Lecture 14 - Feb 28, 2023

- Monte-Carlo Methods
  - First-Visit MC
  - Online MC

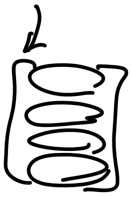
- Temporal Difference Learning
  - TD(0)
  - SARSA
  - Q-Learning
  - On-Policy Vs. Off-Policy

HW3 is posted → Due March 17

Project 2 → Due March 5

TA's office hour: Wendsdays, 2pm-3pm (in-person)
Fridays, 2pm-3pm (virtual)

$$s_k = \overline{C \; s_{k-1}} \oplus a_{k-1} \oplus n_k$$ ⇐ State Space Modeling

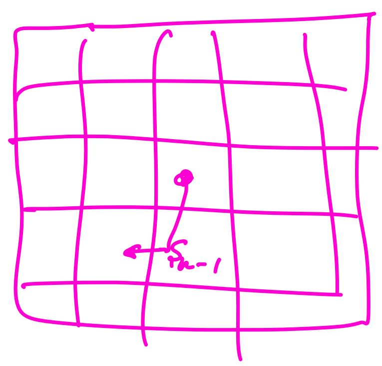


$$s^1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad s^2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad s^{16} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$s_k = s^j$ j

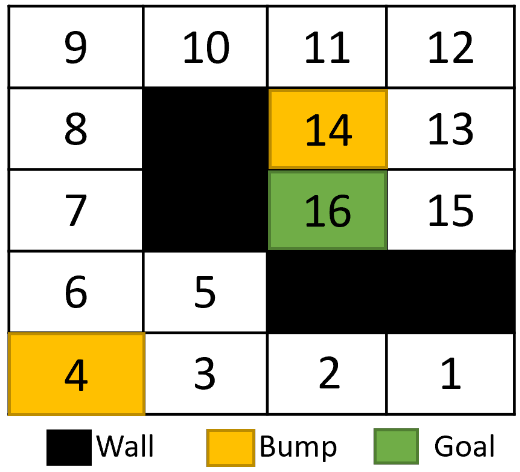
$$(M(a))_{ij} = \begin{bmatrix} & & & \\ - & - & \boxed{\cdot} & \\ & & & \end{bmatrix}_{16\times 16}$$

i $s_{k-1} = s^i$



Monte-Carlo Methods:

$MDP(S, A, \cancel{T}, r)$



$$V_{k+1} = \max_{a \in A} \sum_{s'} P(s'|s,a)\left[R + \gamma V_k(s')\right]$$

$$s_{k-1}, a_{k-1} \rightarrow \boxed{\text{Simulator System}} \rightarrow s_k, r \quad \Rightarrow \quad P(s'|s,a)$$

$$R(s,a,s')$$

$$\pi^*(s) \leftarrow a$$

$$V^\pi(s) = E\left[\overbrace{R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots}^{G_t} \mid S_t = s, \pi\right]$$

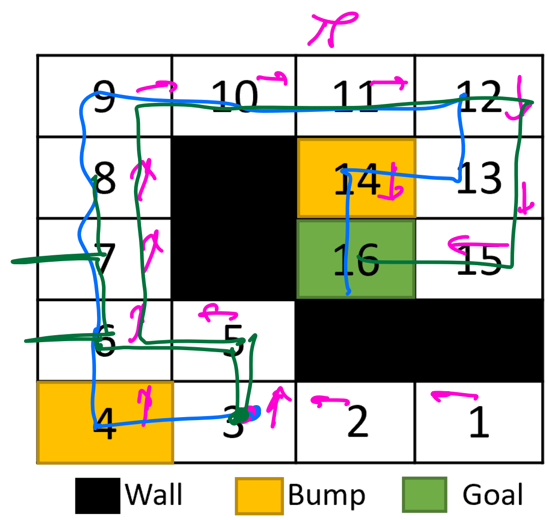
π^* $\quad V^{\pi^*}(s) \geq V^{\pi}(s)$ for all π ✓✓

Bellman Eq

$$V^\pi(s) = E\left[R_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s, \pi\right]$$

$$= \sum_{s'} P(s' \mid s, \pi(s)) \left[R + \gamma V^\pi(s')\right]$$

$S_t = 3$



$$\underbrace{-11, -1, -1, \ldots, -1, -11, 99}_{} \quad G_t$$

$$V^\pi(s) = E\left[G_t \mid S_t = s, \pi\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} G_t^{i}$$

A	B

$\pi = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix} \Longrightarrow V^\pi$

Unknown $\left\{ \begin{array}{l} M(a^1) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \\ M(a^2) = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \end{array} \right.$

episode 1

Random

$A, \pi(A) = a^2, A, r = -1$

$A, \pi(A) = a^2, B, r = 4$

$B, \pi(B) = a^1, B, r = 5$ G(B)

$B, \pi(B) = a^1, B, r = 5$

$B, \pi(B) = a^1, A, r = 0$

$A, \pi(A) = a^2, B, r = 4$

A < 5, −1 B a^2

episod 2

$B, \pi(B) = a^1, B, r = 5$

$B, \pi(B) = a^1, A, r = 0$

$A, \pi(A) = a^2, B, r = 4$

First visit of A

$B, \pi(B) = a^1, B, r = 5$

$B, \pi(B) = a^1, B, r = 5$

$B, \pi(B) = a^1, B, r = 5$

$B, \pi(B) = a^1, B, r = 5$

$G^1(A) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$= -1 + \gamma 4 + \gamma^2 5 + \gamma^3 5 + \gamma^4 0 + \gamma^5 4 =$

$G^2(A) = 4 + \gamma 5 + \gamma^2 5 + \gamma^3 5 + \gamma^4 5$

$G^2(B) = 5 + \gamma 0 + \gamma^2 4 + \gamma^3 5 + \cdots$

$G(B) = 5 + \gamma 5 + \gamma^2 0 + \gamma^3 4$

$V^\pi(A) \approx G^1_A$

$V^\pi(B) \approx G^1_B$

$V^\pi(A) = \dfrac{G^1(A) + G^2(A)}{2}$

$V^\pi(B) = \dfrac{G^1(B) + G^2(B)}{2}$

$$\underset{\text{new estimate}}{\underline{V(s)}} = \underset{\text{old estimate}}{V(s)} + \alpha \left[G(s) - V(s) \right]$$

$$\alpha = \frac{1}{n}$$

No applicability

First-Visit Monte Carlo Policy Evaluation

Initialize:

- $\pi \leftarrow$ policy to be evaluated
- $V \leftarrow$ an arbitrary state-value function
- $Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Repeat forever:

- Generate an episode using π
- For each state s appearing in the episode:
  - $G \leftarrow$ return following the first occurrence of s
  - Append G to $Returns(s)$
  - $V(s) \leftarrow \text{average}(Returns(s))$

$\boxed{A \mid B}$ $\quad A = \{a^1, a^2\}$

$|A|^{|S|} \leftarrow$ policies

$\pi^1 = \begin{bmatrix} a^1 \\ a^1 \end{bmatrix} \quad \pi^2 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix} \quad \pi^3 = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix} \quad \pi^4 = \begin{bmatrix} a^2 \\ a^2 \end{bmatrix}$

$$\underbrace{V^{\pi^1}\begin{bmatrix} \bigcirc \\ \bigcirc \end{bmatrix} \quad V^{\pi^2} \quad V^{\pi^3} \quad V^{\pi^2}}_{\pi^*}$$

Policy Iteration

$$\pi^0 \longrightarrow V^{\pi^0} \longrightarrow \pi^1 \longrightarrow V^{\pi^1} \cdots \longrightarrow \pi^T \qquad \underbrace{\pi^{T+1} \quad \pi^{T}}_{\pi^*}$$

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}} \underbrace{\sum_{s'} \underbrace{P(s'|s,a)}_{\text{unknown}} \left[R + \gamma V_{\pi}(s') \right]}_{Q_{\pi}(s,a)}$$

$$s_t \xrightarrow{\;a\;} s_{t+1} \xrightarrow{\;\pi(s_{t+1})\;} s_{t+2} \xrightarrow{\;\pi(s_{t+2})\;} \bigcirc \cdots$$

$\Longrightarrow$ our Goal is to estimate $Q_{\pi}(s,a)$ Using MC

$\boxed{A \mid B}$

$\pi = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$

$Q_\pi(A, a^1)$ $\quad$ $Q_\pi(B, a^1)$

$Q_\pi(A, a^2)$ $\quad$ $Q_\pi(B, a^2)$

Start with all pairs of s, a

$A, a^1, A, r=0$

follow policy π

$A, \pi(A)=a^2, B, r=4$

$B, \pi(B)=a^1, B, r=5$

$\vdots$

$A, a^2, B, r=4$

$B, \pi(B)=a^1, A, r=0$

$A, \pi(A)=a^2, B, r=4$

$\vdots$

$B, a^1, B, r=5$

$B, \pi(B)=a^1, B, 5$

$\vdots$

$B, a^2, A, -1$

$A, \pi(A)=a^2, B, 4$

$\vdots$

$G_{A,a^1} = 0+\gamma 4+\gamma^2 5+\dots$

$Q^\pi(A, a^1) \approx G_{A,a^1}$

$G^1_{A,a^2} = 4+\gamma 0+\gamma^2 4$

$G^2_{A,a^2} = 4+\gamma 5+\gamma^2 \dots$

$G^3_{A,a^2} = 4+\gamma \dots$

$Q^\pi(A, a^2) = \dfrac{G^1_{A,a^2} + G^2_{A,a^2} + G^3_{A,a^2}}{3}$

$Q^\pi(A, a^1)$ $\quad$ $Q_\pi(B, a^1)$

$Q^\pi(A, a^2)$ $\quad$ $Q_\pi(B, a^2)$

$$\pi^0 \xrightarrow{MC} Q^{\pi}(s,a) \longrightarrow \pi^1$$

$$\pi^1(s) = \underset{a \in A}{\arg\max}\ Q_{\pi^0}(s,a)$$

$$\pi^1(A) = \arg\max\ \underbrace{Q_{\pi}(A,a)}_{\{Q_{\pi}(A,a_1),\ Q_{\pi}(A,a_2)\}}$$

Monte Carlo Policy Iteration

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

Fixed point is optimal policy π^*

Now proven (almost)

Repeat forever:
Choose $S_0 \in \mathcal{S}$ and $A_0 \in \mathcal{A}(S_0)$ s.t. all pairs have probability > 0
Generate an episode starting from S_0, A_0, following π
For each pair s, a appearing in the episode:
$G \leftarrow$ return following the first occurrence of s, a
Append G to $Returns(s, a)$
$Q(s, a) \leftarrow \text{average}(Returns(s, a))$
For each s in the episode:
$\pi(s) \leftarrow \arg\max_a Q(s, a)$

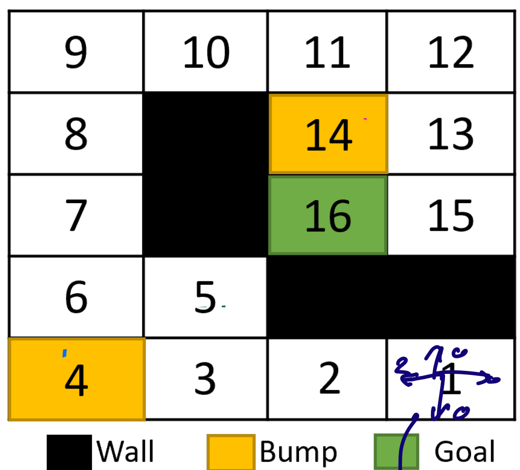
Online Monte Carlo Algorithm

$$\pi(a \mid s)$$

Softmax Policy

Randomness

$$\pi^{\varepsilon\text{-greedy}}(s) = \begin{cases} \underset{a \in A}{\operatorname{argmax}}\ Q(s, a) & 1-\varepsilon \\ \text{Random}\{a^1, a^2, \dots, a^L\} & \varepsilon \end{cases}$$



$$\pi(s) = \begin{cases} a^* & 1-\varepsilon+\dfrac{\varepsilon}{|A|} \\ a \neq a^* & \dfrac{\varepsilon}{|A|} \end{cases}$$

$$a^* = L = \operatorname{argmax}\ Q(s=1, a)$$

$\varepsilon = 0.1$

$|A| = 4$

$$\pi(s) = \begin{cases} a^* & 0.925 \\ a \neq a^* & \dfrac{0.1}{4} = 0.025 \quad a \in A - \{a^*\} \end{cases}$$

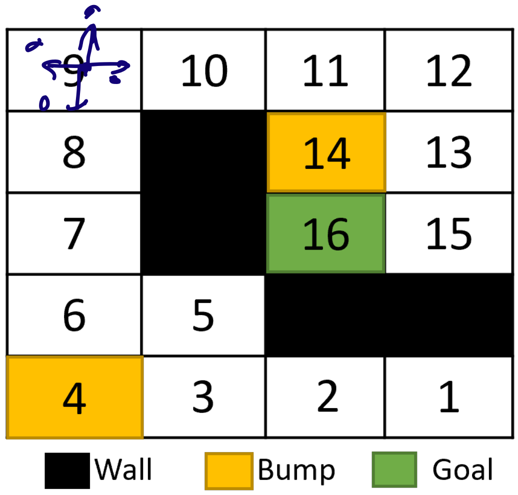
$a^* = \operatorname{argmax} Q(s,a)$

$\pi^{\varepsilon g}(s) = 0.25 \quad \text{all } a \in A$

$a^1 \; a^2 \; a^3 \; a^4$

$\pi^{\varepsilon\text{-greedy}} = \begin{cases} \text{greedy} & 1-\varepsilon \\ \text{random} & \varepsilon \end{cases}$

$\hookrightarrow a^1 \; a^2 \; a^3 \; a^4$



$Q(s,a) = 0$

Random State

$A, \; \pi^{\varepsilon\text{-greedy}}(A) = a^1, \; A, \; r=0 \longrightarrow Q(A,a^1) = 0 + \gamma 4 + \gamma^2 5 + \gamma^3 (-1) + \dots$

$A, \; \pi^{\varepsilon g}(A) = a^2, \; B, \; r=4 \longrightarrow Q(A,a^2) = \_\_ \quad \_\_$

$B, \; \pi^{\varepsilon\text{-greedy}}(B) = a^1, \; B, \; r=5 \qquad Q(B,a) =$

$B, \; \pi(B) = a^3, \; A, \; r=-1 \qquad Q(B,a^2) = \_\_$

$\vdots$

$$\pi^{\text{new } \varepsilon\text{-greedy}}(s) = \begin{cases} \operatorname{argmax} Q(s,a) = a^*_s & 1-\varepsilon+\frac{\varepsilon}{|A|} \\ a \neq a^*_s & \frac{\varepsilon}{|A|} \end{cases}$$

On-Policy Monte Carlo Control

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:

$Q(s, a) \leftarrow$ arbitrary

$Returns(s, a) \leftarrow$ empty list

$\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:

(a) Generate an episode using π

(b) For each pair s, a appearing in the episode:

$G \leftarrow$ return following the first occurrence of s, a

Append G to $Returns(s, a)$

$Q(s, a) \leftarrow$ average($Returns(s, a)$)

(c) For each s in the episode:

$A^* \leftarrow \arg\max_a Q(s, a)$

For all $a \in \mathcal{A}(s)$:

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