Reinforcement Learning

Lecture 5: Offline Reinforcement Learning

Chengchun Shi

Lecture Outline

Lecture Outline

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)

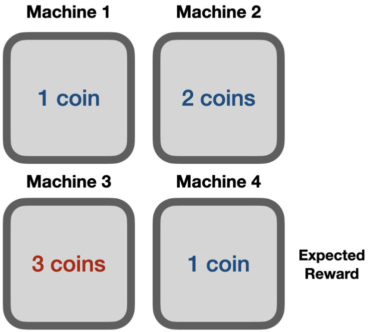
Lecture Outline

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)
- $A_t \in \{1, \cdots, k\}$: 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



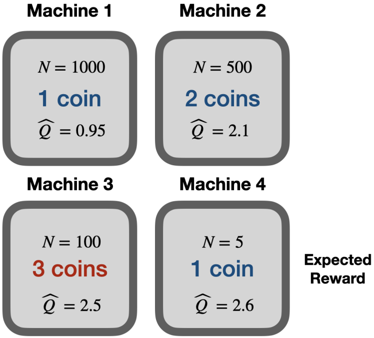
Greedy Action Selection

- **Action-value methods**:

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

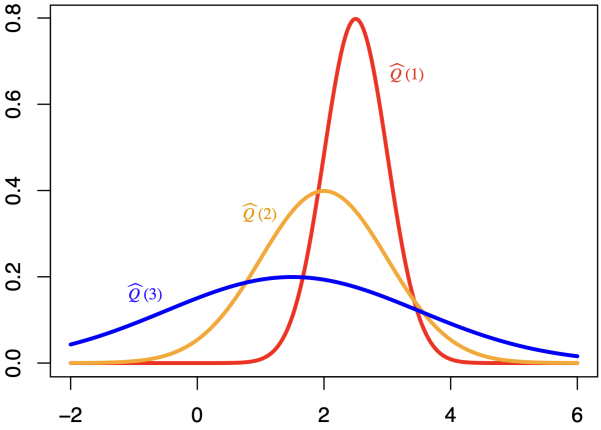
  where $N(a) = \sum_{t=0}^{T-1} \mathbb{I}(A_t = a)$ denotes the action counter

- **Greedy policy**: $\arg\max_a \widehat{Q}(a)$
- Less-explored action $\to$ $N(a)$ is small $\to$ inaccurate $\widehat{Q}(a)$ $\to$ **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** $U_t(a)$ for each action value such that

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

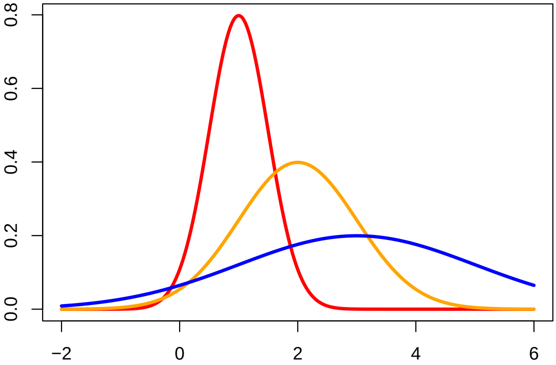
  with high probability.
- $U_t(a)$ quantifies the **uncertainty** and depends on $N_t(a)$ (number of times arm a has been selected up to time t)
  - Large $N_t(a) \rightarrow$ small $U_t(a)$;
  - Small $N_t(a) \rightarrow$ large $U_t(a)$.
- Select actions maximizing upper confidence bound

$$a^* = \arg\max_a [\widehat{Q}_t(a) + U_t(a)].$$

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

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** $L(a)$ for each action value such that

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

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

$$a^* = \arg\max_a [\widehat{Q}(a) - L(a)].$$

Lower Confidence Bound (Cont'd)

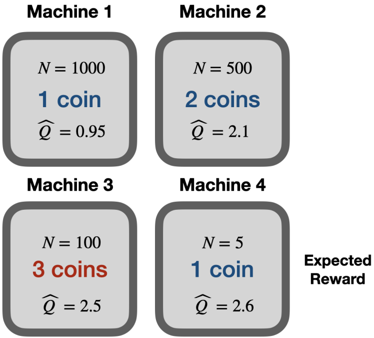
- Set $L(a) = \sqrt{c \log(T)/N(a)}$ for some positive constant c where T is the sample size of historical data
- According to **Hoeffding's inequality** (link), when rewards are bounded between 0 and 1, the event

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

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

Lower Confidence Bound (Cont'd)

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



Algorithm

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

$$\widehat{Q}(A_t) \leftarrow \frac{N(A_t) - 1}{N(A_t)} \widehat{Q}(A_t) + \frac{1}{N(A_t)} R_t.$$

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

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

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 $2 \max_a |\widehat{Q}(a) - Q(a)|$, *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

Offline Policy Optimization and Fitted Q-Iteration

- Offline data: $\{(S_t, A_t, R_t) : 0 \le t \le T\}$
- Fitted Q-Iteration can be naturally applied by repeating
    1. Compute $\widehat{Q}$ as the argmin of

$$\arg\min_{Q} \sum_{t} \left[R_t + \gamma \max_{a} \widetilde{Q}(S_{t+1}, a) - Q(S_t, A_t) \right]^2$$

    2. Set $\widetilde{Q} = \widehat{Q}$
- **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

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

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

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

  where the lower bound satisfies that with probability approaching 1,

$$Q^{\pi^{\text{opt}}}(a,s) \geq \widehat{Q}(a,s) - L(a,s), \qquad \forall a, s.$$

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

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

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

MOPO: The Pessimism Principle

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

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

  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 λ rollout horizon h, rollout batch size b.

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

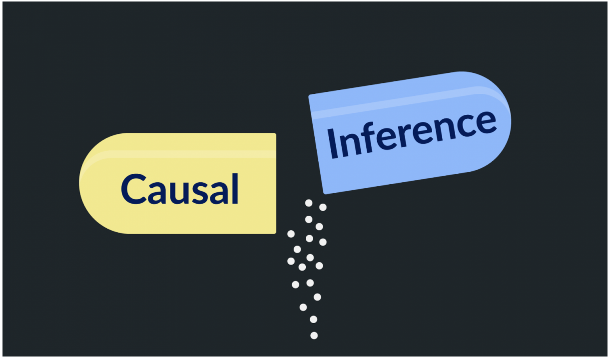
Lecture Outline

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

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 $\{(S_t, A_t, R_t) : 0 \le t < T\}$ generated by a behavior policy b, i.e.,

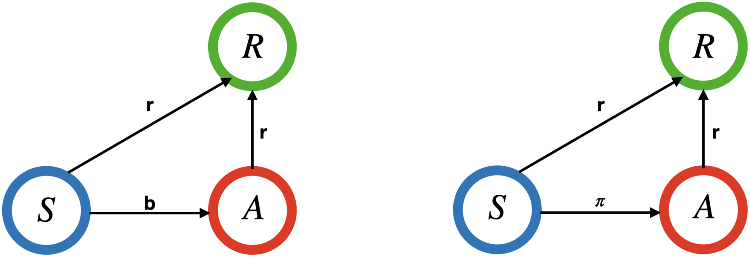
$$\Pr(A_t = a | S_t = s) = b(a|s),$$

we aim to evaluate the mean outcome under a target policy π, 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 π 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

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

- 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 π, 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

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

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

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

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

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

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]

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

for a given estimator $\widehat{b}$

Direct Estimator v.s. IS Estimator

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

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

where $\boldsymbol{\sigma}^{\mathbf{2}} = \mathrm{Var}(\varepsilon)$

Extensions

- When π 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

$$\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right]^{-1}\frac{1}{T}\sum_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}R_t$$

- **Truncated** IS

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

for some $\varepsilon > 0$

Doubly Robust Estimator

- Direct estimator

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

  requires $\widehat{r}$ to be consistent
- IS estimator

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

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

Doubly Robust Estimator (Cont'd)

- Consider the estimating function

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

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

Fact 1: Double Robustness

- The estimating function

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

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

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

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

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

$$= \mathbb{E}\left[\frac{\pi(A|S)}{b(A|S)}R\right] = \text{target policy's value}$$

Fact 2: Efficiency

- When $\widehat{\boldsymbol{b}} = \boldsymbol{b}$, the estimating function

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

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

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

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

Fact 3: Efficiency

- When $\widehat{\boldsymbol{b}}$ 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 $\boldsymbol{b}$ [Tsiatis, 2007]
- Estimating $\widehat{\boldsymbol{b}}$ yields a more efficient estimator, even if we know the oracle $\boldsymbol{b}$
- **Multi-armed bandit** example without context information
  - **Objective**: evaluate $\mathbb{E}(\boldsymbol{R}|\boldsymbol{A} = \boldsymbol{a})$ for a given $\boldsymbol{a}$
  - IS estimator with **known** $\Pr(\boldsymbol{A} = \boldsymbol{a})$

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

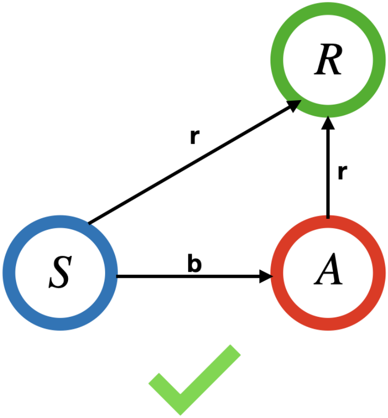
  - IS estimator with **estimated** $\Pr(\boldsymbol{A} = \boldsymbol{a})$ has a **smaller** asymptotic variance

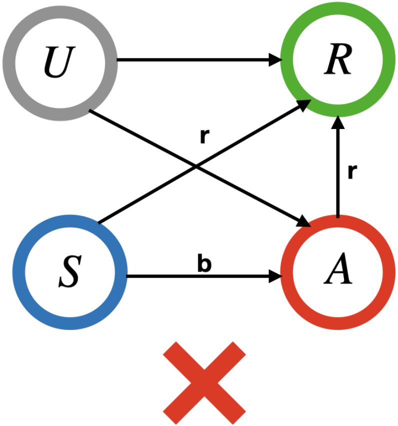
$$\frac{\sum_{t=0}^{T-1} \mathbb{I}(\boldsymbol{A}_t = \boldsymbol{a})\boldsymbol{R}_t}{\sum_{t=0}^{T-1} \mathbb{I}(\boldsymbol{A}_t = \boldsymbol{a})}$$

Fact 4: Asymptotic Normality

- The DR estimator converges at a parametric rate and is asymptotically normal even when both $\widehat{r}$ and $\widehat{b}$ 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 $\widehat{r}$ and $\widehat{b}$ to converge at a rate of $o_p(n^{-1/4})$, due to the double robustness property

Assumption: No Unmeasured Confounders





Lecture Outline

General OPE Problem

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

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

  When $\gamma = 1$, the task is assumed to be episodic
- We focus on the case where both π and b are **stationary** policies
- Challenge: **Distributional shift**
  - In the offline dataset, actions are generated according to b
  - The target policy π we wish to evaluate is different from b

Direct Estimator

- The target policy's value is given by $\mathbb{E} V^{\pi}(S_0)$, or equivalently,

$$\mathbb{E}[\sum_{a} \pi(a|S_0) Q^{\pi}(S_0, a)]$$

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

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{a} \pi(a|S_{i,0}) \widehat{Q}(S_{i,0}, a)$$

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

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

- Bellman equation

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

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

$$\arg\min_{Q} \sum_{t} \left[R_{i,t} + \gamma \sum_{a} \pi(a|S_{i,t+1}) \widetilde{Q}(S_{i,t+1}, a) - Q(S_{i,t}, A_{i,t})\right]^2$$

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

Other Direct Estimators

- Sieve-based estimator [Shi et al., 2020b]
  - Use linear sieves to parametrize $\boldsymbol{Q^\pi}$
  - Estimate regression coefficients by solving the Bellmen equation
- Kernel-based estimator [Liao et al., 2021]
  - Use RHKSs to parametrize $\boldsymbol{Q^\pi}$
  - 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 $\boldsymbol{T}$ is the termination time
- Importance sampling ratio needs to be employed

$$
\begin{aligned}
\mathbb{E}^{\pi} R_0 &= \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)} R_0\right] \\
\mathbb{E}^{\pi} R_1 &= \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)} \frac{\pi(A_1|S_1)}{b(A_1|S_1)} R_1\right] \\
&\vdots \\
\mathbb{E}^{\pi} R_t &= \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)} \cdots \frac{\pi(A_t|S_t)}{b(A_t|S_t)} R_t\right]
\end{aligned}
$$

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

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

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

Limitation

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

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

- Variance of the ratio grows **exponentially** with respect to $\boldsymbol{t}$, 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

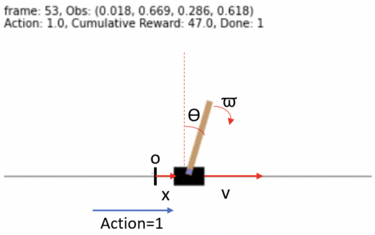
- Bias/Variance Trade-Off
- When $\boldsymbol{b}$ is known, sequential IS is an **unbiased** estimator since

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

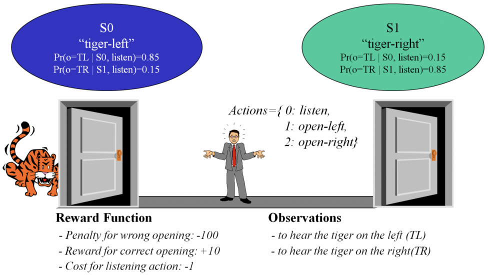
- Direct estimator has **some bias**, since Q^{π} needs to be estimated from data
- sequential IS suffers from **curse of horizon** and a **large variance**
- Direct estimator has a much lower variance

Pros & Cons of Direct v.s. s,a,r IS (Cont'd)

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



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



Marginalized IS Estimator

- As we have discussed, s,a,r 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
- s,a,r IS does **not** exploit these properties

Marginalized IS Estimator (Cont'd)

- s,a,r IS uses the **cumulative** IS ratio

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

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

$$\mathbb{E}^{\pi} R_t = \mathbb{E}^{b}\left[\frac{p_t^{\pi}(S_t, A_t)}{p_t^{b}(S_t, A_t)} R_t\right] \qquad (1)$$

  where p_t^{π} and p_t^{b} are the marginal density functions of (S_t, A_t) under π and b

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

Marginalized IS Estimator

- Under Markovianity and time-homogeneity (MDP),

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

  where p_∞ denotes the limiting state-action distribution under b and the numerator corresponds to the γ-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

$$\text{Direct Estimator} + \text{Augmentation Term}$$

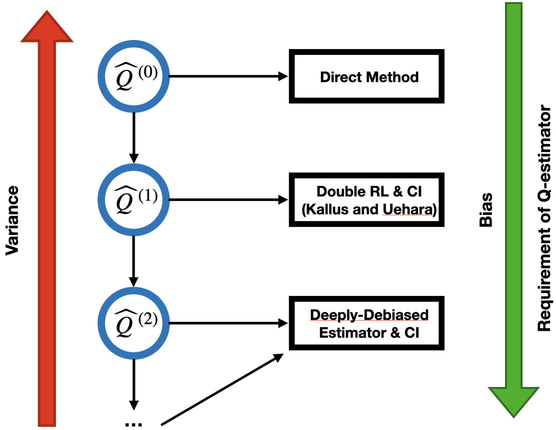
- **Augmentation** term is to **debias** the bias of direct estimator and offer protection against model misspecification of $\boldsymbol{Q^{\pi}}$; it relies on the marginalized IS ratio
- Similar to DR, the estimator is **doubly-robust**, e.g., consistent when either $\boldsymbol{Q^{\pi}}$ or the marginalized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

$$\text{direct estimator} = \text{marginalized IS estimator} = \text{double RL estimator}$$

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

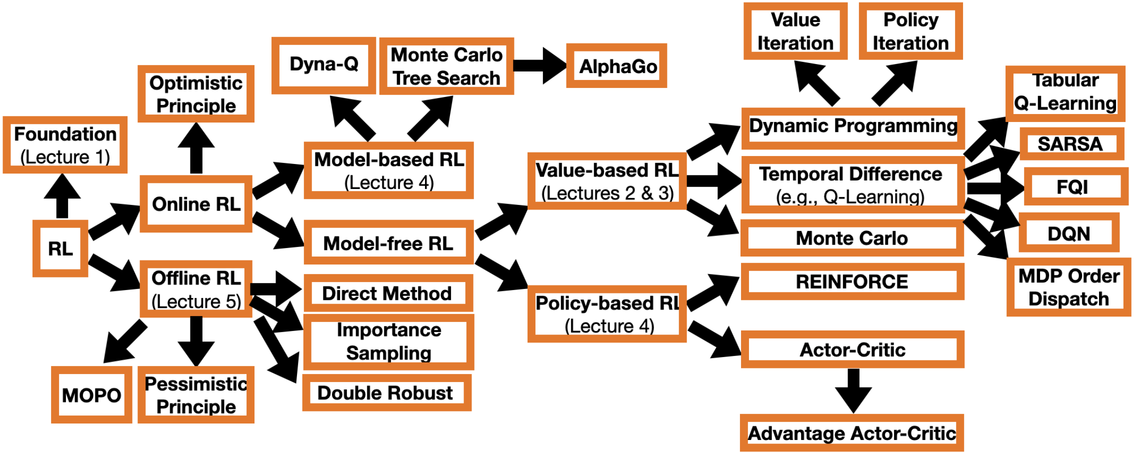


- Constructed based on high-order influence function [Robins et al., 2008, 2017]
- Ensures bias decays much faster than standard deviation
- Allows to provide valid **uncertainty quantification** (e.g., confidence interval)

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



RL
Foundation (Lecture 1)
Online RL
Offline RL (Lecture 5)
Optimistic Principle
MOPO
Pessimistic Principle
Dyna-Q
Monte Carlo Tree Search
AlphaGo
Model-based RL (Lecture 4)
Model-free RL
Direct Method
Importance Sampling
Double Robust
Value-based RL (Lectures 2 & 3)
Policy-based RL (Lecture 4)
Value Iteration
Policy Iteration
Dynamic Programming
Temporal Difference (e.g., Q-Learning)
Monte Carlo
REINFORCE
Actor-Critic
Advantage Actor-Critic
Tabular Q-Learning
SARSA
FQI
DQN
MDP Order Dispatch

References I

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.

Baqun 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 a^* denote the action selected by the greedy policy. By definition, the regret is given by $Q(a^{opt}) - Q(a^*)$. Notice that

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

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

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

Appendix: Proof of Regret (Cont'd)

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

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

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

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

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

Thank you!