Policy Evaluation in Reinforcement Learning

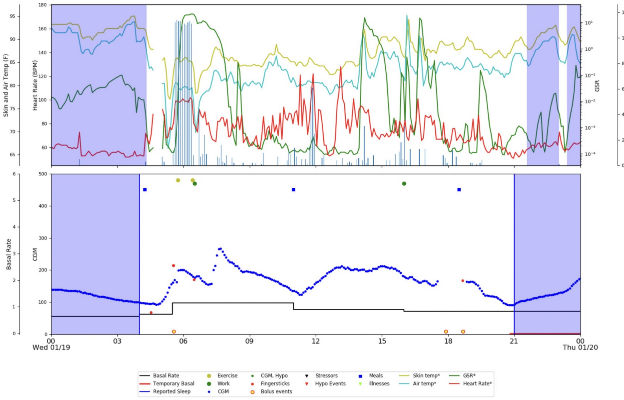
Chengchun Shi

Developing AI with Reinforcement Learning

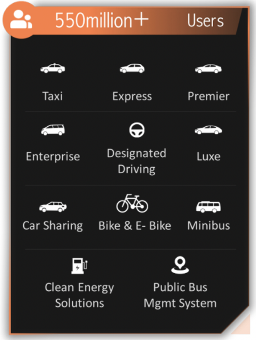
We focus on applications in **mobile health** (mHealth) and **ridesharing**

Applications in mHealth

- Management of **Type-I diabetes**
- **Subject**: Patients with Type-I diabetes
- **Intervention**: Determine whether a patient needs to **inject insulin or not** based on their glucose levels, food intake, exercise intensity
- **Data**: OhioT1DM dataset (Marling and Bunescu, 2018)

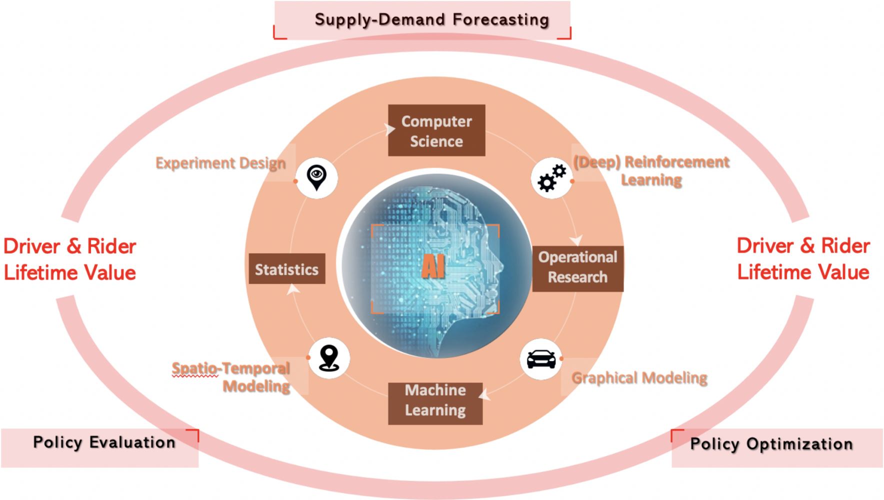


Applications in Ridesharing



106TB+	4875TB+	40billion+	15billion+
vehicle trajectory data/day	data processed/day	routing requests/day	location points/day

Applications in Ridesharing (Cont'd)



What is Off-Policy Evaluation (OPE) and Why OPE

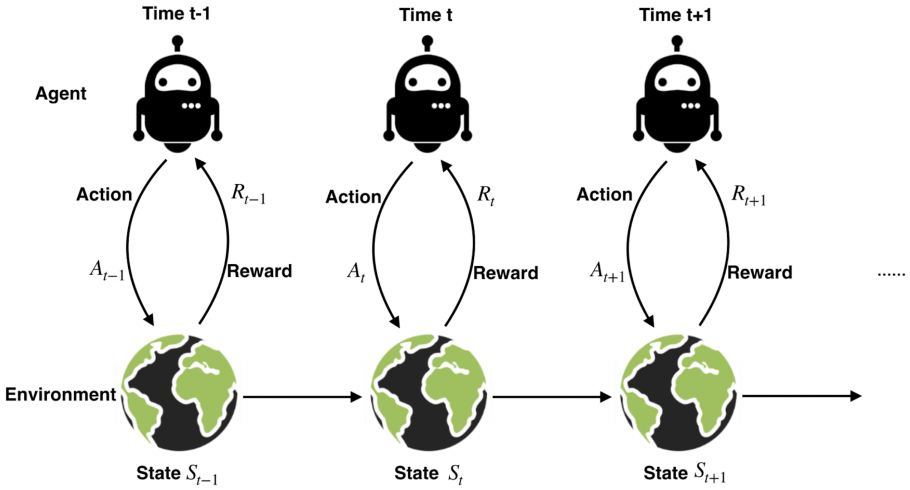
- **Objective**: Evaluate the impact of a **target policy** offline using historical data generated from a different **behavior policy**
- **Motivation**:
  - In many applications, it can be **dangerous** to evaluate a **target policy** by directly running this policy.
    - **Healthcare**: which **medical treatment** to suggest for a patient
    - **Ridesharing**: which **driver** to assign for a call order
  - In additional to a point estimator of the policy value, many applications would benefit significantly from having a **confidence interval** or **p-value** that quantifies the uncertainty of the point estimate.

Project I

Statistical Inference of the Value Function for Reinforcement Learning in Infinite Horizon Settings

Joint work with Sheng Zhang, Wenbin Lu and Rui Song
——JRSSB (2022)

Sequential Decision Making



Objective: infer the expected (discounted) cumulative reward under a target policy

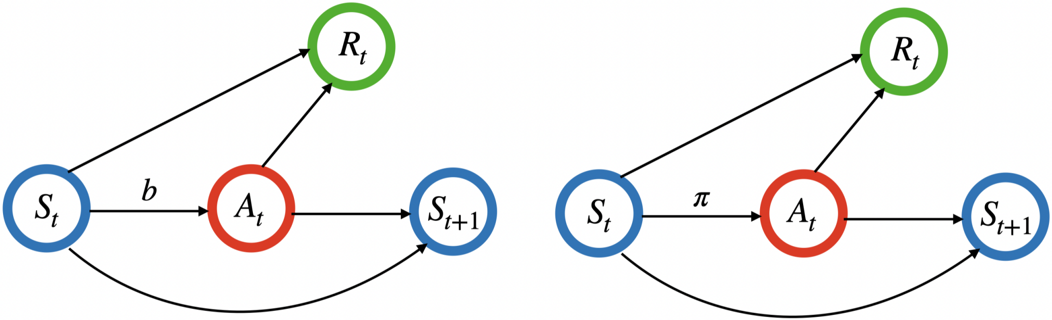
A general framework for inference of the value

- Our proposal:

Policies	Types of values	On/off-policy
Fixed: random ✓ deterministic ✓	CI for the value under a given state ✓	Off-policy ✓
Data-dependent: regular ✓ nonregular ✓	CI for the integrated value with respect to a reference function ✓	On-policy ✓

- Existing literature focus on evaluating a **fixed** policy's **integrated** value in **off-policy** settings.

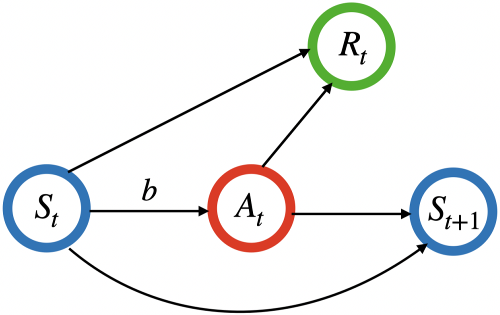
Type-I Inference: fixed off-policy



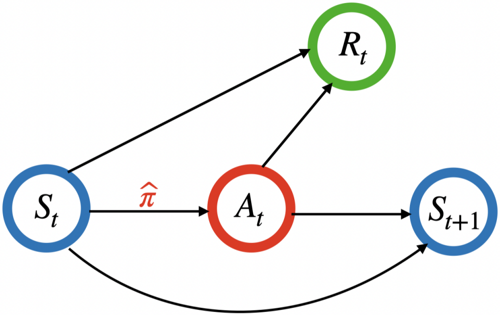
historical data

what we want to evaluate

Type-II Inference: data-dependent off-policy

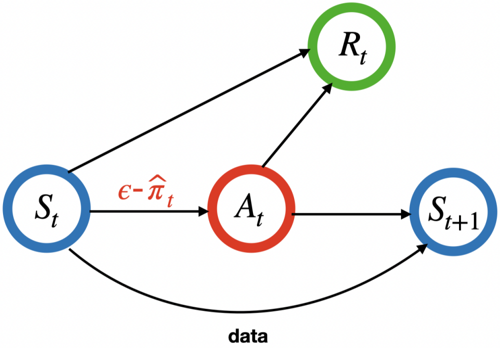


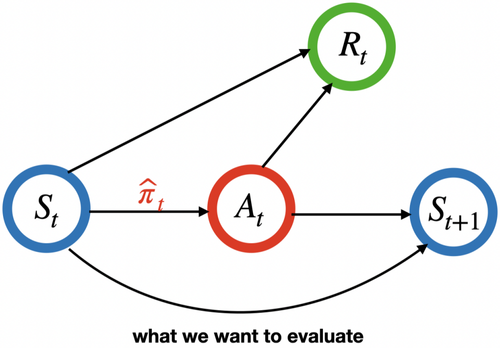
historical data



what we want to evaluate

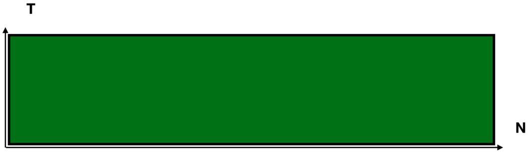
Type-III Inference: data-dependent on-policy



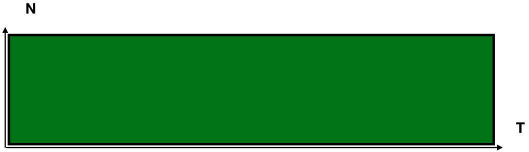


Bidirectional Theory

- N the number of trajectories
- T the number of decision points per trajectory
- **bidirectional asymptotics**: a framework allows either N or $T \to \infty$
- large N, small T (Intern Health Study)

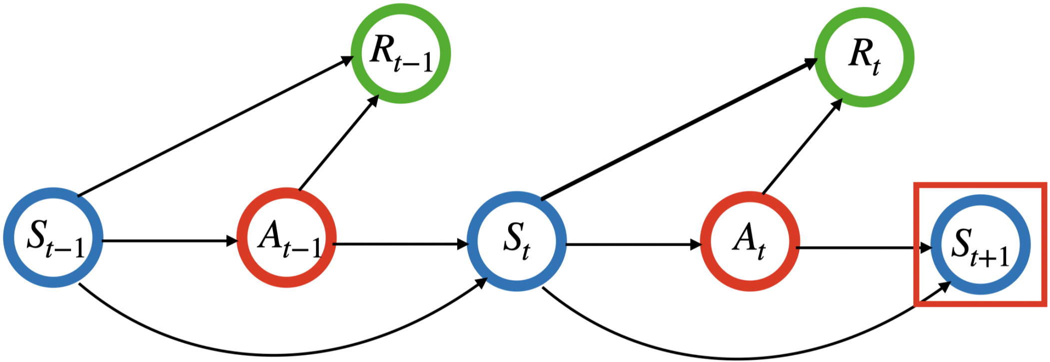


- small N, large T (OhioT1DM dataset)

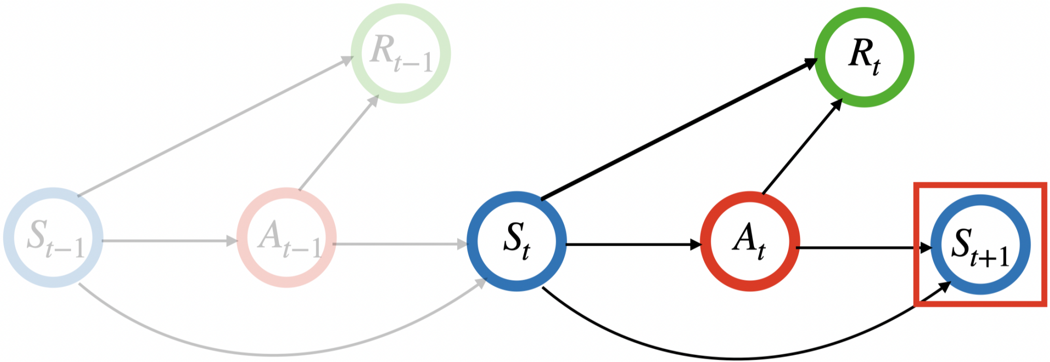


- large N, large T (Games)

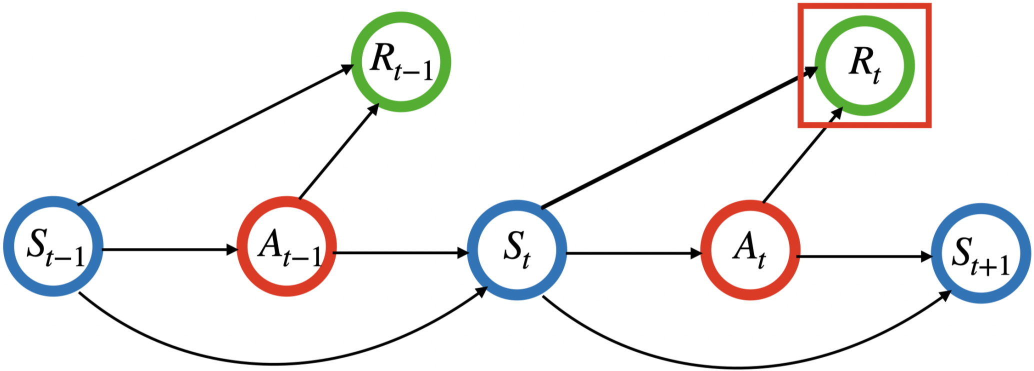
Markov Assumption



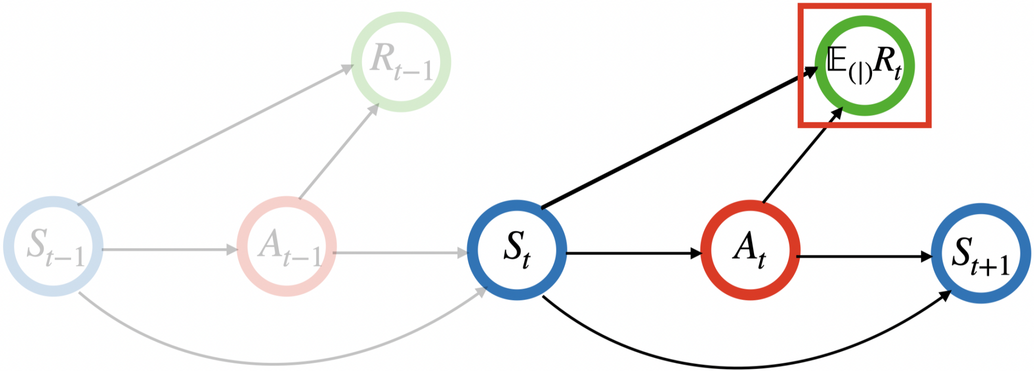
Markov Assumption



Conditional Mean Independence Assumption



Conditional Mean Independence Assumption



Type-I Inference: Method

- Model the **Q-function** via the **sieve** method
  - Directly model the value instead of Q-function poses challenges in performing inference to policies that are **discontinuous** functions of the state
  - Ensure the estimator has a **tractable limiting distribution**
  - Increase the number of sieves to reduce the bias resulting from **model misspecification**
- Derive value estimator based on the estimated Q-function (**direct method**)
- Provide consistent standard error estimators and construct **Wald-type** CI

Type-I Inference: Theory

Theorem (Informal Statement)

Under certain conditions, the proposed CI achieves nominal coverage asymptotically, as either $\boldsymbol{N}$ or $\boldsymbol{T} \to \infty$.

- The proposed estimator is valid under **bidirectional** asymptotics
- Classical augmented inverse propensity score estimator (Zhang et al., 2013) is inefficient and its consistency requires $\boldsymbol{N} \to \infty$.
- **Undersmoothing** is not needed to guarantee that the resulting value estimator has a tractable limiting distribution
  - Sieve estimators of conditional expectations are **idempotent** (Shen et al., 1997)
  - The proposed CI will **not** be overly sensitive to the number of basis functions
- **Cross-validation** can be employed to select the basis functions
- Refer to Section E.2.1 of Shi et al. (2022; Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework)

Type-II Inference: Challenges and Methods

- Considers evaluating the value of a **data-dependent** policy $\widehat{\pi}$ in **off-policy** settings
- Suppose $\widehat{\pi}$ is computed by some Q-learning type algorithms,

$$\widehat{\pi}(a|s) = \begin{cases} 1, & \text{if } \ a = \arg\max_{a' \in \mathcal{A}} \widehat{Q}(s, a'), \\ 0, & \text{otherwise}, \end{cases}$$

  where $\widehat{Q}(\cdot,\cdot)$ denotes some consistent estimator for $Q^{opt}(\cdot,\cdot)$.
- In **nonregular** cases where $\arg\max_a Q^{opt}(s, a)$ is not unique for some s, $\widehat{\pi}$ will not converge to a fixed quantity.
- The variance of the value estimator is difficult to estimate.
- **Our proposal**: SequetiAl Value Evaluation (SAVE)

SAVE

- Our procedure:

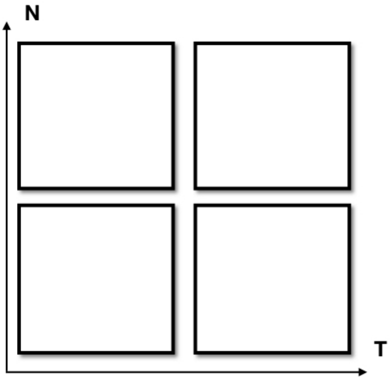
Step 1 Divide the data into $K_N \times K_T$ blocks.
Step 2 Initialize $k = 1$. While $k < K_N K_T$:

1. Use the first k-th blocks of data to estimate the optimal policy and use the $k+1$-th block of data to evaluate its value;
2. Set $k \to k + 1$.

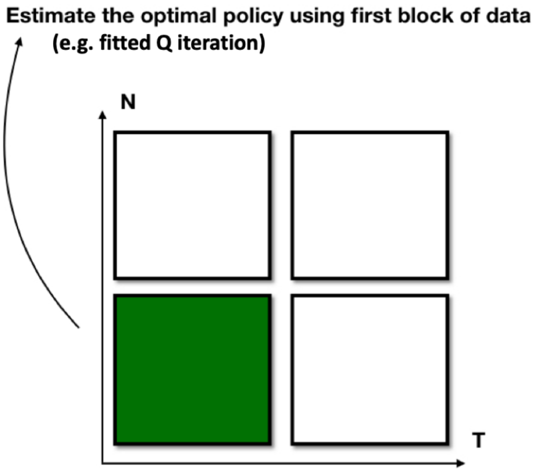
Step 3 Derive the final estimator as a weighted average of all $K - 1$ value estimators.

- Orders of these blocks cannot be arbitrarily determined since observations are time dependent

An illustration of SAVE



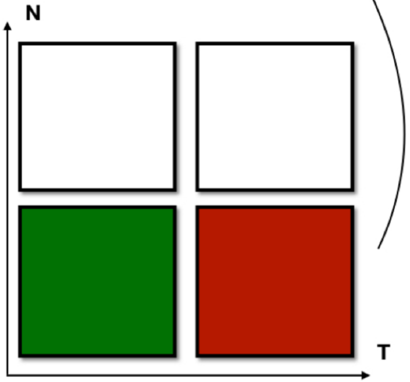
An illustration of SAVE



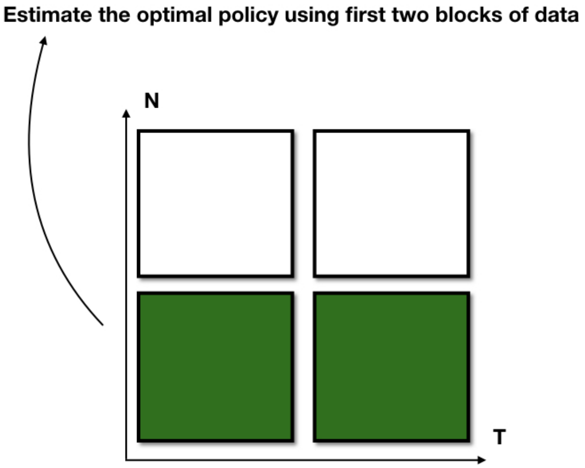
An illustration of SAVE

$$\widehat{V}(\pi;x) = \sum_{a\in\mathcal{A}} \Phi_L^T(x)\widehat{\beta}_{\pi,a}\pi(a|x) = \boldsymbol{U}_\pi^T(x)\widehat{\boldsymbol{\beta}}_\pi. \qquad \widehat{\sigma}^2(\pi;x) = \boldsymbol{U}_\pi^T(x)\widehat{\boldsymbol{\Sigma}}_\pi^{-1}\widehat{\boldsymbol{\Omega}}_\pi(\widehat{\boldsymbol{\Sigma}}_\pi^T)^{-1}\boldsymbol{U}_\pi(x),$$

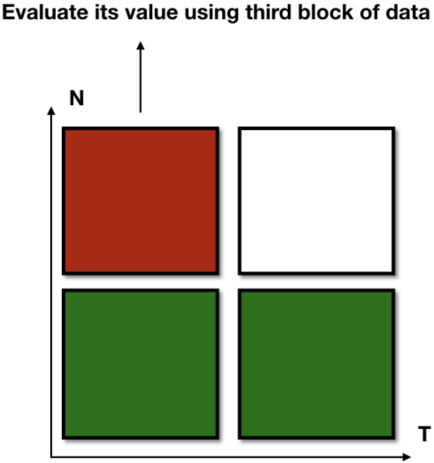
Evaluate its value using second block of data



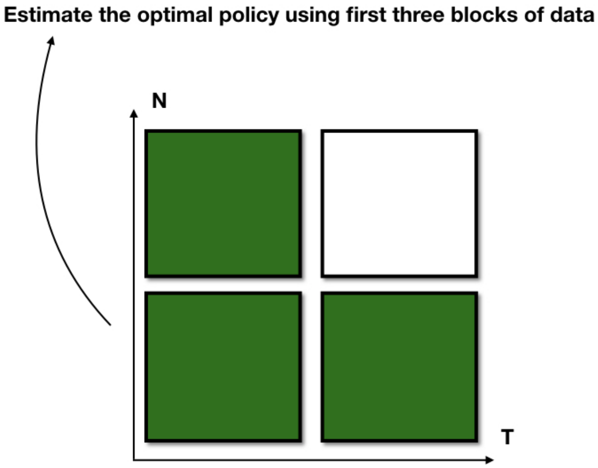
An illustration of SAVE



An illustration of SAVE

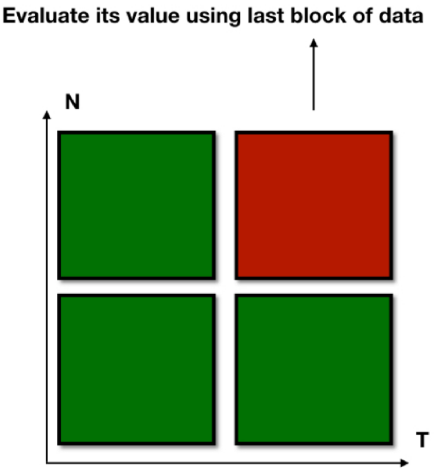


An illustration of SAVE



An illustration of SAVE

$$\widetilde{V}(\mathbb{G}) = \left\{\sum_{k=2}^{K} \frac{1}{\widehat{\sigma}_{\mathcal{I}(k)}(\widehat{\pi}_{\bar{\mathcal{I}}(k-1)};\mathbb{G})}\right\}^{-1}\left\{\sum_{k=2}^{K} \frac{\widehat{V}_{\mathcal{I}(k)}(\widehat{\pi}_{\bar{\mathcal{I}}(k-1)};\mathbb{G})}{\widehat{\sigma}_{\mathcal{I}(k)}(\widehat{\pi}_{\bar{\mathcal{I}}(k-1)};\mathbb{G})}\right\}, \quad \widetilde{\sigma}(\mathbb{G}) = (K-1)\{\textstyle\sum_{k=2}^{K} \widehat{\sigma}_k^{-1}(\widehat{\pi}_{\bar{\mathcal{I}}(k-1)};\mathbb{G})\}^{-1}.$$



Type-II Inference: Theory

Theorem (Informal Statement)

Suppose we Q-learning type estimators to compute $\widehat{\pi}$ and the estimated Q-function converges at certain nonparametric rate. Then under certain other regularity conditions, the proposed CI achieves nominal coverage as either $\boldsymbol{N}$ or $\boldsymbol{T} \to \infty$.

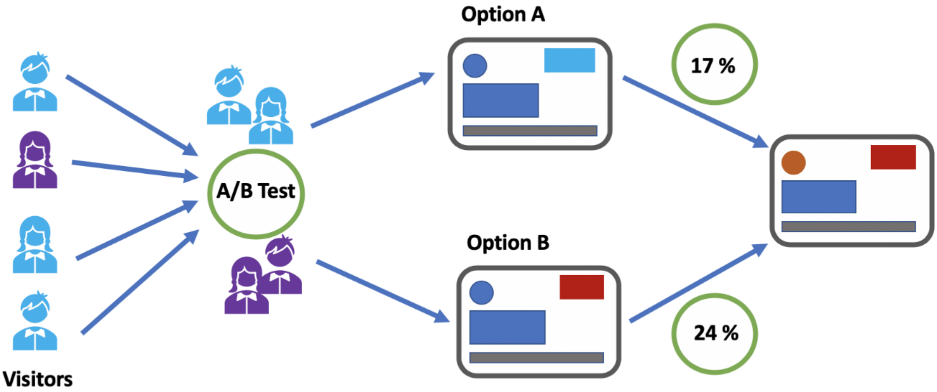
- The value of an estimated optimal policy converges to the optimal value at a **faster** rate than the estimated Q-function under certain margin type conditions
- Similar results have been established in the **classification** literature (Tsybakov, 2004; Audibert and Tsybakov, 2007) and the **DTR** literature (Qian and Murphy, 2011; Luedtke and van der Laan, 2016)
- We extend these results to the RL setting with infinite horizons

Project II

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

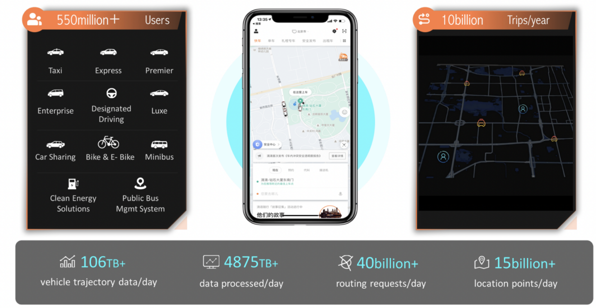
Joint work with Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye and Rui Song
——JASA, accepted

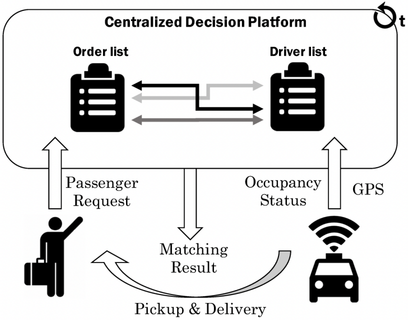
A/B Testing



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

Motivation: Order Dispatch

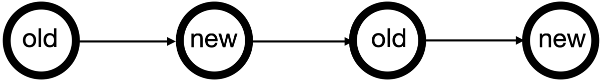




Our project is motivated by the need for comparing the **long-term rewards** of different **order dispatching** policies in **ridesharing platforms**

Challenges

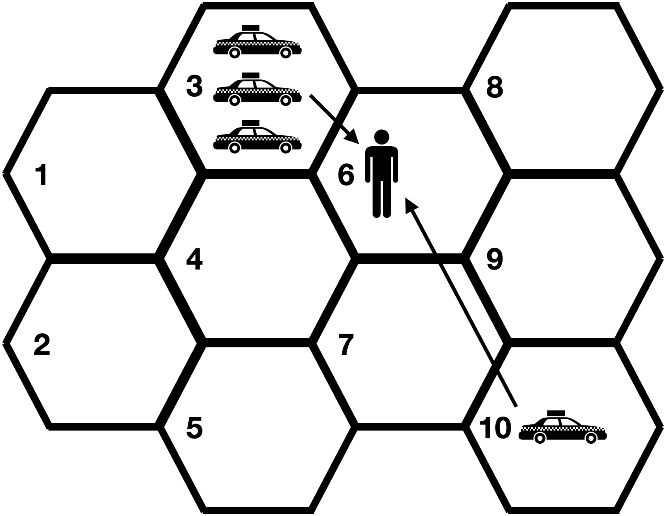
1. The existence of **carryover effects**:
   - Under the alternating-time-interval (or switchback) design



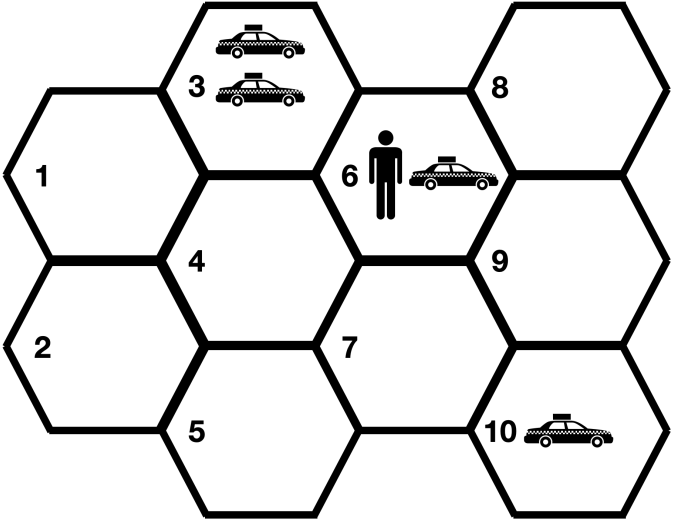
   - Past actions will affect future outcomes
2. The need for **early termination**:
   - Each experiment takes a considerable time (at most 2 weeks)
   - Early termination to save time and budget
3. The need for **adaptive randomization**:
   - Maximize the total reward (e.g., epsilon-greedy)
   - Detect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

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

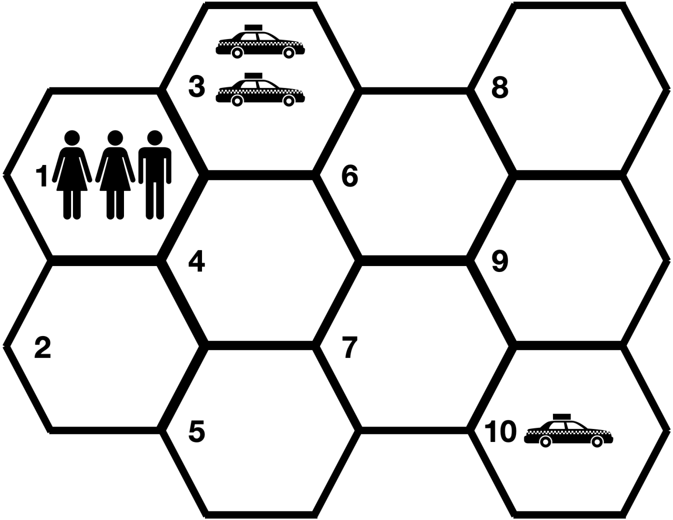
Illustration of the Carryover Effects



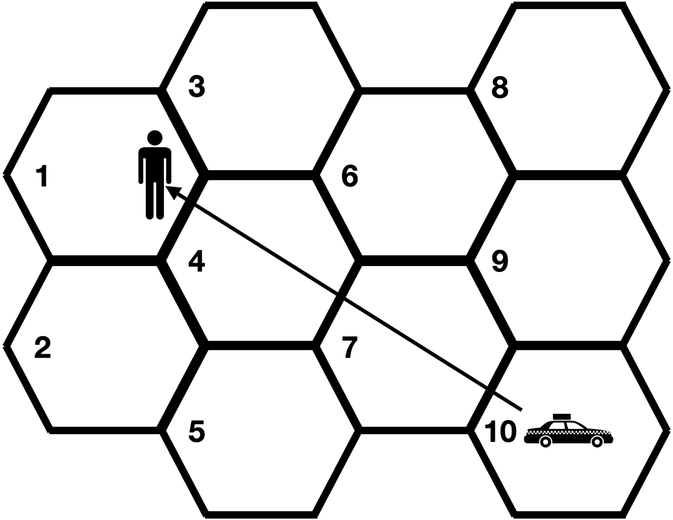
Adopting the Closest Driver Policy



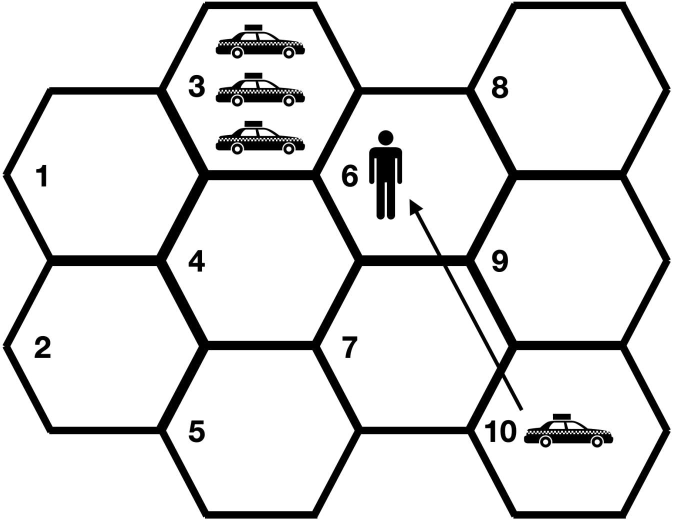
Some Time Later · · ·



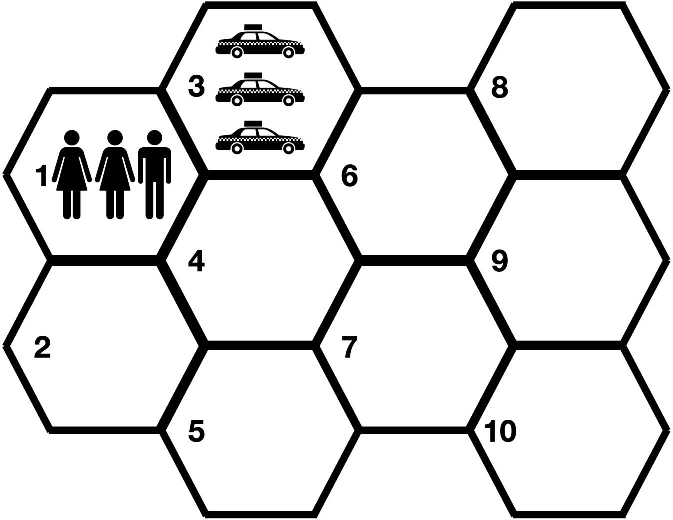
Miss One Order



Consider a Different Action



Able to Match All Orders

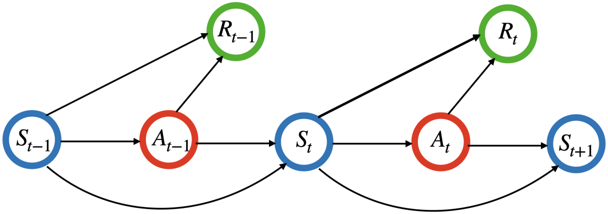


Existence of Carryover Effects

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

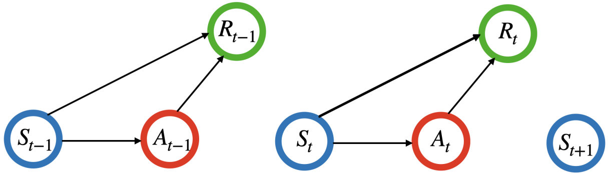
Contributions and Advances of Our Proposal

- Introduce an RL framework for A/B testing



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 A/B tests require the independence assumption



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

Methodology

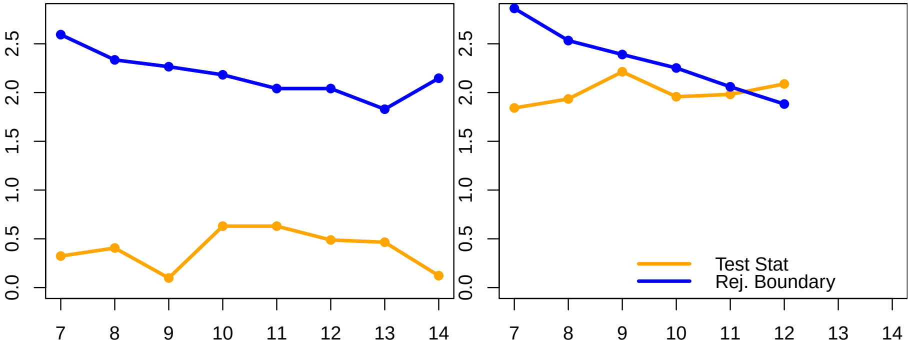
- Apply **temporal difference learning** with **sieve** method to evaluate value difference and provide **uncertainty quantification**
- Adopt the α-**spending approach** (Lan & DeMets, 1983) for sequential monitoring
- Develop a **bootstrap-assisted procedure** for determining the stopping boundary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

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

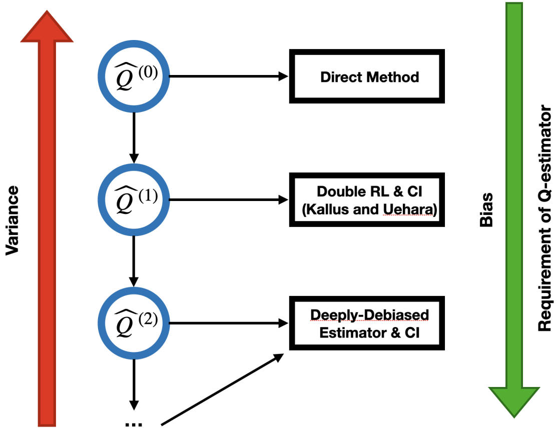
- t-test: **fail** to reject $\mathcal{H}_0$ in A/B experiment with p-value 0.18

Project III

Deeply-Debiased Off-Policy Interval Estimation

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

Deeply-Debiased OPE



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

Thank You!

☺ Papers and softwares can be found on my personal website

`callmespring.github.io`