REINFORCEMENT LEARNING: THEORY AND PRACTICE

Ch. 5: Monte Carlo Methods

Profs. Amy Zhang and Peter Stone

Questions?

Previously

Chapter 4: Computed value functions using knowledge of the MDP

Today

Chapter 5: Learn value functions from sample (Monte Carlo) returns

Similar to Chapter 2 on bandits, but now compute average *returns* rather than average *rewards*

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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow$ average$(Returns(S_t))$

Episode 0:

State, Action, Reward

0, Clap, 0

1, Clap, 0

2, Clap, 1

3, Wave, 0

7, Clap, 10

0 0	1 0	2 0	3 0
4 0	5 0	6 0	7 0
8 0	9 0	10 0	11 0

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: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

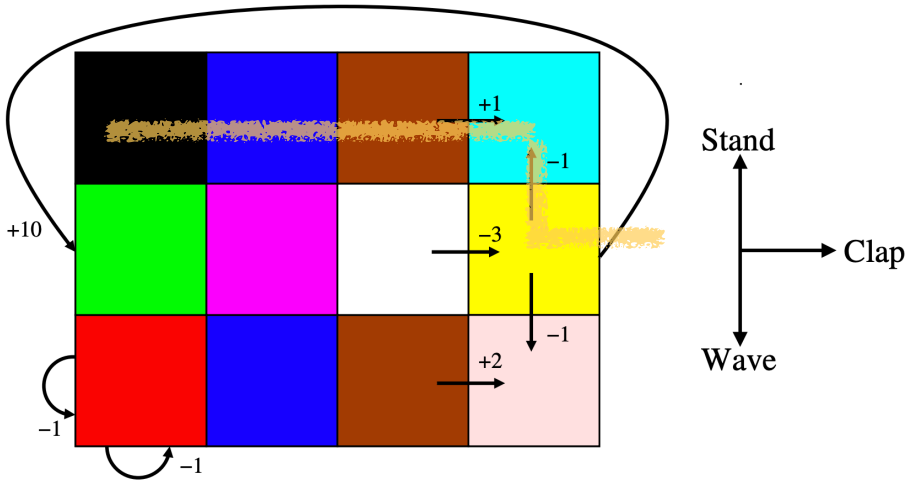
Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow$ average$(Returns(S_t))$



0 11	1 11	2 11	3 10
4 0	5 0	6 0	7 10
8 0	9 0	10 0	11 0

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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow \text{average}(Returns(S_t))$

Episode 1:

State, Action, Reward

0, Wave, 0

4, Wave, 0

8, Clap, 0

9, Stand, 0

5, Clap, 0

6, Wave, 0

10, Clap, 2

11, Stand, 0

7, Clap, 10

0 11	1 11	2 11	3 10
4 0	5 0	6 0	7 10
8 0	9 0	10 0	11 0

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: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

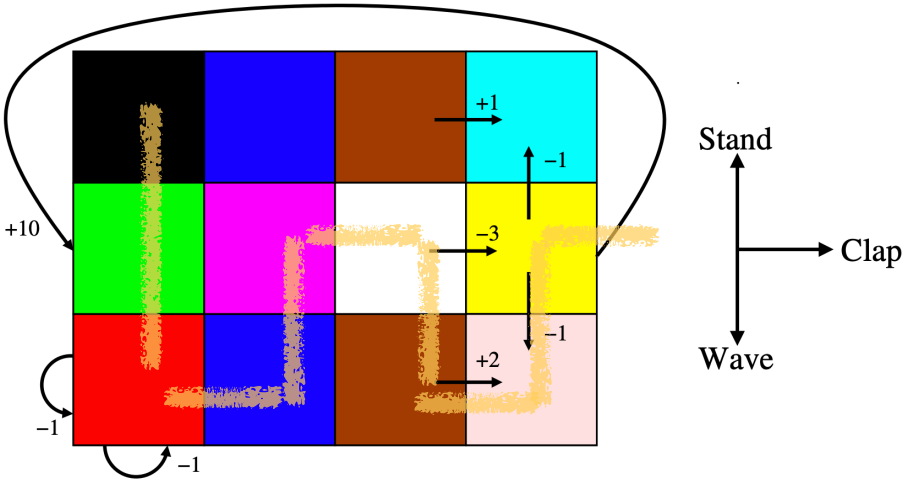
Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow$ average$(Returns(S_t))$



0 11.5	1 11	2 11	3 10
4 12	5 12	6 12	7 10
8 12	9 12	10 12	11 10

First visit vs. every visit

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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow \text{average}(Returns(S_t))$

First visit vs. every visit

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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

~~Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:~~

Append G to $Returns(S_t)$

$V(S_t) \leftarrow$ average$(Returns(S_t))$

First visit vs. every visit

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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

~~Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:~~

Append G to $Returns(S_t)$

$V(S_t) \leftarrow$ average$(Returns(S_t))$

Difference between DP and MC methods

DP methods:

- Require environment dynamics $p(s', r \mid s, a)$
- Difficult to acquire in practice

Difference between DP and MC methods

DP methods:

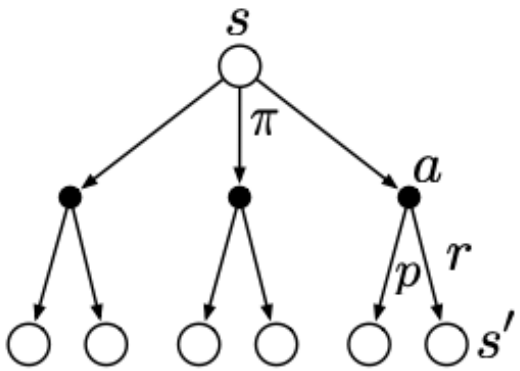
- Require environment dynamics $p(s', r \mid s, a)$
- Difficult to acquire in practice

MC methods:

- Don't need environment dynamics $p(s', r \mid s, a)$
- Only need environment samples!

Generalizing Backup Diagrams to Monte Carlo Algorithms

Whereas the DP diagram shows all possible transitions, the Monte Carlo diagram shows only those sampled on the one episode.



Backup diagram for v_π

Whereas the DP diagram includes only one-step transitions, the Monte Carlo diagram goes all the way to the end of the episode.

Draw the backup diagram for a Monte Carlo algorithm

Episode 1:

State, Action, Reward

0, Wave, 0

4, Wave, 0

8, Clap, 0

9, Stand, 0

5, Clap, 0

6, Wave, 0

10, Clap, 2

11, Stand, 0

7, Clap, 10

Draw the backup diagram for a Monte Carlo algorithm

Episode 1:

State, Action, Reward

0, Wave, 0

4, Wave, 0

8, Clap, 0

9, Stand, 0

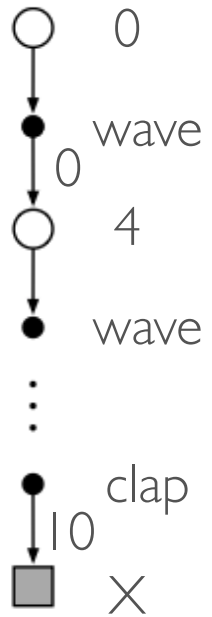
5, Clap, 0

6, Wave, 0

10, Clap, 2

11, Stand, 0

7, Clap, 10



Maintaining Exploration

What happens if a state-action pair is never visited?

Maintaining Exploration

What happens if a state-action pair is never visited?

For policy evaluation to work, we need to see all state-action pairs.

Exploring starts: every state-action pair has a nonzero probability of being selected as the start

Maintaining Exploration

What are other ways to ensure continual exploration?

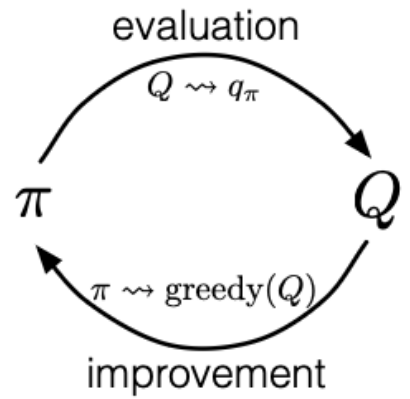
Maintaining Exploration

What are other ways to ensure continual exploration?

Consider only stochastic policies with nonzero probability of selecting all actions in each state

Monte Carlo Control

Generalized Policy Iteration (GPI): maintains both an approximate policy and an approximate value function. The value function is repeatedly altered to more closely approximate the value function for the current policy, and the policy is repeatedly improved with respect to the current value function.



Monte Carlo Control

Monte Carlo version of policy iteration:

$$\pi_0 \xrightarrow{\text{E}} q_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} q_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} q_*$$

Assumptions required for convergence?

Monte Carlo Control

Monte Carlo version of policy iteration:

$$\pi_0 \xrightarrow{\text{E}} q_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} q_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} q_*$$

Assumptions required for convergence?

Exploring starts
Infinite number of episodes for policy evaluation

Monte Carlo Control

Monte Carlo version of policy iteration:

$$\pi_0 \xrightarrow{\text{E}} q_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} q_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} q_*$$

Assumptions required for convergence?

Exploring starts

Infinite number of episodes for policy evaluation

Easy to remove - value iteration!

Monte Carlo Control

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 $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}(S_0)$ such that all pairs have probability > 0 (exploring starts)

Generate an episode starting 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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

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 $Returns(S_t, A_t)$

$Q(S_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

alternate between evaluation and improvement on an episode-by-episode basis

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 $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}(S_0)$ such that all pairs have probability > 0 (exploring starts)

Generate an episode starting 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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

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 $Returns(S_t, A_t)$

$Q(S_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

Episode 0:

State, Action, Reward

0, Clap, 0

1, Clap, 0

2, Clap, 1

3, Wave, 0

7, Clap, 10

	Stand	Clap	Wave
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0
11	0	0	0

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 $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}(S_0)$ such that all pairs have probability > 0 (exploring starts)

Generate an episode starting 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, $t = T-1, T-2, \ldots, 0$:

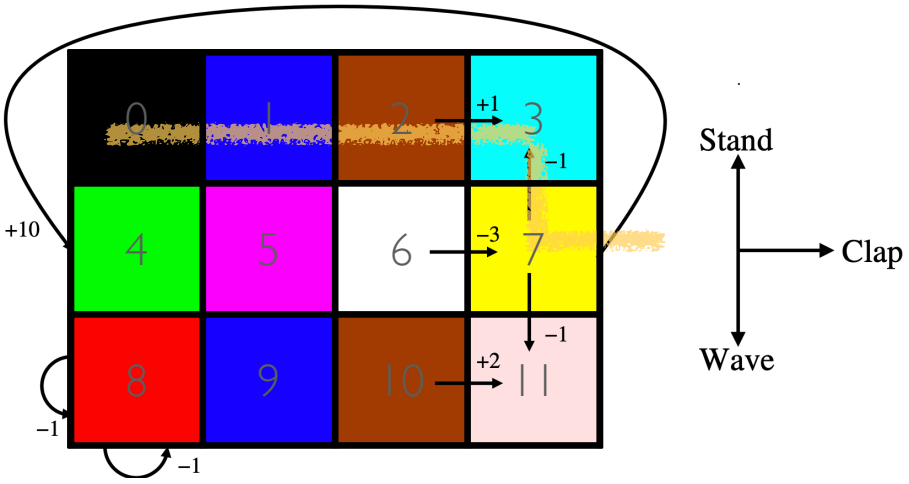
$G \leftarrow G + R_{t+1}$

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 $Returns(S_t, A_t)$

$Q(S_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$



	Stand	Clap	Wave
0	0	11	0
1	0	11	0
2	0	11	0
3	0	0	10
4	0	0	0
5	0	0	0
6	0	0	0
7	0	10	0
8	0	0	0
9	0	0	0
10	0	0	0
11	0	0	0

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 $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}(S_0)$ such that all pairs have probability > 0 (exploring starts)

Generate an episode starting 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, $t = T-1, T-2, \ldots, 0$:

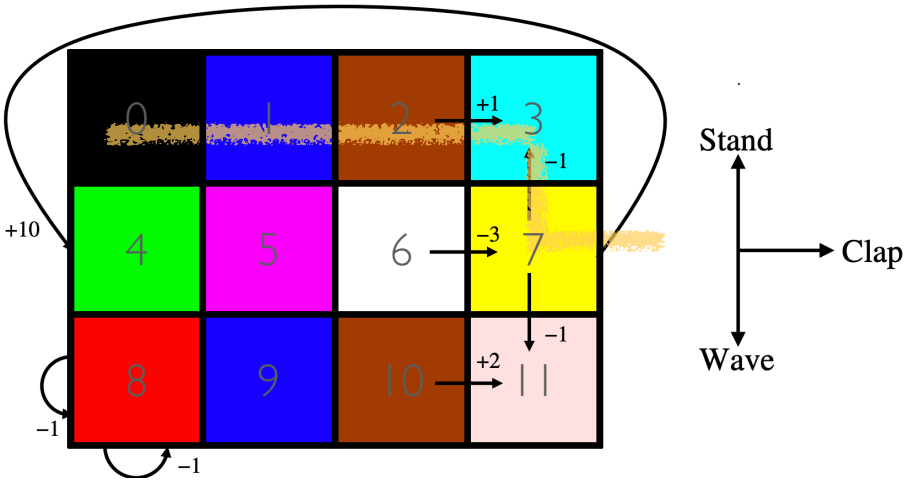
$G \leftarrow G + R_{t+1}$

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 $Returns(S_t, A_t)$

$Q(S_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$



Next episode starts from state 5

	Stand	Clap	Wave
0	0	11	0
1	0	11	0
2	0	11	0
3	0	0	10
4	0	0	0
5	0	0	0
6	0	0	0
7	0	10	0
8	0	0	0
9	0	0	0
10	0	0	0
11	0	0	0

How do we remove the exploring starts assumption?

How do we remove the exploring starts assumption?

On-policy methods:

Policy is generally *soft:* probability of taking any action at any state > 0.
Gradually shift closer and closer to deterministic optimal policy.

Ex: eps-greedy policies

Monte Carlo Control without Exploring Starts

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, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow G + R_{t+1}$

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 $a \in \mathcal{A}(S_t)$:

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

How do we remove the exploring starts assumption?

On-policy methods:

Policy is generally *soft:* probability of taking any action at any state > 0.
Gradually shift closer and closer to deterministic optimal policy.

Ex: eps-greedy policies

Policy iteration works with eps-soft policies!
-> removed the need for exploring starts.

How do we remove the exploring starts assumption?

The on-policy approach in the preceding section is actually a compromise—it learns action values not for the optimal policy, but for a near-optimal policy that still explores.

Off-policy methods:

More straightforward approach: two policies, one that is learned about and that becomes the optimal policy, and one that is more exploratory and is used to generate behavior.

Target policy: Policy being learned
Behavior policy: Policy used to generate behavior

Trade-offs of on-policy vs. off-policy

Trade-offs of on-policy vs. off-policy

On-policy method are simpler

Off-policy methods:
Because the data is due to a different policy, off-policy methods are often of greater variance and are slower to converge.

More powerful and general. They include on-policy methods as the special case in which the target and behavior policies are the same.

Off-policy Prediction via Importance Sampling

Simplest setting:

1. *prediction* problem
2. Both target and behavior policies are fixed

Required assumptions:

Coverage assumption: Every action taken under the target policy is also taken under the behavior policy.

Off-policy Prediction via Importance Sampling

Importance sampling: technique for estimating expected values under one distribution given samples from another.

Probability of a state-action trajectory under any policy pi:

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

What is the relative probability of the trajectory under target and behavior policies?

Off-policy Prediction via Importance Sampling

Importance sampling: technique for estimating expected values under one distribution given samples from another.

Probability of a state-action trajectory under any policy pi:

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

What is the relative probability of the trajectory under target and behavior policies?

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

Dynamics cancel out - doesn't depend on MDP!

Importance sampling ratio

Target policy: 0.1 for stand, 0.8 for clap, 0.1 for wave at each state.
Behavior policy: 0.8 for actions taken, 0.1 otherwise.

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

Episode 0:

State, Action, Reward
0, Clap, 0
1, Clap, 0
2, Clap, 1
3, Wave, 0
7, Clap, 10

t	rho
0	
1	
2	
3	
4	

Importance sampling ratio

Target policy: 0.1 for stand, 0.8 for clap, 0.1 for wave at each state.
Behavior policy: 0.8 for actions taken, 0.1 otherwise.

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

Episode 0:

State, Action, Reward
0, Clap, 0
1, Clap, 0
2, Clap, 1
3, Wave, 0
7, Clap, 10

t	rho
0	0.8/0.8 * 0.125 = 0.125
1	0.8/0.8 * 0.125 = 0.125
2	0.8/0.8 * 0.125 = 0.125
3	1 * (0.1/0.8) = 0.125
4	1

Off-policy Prediction via Importance Sampling

Recall: wish to estimate expected returns under target policy, but only have returns under behavior policy.

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

$$\mathbb{E}[\rho_{t:T-1} G_t \mid S_t] = v_\pi(S_t)$$

Computing this expectation in practice requires a scaling term

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

Number of steps

Ordinary importance sampling

Off-policy Prediction via Importance Sampling

Recall: wish to estimate expected returns under target policy, but only have returns under behavior policy.

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

$$\mathbb{E}[\rho_{t:T-1} G_t \mid S_t] = v_\pi(S_t)$$

Computing this expectation in practice requires a scaling term

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

Weighted importance sampling

Ordinary vs. weighted importance sampling

Target policy: 0.1 for stand, 0.8 for clap, 0.1 for wave at each state.
Behavior policy: 0.8 for actions taken, 0.1 otherwise.

Episode 0:

State, Action, Reward

0, Clap, 0

1, Clap, 0

2, Clap, 1

3, Wave, 0

7, Clap, 10

t	rho
0	0.125
1	0.125
2	0.125
3	0.125
4	1

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

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

Ordinary importance sampling

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

Weighted importance sampling

Ordinary vs. weighted importance sampling

Target policy: 0.1 for stand, 0.8 for clap, 0.1 for wave at each state.
Behavior policy: 0.8 for actions taken, 0.1 otherwise.

Episode 0:
State, Action, Reward
0, Clap, 0
1, Clap, 0
2, Clap, 1
3, Wave, 0
7, Clap, 10

t	rho
0	0.125
1	0.125
2	0.125
3	0.125
4	1

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

V(s=0) = 0.125 * 11 / 1 = 1.375

Ordinary importance sampling

V(s=0) = (0.125 * 11) / 0.04096 = 11

Weighted importance sampling

Incremental Implementation of 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, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$
        If $W = 0$ then exit For loop

Extension of incremental implementation for bandits (Section 2.4)
+ off-policy importance sampling

Off-policy Monte Carlo 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)
$C(s,a) \leftarrow 0$
$\pi(s) \leftarrow \arg\max_a Q(s,a)$ (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 For loop
$W \leftarrow W \frac{1}{b(A_t|S_t)}$

Add a step of policy improvement

A Summary

1. Sampling episodes over dynamic programming
2. On-policy vs. off-policy
3. Off-policy prediction
4. Ordinary vs. weighted importance sampling

Next Time

Incorporate bootstrapping through temporal-difference learning - a combination of MC and TD!

Final Logistics

Next lecture:
Chapter 6: Temporal-Difference Learning
Reading assignments due **2PM Monday**

Office hours:

Mon: Michael 1-2PM GDC Basement TA Station #5
Tues: Peter 11-11:50AM GDC 3.508; Caroline 11:15-12:15PM
Wed: Amy 2-3PM EER 6.878
Thurs: Haoran 11-12PM; Siddhant 5-6PM
Fri: Shuozhe 4-5PM

Final Logistics

Coding assignment for Chapter 2 on edX **due Feb 4th 11:59PM CST**

Chapter 5+6 Homework on edX **due Friday 11:59PM CST**