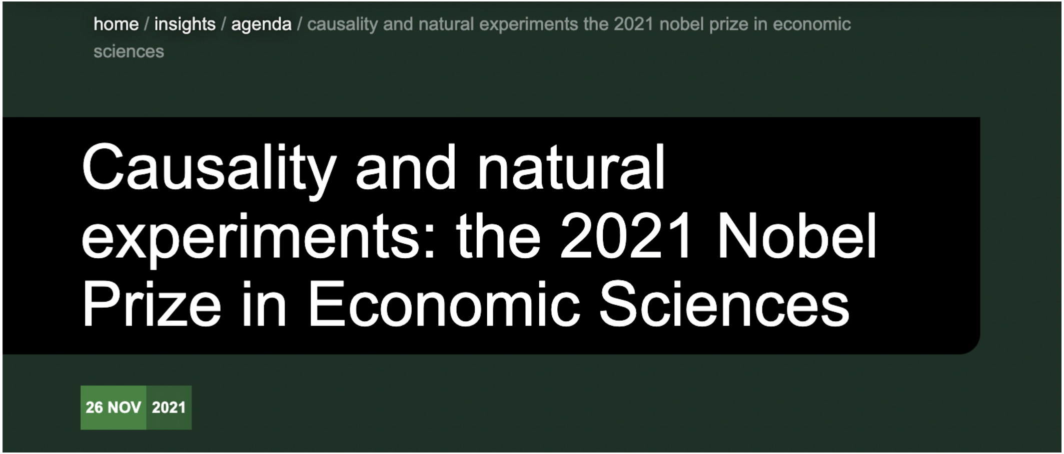
A Reinforcement Learning Framework for Dynamic Causal Effects Evaluation

Chengchun Shi

Assistant Professor of Data Science

London School of Economics and Political Science

Causal Inference


home / insights / agenda / causality and natural experiments the 2021 nobel prize in economic sciences

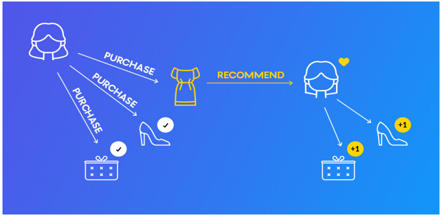
Causality and natural experiments: the 2021 Nobel Prize in Economic Sciences

26 NOV 2021

Causal Inference Applications

(a) Economics

(b) Health Care

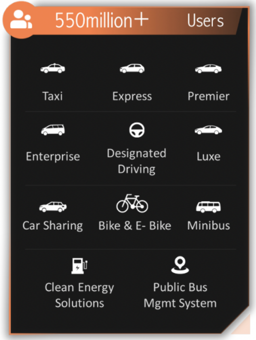


(c) E-commerce Platforms

(d) Ridesharing

We focus on applications in **ridesharing**

Applications in Ridesharing



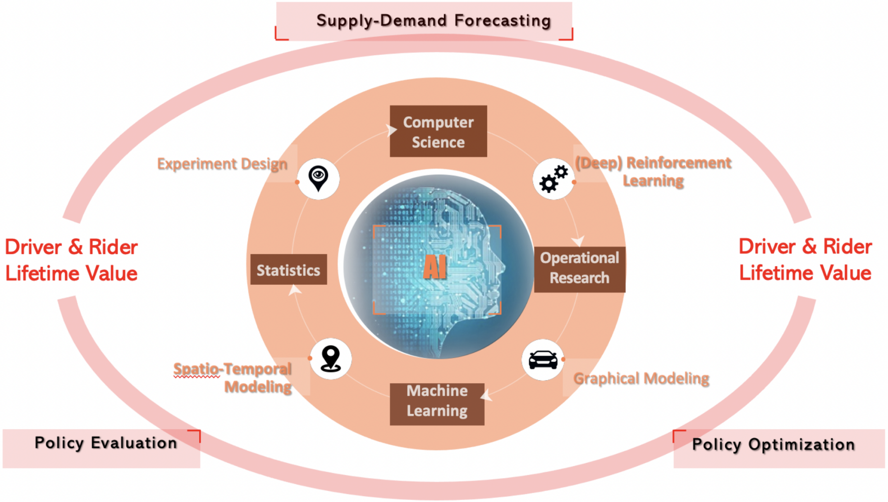
106TB+ vehicle trajectory data/day

4875TB+ data processed/day

40billion+ routing requests/day

15billion+ location points/day

Applications in Ridesharing (Cont'd)

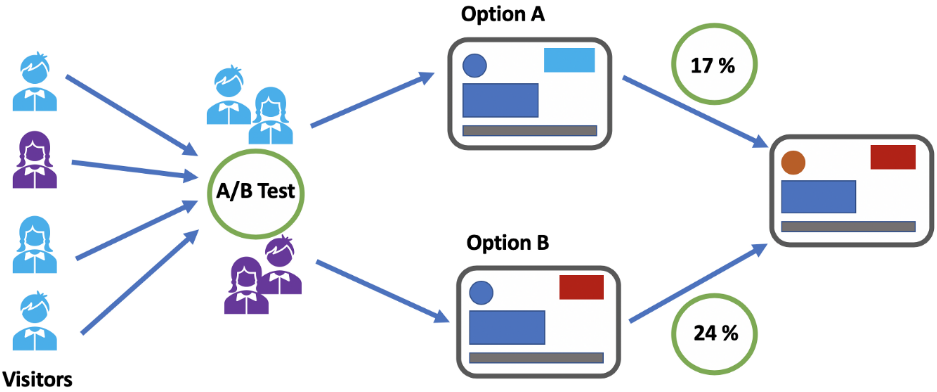


Project I

Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework

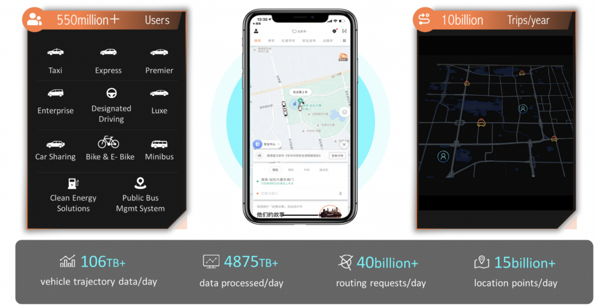
Joint work with Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye and Rui Song
——JASA, accepted

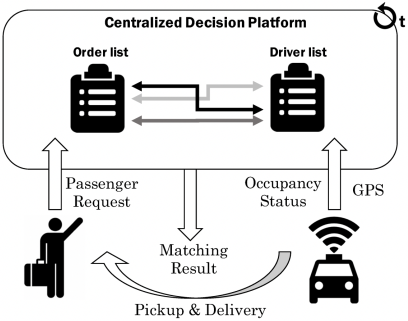
A/B Testing



Taken from
`https://towardsdatascience.com/how-to-conduct-a-b-testing-3076074a8458`

Motivation: Order Dispatch

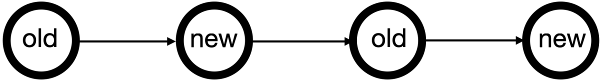




Our project is motivated by the need for comparing the **long-term rewards** of different **order dispatching** policies in **ridesharing platforms**

Data

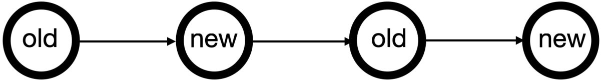
- Data from an **online experiment** that last for two weeks
- **30 minutes/1 hour** as one time unit
- **Time-varying variables S_t**: e.g., number of drivers (supply), number of call orders (demand)
- **Treatment A_t**: new policy v.s. old policy; adopts an alternating-time-interval (switchback) design



- **Outcome R_t**:
  - **Answer rate** (percentage of call orders being responded by drivers)
  - **Completion rate** (percentage of call orders being completed)
  - **Drivers' income**

Challenges

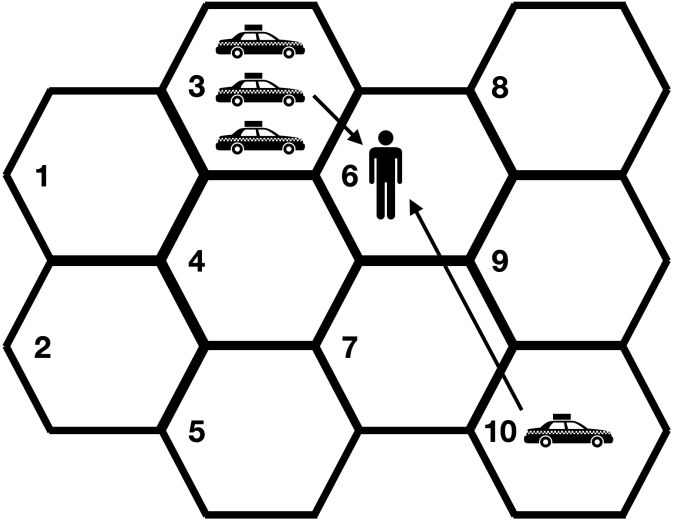
1. The existence of **carryover effects**:
    - Under the alternating-time-interval (or switchback) design



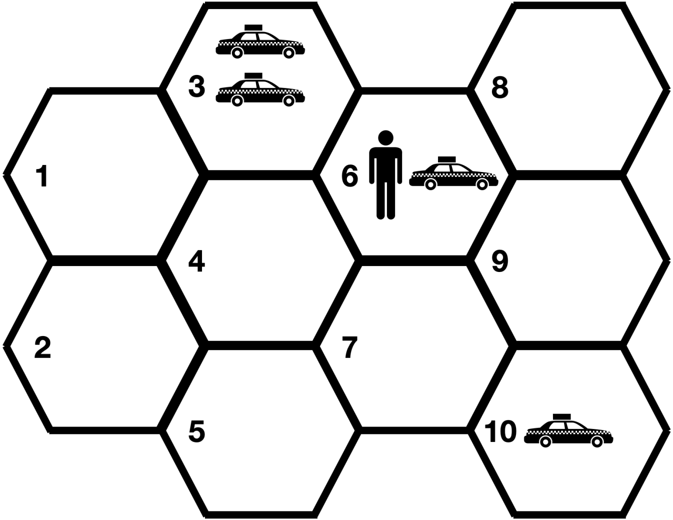
    - Past actions will affect future outcomes
2. The need for **early termination**:
    - Each experiment takes a considerable time (at most 2 weeks)
    - Early termination to save time and budget
3. The need for **adaptive randomization**:
    - Maximize the total reward (e.g., epsilon-greedy)
    - Detect the alternative faster

To our knowledge, **no** existing test has addressed three challenges simultaneously

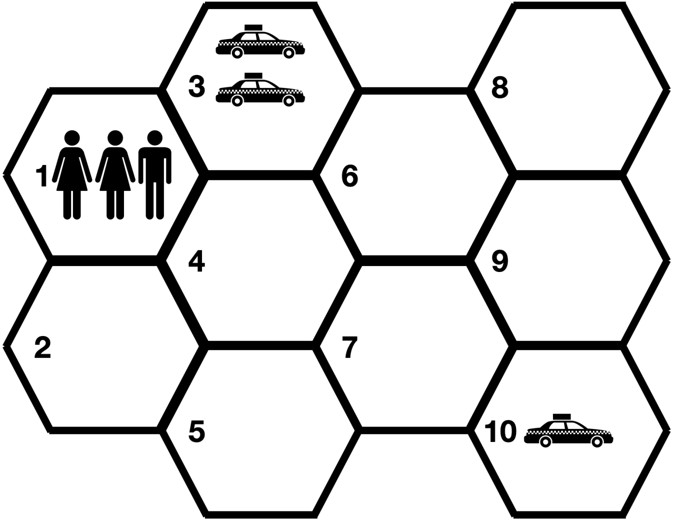
Illustration of the Carryover Effects



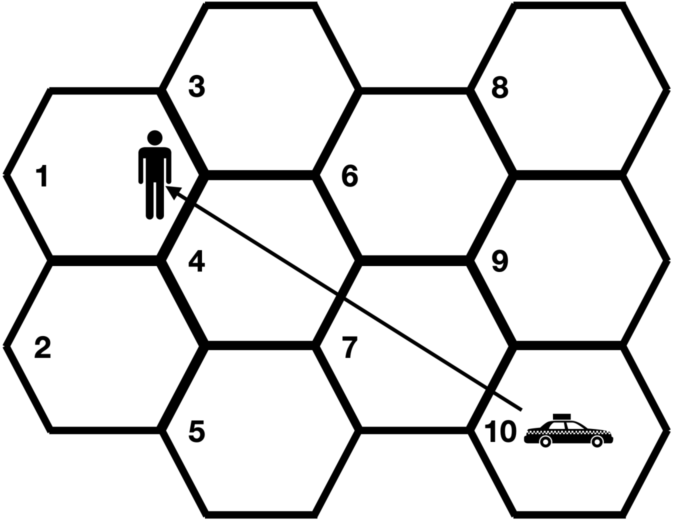
Adopting the Closest Driver Policy



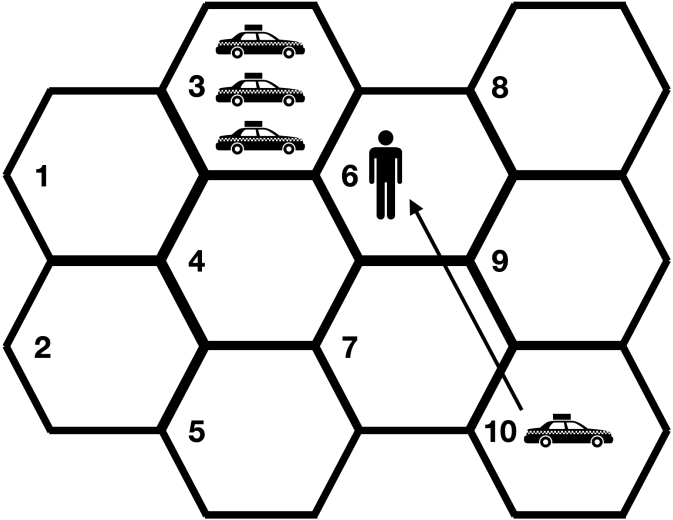
Some Time Later · · ·



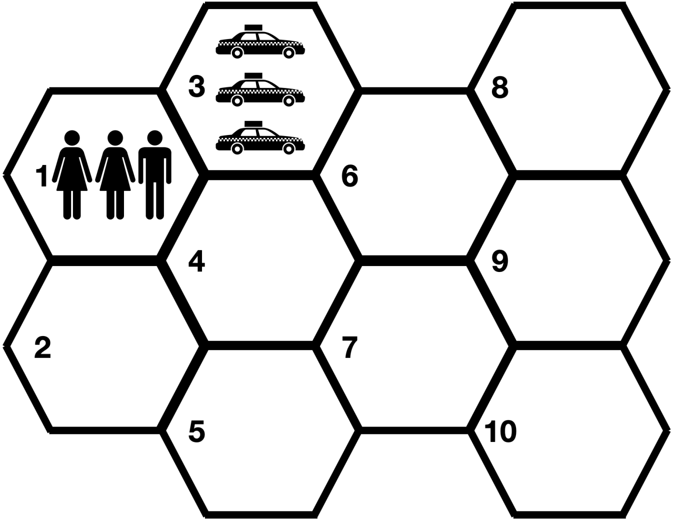
Miss One Order



Consider a Different Action



Able to Match All Orders



Existence of Carryover Effects

past actions $\rightarrow$ distribution of drivers $\rightarrow$ future rewards

Limitations of Existing A/B tests

- Most existing tests **cannot** detect carryover effects
- $\mathcal{H}_0$: The old policy ($\boldsymbol{A} = 0$) has larger cumulative rewards
- $\mathcal{H}_1$: The new policy ($\boldsymbol{A} = 1$) has larger cumulative rewards
- **Example 1**. $\boldsymbol{S_t} \sim N(0, 0.25), \boldsymbol{R_t} = \boldsymbol{S_t} + \delta \boldsymbol{A_t}$
- **Example 2**. $\boldsymbol{S_t} = 0.5\boldsymbol{S_{t-1}} + \boldsymbol{A_{t-1}} + N(0, 0.25), \boldsymbol{R_t} = \boldsymbol{S_t}$

Example 1	t-test 0.76	DML-based test **1.00**	our test **0.98**
Example 2	t-test 0.04	DML-based test 0.06	our test **0.73**

Table: Powers of t-test, DML-based test (Chernozhukov et al., 2018) and the proposed test with $T = 500, \delta = 0.1$

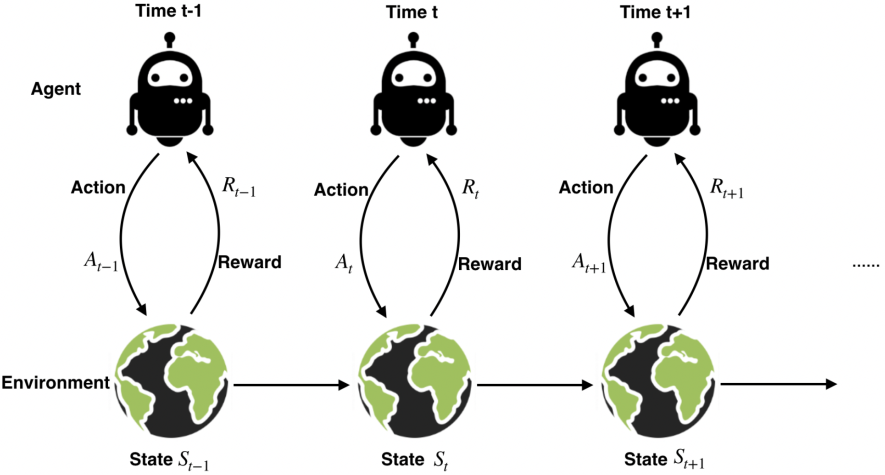
Contributions and Advances of Our Proposal

- Introduce an RL framework for A/B testing
    1. Allow to measure long-term rewards using **value function**
    2. Model carryover effects using the **dynamic system transitions** (address Challenge 1)
    3. Enable **consistent** estimation with a **single** time series
- Propose an original test procedure for comparing long-term rewards of two policies
    1. allows for **sequential monitoring** (address Challenge 2)
    2. allows for **online updating**
    3. applicable to a wide range of designs, including the **Markov** design, **alternating-time-interval** design and **adaptive** design (address Challenge 3)

An RL framework for A/B Testing

- **What** is the RL framework
- **Why** use the RL framework

What is the RL Framework



Objective: find an optimal policy that maximizes the cumulative reward

RL Designed for Sequential Decision Making

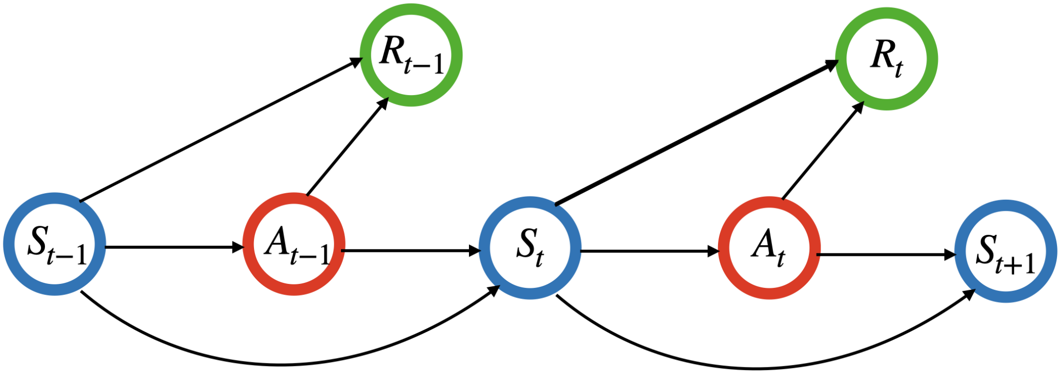
Markov Decision Process

- **RL algorithms**: trust region policy optimization (Schulman et al., 2015), deep Q-network (DQN, Mnih et al., 2015), asynchronous advantage actor-critic (Minh et al., 2016), quantile regression DQN (Dabney et al., 2018).
- **Foundations** of RL:
  - **Markov decision process** (MDP, Puterman, 1994)
  - **Markov assumption**: conditional on the present, the future and the past are independent,

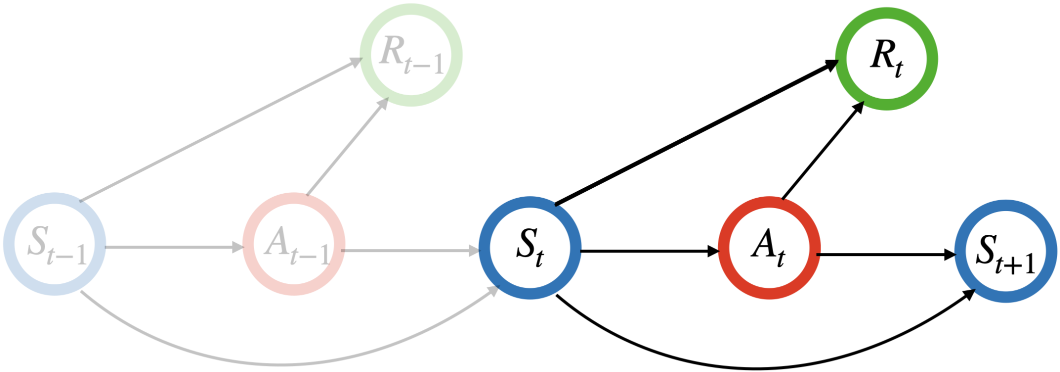
$$S_{t+1}, R_t \perp\!\!\!\perp \{(S_j, A_j, R_j)\}_{j<t} | S_t, A_t.$$

  - **Stationarity assumption**: The Markov transition function is stationary over time.
- By introducing an RL framework, we use the MDP model to formulate the A/B testing problem

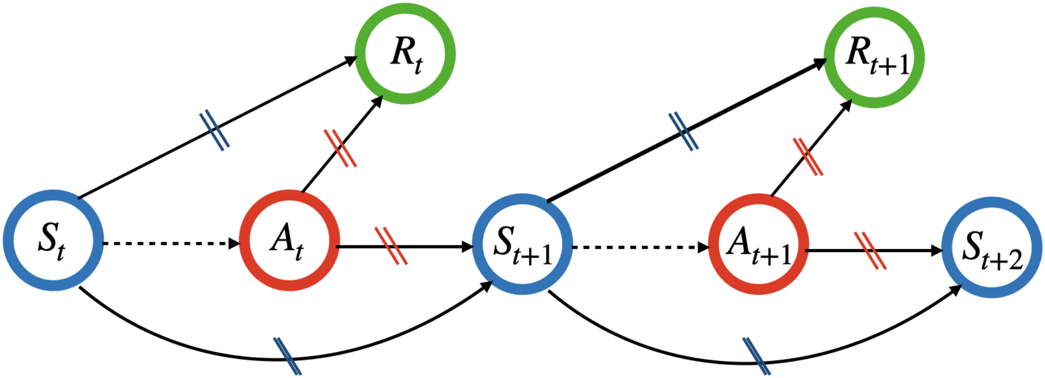
Markov Assumption



Markov Assumption



Stationarity Assumption



Why use the RL framework

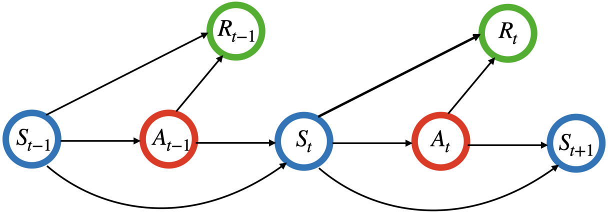
- Allows to measure the long-term rewards using the **value function** in RL

$$V^{a}(s) = \sum_{t \geq 0} \gamma^{t} \mathbb{E}^{a}(R_t | S_0 = s),$$

- The expectation is taken by assuming treatment a is repeatedly assigned all the time
- The **discounted factor** $0 \leq \gamma < 1$ represents the **trade-off** between **immediate** and **future** rewards
- $\gamma = 0$ leads to "**myopic**" evaluation
- γ close to 1 leads to "**far-sighted**" evaluation

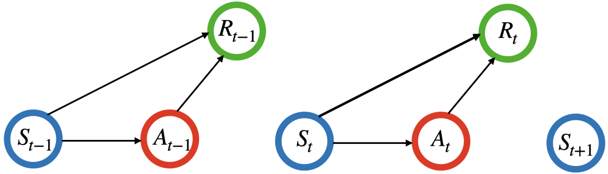
Why use the RL framework (Cont'd)

- Allows to model the carryover effects using the **dynamic state transitions**



  1. $\boldsymbol{A_{t-1}}$ impacts $\boldsymbol{R_t}$ indirectly through its effect on $\boldsymbol{S_t}$
  2. $\boldsymbol{S_t}$ shall include important **mediators** between $\boldsymbol{A_{t-1}}$ and $\boldsymbol{R_t}$

- Most existing works require the independence assumption



Why use the RL framework (Cont'd)

- **Markov** and **stationarity** assumptions allow us to **consistently** estimate the policy's value based on a **single** time series
- These assumptions are **mild**
  - **Concatenate** observations over multiple decision points to meet Markovanity
  - **Include** dummy variables (e.g., peak/off-peak hours) in the state to meet stationarity

Contributions and Advances (Cont'd)

Propose a test procedure for comparing long-term rewards of two policies

1. allows for **sequential monitoring**
2. allows for **online updating**
3. applicable to a wide range of designs, including the **Markov** design, **alternating-time-interval** design and **adaptive** design

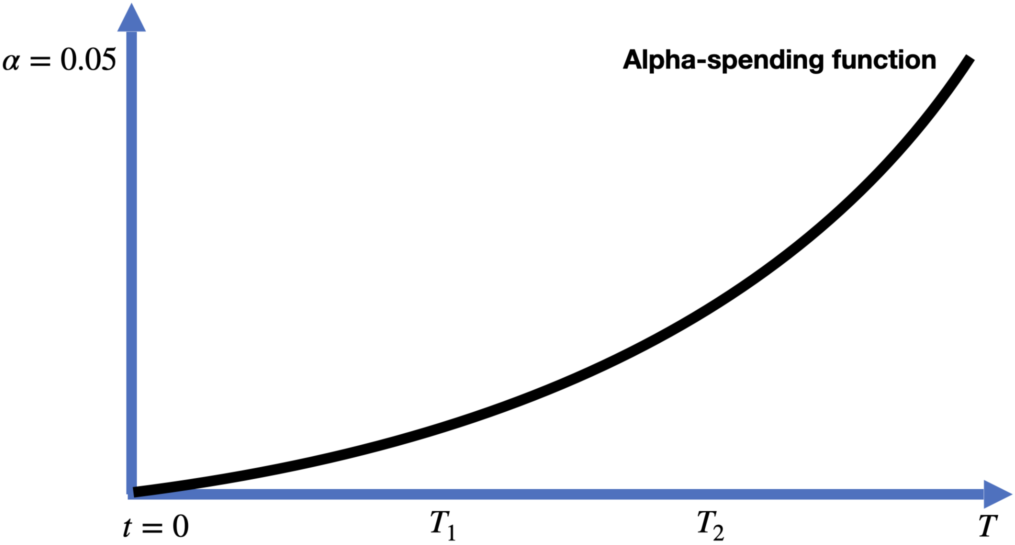
Contributions and Advances (Cont'd)

Algorithm	Carryover effects	Sequential monitoring	Switchback design
Two-sample t-test	✗	✗	✓
Classical sequential tests	✗	✓	✓
Bojinov & Shephard (2019)	✓	✗	✗
V-learning (Luckett et al., 2020)	✓	✗	✗
Double RL (Kallus & Uehara, 2019)	✓	✗	✗
CausalRL (our proposal)	✓	✓	✓

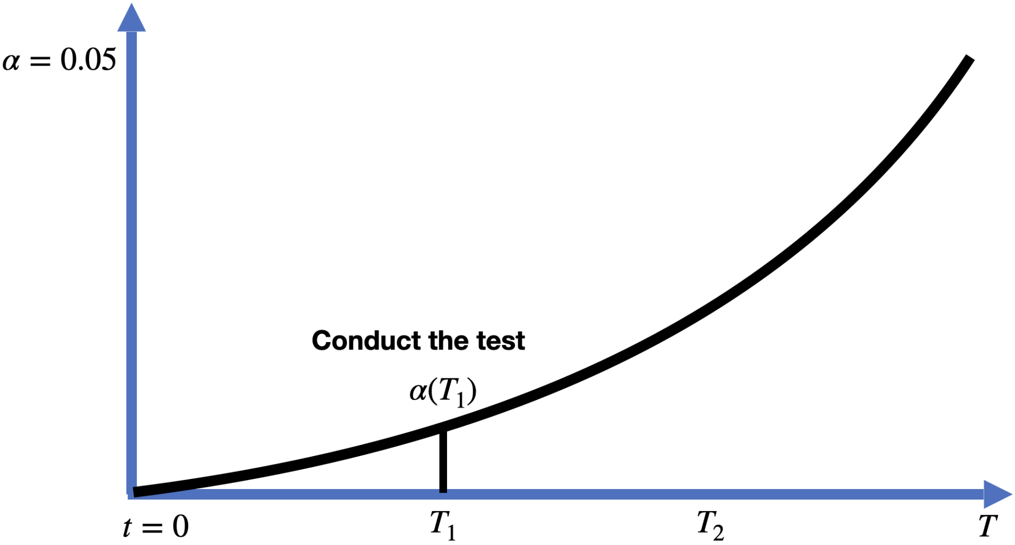
Methodology

- Apply **temporal difference learning** with **sieve** method to evaluate value difference and provide **uncertainty quantification** (Shi et al., 2021, JRSSB)
- Adopt the α-**spending approach** (Lan & DeMets, 1983) for sequential monitoring
- Develop a **bootstrap-assisted procedure** for determining the stopping boundary
  - The numerical integration method designed for classical sequential tests is **not** applicable in adaptive design, due to the carryover effects

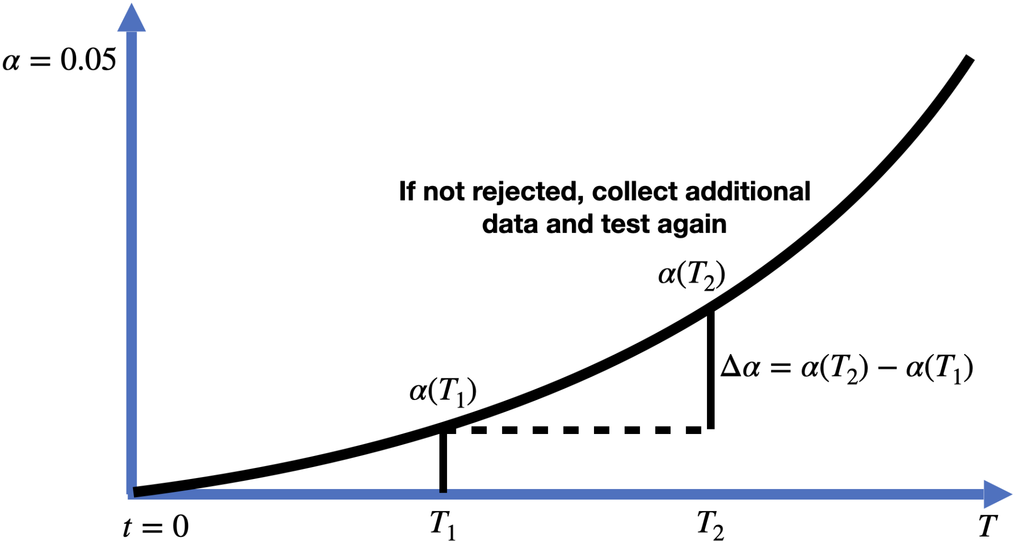
α-Spending Approach



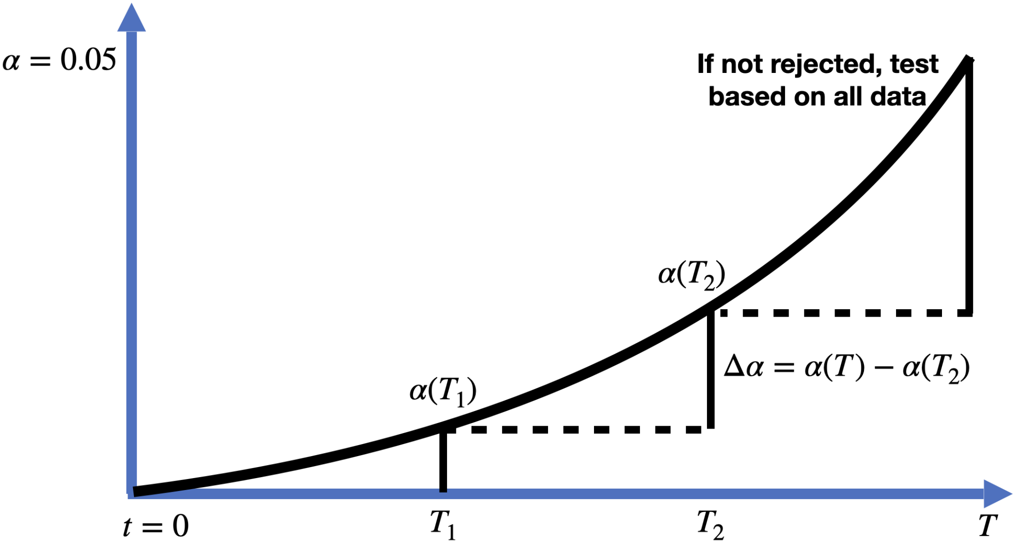
α-Spending Approach



α-Spending Approach



α-Spending Approach



Theory

Theorem (Validity and Consistency)

Under the Markov, alternating-time-interval or adaptive design, the proposed test can **control type-I error** *and is* **consistent** *against alternatives that converge to the null at the parametric rate*

Theory (Cont'd)

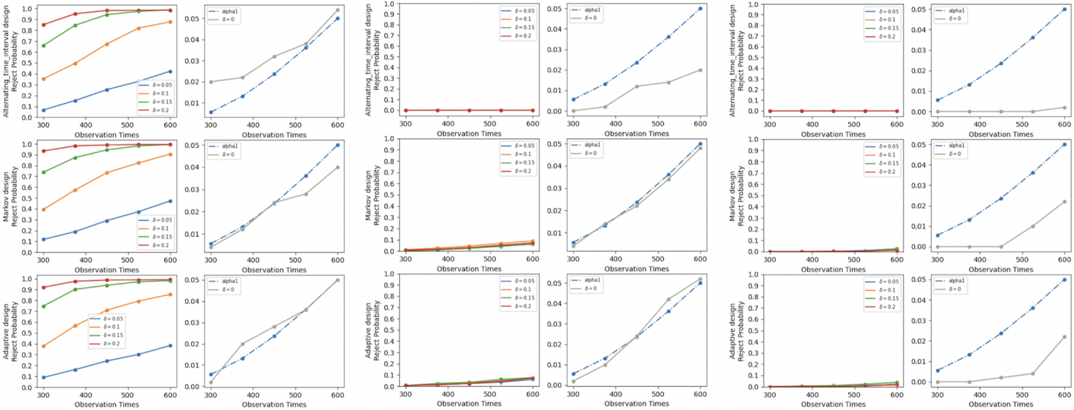
Theorem (Undersmoothing and Efficiency)

Suppose **sieve** *method is used for function approximation in temporal difference learning.*

1. **Undersmoothing** *is not needed to guarantee that the resulting value estimator has a tractable limiting distribution.*
2. *The value estimator is* **semiparametrically efficient***.*

- Sieve estimators of conditional expectations are **idempotent** (Shen et al., 1997)
- The proposed test will **not** be overly sensitive to the number of basis functions
- **Cross-validation** can be employed to select the basis functions

Simulation



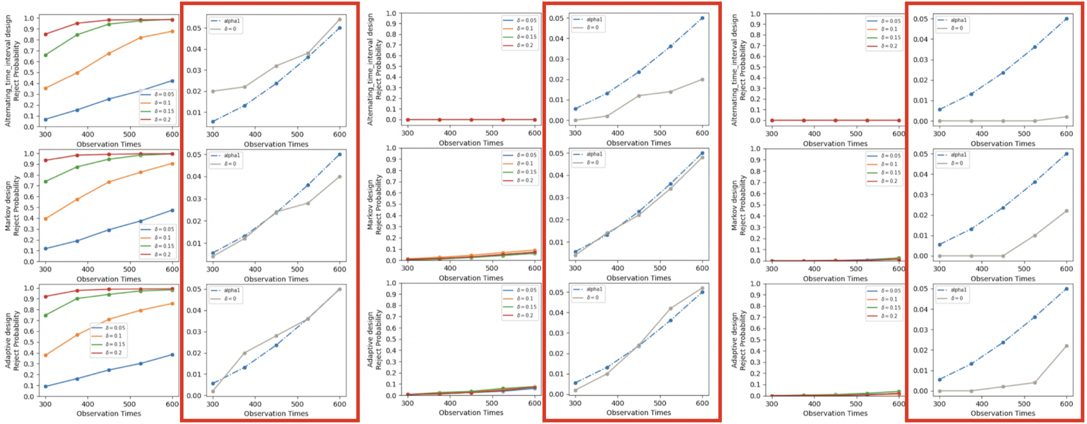
(a) Power and size of our test

(b) Power and size of t test

(c) Power and size of a version of the O'Brien Fleming sequential test

Simulation

Under the null, the blue line denotes the alpha-spending function and the grey line denotes the empirical size



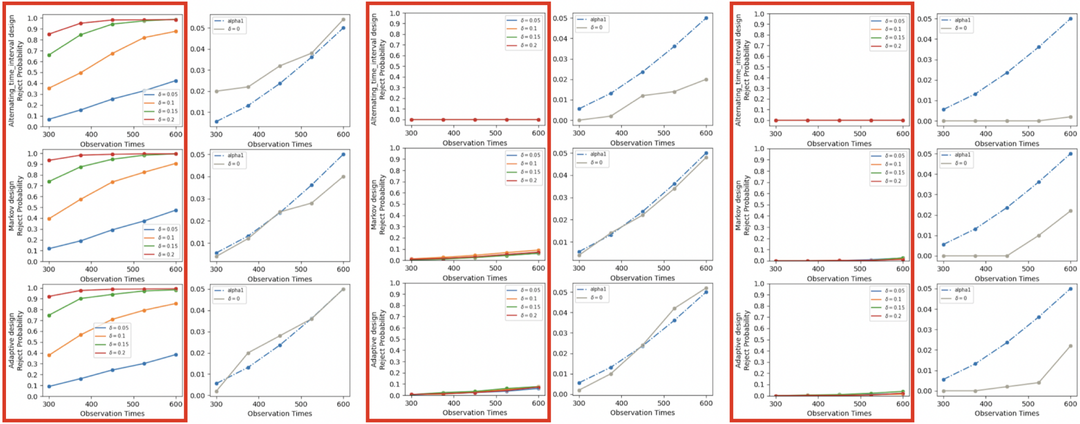
(a) Power and size of our test

(b) Power and size of t test

(c) Power and size of a version of the O'Brien Fleming sequential test

Simulation

Under the alternative, empirical powers



(a) Power and size of our test

(b) Power and size of t test

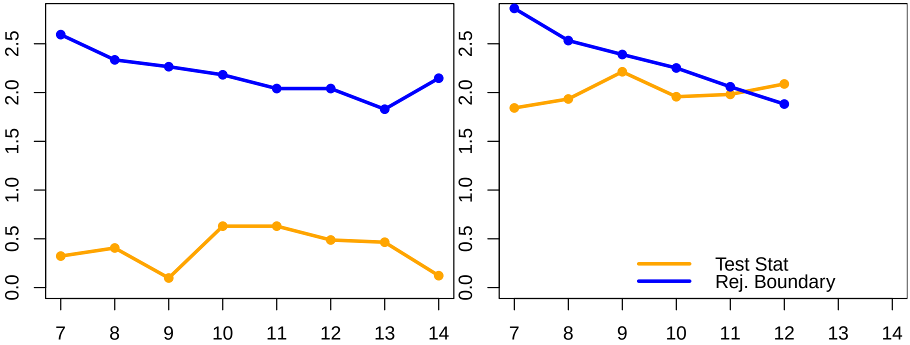
(c) Power and size of a version of the O'Brien Fleming sequential test

Application to Ridesharing Platform

- **Data**: a given city from December 3rd to 16th (two weeks)
- **30 minutes** as one time unit, sample size = **672**
- **State**:
  1. number of drivers (supply)
  2. number of requests (demand)
  3. supply and demand equilibrium metric (mediator)
- **Action**: new policy $\boldsymbol{A} = \mathbf{1}$ v.s. old $\boldsymbol{A} = \mathbf{0}$
- **Reward**: drivers' income
- The new policy is expected to have **better** performance

Application to Ridesharing Platform (Cont'd)

- The proposed test



(a) AA Experiment: Day (b) AB Experiment: Day

- t-test: **fail** to reject $\mathcal{H}_0$ in A/B experiment with p-value 0.18

Project II

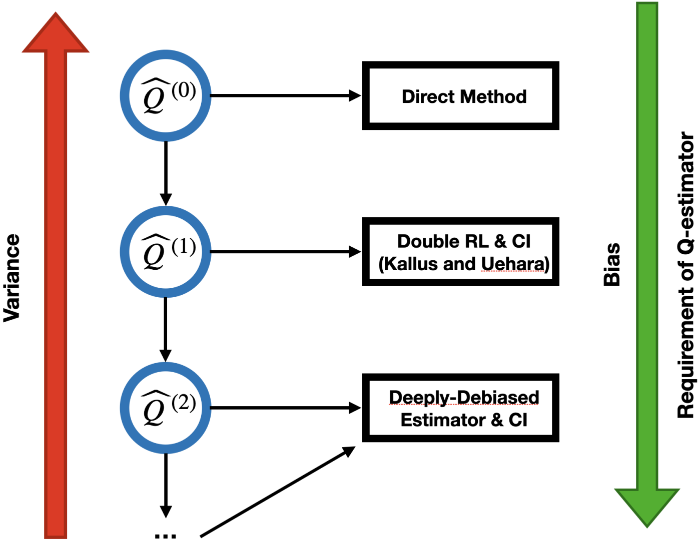
Deeply-Debiased Off-Policy Interval Estimation

joint work with Runzhe Wan, Victor Chernozhukov, and Rui Song
——ICML, 2021 (long talk, top 3% of submissions)

Off-Policy Interval Estimation

- **Objective**: Evaluate the impact of a target policy **offline** using historical data generated from a different behavior policy and provide rigorous **uncertainty quantification** (healthcare, automated driving, ridesharing, robotics, e.g.)
- Consider the reinforcement learning (e.g., MDP) setting
- Most existing methods focus on providing point estimators
- **Main idea**: Develop a **deeply-debiasing** process using higher order influence function (Robins et al., 2017)

Method



Variance
$\widehat{Q}^{(0)}$
Direct Method
$\widehat{Q}^{(1)}$
Double RL & CI
(Kallus and Uehara)
$\widehat{Q}^{(2)}$
Deeply-Debiased
Estimator & CI
...
Bias
Requirement of Q-estimator

Theory

Theorem

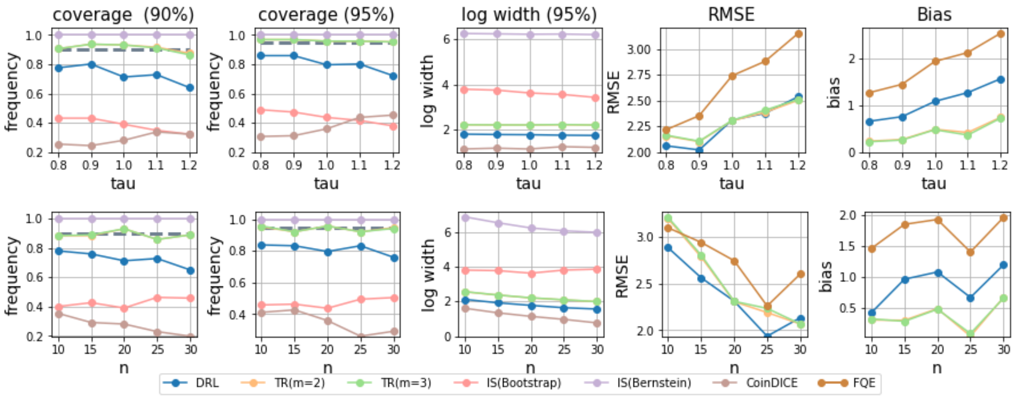
Under certain mild conditions, the proposed method is:

- **robust** *as the value estimator is consistent when one of the three nuisance functions is correct;*
- **efficient** *as it achieves the semiparametric efficiency bound;*
- **flexible** *as it achieves nominal coverage allowing nuisance function to converge at any rate.*

Comparison

Algorithm	Allow High-D?	Semiparametric efficient under MDP?	Rate requirement on nuisance function
Jiang & Li (2016)	✓	✗	$o_p(n^{-1/4})$
Sieve method (Shi et al. 2021)	✗	✓	$o_p(n^{-1/4})$
Double RL (Kallus & Uehara, 2019)	✓	✓	$o_p(n^{-1/4})$
Deeply-Debiased OPE (our proposal)	✓	✓	$O_p(n^{-\kappa})$ for any $\kappa > 0$

Simulation



- Proposed methods are colored in yellow and green (the two lines largely overlapped)
- Competing method either cannot achieve nominal coverage, or is wider than our CI

Thank You!

☺ Papers and softwares can be found on my personal website

`callmespring.githuo.io`

Hiring! I have a postdoc position. More information can be found

`https://jobs.lse.ac.uk/Vacancies/W/3537/0/335760/15539/`

`research-officer-in-statistics`