lecture 3

Policy evaluation

1. DP_Dynamic Programming
2. MC_Monte Carlo
3. TD_Temporal Difference

lecture 3

Policy evaluation

1. DP_Dynamic Programming

transition model

reward model

= model given

model used to
compute one timestep

$$V_k^\pi(s) = r(s,\pi(s)) + \gamma \sum_{s' \in S} p(s'|s,\pi(s)) V_{k-1}^\pi(s')$$

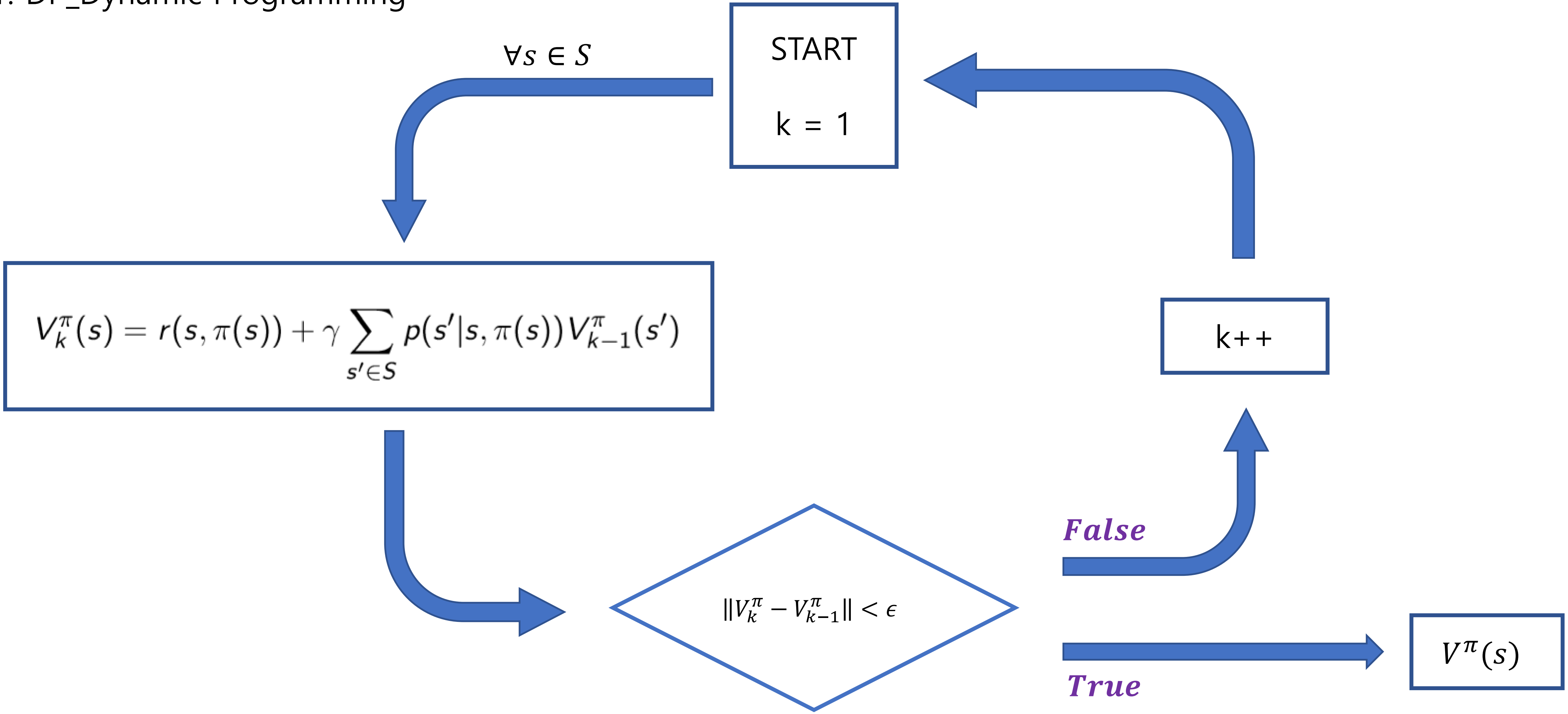
<u>Bootstrapping :</u>

1. Using the value estimate for $V_{k-1}^\pi(s')$
2. Using previously calculated $V_{k-1}^\pi(s')$
3. The expected value starting from state s' is identical no matter what timestep it starts from when considering infinite horizon

lecture 3

Policy evaluation

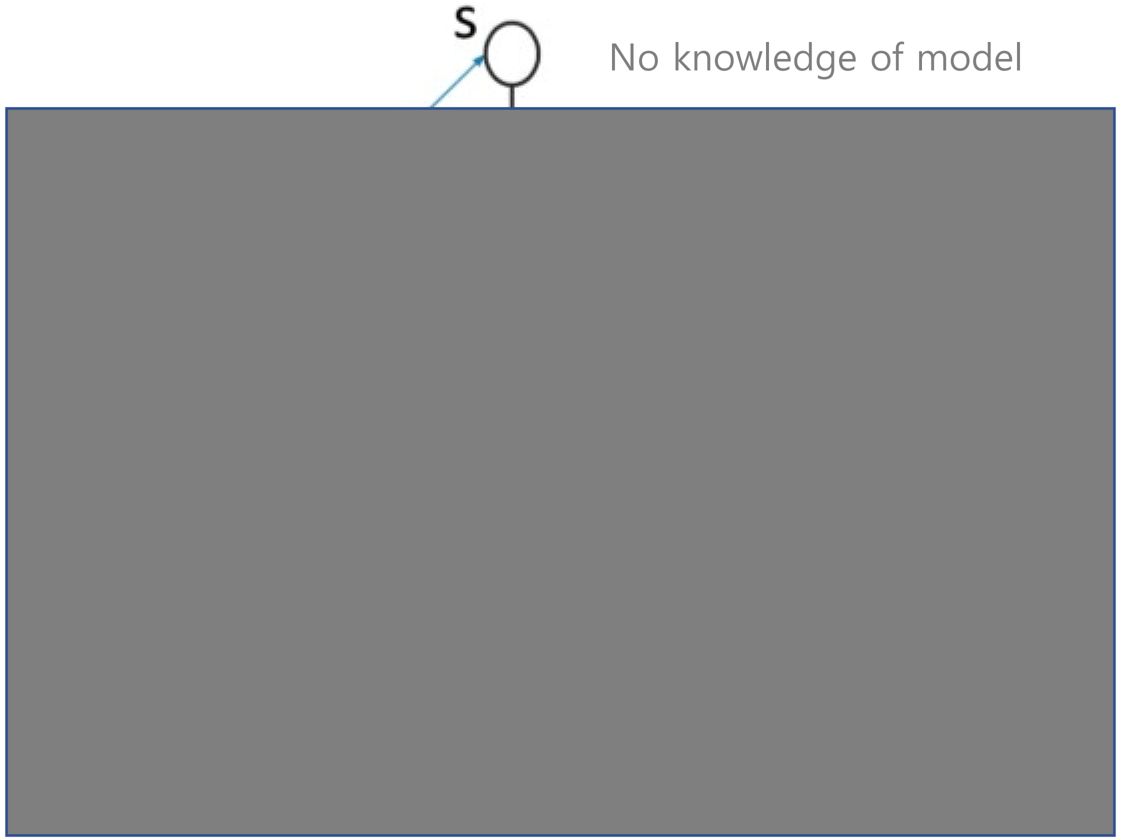
1. DP_Dynamic Programming



lecture 3

Policy evaluation

2. MC_Monte Carlo

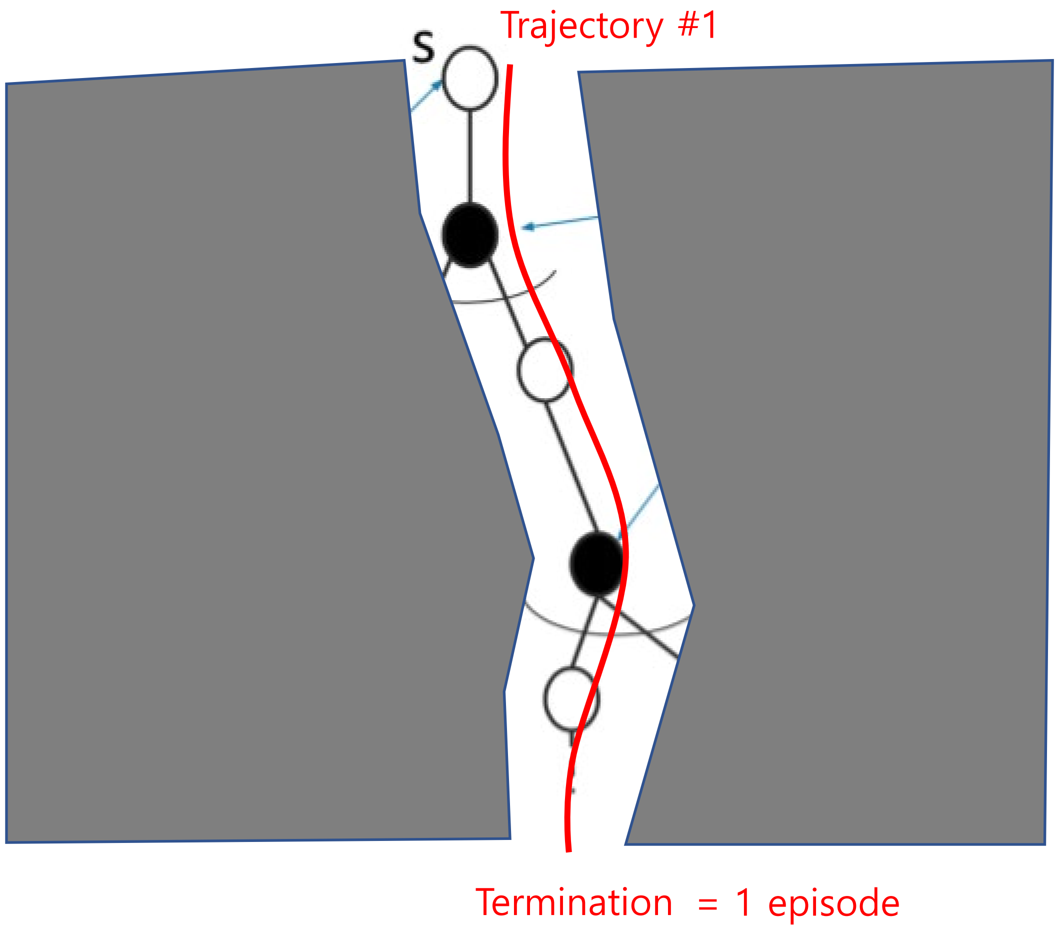


No knowledge of model

lecture 3

Policy evaluation

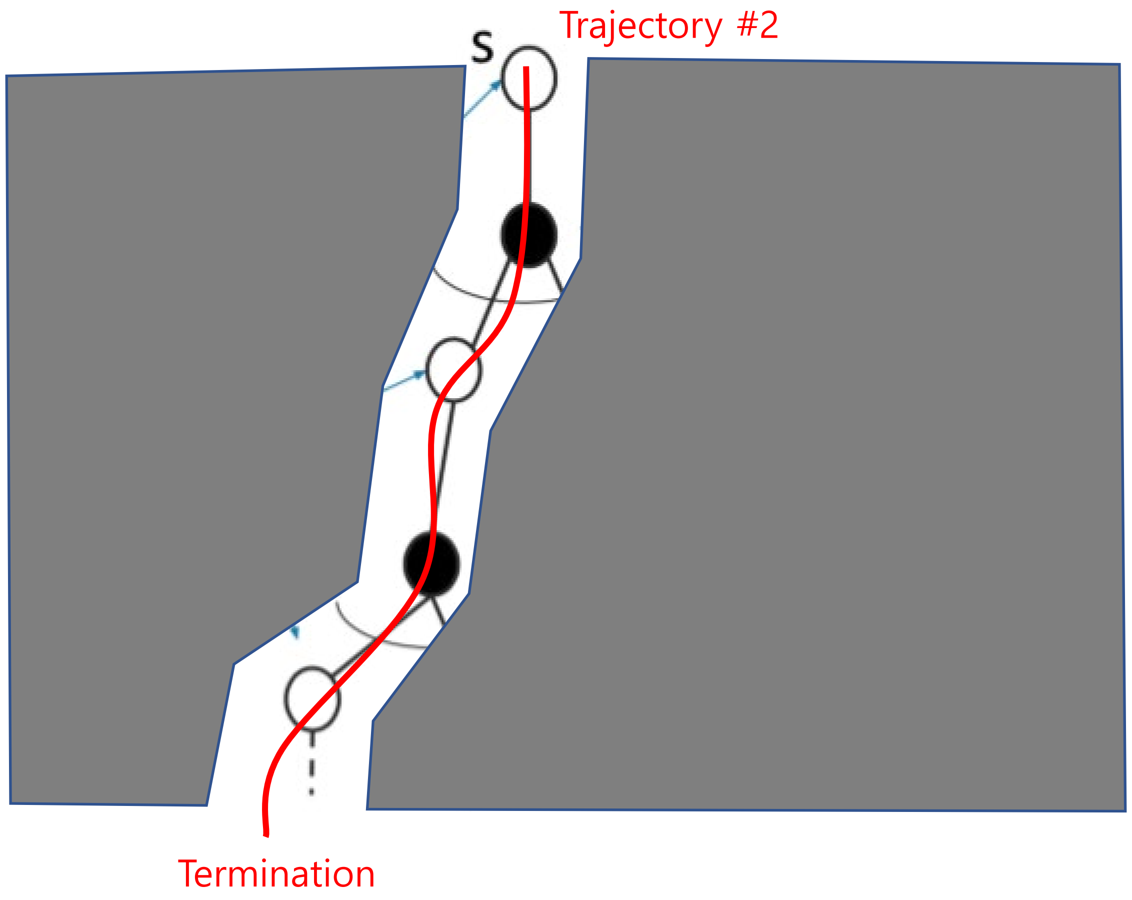
2. MC_Monte Carlo



lecture 3

Policy evaluation

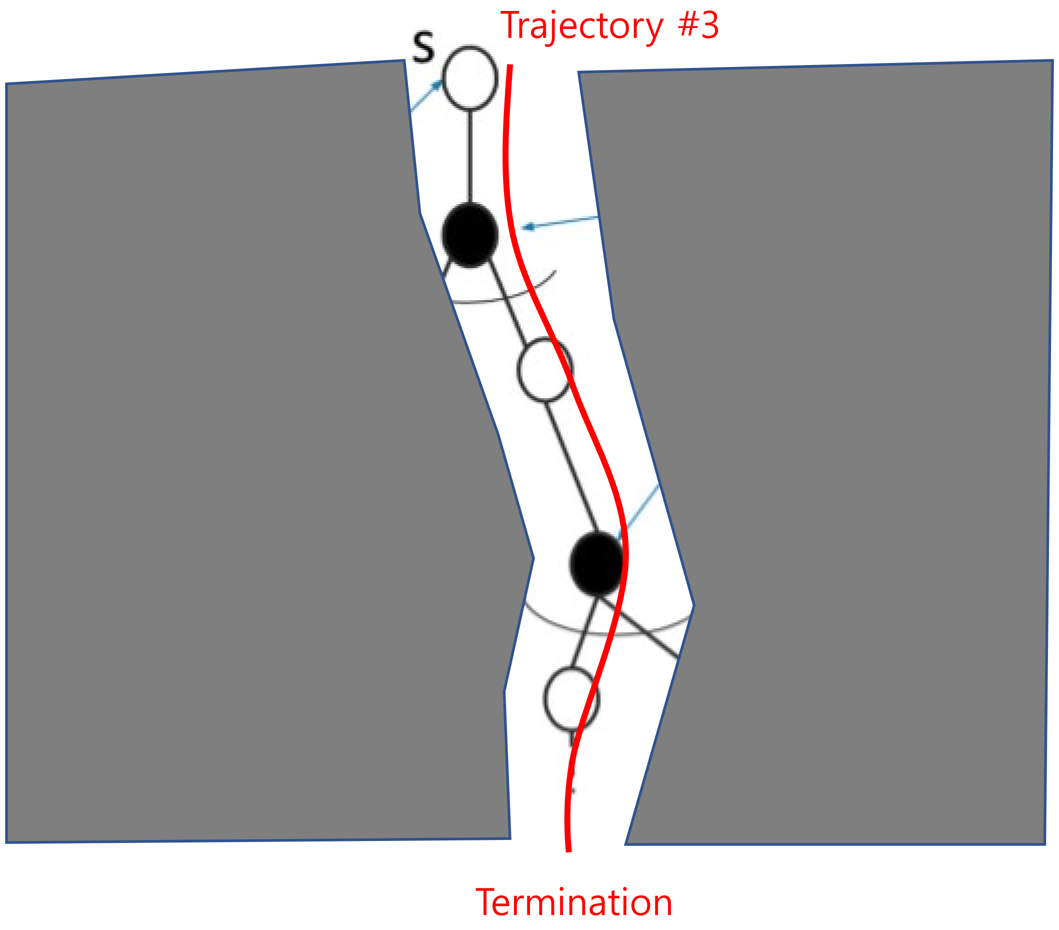
2. MC_Monte Carlo



lecture 3

Policy evaluation

2. MC_Monte Carlo



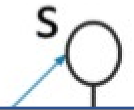
Value = average of
return from
all trajectories

Trajectory #1 … #N

lecture 3

Policy evaluation

2. MC_Monte Carlo



- Condition = episodic MDP = each episode must terminate

- Does not assume state is Markov = current state is all that is required to know what will happen next

- No bootstrapping

lecture 3

Policy evaluation

2. MC_Monte Carlo

- First-visit
- Every-visit
- Incremental

When α = 1/N(s) : Incremental = every-visit

When α = 1/N(s) : forget older data

lecture 3

Policy evaluation

2. MC_Monte Carlo

- First-visit
  - Unbiased
- Every-visit
  - Biased : during the same episode return for different states are correlated
  - Lower variance than first-visit : more data points
- But still requires a lot of data to reduce variance

Q1. Exactly why first-visit Monte Carlo is unbiased?

During one episode there can be multiple states encountered. And they will share similar return depending on the discount factor.

lecture 3

Policy evaluation

3. TD_Temporal Difference

Temporal Difference Learning :

Combination of MC & DP
= Bootstraps and samples

Bootstrapping = relying on previous data results that may not be true = biased

Available for both episodic or infinite horizon settings

Updates value each timestep

lecture 3

Policy evaluation

2. MC_Monte Carlo
3. TD_Temporal Difference

2. MC_Monte Carlo

$$G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \cdots \gamma^{T_i-1} r_i$$

$$N(s) = N(s) + 1$$

$$G(s) = G(s) + G_{i,t}$$

$$V^\pi(s) = G(s)/N(s) \Longleftrightarrow V^\pi(s) = V^\pi(s) + \alpha(G_{i,t} - V^\pi(s))$$

Incremental factor

3. TD_Temporal Difference

$$V^\pi(s_t) = V^\pi(s_t) + \alpha([r_t + \gamma V^\pi(s_{t+1})] - V^\pi(s_t))$$

Update every timestep

lecture 3

Policy evaluation

1. DP_Dynamic Programming
2. MC_Monte Carlo
3. TD_Temporal Difference

	DP	MC	TD
Require model	Yeah	No	No
Require episodic	No	Yah	no
Require mark	No	Yahj	no
Consistent	ye	ye	ye
Unbiased	Biased	Unbiased	biased

With enough data

Q2. Difference between consistency and being unbiased

A2. With small amount of data the estimated value may be off from true value = biased

But with large amounts of data the expected value will converge to the true value

Q3. What is...
- Tabular representation
- Functional approximation

lecture 3

Q4. Explain how TD
exploits Markov structure.

lecture 3

Q5. Help on Certainty Equivalence...

lecture 4

Control

Control

1. Making decisions
2. Optimization : identify policy with high expected rewards
3. Explore : try different actions

Control

1. Making decisions

2. Optimization : identify policy with high expected rewards

3. Explore : try different actions

On-policy

Off-policy

Generalized policy improvement

$$\pi_{i+1}(s) = \arg\max_a Q^{\pi_i}(s, a)$$

ϵ-greedy policy improvement

$$\pi(a|s) = [\arg\max_a Q(s,a), \text{ w. prob } 1-\epsilon; \text{ a w. prob } \frac{\epsilon}{|A|}]$$

GLIE_Greedy in the Limit of Infinite Exploration

- All (s, a) is visited infinite number of times
- $\lim_{i \to \infty} \pi(a|s) \to \arg\max_a Q(s, a)$

SARSA Algorithm

ϵ-greedy policy improvement done for TD methods

Robbins-Munro sequence

SARSA → Q-Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha((r_t + \gamma Q(s_{t+1}, a_{t+1})) - Q(s_t, a_t))$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha((r_t + \gamma \max_{a'} Q(s_{t+1}, a')) - Q(s_t, a_t))$$

Bootstrapping

Q

SARSA → Q-Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha((r_t + \gamma Q(s_{t+1}, a_{t+1})) - Q(s_t, a_t))$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha((r_t + \gamma \boxed{\max_{a'}} Q(s_{t+1}, a')) - Q(s_t, a_t))$$

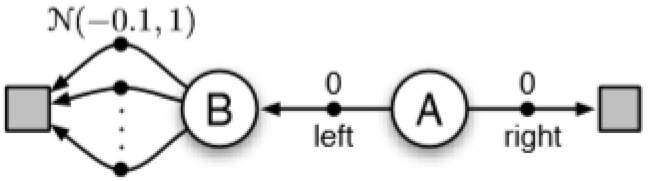
Leads to positive bias

= Maximization Bias

Q-Learning → Double Q-Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha((r_t + \gamma \max_{a'} Q(s_{t+1}, a')) - Q(s_t, a_t))$$

$$Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) + \alpha\Big[R_{t+1} + \gamma Q_2\big(S_{t+1}, \arg\max_a Q_1(S_{t+1}, a)\big) - Q_1(S_t, A_t)\Big]$$



Q7. What is double Q-learning? And how does it overcome maximization bias?

Q8. For the example on the left, using Q-learning would always lead to "left" action from A?

Q9. Is double Q-learning basically bootstrapping from each other samples?

Q1. Exactly why first-visit Monte Carlo is unbiased and every-visit is biased.

A1.

First-visit 은 순수한 평균값이므로 unbiased

Every-visit은 혼합된 (불순한) 평균값이므로 biased

Q2. Difference between consistency and being unbiased

A2. With small amount of data the estimated value may be off from true value = biased

But with large amounts of data the expected value will converge to the true value

Q3. Bootstrapping

A3.

- 다음 값 계산 보다 예측 값을 갖고 옴
- 예측 값은 이전의 episode/trial에서 계산되었던 값 활용

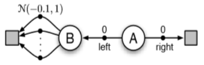
Q4. Explain how TD exploits Markov structure.

A4. Bootstrapping

Q7. What is double Q-learning? And how does it overcome maximization bias?

A7.

Maximization bias: 필연적인 게 아니라는 점 … 해당 예시에서



한 번 왼쪽 action 에서 0보다 큰 값이 나오면 지속적으로 argmax가 left action이 되는 문제가 생김 (true action(?) = right)

Double Q-learning 은 해결이 아닌 완화 method으로 이해…

Q5. difference between SARSA and Q-Learning

A5.

SARSA : On-policy
Q-Learning : Off-policy

Q6. What is Markov?

A6. 현재에 대한 정보로 미래를 예측할 수 있음.

답변 불충분…더 생각해 보기로

답변 불충분…더 생각해 보기로

답변 불충분…더 생각해 보기로

파생 질문

이전 강의 내용 복습