

# Reinforcement Learning

## Lecture 5: Offline Reinforcement Learning

Chengchun Shi

# Lecture Outline

---

## 1. Introduction to Offline RL

## 2. Offline Policy Optimization

- 2.1 The Pessimistic Principle
- 2.2 Model-based Offline Policy Optimization (MOPO)

## 3. Off-Policy Evaluation (OPE)

- 3.1 Introduction to OPE
- 3.2 OPE in Contextual Bandits
- 3.3 OPE in Reinforcement Learning

# Lecture Outline

---

## 1. Introduction to Offline RL

## 2. Offline Policy Optimization

- 2.1 The Pessimistic Principle
- 2.2 Model-based Offline Policy Optimization (MOPO)

## 3. Off-Policy Evaluation (OPE)

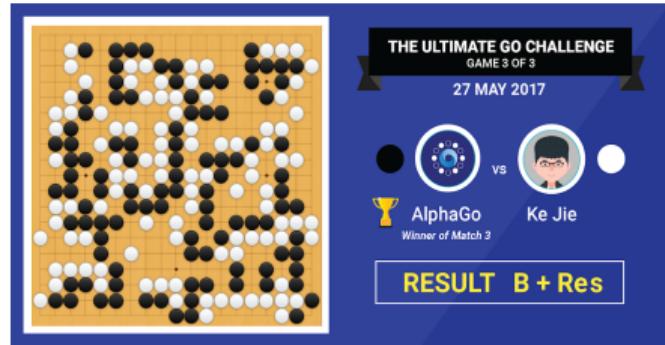
- 3.1 Introduction to OPE
- 3.2 OPE in Contextual Bandits
- 3.3 OPE in Reinforcement Learning

# So Far, We Focused on Online RL Applications

---



(a) Video Games



(b) AlphaGo

# This Lecture Considers Offline Settings

---



(a) Health Care



(b) Robotics



(c) Ridesharing



(d) Auto-driving

# This Lecture Considers Offline Settings (Cont'd)

---

- What is offline RL?
  - RL with a pre-collected historical dataset
- Why offline RL?
  - Online interaction with the environment is **impractical**
  - Either because online data collection is **expensive** (e.g., robotics or healthcare); rely on historical data
  - Or **dangerous** (e.g., healthcare, ridesharing or auto-driving)

# Online v.s. Offline RL

---

## Online RL:

- Data are **adaptively** generated, i.e., able to select **any** action at each time
- Data are **cheap** to generate, i.e., able to simulate **numerous** observations
- Likely to **satisfy** MDP assumption (Markovianity & time-homogeneity)

## Offline RL:

- Data are **pre-collected**, i.e., from an observational study
- Size of data is **limited**
- MDP assumption likely to be **violated** (Non-Markovianity or Non-stationarity)

# Offline RL Challenges and Solutions

---

- Data are **pre-collected**
  - Learning relies entirely on the historical data
  - Not possible to improve exploration
  - For actions that are less-explored, difficult to accurately learn their values
  - **Solution:** the pessimistic principle (first part of this lecture)
- Size of data is **limited**
  - **Solution:** develop sample-efficient RL algorithms (second part of this lecture)

# Lecture Outline

---

## 1. Introduction to Offline RL

## 2. Offline Policy Optimization

2.1 The Pessimistic Principle

2.2 Model-based Offline Policy Optimization (MOPO)

## 3. Off-Policy Evaluation (OPE)

3.1 Introduction to OPE

3.2 OPE in Contextual Bandits

3.3 OPE in Reinforcement Learning

# Recap: Multi-Armed Bandit Problem

---

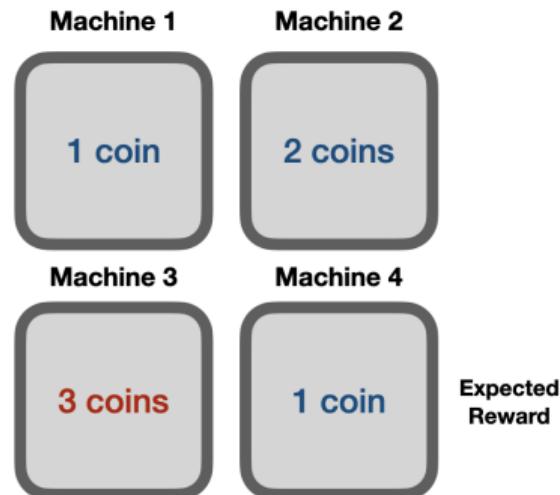


- The **simplest** RL problem
- A casino with **multiple** slot machines
- Playing each machine yields an independent **reward**.
- Limited knowledge (unknown reward distribution for each machine) and resources (**time**)
- **Objective:** determine which machine to pick at each time to maximize the expected **cumulative rewards**

# Offline Multi-Armed Bandit Problem

---

- $k$ -armed bandit problem ( $k$  machines)
- $A_t \in \{1, \dots, k\}$ : arm (machine) pulled (experimented) at time  $t$
- $R_t \in \mathbb{R}$ : reward at time  $t$
- $Q(a) = \mathbb{E}(R_t | A_t = a)$  expected reward for each arm  $a$  (**unknown**)
- **Objective**: Given  $\{A_t, R_t\}_{0 \leq t < T}$ , identify the best arm



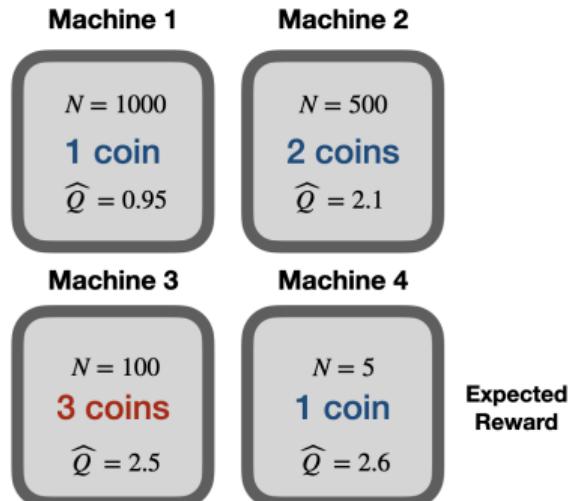
# Greedy Action Selection

- Action-value methods:

$$\hat{Q}(a) = N^{-1}(a) \sum_{t=0}^{T-1} R_t \mathbb{I}(A_t = a)$$

where  $N(a) = \sum_{t=0}^{T-1} \mathbb{I}(A_t = a)$   
denotes the action counter

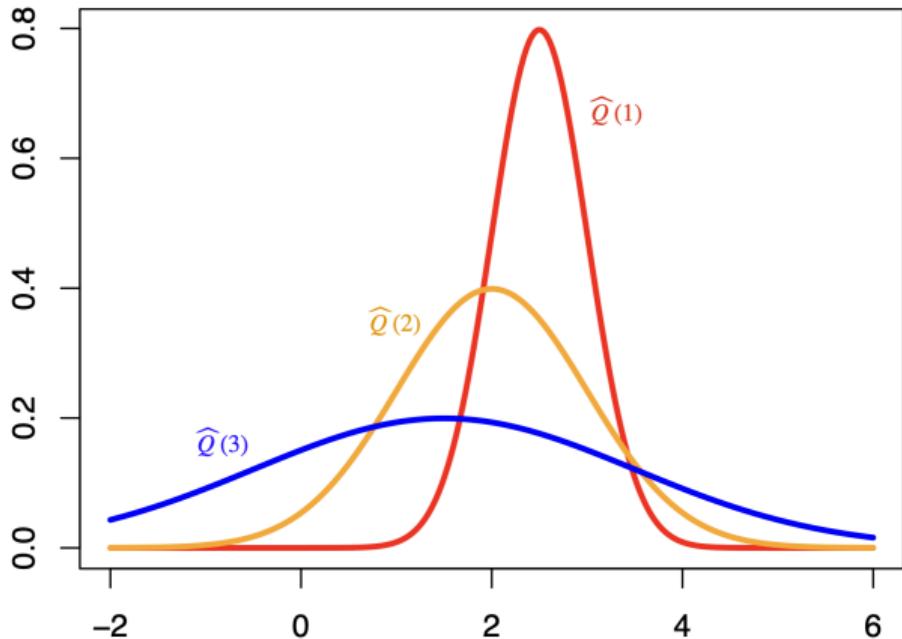
- Greedy policy:  $\arg \max_a \hat{Q}(a)$
- Less-explored action  $\rightarrow N(a)$  is small  
 $\rightarrow$  inaccurate  $\hat{Q}(a)$   $\rightarrow$  suboptimal  
policy (see the plot on the right)



# Recap: The Optimistic Principle

---

- Used in **online** settings to balance exploration-exploitation tradeoff
- The more **uncertain** we are about an action-value
- The more **important** it is to explore that action
- It could be the **best** action
- Likely to pick blue action
- Forms the basis for **upper confidence bound** (UCB)



# Recap: Upper Confidence Bound

---

- Estimate an **upper confidence**  $U_t(\mathbf{a})$  for each action value such that

$$Q(\mathbf{a}) \leq \hat{Q}_t(\mathbf{a}) + U_t(\mathbf{a}),$$

with high probability.

- $U_t(\mathbf{a})$  quantifies the **uncertainty** and depends on  $N_t(\mathbf{a})$  (number of times arm  $\mathbf{a}$  has been selected up to time  $t$ )
  - Large  $N_t(\mathbf{a}) \rightarrow$  small  $U_t(\mathbf{a})$ ;
  - Small  $N_t(\mathbf{a}) \rightarrow$  large  $U_t(\mathbf{a})$ .
- Select actions maximizing upper confidence bound

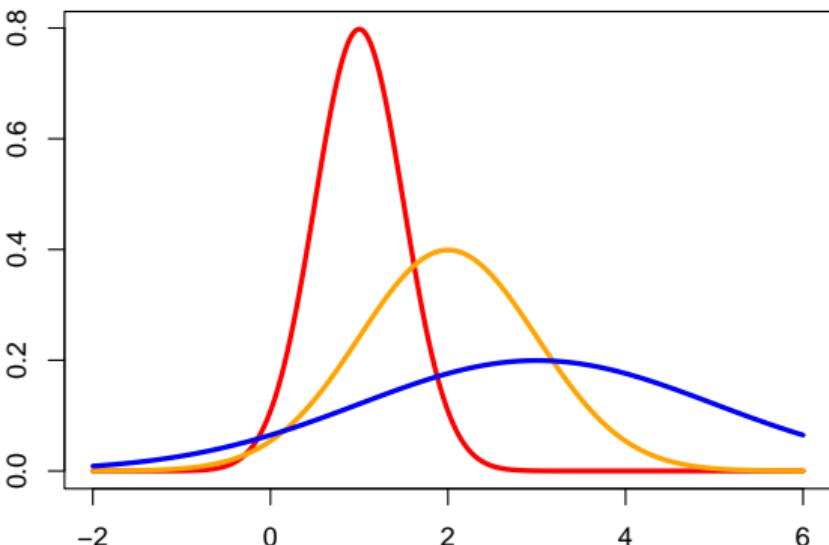
$$\mathbf{a}^* = \arg \max_{\mathbf{a}} [\hat{Q}_t(\mathbf{a}) + U_t(\mathbf{a})].$$

- Combines **exploration** ( $U_t(\mathbf{a})$ ) and **exploitation** ( $\hat{Q}_t(\mathbf{a})$ ).

# The Pessimistic Principle

---

- In **offline** settings
- The less **uncertain** we are about an action-value
- The more **important** it is to use that action
- It could be the **best** action
- Likely to pick red action
- Yields the **lower confidence bound** (LCB) algorithm



# Lower Confidence Bound

---

- Estimate an **lower confidence**  $L(\mathbf{a})$  for each action value such that

$$Q(\mathbf{a}) \geq \hat{Q}(\mathbf{a}) - L(\mathbf{a}),$$

with high probability.

- $L(\mathbf{a})$  quantifies the **uncertainty** and depends on  $N(\mathbf{a})$  (number of times arm  $\mathbf{a}$  has been selected in the historical data)
  - Large  $N(\mathbf{a}) \rightarrow$  small  $L(\mathbf{a})$ ;
  - Small  $N(\mathbf{a}) \rightarrow$  large  $L(\mathbf{a})$ .
- Select actions maximizing lower confidence bound

$$\mathbf{a}^* = \arg \max_{\mathbf{a}} [\hat{Q}(\mathbf{a}) - L(\mathbf{a})].$$

# Lower Confidence Bound (Cont'd)

---

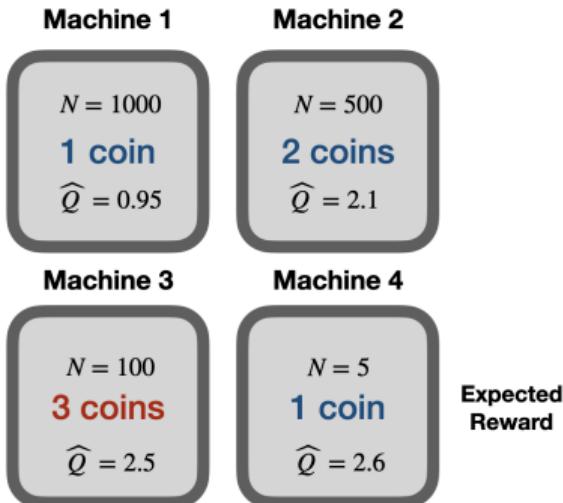
- Set  $L(a) = \sqrt{c \log(T)/N(a)}$  for some positive constant  $c$  where  $T$  is the sample size of historical data
- According to **Hoeffding's inequality** ([link](#)), when rewards are bounded between  $0$  and  $1$ , the event

$$|Q(a) - \hat{Q}(a)| \leq L(a),$$

holds with probability at least  $1 - 2T^{-2c}$  (converges to  $1$  as  $T \rightarrow \infty$ ).

# Lower Confidence Bound (Cont'd)

- $\hat{Q}(4) > \hat{Q}(3)$
- $T = 1605$ . Set  $c = 1$
- $L(3) = \sqrt{\log(T)/N(3)} = 0.272$
- $L(4) = \sqrt{\log(T)/N(4)} = 1.215$
- $\hat{Q}(3) - L(3) > \hat{Q}(4) - L(4)$
- $\hat{Q}(3) - L(3) > \max(\hat{Q}(1), \hat{Q}(2))$
- Correctly identify optimal action



# Algorithm

---

- **Input:** some positive constant  $c$ , offline data  $\{(\mathbf{A}_t, \mathbf{R}_t)\}_{0 \leq t < T}$ .
- **Initialization:**  $t = 0$ ,  $\widehat{\mathbf{Q}}(\mathbf{a}) = \mathbf{0}$ ,  $\mathbf{N}(\mathbf{a}) = \mathbf{0}$ , for  $a = 1, 2, \dots, k$ .
- **While**  $t < T$ :
  - **Update  $N$ :**  $\mathbf{N}(\mathbf{A}_t) \leftarrow \mathbf{N}(\mathbf{A}_t) + 1$ .
  - **Update  $\widehat{Q}$ :**

$$\widehat{\mathbf{Q}}(\mathbf{A}_t) \leftarrow \frac{\mathbf{N}(\mathbf{A}_t) - 1}{\mathbf{N}(\mathbf{A}_t)} \widehat{\mathbf{Q}}(\mathbf{A}_t) + \frac{1}{\mathbf{N}(\mathbf{A}_t)} \mathbf{R}_t.$$

- **Update  $t$ :**  $t \leftarrow t + 1$ .
- **LCB action selection:**

$$\mathbf{a}^* \leftarrow \arg \max_{\mathbf{a}} [\widehat{\mathbf{Q}}(\mathbf{a}) - \sqrt{c \log(T) / \mathbf{N}(\mathbf{a})}].$$

# Theory

Define the regret, as the difference between the expected reward under the **best arm** and that under the **selected arm**.

## Theorem (Greedy Action Selection)

*Regret of greedy action selection is upper bounded by  $2 \max_a |\hat{Q}(a) - Q(a)|$ , whose value is bounded by  $2\sqrt{c \log(T) / \min_a N(a)}$  (according to Hoeffding's inequality) with probability approaching 1*

- The upper bound depends on the estimation error of **each** Q-estimator
- The regret is small when **each** arm has sufficiently many observations
- However, it would yield a large regret when one arm is **less-explored**
- This reveals the **limitation** of greedy action selection
- Proof is simple (see Appendix)

# Theory (Cont'd)

---

Theorem (LCB; see also Jin et al. [2021])

*Regret of the LCB algorithm is upper bounded by  $2\sqrt{c \log(T)/N(a^{opt})}$  where  $a^{opt}$  denotes the best arm with probability approaching 1*

- The upper bound depends on the estimation error of best arm's Q-estimator **only**
- The regret is small when the **best** arm has sufficiently many observations
- This is much weaker than requiring **each** arm to have sufficiently many observations
- This reveals the **advantage** of LCB algorithm
- Proof given in the Appendix

# Lecture Outline

---

## 1. Introduction to Offline RL

## 2. Offline Policy Optimization

2.1 The Pessimistic Principle

2.2 Model-based Offline Policy Optimization (MOPO)

## 3. Off-Policy Evaluation (OPE)

3.1 Introduction to OPE

3.2 OPE in Contextual Bandits

3.3 OPE in Reinforcement Learning

# Offline Policy Optimization and Fitted Q-Iteration

---

- Offline data:  $\{(\mathbf{S}_t, \mathbf{A}_t, \mathbf{R}_t) : 0 \leq t \leq T\}$
- Fitted Q-Iteration can be naturally applied by repeating
  1. Compute  $\hat{\mathbf{Q}}$  as the argmin of

$$\arg \min_{\mathbf{Q}} \sum_t \left[ \mathbf{R}_t + \gamma \max_{\mathbf{a}} \tilde{\mathbf{Q}}(\mathbf{S}_{t+1}, \mathbf{a}) - \mathbf{Q}(\mathbf{S}_t, \mathbf{A}_t) \right]^2$$

2. Set  $\tilde{\mathbf{Q}} = \hat{\mathbf{Q}}$
- **Limitation:** for less-explored state-action pairs, their Q-values **cannot** be learned accurately
  - **Solution:** the pessimistic principle

# Pessimistic Principle in RL

---

- In multi-armed bandit, we select action to maximize lower confidence bound

$$\mathbf{a}^* = \arg \max_{\mathbf{a}} [\hat{Q}(\mathbf{a}) - L(\mathbf{a})]$$

- In more general RL, we can adopt a similar principle by setting

$$\pi(\mathbf{a}|\mathbf{s}) = \begin{cases} 1, & \text{if } \mathbf{a} = \arg \max \hat{Q}(\mathbf{a}, \mathbf{s}) - L(\mathbf{a}, \mathbf{s}) \\ 0, & \text{otherwise} \end{cases}$$

where the lower bound satisfies that with probability approaching 1,

$$Q^{\pi^{\text{opt}}}(\mathbf{a}, \mathbf{s}) \geq \hat{Q}(\mathbf{a}, \mathbf{s}) - L(\mathbf{a}, \mathbf{s}), \quad \forall \mathbf{a}, \mathbf{s}.$$

- Many offline algorithms [see e.g., Wu et al., 2019, Kumar et al., 2020, Levine et al., 2020] adopt similar ideas, but do not directly use the above formula

# Model-based Offline Policy Optimisation (MOPO)

---

- As we discussed in Lecture 4, **model-based** method is preferred in offline settings
- Online RL algorithms are **not** applicable, as adaptive interaction is not feasible
- Model-based method
  - learns a model using the **offline** data
  - allows to **adaptively** generate data based on the model
  - applies **online** RL algorithms to simulated data for policy optimisation
  - embraces the power of online RL algorithms for offline policy optimisation
- MOPO [Yu et al., 2020] integrates model-based method with **pessimistic** principle

# MOPO: Offline Model Learning

---

- Learn the conditional distribution of  $(S_{t+1}, R_t)$  given  $(A_t, S_t)$
- Approximate the conditional distribution using Gaussian, i.e.,

$$(S_{t+1}, R_t) | (A_t, S_t) \sim N(\mu_\theta(A_t, S_t), \Sigma_\phi(A_t, S_t))$$

- Parametrize  $\mu_\theta$  and  $\Sigma_\phi$  using e.g., neural networks
- Use bootstrap to learn  $N$  different models  $\{\mathcal{M}_i\}_{i=1,\dots,N}$

# MOPPO: The Pessimism Principle

---

- Penalize reward to incorporate pessimism
- Simulate reward  $r$  given the state-action pair  $(s, a)$  from model
- Define the **transformed** reward

$$\tilde{r} = r - L(a, s),$$

for some lower bound  $L(a, s)$  that quantifies the **uncertainty** of model

- More uncertain  $\rightarrow$  smaller transformed reward
- Less uncertain  $\rightarrow$  larger transformed reward
- Apply online RL to transformed data (see next slide)

# MOPO: Pseudocode

---

**Algorithm 2** MOPO instantiation with regularized probabilistic dynamics and ensemble uncertainty

**Require:** reward penalty coefficient  $\lambda$  rollout horizon  $h$ , rollout batch size  $b$ .

- 1: Train on batch data  $\mathcal{D}_{\text{env}}$  an ensemble of  $N$  probabilistic dynamics  $\{\hat{T}^i(s', r | s, a) = \mathcal{N}(\mu^i(s, a), \Sigma^i(s, a))\}_{i=1}^N$ .
  - 2: Initialize policy  $\pi$  and empty replay buffer  $\mathcal{D}_{\text{model}} \leftarrow \emptyset$ .
  - 3: **for** epoch 1, 2, . . . **do** ▷ This for-loop is essentially one outer iteration of MBPO
  - 4:   **for** 1, 2, . . . ,  $b$  (in parallel) **do**
  - 5:     Sample state  $s_1$  from  $\mathcal{D}_{\text{env}}$  for the initialization of the rollout.
  - 6:     **for**  $j = 1, 2, \dots, h$  **do**
  - 7:       Sample an action  $a_j \sim \pi(s_j)$ .
  - 8:       Randomly pick dynamics  $\hat{T}$  from  $\{\hat{T}^i\}_{i=1}^N$  and sample  $s_{j+1}, r_j \sim \hat{T}(s_j, a_j)$ .
  - 9:       Compute  $\tilde{r}_j = r_j - \lambda \max_{i=1}^N \|\Sigma^i(s_j, a_j)\|_{\text{F}}$ .
  - 10:      Add sample  $(s_j, a_j, \tilde{r}_j, s_{j+1})$  to  $\mathcal{D}_{\text{model}}$ .
  - 11:     Drawing samples from  $\mathcal{D}_{\text{env}} \cup \mathcal{D}_{\text{model}}$ , use SAC to update  $\pi$ .
-

# Lecture Outline

---

## 1. Introduction to Offline RL

## 2. Offline Policy Optimization

2.1 The Pessimistic Principle

2.2 Model-based Offline Policy Optimization (MOPO)

## 3. Off-Policy Evaluation (OPE)

3.1 Introduction to OPE

3.2 OPE in Contextual Bandits

3.3 OPE in Reinforcement Learning

# What is OPE and Why OPE

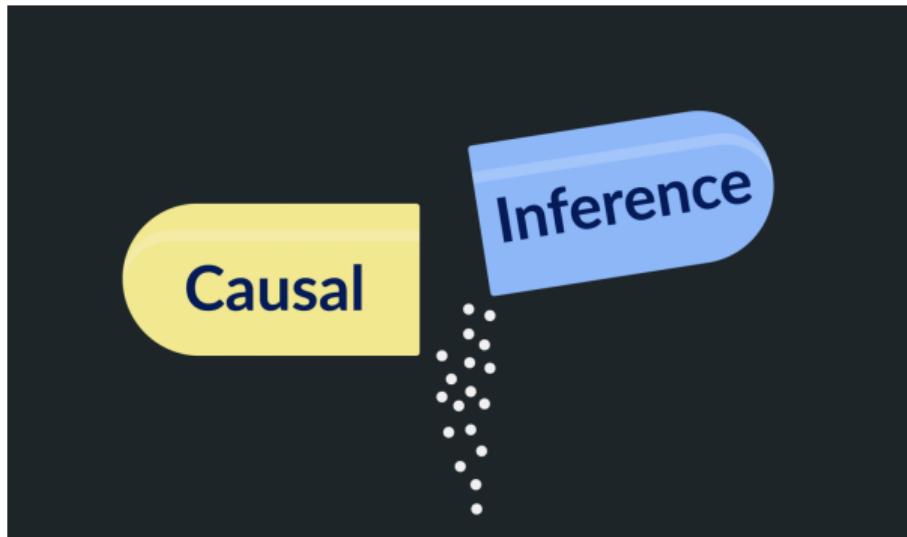
---

- **Objective:** Evaluate the impact of a **target policy** offline using historical data generated from a different **behavior policy**
- **Motivation:** In many applications, it can be **dangerous** to evaluate a **target policy** by directly running this policy
- **Healthcare:** which **medical treatment** to suggest for a patient
- **Ridesharing:** which **driver** to assign for a call order

# Causal Inference

---

Off-policy evaluation is closely related to **causal inference**, whose objective is to learn the difference between a new treatment and a standard treatment



# Outline

---

## 1. Introduction to Offline RL

## 2. Offline Policy Optimization

2.1 The Pessimistic Principle

2.2 Model-based Offline Policy Optimization (MOPO)

## 3. Off-Policy Evaluation (OPE)

3.1 Introduction to OPE

3.2 OPE in Contextual Bandits

3.3 OPE in Reinforcement Learning

# Contextual Bandits

---

- Extension of MAB with **contextual** information.
- A **widely-used** model in medicine and technological industries.
- At time  $t$ , the agent
  - Observe a context  $S_t$ ;
  - Select an action  $A_t$ ;
  - Receives a reward  $R_t$  (depends on both  $S_t$  and  $A_t$ ).
- **Objective:** Given an i.i.d. offline dataset  $\{(S_t, A_t, R_t) : 0 \leq t < T\}$  generated by a behavior policy  $b$ , i.e.,

$$\Pr(A_t = a | S_t = s) = b(a|s),$$

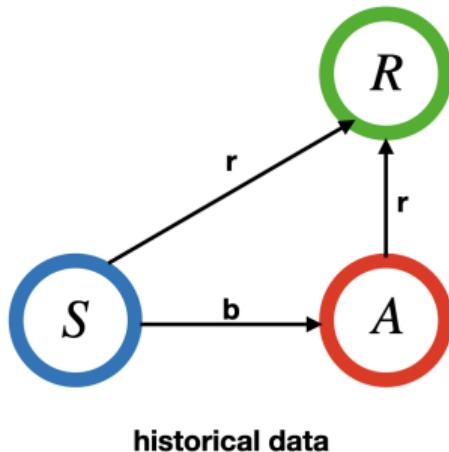
we aim to evaluate the mean outcome under a target policy  $\pi$ , i.e.,

$$\Pr(A_t = a | S_t = s) = \pi(a|s).$$

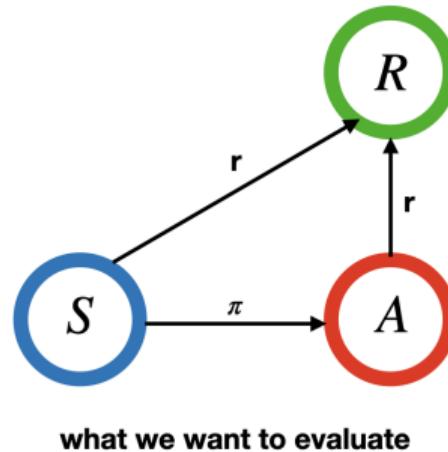
# Challenge

---

- **Confounding:** State serves as confounding variables that confound the action-reward pair
- **Distributional shift:** The target policy generally differs from the behavior policy



historical data



what we want to evaluate

## Challenge (Cont'd)

---

- Suppose  $\pi$  is a nondynamic policy, i.e., there exists some  $a$  such that  $\pi(a|s) = 1$  for any  $s$ . We aim to evaluate the value under a given action  $a$ . A naive estimator is

$$\frac{\sum_{t=0}^{T-1} R_t \mathbb{I}(A_t = a)}{\sum_{t=0}^{T-1} \mathbb{I}(A_t = a)} \xrightarrow{P} \mathbb{E}(R|A = a)$$

- This estimator is valid only when no confounding variables exist
- According to the causal diagram, the target policy's value equals

$$\mathbb{E}[\mathbb{E}(R|A = a, S)] \neq \mathbb{E}(R|A = a)$$

# OPE Estimators

---

- With a general target policy  $\pi$ , the target policy's value equals

$$\sum_a \mathbb{E}[\pi(a|S)\mathbb{E}(R|A = a, S)] = \sum_a \mathbb{E}[\pi(a|S)r(S, a)],$$

where  $r(s, a) = \mathbb{E}(R|A = a, S = s)$

- Direct estimator
- Importance sampling estimator
- Doubly robust estimator

# Direct Estimator

---

- Given that the target policy's value is given by

$$\sum_{\color{red}a\color{black}} \mathbb{E}[\pi(\color{red}a\color{black}|\color{blue}S\color{black})r(\color{blue}S\color{black}, \color{red}a\color{black})]$$

- The expectation can be approximated by the sample average, i.e.,

$$\frac{1}{T} \sum_{\color{red}a\color{black}} \sum_{t=0}^{T-1} [\pi(\color{red}a\color{black}|\color{blue}S_t\color{black})r(\color{blue}S_t\color{black}, \color{red}a\color{black})]$$

- The reward function can be replaced with some estimator  $\hat{r}$ . This yields the direct estimator

$$\frac{1}{T} \sum_{\color{red}a\color{black}} \sum_{t=0}^{T-1} [\pi(\color{red}a\color{black}|\color{blue}S_t\color{black})\hat{r}(\color{blue}S_t\color{black}, \color{red}a\color{black})]$$

# Importance Sampling Estimator

---

- Given that the target policy's value is given by

$$\sum_a \mathbb{E}[\pi(a|S)r(S, a)]$$

- By the change of measure theory, it equals

$$\sum_a \mathbb{E} \left[ b(a|S) \frac{\pi(a|S)}{b(a|S)} r(S, a) \right] = \mathbb{E} \left[ \frac{\pi(A|S)}{b(A|S)} r(S, A) \right] = \mathbb{E} \left[ \frac{\pi(A|S)}{b(A|S)} R \right]$$

- This yields the following importance sampling (IS) estimator [Zhang et al., 2012]

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{\pi(A_t|S_t)}{\hat{b}(A_t|S_t)} R_t,$$

for a given estimator  $\hat{b}$

# Direct Estimator v.s. IS Estimator

---

- Bias/Variance Trade-Off
- The direct estimator has **some bias**, since  $r$  needs to be estimated from data
- The IS estimator has **zero bias** when  $b$  is known as in randomized studies
- The IS estimator might have a **large variance** when  $\pi$  differs significantly from  $b$
- Suppose  $R = r(S, A) + \varepsilon$  for some  $\varepsilon$  independent of  $(S, A)$ ,

$$\begin{aligned}\text{Var} \left[ \frac{\pi(A|S)}{b(A|S)} R \right] &= \mathbb{E} \left[ \frac{\pi(A|S)}{b(A|S)} \{R - r(S, A)\} \right]^2 + \text{some term} \\ &= \sigma^2 \mathbb{E} \left[ \frac{\pi^2(A|S)}{b^2(A|S)} \right] + \text{some term},\end{aligned}$$

where  $\sigma^2 = \text{Var}(\varepsilon)$

# Extensions

---

- When  $\pi$  differs from  $b$  significantly, IS estimator suffers from **large variance** and becomes **unstable**
- Solutions sought by using **self-normalized** and/or **truncated** IS
- **Self-normalized** IS

$$\left[ \frac{1}{T} \sum_{t=0}^{T-1} \frac{\pi(\mathbf{A}_t | \mathbf{S}_t)}{b(\mathbf{A}_t | \mathbf{S}_t)} \right]^{-1} \frac{1}{T} \sum_{t=0}^{T-1} \frac{\pi(\mathbf{A}_t | \mathbf{S}_t)}{b(\mathbf{A}_t | \mathbf{S}_t)} \mathbf{R}_t$$

- **Truncated** IS

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{\pi(\mathbf{A}_t | \mathbf{S}_t)}{\max(\hat{b}(\mathbf{A}_t | \mathbf{S}_t), \varepsilon)} \mathbf{R}_t,$$

for some  $\varepsilon > 0$

# Doubly Robust Estimator

---

- Direct estimator

$$\frac{1}{T} \sum_{\color{red}\mathbf{a}} \sum_{t=0}^{T-1} [\pi(\color{red}\mathbf{a}|\color{blue}\mathcal{S}_t) \hat{r}(\color{blue}\mathcal{S}_t, \color{red}\mathbf{a})]$$

requires  $\hat{r}$  to be consistent

- IS estimator

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{\pi(\color{red}\mathbf{A}_t|\color{blue}\mathcal{S}_t)}{\hat{b}(\color{red}\mathbf{A}_t|\color{blue}\mathcal{S}_t)} \color{green}\mathcal{R}_t,$$

requires  $\hat{b}$  to be consistent

- Doubly robust (DR) estimator combines both, and requires either  $\hat{r}$  or  $\hat{b}$  to be consistent (“**doubly-robustness**” property)

# Doubly Robust Estimator (Cont'd)

---

- Consider the estimating function

$$\phi(\mathbf{S}, \mathbf{A}, \mathbf{R}) = \sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{S}) \hat{r}(\mathbf{S}, \mathbf{a}) + \frac{\pi(\mathbf{A}|\mathbf{S})}{\hat{b}(\mathbf{A}|\mathbf{S})} [\mathbf{R} - \hat{r}(\mathbf{S}, \mathbf{A})]$$

- First term on the RHS is the estimating function of the direct estimator
- Second term corresponds to the **augmentation term**
  - Zero mean when  $\hat{r} = r$
  - Debias the bias of the direct estimator
  - Offering additional robustness against model misspecification of  $\hat{r}$
- DR estimator given by  $\mathbf{T}^{-1} \sum_{t=0}^{T-1} \phi(\mathbf{S}_t, \mathbf{A}_t, \mathbf{R}_t)$

# Fact 1: Double Robustness

---

- The estimating function

$$\phi(\mathbf{S}, \mathbf{A}, \mathbf{R}) = \sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{S}) \hat{r}(\mathbf{S}, \mathbf{a}) + \frac{\pi(\mathbf{A}|\mathbf{S})}{\hat{b}(\mathbf{A}|\mathbf{S})} [\mathbf{R} - \hat{r}(\mathbf{S}, \mathbf{A})]$$

- In large sample size, DR estimator converges to  $\mathbb{E}\phi(\mathbf{S}, \mathbf{A}, \mathbf{R})$
- When  $\hat{r} = r$ , the augmentation term has zero mean. It follows that

$$\mathbb{E}\phi(\mathbf{S}, \mathbf{A}, \mathbf{R}) = \sum_{\mathbf{a}} \mathbb{E}[\pi(\mathbf{a}|\mathbf{S}) r(\mathbf{S}, \mathbf{a})] = \text{target policy's value}$$

- When  $\hat{b} = b$ , it has the same mean as the IS estimator

$$\begin{aligned} \mathbb{E}\phi(\mathbf{S}, \mathbf{A}, \mathbf{R}) &= \mathbb{E} \left[ \frac{\pi(\mathbf{A}|\mathbf{S})}{b(\mathbf{A}|\mathbf{S})} \mathbf{R} \right] + \mathbb{E} \left[ \sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{S}) \hat{r}(\mathbf{S}, \mathbf{a}) - \frac{\pi(\mathbf{A}|\mathbf{S})}{b(\mathbf{A}|\mathbf{S})} \hat{r}(\mathbf{S}, \mathbf{A}) \right] \\ &= \mathbb{E} \left[ \frac{\pi(\mathbf{A}|\mathbf{S})}{b(\mathbf{A}|\mathbf{S})} \mathbf{R} \right] = \text{target policy's value} \end{aligned}$$

## Fact 2: Efficiency

---

- When  $\hat{\mathbf{b}} = \mathbf{b}$ , the estimating function

$$\phi(\mathbf{S}, \mathbf{A}, \mathbf{R}) = \sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{S}) \hat{r}(\mathbf{S}, \mathbf{a}) + \frac{\pi(\mathbf{A}|\mathbf{S})}{\mathbf{b}(\mathbf{A}|\mathbf{S})} [\mathbf{R} - \hat{r}(\mathbf{S}, \mathbf{A})]$$

- The MSE of DR estimator is proportional to the variance of  $\phi(\mathbf{S}, \mathbf{A}, \mathbf{R})$

$$\text{Var}(\phi(\mathbf{S}, \mathbf{A}, \mathbf{R})) = \mathbb{E}[\text{Var}(\phi(\mathbf{S}, \mathbf{A}, \mathbf{R}) | \mathbf{S}, \mathbf{A})] + \text{Var}[\mathbb{E}(\phi(\mathbf{S}, \mathbf{A}, \mathbf{R}) | \mathbf{S}, \mathbf{A})]$$

- The first term on the RHS is independent of  $\hat{r}$
- The second term is minimized when  $\hat{r} = r$
- A good working model for  $r$  improves the estimator's efficiency
- When  $\hat{r} = r$ , the estimator achieves the **efficiency bound** [e.g., smallest MSE among a class of regular estimators; see Tsiatis, 2007]

## Fact 3: Efficiency

---

- When  $\hat{\mathbf{b}}$  is estimated from data and the model is **correctly specified**, the IS estimator's MSE would be **generally smaller than** the one that uses the oracle behavior policy  $\mathbf{b}$  [Tsiatis, 2007]
- Estimating  $\hat{\mathbf{b}}$  yields a more efficient estimator, even if we know the oracle  $\mathbf{b}$
- **Multi-armed bandit** example without context information
  - **Objective:** evaluate  $\mathbb{E}(\mathbf{R}|\mathbf{A} = \mathbf{a})$  for a given  $\mathbf{a}$
  - IS estimator with **known**  $\Pr(\mathbf{A} = \mathbf{a})$

$$\frac{\sum_{t=0}^{T-1} \mathbb{I}(\mathbf{A}_t = \mathbf{a}) \mathbf{R}_t}{T \Pr(\mathbf{A}_t = \mathbf{a})}$$

- IS estimator with **estimated**  $\Pr(\mathbf{A} = \mathbf{a})$  has a **smaller** asymptotic variance

$$\frac{\sum_{t=0}^{T-1} \mathbb{I}(\mathbf{A}_t = \mathbf{a}) \mathbf{R}_t}{\sum_{t=0}^{T-1} \mathbb{I}(\mathbf{A}_t = \mathbf{a})}$$

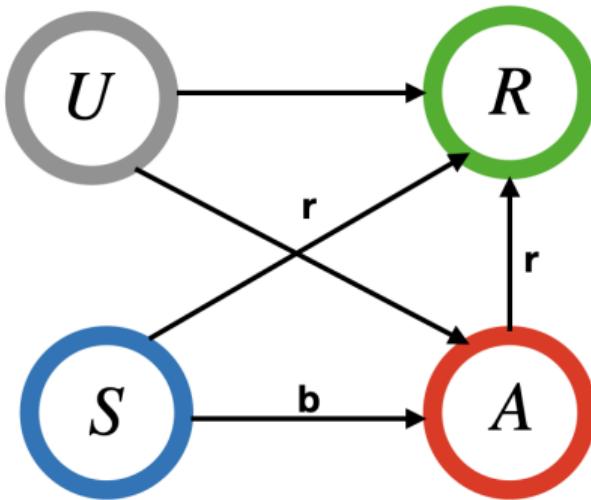
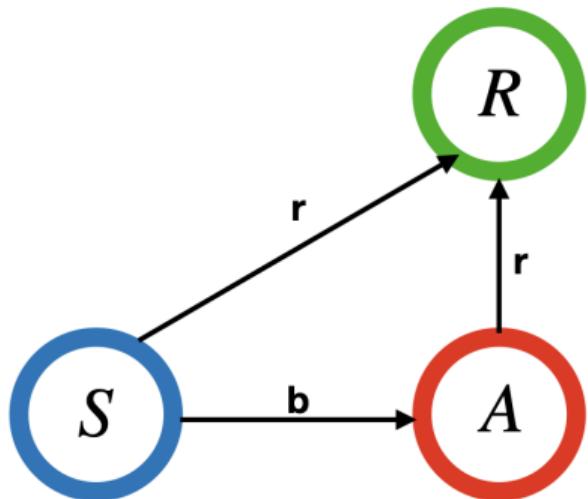
## Fact 4: Asymptotic Normality

---

- The DR estimator converges at a parametric rate and is asymptotically normal even when both  $\hat{r}$  and  $\hat{b}$  converge **slower** than the parameter rate (i.e., root- $n$  rate)
- This observation allows us to apply machine learning methods to estimate both nuisance functions, leading to the **double machine learning** estimator [Chernozhukov et al., 2017]
- Indeed, it only requires  $\hat{r}$  and  $\hat{b}$  to converge at a rate of  $o_p(n^{-1/4})$ , due to the double robustness property

# Assumption: No Unmeasured Confounders

---



# Lecture Outline

---

## 1. Introduction to Offline RL

## 2. Offline Policy Optimization

2.1 The Pessimistic Principle

2.2 Model-based Offline Policy Optimization (MOPO)

## 3. Off-Policy Evaluation (OPE)

3.1 Introduction to OPE

3.2 OPE in Contextual Bandits

3.3 OPE in Reinforcement Learning

# General OPE Problem

---

- **Objective:** Given an offline dataset  $\{(S_{i,t}, A_{i,t}, R_{i,t}) : 1 \leq i \leq N, 0 \leq t \leq T\}$  generated by a behavior policy  $b$ , where  $i$  indexes the  $i$ th episode and  $t$  indexes the  $t$ th time point, we aim to evaluate the mean return under a target policy  $\pi$

$$\mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_t \right] = \mathbb{E} V^\pi(S_0)$$

When  $\gamma = 1$ , the task is assumed to be episodic

- We focus on the case where both  $\pi$  and  $b$  are **stationary** policies
- Challenge: **Distributional shift**
  - In the offline dataset, actions are generated according to  $b$
  - The target policy  $\pi$  we wish to evaluate is different from  $b$

# Direct Estimator

---

- The target policy's value is given by  $\mathbb{E} V^\pi(\mathbf{S}_0)$ , or equivalently,

$$\mathbb{E}\left[\sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{S}_0) Q^\pi(\mathbf{S}_0, \mathbf{a})\right]$$

- The expectation can be approximated via the **empirical initial state distribution**
- Q-learning is an **off-policy** algorithm. Can be applied to learn  $Q^\pi$  offline
- This yields the direct estimator

$$\frac{1}{N} \sum_{i=1}^N \sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{S}_{i,0}) \hat{Q}(\mathbf{S}_{i,0}, \mathbf{a})$$

- It remains to compute  $\hat{Q}$

# Fitted Q-Evaluation [Le et al., 2019]

---

- Bellman equation

$$\mathbb{E} [R_t + \gamma \pi(a|S_{t+1}) Q^\pi(S_{t+1}, a) | S_t, A_t] = Q^\pi(S_t, A_t)$$

- Both LHS and RHS involves  $Q^\pi$
- Repeat the following procedure
  1. Compute  $\hat{Q}$  as the argmin of

$$\arg \min_Q \sum_t \left[ R_{i,t} + \gamma \sum_a \pi(a|S_{i,t+1}) \tilde{Q}(S_{i,t+1}, a) - Q(S_{i,t}, A_{i,t}) \right]^2$$

2. Set  $\tilde{Q} = \hat{Q}$
- Designed for learning  $Q^\pi$
  - Do **not** require actions to follow the target policy

# Other Direct Estimators

---

- Sieve-based estimator [Shi et al., 2020b]
  - Use linear sieves to parametrize  $Q^\pi$
  - Estimate regression coefficients by solving the Bellmen equation
- Kernel-based estimator [Liao et al., 2021]
  - Use RKHSs to parametrize  $Q^\pi$
  - Estimate parameters by solving a coupled optimization [Farahmand et al., 2016]
- Limiting distributions of value estimators are derived in the two papers

# Sequential IS Estimator [Zhang et al., 2013]

---

- Consider episodic task where  $T$  is the termination time
- Importance sampling ratio needs to be employed

$$\mathbb{E}^\pi R_0 = \mathbb{E}^b \left[ \frac{\pi(A_0|S_0)}{b(A_0|S_0)} R_0 \right]$$

$$\mathbb{E}^\pi R_1 = \mathbb{E}^b \left[ \frac{\pi(A_0|S_0)}{b(A_0|S_0)} \frac{\pi(A_1|S_1)}{b(A_1|S_1)} R_1 \right]$$

$$\vdots$$

$$\mathbb{E}^\pi R_t = \mathbb{E}^b \left[ \frac{\pi(A_0|S_0)}{b(A_0|S_0)} \dots \frac{\pi(A_t|S_t)}{b(A_t|S_t)} R_t \right]$$

# Sequential IS Estimator (Cont'd)

---

- According to this logic, the target policy's value can be represented by

$$\mathbb{E} \left[ \sum_{t=0}^T \gamma^t \left\{ \prod_{j=0}^t \frac{\pi(A_j | S_j)}{b(A_j | S_j)} \right\} R_t \right]$$

- This yields the sequential IS estimator

$$\frac{1}{N} \sum_{i=1}^N \left[ \sum_{t=0}^T \gamma^t \left\{ \prod_{j=0}^t \frac{\pi(A_{i,j} | S_{i,j})}{\hat{b}(A_{i,j} | S_{i,j})} \right\} R_{i,t} \right]$$

for a given estimator  $\hat{b}$  computed using supervised learning algorithms

# Limitation

---

- Sequential IS suffers from a **large variance**
- In particular, the IS ratio at time  $t$  is the product of individual ratios from the **initial** time to time  $t$

$$\prod_{j=0}^t \frac{\pi(\textcolor{red}{A_j}|\textcolor{blue}{S_j})}{b(\textcolor{red}{A_j}|\textcolor{blue}{S_j})}$$

- Variance of the ratio grows **exponentially** with respect to  $t$ , referred to as the **curse of horizon** [Liu et al., 2018]
- Extension: **Doubly-robust** estimator by [Jiang and Li, 2016]

# Pros & Cons of Direct v.s. Sequential IS

---

- Bias/Variance Trade-Off
- When  $b$  is known, sequential IS is an **unbiased** estimator since

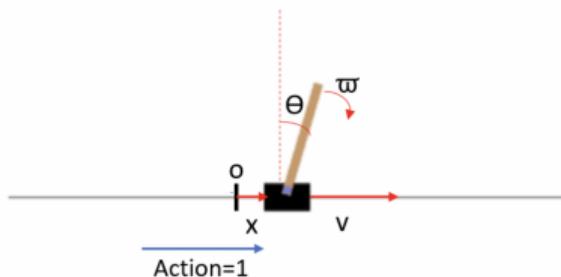
$$\mathbb{E}^\pi \textcolor{green}{R}_t = \mathbb{E}^b \left[ \frac{\pi(\textcolor{red}{A}_0 | \textcolor{blue}{S}_0)}{b(\textcolor{red}{A}_0 | \textcolor{blue}{S}_0)} \dots \frac{\pi(\textcolor{red}{A}_t | \textcolor{blue}{S}_t)}{b(\textcolor{red}{A}_t | \textcolor{blue}{S}_t)} \textcolor{green}{R}_t \right]$$

- Direct estimator has **some bias**, since  $Q^\pi$  needs to be estimated from data
- Sequential IS suffers from **curse of horizon** and a **large variance**
- Direct estimator has a much lower variance

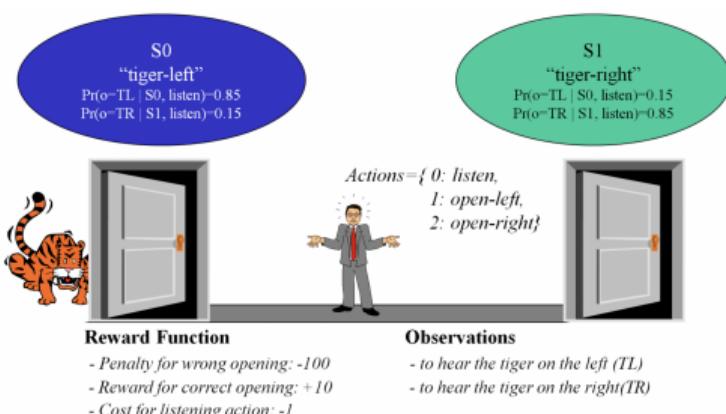
# Pros & Cons of Direct v.s. Sequential IS (Cont'd)

- Direct estimator exploits **Markov** & **stationary** properties
- Relies on the **Bellman equation**
- More **efficient** in MDP environments

frame: 53, Obs: (0.018, 0.669, 0.286, 0.618)  
Action: 1.0, Cumulative Reward: 47.0, Done: 1



- SIS does **not** exploit these properties
- More **flexible** in non-MDP environments (e.g., POMDP)



# Marginalized IS Estimator

---

- As we have discussed, sequential IS suffers from **curse of horizon**
- Curse of horizon is **unavoidable** in general **Non-Markov decision processes** (e.g., POMDP)
- Under some additional model assumptions (e.g., Markovianity & time-homogeneity), it is possible to break the curse of horizon using **marginalized IS** estimator
- Sequential IS does **not** exploit these properties

# Marginalized IS Estimator (Cont'd)

---

- Sequential IS uses the **cumulative** IS ratio

$$\mathbb{E}^\pi \mathbf{R}_t = \mathbb{E}^{\mathbf{b}} \left[ \frac{\pi(\mathbf{A}_0 | \mathbf{S}_0)}{b(\mathbf{A}_0 | \mathbf{S}_0)} \cdots \frac{\pi(\mathbf{A}_t | \mathbf{S}_t)}{b(\mathbf{A}_t | \mathbf{S}_t)} \mathbf{R}_t \right]$$

- Under Markovianity (TMDP), marginalized IS uses the **marginalized** IS ratio

$$\mathbb{E}^\pi \mathbf{R}_t = \mathbb{E}^{\mathbf{b}} \left[ \frac{\mathbf{p}_t^\pi(\mathbf{S}_t, \mathbf{A}_t)}{\mathbf{p}_t^{\mathbf{b}}(\mathbf{S}_t, \mathbf{A}_t)} \mathbf{R}_t \right] \quad (1)$$

where  $\mathbf{p}_t^\pi$  and  $\mathbf{p}_t^{\mathbf{b}}$  are the marginal density functions of  $(\mathbf{S}_t, \mathbf{A}_t)$  under  $\pi$  and  $\mathbf{b}$

- The resulting marginalized IS estimator can be derived from (1)

# Marginalized IS Estimator

---

- Under Markovianity and time-homogeneity (MDP),

$$\mathbb{E} V^\pi(S_0) = \mathbb{E}^{\boldsymbol{b}} \left[ \frac{\sum_{t=0}^{\infty} \gamma^t p_t^\pi(\mathbf{S}, \mathbf{A}) R}{p_\infty(\mathbf{S}, \mathbf{A})} \right] \quad (2)$$

where  $p_\infty$  denotes the limiting state-action distribution under  $\boldsymbol{b}$  and the numerator corresponds to the  $\gamma$ -discounted state-action visitation probability

- The resulting marginalized IS estimator can be derived from (2)
- Marginal IS ratio can be estimated via **minimax learning** [Uehara et al., 2019]
- Closed-form expression is available when using **linear sieves**
- Coupled optimization can also be employed when using **RKHSs** [Liao et al., 2020]
- Alternatively, we can use **RKHSs** to parametrize the discriminator class, use **neural networks** to parametrize the ratio and apply SGD for parameter estimation

# Double RL [Kallus and Uehara, 2019]

---

- Double RL extends DR in **contextual bandits** to the general RL problem
- Similar to DR, the estimator can be represented as

Direct Estimator + Augmentation Term

- **Augmentation** term is to **debias** the bias of direct estimator and offer protection against model misspecification of  $Q^\pi$ ; it relies on the marginalized IS ratio
- Similar to DR, the estimator is **doubly-robust**, e.g., consistent when either  $Q^\pi$  or the marginalized IS ratio is correct
- Similar to DR, the estimator achieves the **efficiency bound** in MDPs

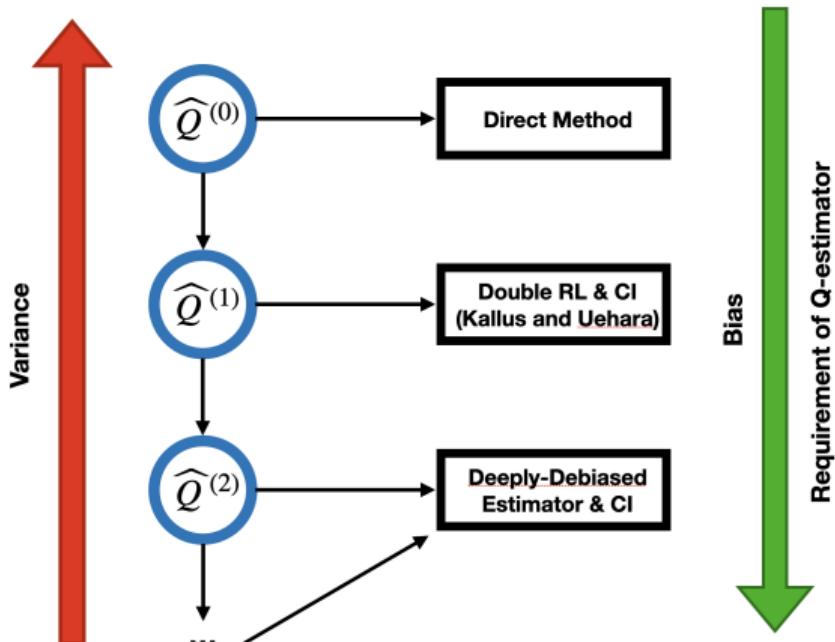
## Fact 5: Efficiency

---

- Direct estimators (based on linear sieves or RKHSs) also achieve the **efficiency bound** in MDPs [Liao et al., 2021, Shi et al., 2022a]
- Marginalized IS estimators (based on linear sieves) also achieve the **efficiency bound** in MDPs
- When using linear sieves,

direct estimator = marginalized IS estimator = double RL estimator

# Deeply-Debiased OPE [Shi et al., 2021b]



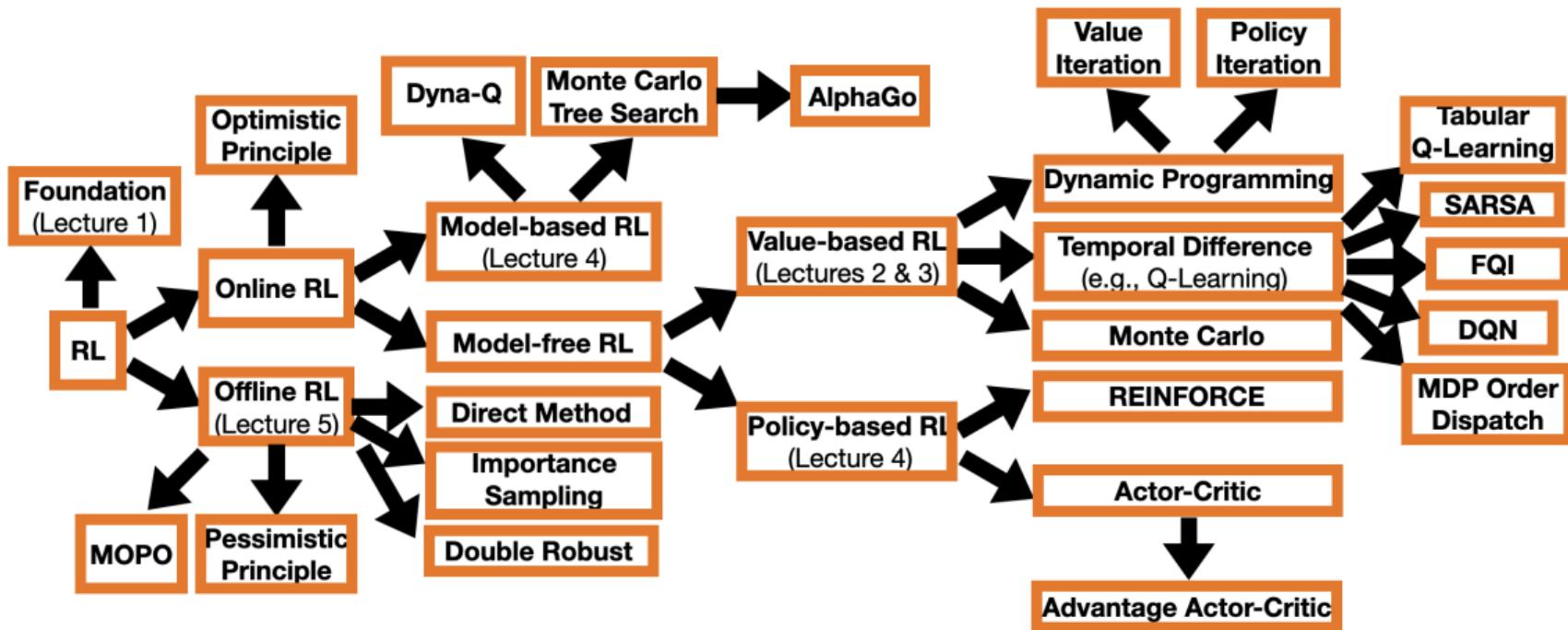
- Constructed based on high-order influence function [Robins et al., 2008, 2017]
- Ensures bias decays much faster than standard deviation
- Allows to provide valid **uncertainty quantification** (e.g., confidence interval)

# Other Topics

---

- Evaluation of the expected return under optimal policy
  - Inference is challenging in **nonregular** settings where the optimal policy is not unique
  - $m$ -out-of- $n$  bootstrap [Chakraborty et al., 2013]
  - Martingale-based method [Luedtke and Van Der Laan, 2016, Shi et al., 2020b]
  - Subagging-based method [Shi et al., 2020a]
- Confounded OPE
  - Confounded POMDP [Tennenholtz et al., 2020, Bennett and Kallus, 2021, Shi et al., 2021a]
  - Confounded MDPs [Zhang and Bareinboim, 2016, Wang et al., 2021, Fu et al., 2022, Shi et al., 2022b]

# Summary



# References |

---

- Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*, 2021.
- Bibhas Chakraborty, Eric B Laber, and Yingqi Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3): 714–723, 2013.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.
- Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

## References II

---

- Zuyue Fu, Zhengling Qi, Zhaoran Wang, Zhuoran Yang, Yanxun Xu, and Michael R Kosorok. Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*, 2022.
- Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.
- Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.
- Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

## References III

---

- Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.
- Hoang M Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.
- Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.
- Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.
- Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391, 2021.

## References IV

---

- Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.
- Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.
- James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.
- James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, Aad van der Vaart, et al. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

## References V

---

- Chengchun Shi, Wenbin Lu, and Rui Song. Breaking the curse of nonregularity with subagging: inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21, 2020a.
- Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020b.
- Chengchun Shi, Masatoshi Uehara, and Nan Jiang. A minimax learning approach to off-policy evaluation in partially observable markov decision processes. *arXiv preprint arXiv:2111.06784*, 2021a.
- Chengchun Shi, Runzhe Wan, Victor Chernozhukov, and Rui Song. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pages 9580–9591. PMLR, 2021b.

## References VI

---

- Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, (just-accepted):1–29, 2022a.
- Chengchun Shi, Jin Zhu, Ye Shen, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *arXiv preprint arXiv:2202.10589*, 2022b.
- Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.
- Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

## References VII

---

- Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.
- Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34:21164–21175, 2021.
- Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.
- Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.
- Baquin Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

## References VIII

---

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical report, Technical Report R-23, Purdue AI Lab, 2016.

## Appendix: Proof of Regret

---

Consider the regret of greedy action selection first. Let  $\mathbf{a}^*$  denote the action selected by the greedy policy. By definition, the regret is given by  $\mathbf{Q}(\mathbf{a}^{opt}) - \mathbf{Q}(\mathbf{a}^*)$ . Notice that

$$\begin{aligned}\mathbf{Q}(\mathbf{a}^{opt}) - \mathbf{Q}(\mathbf{a}^*) &= \mathbf{Q}(\mathbf{a}^{opt}) - \widehat{\mathbf{Q}}(\mathbf{a}^{opt}) + \widehat{\mathbf{Q}}(\mathbf{a}^{opt}) - \widehat{\mathbf{Q}}(\mathbf{a}^*) + \widehat{\mathbf{Q}}(\mathbf{a}^*) - \mathbf{Q}(\mathbf{a}^*) \\ &\leq \mathbf{Q}(\mathbf{a}^{opt}) - \widehat{\mathbf{Q}}(\mathbf{a}^{opt}) + \widehat{\mathbf{Q}}(\mathbf{a}^*) - \mathbf{Q}(\mathbf{a}^*),\end{aligned}$$

as  $\mathbf{a}^*$  maximizes  $\arg \max_{\mathbf{a}} \widehat{\mathbf{Q}}(\mathbf{a})$  by definition.

It is immediate to see that the right-hand-side is upper bounded by  $2 \max_{\mathbf{a}} |\widehat{\mathbf{Q}}(\mathbf{a}) - \mathbf{Q}(\mathbf{a})|$ . The proof is thus completed.

## Appendix: Proof of Regret (Cont'd)

---

Next, consider the regret of the LCB algorithm. Let  $a^*$  denote the action selected by the LCB algorithm. By definition of  $L(a^*)$ , we have with probability approaching 1 that

$$Q(a^{opt}) - Q(a^*) \leq Q(a^{opt}) - \hat{Q}(a^*) + L(a^*).$$

According to the LCB algorithm,  $\hat{Q}(a^*) - L(a^*) \geq \hat{Q}(a^{opt}) - L(a^{opt})$ . It follows that the right-hand-side is upper bounded by

$$Q(a^{opt}) - \hat{Q}(a^{opt}) + L(a^{opt}),$$

which is further bounded by  $2L(a^{opt})$ , by definition. The proof is completed by directly applying Hoeffding's inequality.

**Thank you!**