10-301/601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Reinforcement Learning: Value Iteration & Policy Iteration

Matt Gormley & Henry Chai

Lecture 22

Nov. 10, 2021

Reminders

- **Homework 7: HMMs**
  - Out: Wed, Nov. 03
  - Due: Fri, Nov. 12 at 11:59pm
- **Homework 8: RL**
  - Out: Fri, Nov. 12
  - Due: Sun, Nov. 21 at 11:59pm

Markov Decision Processes (MDPs)

- In RL, the model for our data is an MDP:

1. Start in some initial state s_0
2. For time step t:
   1. Agent observes state s_t
   2. Agent takes action $a_t = \pi(s_t)$
   3. Agent receives reward $r_t = R(s_t, a_t)$
   4. Agent transitions to state $s_{t+1} \sim p(s' \mid s_t, a_t)$
3. Total reward is $\sum_{t=0}^{\infty} \gamma^t r_t$

- Makes the same Markov assumption we used for HMMs! The next state only depends on the current state and action.

MDP Example: Multi-armed bandit

- Single state:
  $$|\mathcal{S}| = 1$$
- Three actions:
  $$\mathcal{A} = \{1, 2, 3\}$$
- Rewards are stochastic

MDP Example: Multi-armed bandit

Bandit 1	Bandit 2	Bandit 3
1	???	???
1	???	???
1	???	???
???	???	???
???	???	???
???	???	???
???	???	???
???	???	???
???	???	???
???	???	???
???	???	???
???	???	???

RL: Value Function

- $V^{\pi}(s) = \mathbb{E}[$discounted total reward of starting in state s and *executing* policy π forever$]$

$$= \mathbb{E}[R(s_0, \pi(s_0)) + \gamma R(s_1, \pi(s_1)) + \gamma^2 R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$\sim p(s' \mid s, a)$

$$= R(s, \pi(s)) + \gamma \mathbb{E}[R(s_1, \pi(s_1)) + \gamma R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$$= R(s, \pi(s)) + \gamma \sum_{s_1 \in \mathcal{S}} p(s_1 \mid s, \pi(s)) \big(R(s_1, \pi(s_1)) + \gamma \mathbb{E}[R(s_2, \pi(s_2)) + \cdots \mid s_1]\big)$$

RL: Value Function

- $V^{\pi}(s) = \mathbb{E}[$discounted total reward of starting in state s and *executing* policy π forever$]$

$$= \mathbb{E}[R(s_0, \pi(s_0)) + \gamma R(s_1, \pi(s_1)) + \gamma^2 R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$$= R(s, \pi(s)) + \gamma \mathbb{E}[R(s_1, \pi(s_1)) + \gamma R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$$= R(s, \pi(s)) + \gamma \sum_{s_1 \in \mathcal{S}} p(s_1 \mid s, \pi(s)) (R(s_1, \pi(s_1)) + \gamma \mathbb{E}[R(s_2, \pi(s_2)) + \cdots \mid s_1])$$

$0 < \gamma < 1$

RL: Value Function

- $V^{\pi}(s) = \mathbb{E}[$discounted total reward of starting in state s and *executing* policy π forever$]$

$$= \mathbb{E}[R(s_0, \pi(s_0)) + \gamma R(s_1, \pi(s_1)) + \gamma^2 R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$$= R(s, \pi(s)) + \gamma \mathbb{E}[R(s_1, \pi(s_1)) + \gamma R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$$= R(s, \pi(s)) + \gamma \sum_{s_1 \in \mathcal{S}} p(s_1 \mid s, \pi(s))(R(s_1, \pi(s_1)) + \gamma \mathbb{E}[R(s_2, \pi(s_2)) + \cdots \mid s_1])$$

RL: Value Function

- $V^{\pi}(s) = \mathbb{E}[$discounted total reward of starting in state s and *executing* policy π forever$]$

$$= \mathbb{E}[R(s_0, \pi(s_0)) + \gamma R(s_1, \pi(s_1)) + \gamma^2 R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$$= R(s, \pi(s)) + \gamma \mathbb{E}[R(s_1, \pi(s_1)) + \gamma R(s_2, \pi(s_2)) + \cdots \mid s_0 = s]$$

$$= R(s, \pi(s)) + \gamma \sum_{s_1 \in \mathcal{S}} p(s_1 \mid s, \pi(s)) \big(R(s_1, \pi(s_1)) + \gamma \mathbb{E}[R(s_2, \pi(s_2)) + \cdots \mid s_1]\big)$$

RL: Value Function

- $V^{\pi}(s) = \mathbb{E}[$discounted total reward of starting in state s and *executing* policy π forever$]$

$$= \mathbb{E}\big[R\big(s_0, \pi(s_0)\big) + \gamma R\big(s_1, \pi(s_1)\big) + \gamma^2 R\big(s_2, \pi(s_2)\big) + \cdots \mid s_0 = s\big]$$

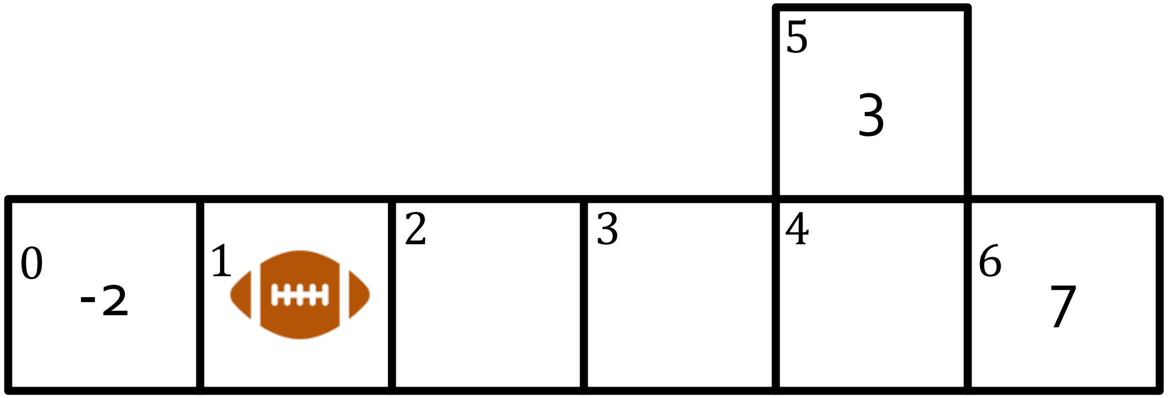
$$= R\big(s, \pi(s)\big) + \gamma \mathbb{E}\big[R\big(s_1, \pi(s_1)\big) + \gamma R\big(s_2, \pi(s_2)\big) + \cdots \mid s_0 = s\big]$$

$$= R\big(s, \pi(s)\big) + \gamma \sum_{s_1 \in \mathcal{S}} p\big(s_1 \mid s, \pi(s)\big)\Big(R\big(s_1, \pi(s_1)\big) + \gamma \mathbb{E}\big[R\big(s_2, \pi(s_2)\big) + \cdots \mid s_1\big]\Big)$$

$$V^{\pi}(s) = R\big(s, \pi(s)\big) + \gamma \sum_{s_1 \in \mathcal{S}} p\big(s_1 \mid s, \pi(s)\big) V^{\pi}(s_1)$$

Bellman equations

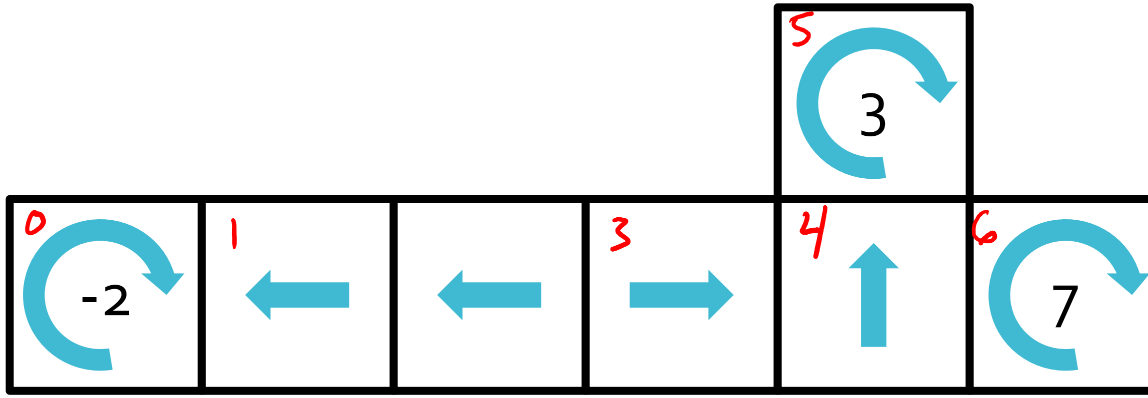
RL: Value Function Example



$$R(s,a) = \begin{cases} -2 \text{ if entering state 0 (safety)} \\ 3 \text{ if entering state 5 (field goal)} \\ 7 \text{ if entering state 6 (touch down)} \\ 0 \text{ otherwise} \end{cases}$$

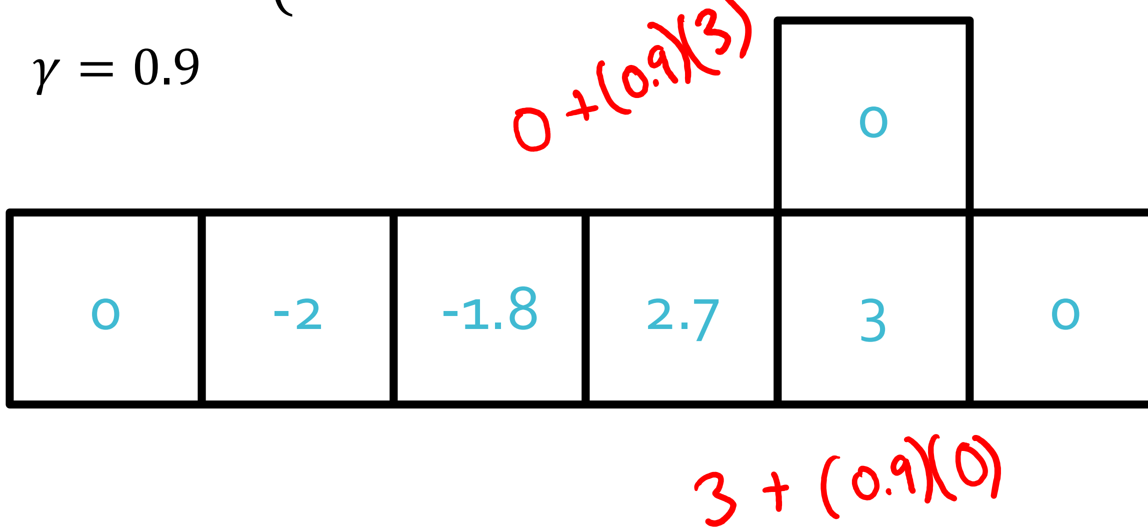
$$\gamma = 0.9$$

RL: Value Function Example

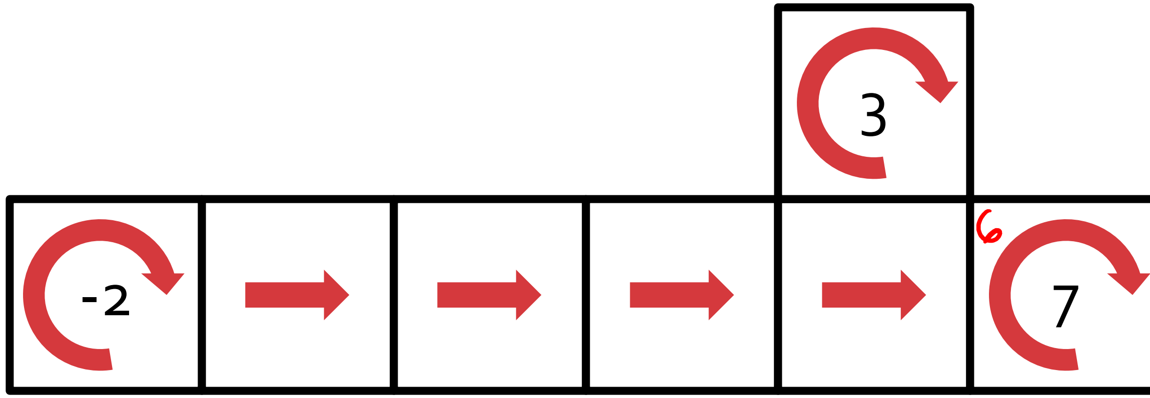


$$R(s,a) = \begin{cases} -2 \text{ if entering state 0 (safety)} \\ 3 \text{ if entering state 5 (field goal)} \\ 7 \text{ if entering state 6 (touch down)} \\ 0 \text{ otherwise} \end{cases}$$

$\gamma = 0.9$



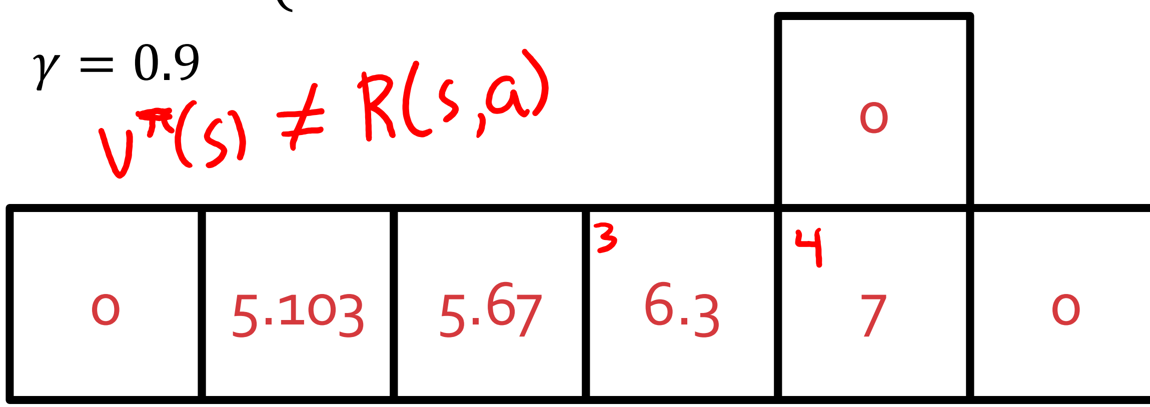
RL: Value Function Example



$$R(s,a) = \begin{cases} -2 \text{ if entering state 0 (safety)} \\ 3 \text{ if entering state 5 (field goal)} \\ 7 \text{ if entering state 6 (touch down)} \\ 0 \text{ otherwise} \end{cases}$$

$\gamma = 0.9$

$V^{\pi}(s) \neq R(s,a)$



$0 + 0.9(7)$ $\quad 7 + (0.9)(0)$

RL: Optimal Value Function & Policy

- Optimal value function:

$$V^*(s) = \max_{a \in \mathcal{A}} R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) V^*(s')$$

  - System of $|\mathcal{S}|$ equations and $|\mathcal{S}|$ variables

- Optimal policy:

$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \underbrace{R(s,a)}_{\text{Immediate reward}} + \underbrace{\gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) V^*(s')}_{\text{(Discounted) Future reward}}$$

Fixed Point Iteration

- Iterative method for solving a system of equations
- Given some equations and initial values

$$x_1 = f_1(x_1, \dots, x_n)$$
$$\vdots$$
$$x_n = f_n(x_1, \dots, x_n)$$
$$x_1^{(0)}, \dots, x_n^{(0)}$$

- While not converged, do

$$x_1^{(t+1)} \leftarrow f_1\left(x_1^{(t)}, \dots, x_n^{(t)}\right)$$
$$\vdots$$
$$x_n^{(t+1)} \leftarrow f_n\left(x_1^{(t)}, \dots, x_n^{(t)}\right)$$

Fixed Point Iteration: Example

$-1/6 + 1/2 = 1/3$ $\quad\quad$ $\left(-\frac{3}{2}\right)\left(\frac{1}{3}\right) = -\frac{1}{2}$

$$x_1 = x_1 x_2 + \frac{1}{2} \qquad x_2 = -\frac{3x_1}{2}$$

$x_1^{(1)} = 0 \cdot 0 + 1/2 = 1/2$ $\quad$ $x_2^{(1)} = -\frac{3}{2}(0) = 0$

$$x_1^{(0)} = x_2^{(0)} = 0$$

$$x_1 = \frac{1}{3}, x_2 = -\frac{1}{2}$$

t	$x_1^{(t)}$	$x_2^{(t)}$
0	0	0
1	0.5	0
2	0.5	-0.75
3	0.125	-0.75
4	0.4063	-0.1875
5	0.4238	-0.6094
6	0.2417	-0.6357
7	0.3463	-0.3626
8	0.3744	-0.5195
9	0.3055	-0.5616
10	0.3284	-0.4582
11	0.3495	-0.4926
12	0.3278	-0.5243
13	0.3281	-0.4917
14	0.3386	-0.4922
15	0.3333	-0.5080
	$\approx 1/3$	$\approx -1/2$

Value Iteration

- Inputs: reward function $R(s,a)$,

  transition probabilities $p(s' \mid s,a)$

- Initialize $V^{(0)}(s) = 0 \ \forall\, s \in \mathcal{S}$ (or randomly) and set $t = 0$
- While not converged, do:
  - For $s \in \mathcal{S}$

$$V^{(t+1)}(s) \leftarrow \max_{a \in \mathcal{A}} \underbrace{R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) V^{(t)}(s')}_{Q(s,a)}$$

$S = \{1, 2, \ldots, |S|\}$

  - $t = t + 1$

- For $s \in \mathcal{S}$

$$\pi^*(s) \leftarrow \operatorname*{argmax}_{a \in \mathcal{A}} R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) V^{(t)}(s')$$

- Return π^*

Value Iteration

- Inputs: reward function $R(s, a)$,

  transition probabilities $p(s' \mid s, a)$
- Initialize $V^{(0)}(s) = 0 \; \forall \; s \in \mathcal{S}$ (or randomly) and set $t = 0$
- While not converged, do:
  - For $s \in \mathcal{S}$
    - For $a \in \mathcal{A}$

      $$Q(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V^{(t)}(s')$$
    - $V^{(t+1)}(s) \leftarrow \max_{a \in \mathcal{A}} Q(s, a)$
  - $t = t + 1$
- For $s \in \mathcal{S}$

  $$\pi^*(s) \leftarrow \operatorname*{argmax}_{a \in \mathcal{A}} R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V^{(t)}(s')$$
- Return π^*

Asynchronous Value Iteration

- Inputs: reward function $R(s, a)$,

  transition probabilities $p(s' \mid s, a)$
- Initialize $V(s) = 0 \; \forall \; s \in \mathcal{S}$ (or randomly)
- While not converged, do:
  - For $s \in \mathcal{S}$
    - For $a \in \mathcal{A}$

      $$Q(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V(s')$$
    - $V(s) \leftarrow \max_{a \in \mathcal{A}} Q(s, a)$
- For $s \in \mathcal{S}$

  $$\pi^*(s) \leftarrow \operatorname*{argmax}_{a \in \mathcal{A}} R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V(s')$$
- Return π^*

Poll Q1: How much computation does one iteration require?

- Inputs: reward function $R(s, a)$,

  transition probabilities $p(s' \mid s, a)$
- Initialize $V(s) = 0 \; \forall \; s \in \mathcal{S}$ (or randomly)
- While not converged, do:
  - For $s \in \mathcal{S}$
    - For $a \in \mathcal{A}$

      $$Q(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V(s')$$
    - $V(s) \leftarrow \max_{a \in \mathcal{A}} Q(s, a)$

$O(|S||A||S| + |S||A|)$

$= O(|S|^2|A|)$

- For $s \in \mathcal{S}$

  $$\pi^*(s) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V(s')$$
- Return π^*

When poll is active, respond at **pollev.com/10301601polls**

Question 1

A

B

C

D

E

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Value Iteration: Theory

- **Theorem 1**: Value function convergence

  V will converge to V^* if each state is "visited" infinitely often (Bertsekas, 1989)

- **Theorem 2**: Convergence criterion

  if $\max_{s \in \mathcal{S}} \left|V^{(t+1)}(s) - V^{(t)}(s)\right| < \epsilon$, then

  $\max_{s \in \mathcal{S}} \left|V^{(t+1)}(s) - V^{*}(s)\right| < \frac{2\epsilon\gamma}{1-\gamma}$ (Williams & Baird, 1993)

- **Theorem 3**: Policy convergence

  The "greedy" policy, $\pi(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q(s, a)$, converges to the optimal π^* in a finite number of iterations, often before the value function has converged! (Bertsekas, 1987)

Policy Iteration

- Inputs: reward function $R(s, a)$,
  transition probabilities $p(s' \mid s, a)$
- Initialize π randomly
- While not converged, do:
  - Solve the Bellman equations defined by policy π

$$V^{\pi}(s) = R\big(s, \pi(s)\big) + \gamma \sum_{s' \in \mathcal{S}} p\big(s' \mid s, \pi(s)\big) V^{\pi}(s')$$

  - Update π

$$\pi(s) \leftarrow \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ R(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V^{\pi}(s')$$

- Return π

system of $|S|$ linear equations
$|S|$ variables

$|S||A||S| = |S|^2|A|$

Policy Iteration: Theory

- Poll Q2: Given finite state and action spaces, how many possible policies are there?

When poll is active, respond at **pollev.com/10301601polls**

Question 2

A

B

C

D

E

Policy Iteration: Theory

- Poll Q2: Given finite state and action spaces, how many possible policies are there?

$|A| \cdot |A| \cdots |A| = |A|^{|S|}$

- In policy iteration, the policy improves in each iteration. Thus, the number of iterations needed to converge is bounded!
- Value iteration takes $O(|\mathcal{S}|^2|\mathcal{A}|)$ time / iteration
- Policy iteration takes $O(|\mathcal{S}|^2|\mathcal{A}| + |\mathcal{S}|^3)$ time / iteration
  - However, empirically policy iteration requires fewer iterations to converge

RL Learning Goals: Value & Policy Iteration

a. Compare the reinforcement learning paradigm to other learning paradigms
b. Cast a real-world problem as a Markov Decision Process
c. Depict the exploration vs. exploitation tradeoff via MDP examples
d. Explain how to solve a system of equations using fixed point iteration
e. Define the Bellman equations
f. Show how to compute the optimal policy in terms of the optimal value function
g. Explain the relationship between a value function mapping states to expected rewards and a value function mapping state-action pairs to expected rewards
h. Implement value iteration
i. Implement policy iteration
j. Contrast the computational complexity and empirical convergence of value iteration vs. policy iteration
k. Identify the conditions under which the value iteration algorithm will converge to the true value function
l. Describe properties of the policy iteration algorithm

Q:

What can we do if we don’t know the reward function / transition probabilities?