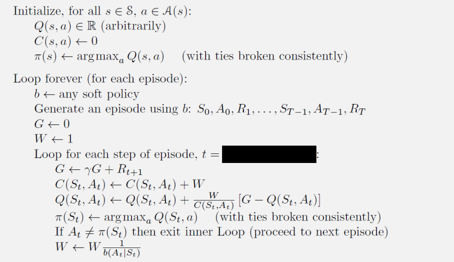


```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$    (with ties broken consistently)

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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

3. What is stored in $C(S_t, A_t)$?

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:



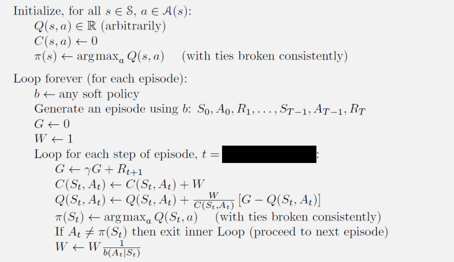
```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$    (with ties broken consistently)

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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

3. What is stored in $C(S_t, A_t)$?

   The cumulative importance weight of all visits to the state-action pair.

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:

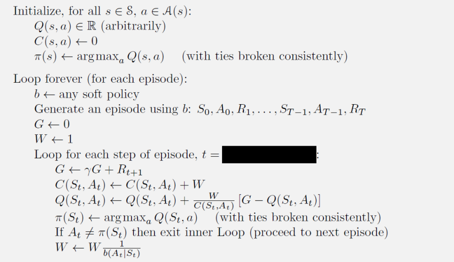


```
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$    (with ties broken consistently)

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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

4. Why is the inner loop stopped when $A_t \neq \pi(S_t)$?

Q 3.3 Exam question: Monte Carlo for control

You are given the algorithm pseudocode:



```
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, $t =$ [redacted]:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$    (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$
```

4. Why is the inner loop stopped when $A_t \neq \pi(S_t)$?

   Since the probability of taking this action under the greedy policy is 0, the importance weight of all histories that include this state-action pair is 0 (realization that importance weight is zero, realization that this has to hold for longer sequences too).

WELL DONE!

Tutorial 3 Overview

1. Monte Carlo exercises
2. Ask anything about HW1 or 3.4 (HW2)
    - Questions 2.3-2.4, 3.4

Ask anything about HW1 or 3.4 (HW2)

1. 2.3: Coding
2. 2.4: Theory
3. 3.4: Coding

That's it!

Good luck with the HW and see you on Monday!