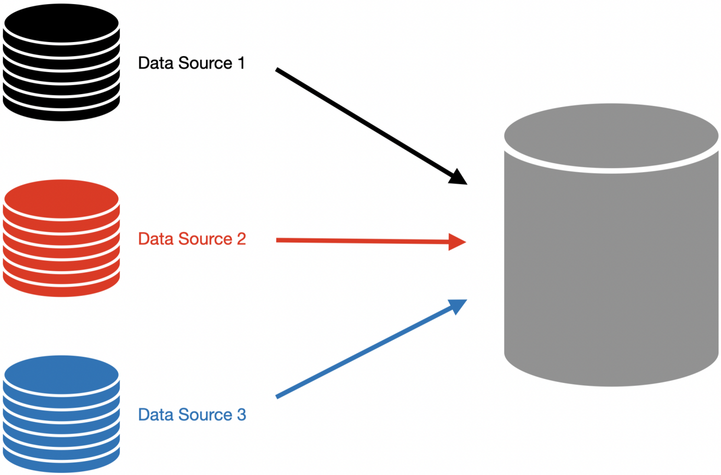
Combining Experimental and Historical Data for Policy Evaluation

Ting Li, **Chengchun Shi**, Qianglin Wen, Yang Sui, Yongli Qin, Chunbo Lai & Hongtu Zhu

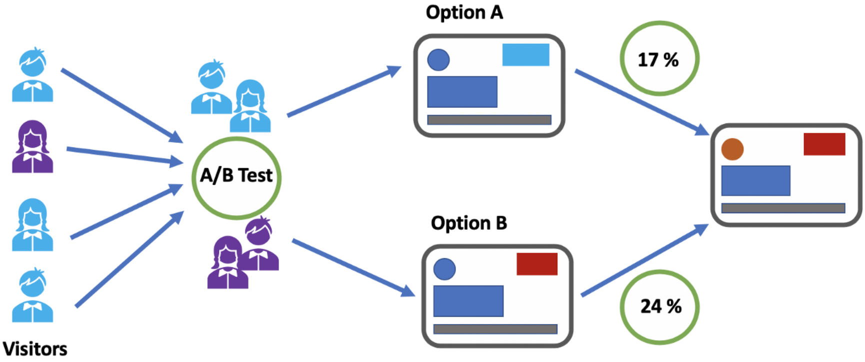
Associate Professor of Data Science

London School of Economics and Political Science

Data Integration



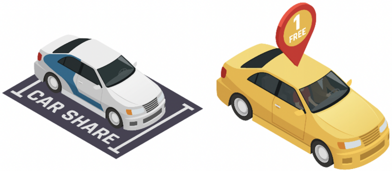
Example I: A/B Testing



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

Example I: A/B Testing with Historical Data

Experiment data



Policy **Control**

- limited duration
- weak treatment effect

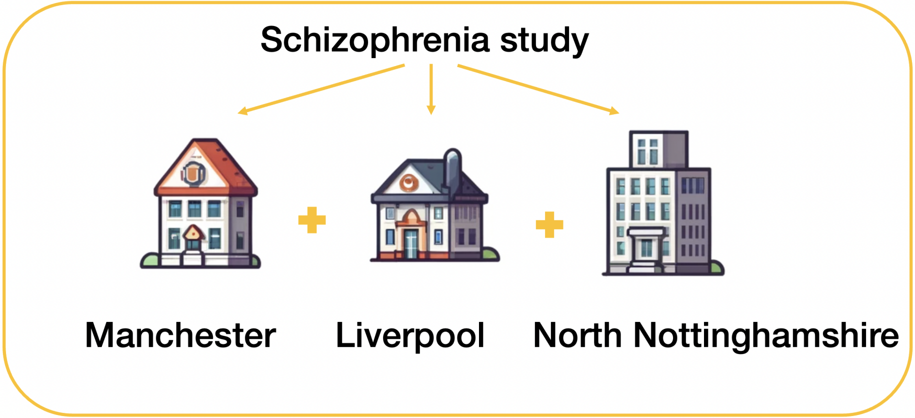
+

Historical data

Control

- substantial volume

Example II: Meta Analysis [Shi et al., 2018]



Example III: Combining Observational Data

RCT

- high cost
- time constraint

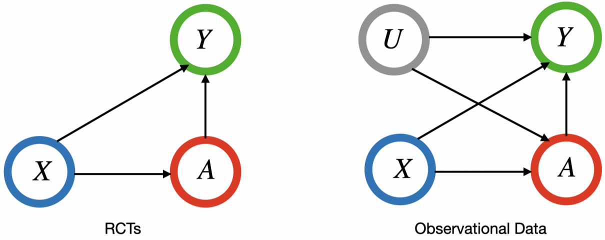
+

Observational data

- large sample size

Challenge: Distributional Shift

- **Example I**: In ridesharing, the **nonstationarity** of the environment $\rightarrow$ distributional shift between experimental and historical datasets [Wan et al., 2021]
- **Example II**: In medicine, the **heterogeneity** in characteristics of treatment setting $\rightarrow$ distributional shift among different data sources [Shi et al., 2018]
- **Example III**: The observational data is subject to **unmeasured confounding** $\rightarrow$ distributional shift between RCT and observational data

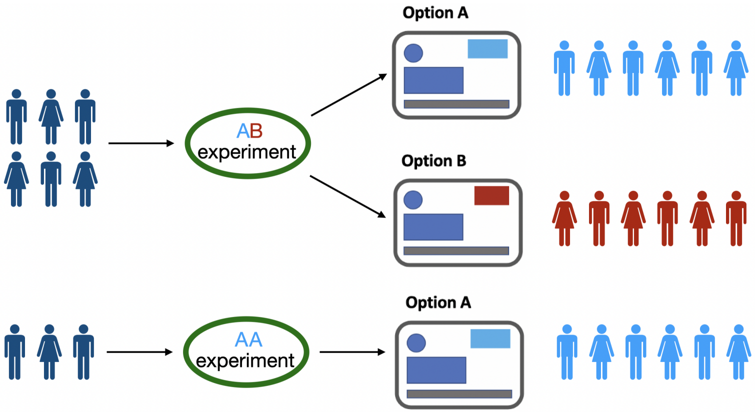


Related Works

- Data integration for causal inference
    - **Example I**: Leverage historical datasets under control [Li et al., 2023]
    - **Example II**: Federated causal inference [Han et al., 2021, 2023]
    - **Example III**: Combining RCT and observational data [Kallus et al., 2018, Yang and Ding, 2020]
- Other related works
    - **Meta** analysis & **meta** learning [DerSimonian and Laird, 1986]
    - **Transfer** & **federated** learning [Li et al., 2022]
    - **Heterogeneous RL** [Shi et al., 2018, Chen et al., 2024]
    - **Off-policy evaluation**

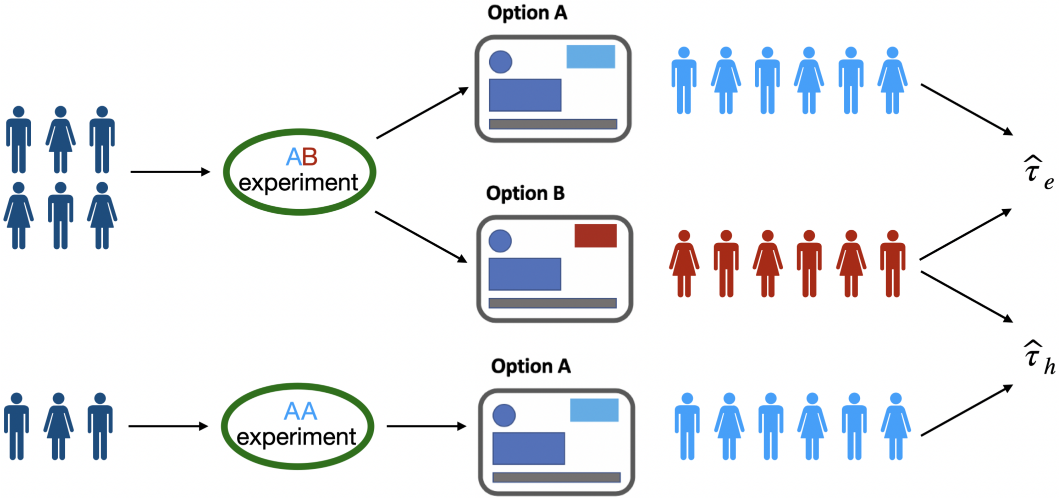
A/B Testing with Historical Data

Objective: combine **experimental data** with **historical data** to improve ATE estimation



Challenge: **distributional shift** between experimental and historical data

Two Base Estimators



A Naive Weighted Estimator

- Consider the weighted estimator

$$\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}} = \boldsymbol{w}\widehat{\boldsymbol{\tau}}_{\boldsymbol{e}} + (\mathbf{1} - \boldsymbol{w})\widehat{\boldsymbol{\tau}}_{\boldsymbol{h}},$$

  for some properly chosen weight $\boldsymbol{w} \in [0, 1]$ to minimize its $\mathrm{MSE}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}})$.
- The weight $\boldsymbol{w}$ reflects a bias-variance tradeoff. A large $\boldsymbol{w}$ can:
  - Reduce **bias** of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}}$ caused by the distributional shift between the datasets
  - Increase **variance** of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}}$ as a result of not fully leveraging the historical data
- Natural to consider the following naive estimator that minimizes an estimated MSE:

$$\widehat{\mathrm{MSE}}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}}) = \widehat{\mathrm{Bias}}^2(\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}}) + \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}}).$$

  We refer to this estimator as the **non-pessimistic** estimator.

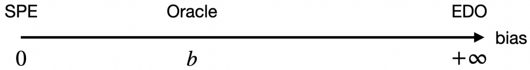
Theoretical Analysis

Three scenarios, depending on the bias $\boldsymbol{b} = \mathbb{E}(\widehat{\boldsymbol{b}}) = \mathbb{E}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{h}} - \widehat{\boldsymbol{\tau}}_{\boldsymbol{e}})$

1. **Small bias**: $\boldsymbol{b}$ is much smaller than the standard deviation of its estimator;
2. **Moderate bias**: $\boldsymbol{b}$ is comparable to or larger than the standard deviation, yet falls within the high confidence bounds of $\widehat{\boldsymbol{b}}$;
3. **Large bias**: $\boldsymbol{b}$ is much larger than the estimation error.

Three competing estimators:

1. **EDO** (experimental-data-only) estimator which sets $\boldsymbol{w} = \mathbf{1}$;
2. **SPE** (semi-parametrically efficient) estimator [Li et al., 2023] developed under the assumption of no bias;
3. **Oracle** estimator which optimizes $\boldsymbol{w}$ to minimize $\mathrm{MSE}(\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}})$;



Theoretical Analysis (Cont'd)

Bias	Non-pessimistic estimator	Optimal estimator
Zero	Close to **efficiency** bound	SPE/Oracle
Small	Close to **oracle** MSE	SPE/Oracle
Moderate	**May suffer a large MSE**	**Oracle**
Large	**Oracle** property	EDO/Oracle

- The **oracle** MSE denotes MSE of the oracle estimator
- The **efficiency bound** is the smallest achievable MSE among a broad class of regular estimators [Tsiatis, 2006].

Our Motivating Question

Can we develop an estimator that works well with moderate bias?

Our Proposal

Main idea: reformulate the weight selection as an **offline bandit** problem

- Each weight $\boldsymbol{w} \in [0, 1] \rightarrow$ an **arm** in bandit
- Negative MSE of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}} \rightarrow$ **reward** of selecting an arm

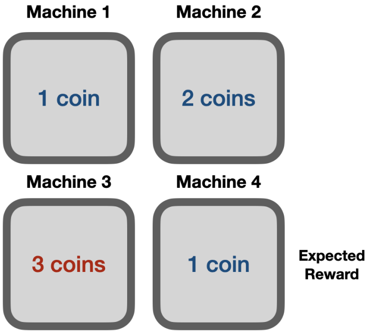
Objective in bandit: choose the **optimal** arm that maximizes its reward.

Multi-Armed Bandit

- 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**

Multi-Armed Bandit (Con't)

- 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**: maximize $\sum_{t=1}^{T} \mathbb{E} R_t$.



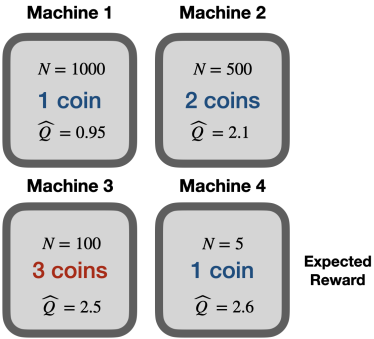
Greedy Action Selection

- **Action-value methods**:

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

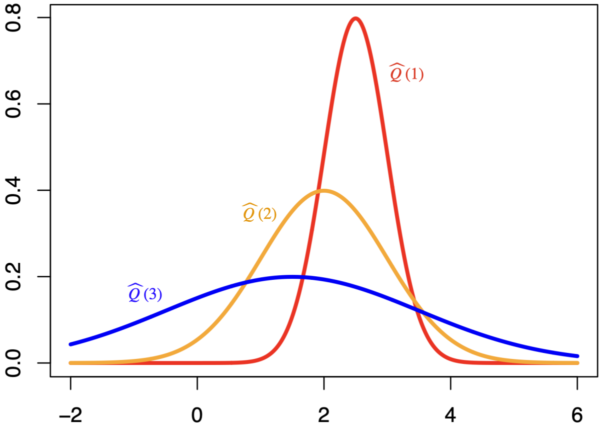
where $N(a) = \sum_{t=0}^{T-1} \mathbb{I}(A_t = a)$ denotes the action counter

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



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)



Upper Confidence Bound

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

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

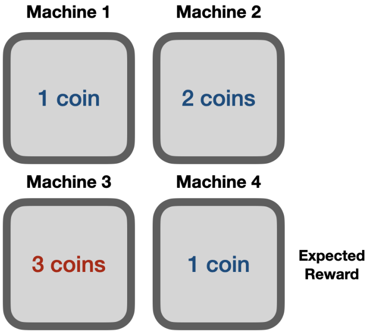
  with high probability.
- $U_t(a)$ quantifies the **uncertainty** and depends on $N_t(a)$ (number of times arm a has been selected up to time t)
  - Large $N_t(a) \rightarrow$ small $U_t(a)$;
  - Small $N_t(a) \rightarrow$ large $U_t(a)$.
- Select actions maximizing upper confidence bound

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

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

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



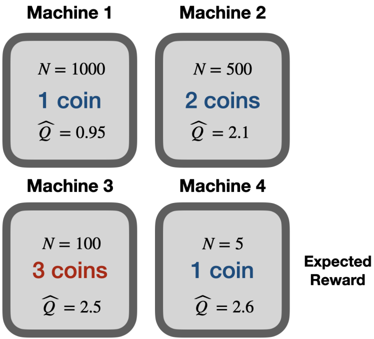
Greedy Action Selection (Non-pessimistic Estimator)

- **Action-value methods**:

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

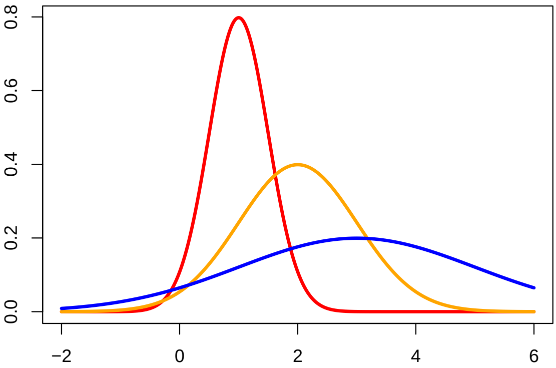
  where $N(a) = \sum_{t=0}^{T-1} \mathbb{I}(A_t = a)$ denotes the action counter

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



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)

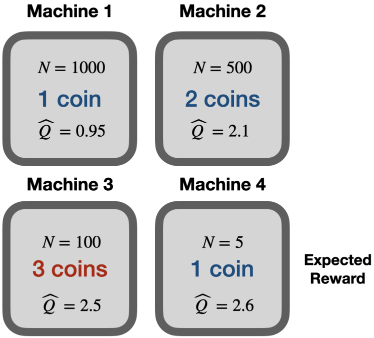
- Set $L(a) = \sqrt{c \log(T)/N(a)}$ for some positive constant c where T is the sample size of historical data
- According to **Hoeffding's inequality** (link), when rewards are bounded between 0 and 1, the event

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

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

Lower Confidence Bound (Cont'd)

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



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

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

Back to Our Problem

Main idea: reformulate the weight selection as an **offline bandit** problem

- Each weight $\boldsymbol{w} \in [0, 1] \to$ an **arm** in bandit
- Negative MSE of $\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}} \to$ **reward** of selecting an arm

Nonpessimistic estimator chooses the arm that maximizes an estimated negative MSE

- It requires a **uniform consistency** condition: the estimated MSE converges to its oracle value uniformly across all weights
- Underestimate the bias $\boldsymbol{b} \to$ low estimated MSE for small weights $\to$ estimated weight tends to be smaller than the ideal value $\to$ a significant bias in $\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}}$
- This reveals the limitation of the nonpessimistic estimator when $\boldsymbol{b}$ is moderate or large.

Pessimistic Estimator

Main idea: select the arm that maximizes a lower bound of the negative MSE, or equivalently, an upper bound of the MSE

- **Uncertainty quantification**: compute an uncertainty quantifier $\boldsymbol{U}$ for the estimated error such that $|\widehat{\boldsymbol{b}} - \boldsymbol{b}| \leq \boldsymbol{U}$ with large probability.
- **MSE estimation**: use $|\widehat{\boldsymbol{b}}| + \boldsymbol{U}$ as a pessimistic estimator for the bias $\boldsymbol{b}$ and plug this estimator into the MSE formula to construct an upper bound of the MSE $\widehat{\mathrm{MSE}}_U(\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}})$.
- **Weight selection**: select $\boldsymbol{w}$ that minimizes the upper bound $\widehat{\mathrm{MSE}}_U(\widehat{\boldsymbol{\tau}}_{\boldsymbol{w}})$.

Theoretical Analysis

Bias	Non-pessimistic estimator	Pessimistic estimator	Optimal estimator
Zero	Close to **efficiency bound**	Same order to **oracle** MSE	SPE/Oracle
Small	Close to **oracle** MSE	Same order to **oracle** MSE	SPE/Oracle
Moderate	**May suffer a large MSE**	**Oracle property**	**Oracle**
Large	**Oracle** property	**Oracle** property	EDO/Oracle

- The **oracle** MSE denotes MSE of the oracle estimator.
- The **efficiency bound** is the smallest achievable MSE among a broad class of regular estimators [Tsiatis, 2006].

Simulation Study

The effectiveness of different estimators is determined by the magnitude of the bias. To validate our theory, we further classify $\boldsymbol{b}$ into different regimes as follows

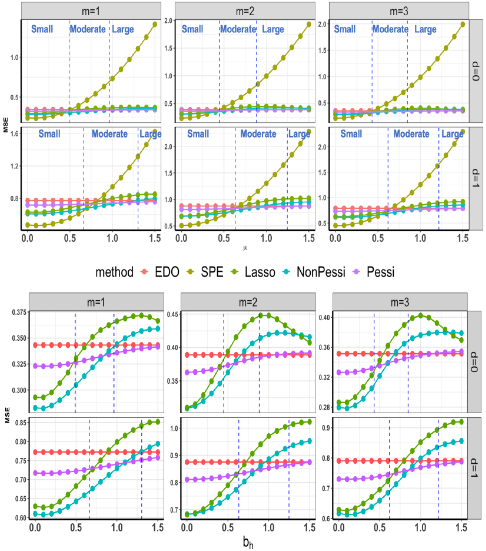
- **Small bias** regime (SPE estimator is expected to be optimal): $|\boldsymbol{b}| \leq c_1\sqrt{\mathrm{Var}(\widehat{\boldsymbol{b}})}$;
- **Moderatel bias** regime (the proposed pessimistic estimator is expected to be optimal): $c_1 < \frac{|\boldsymbol{b}|}{\sqrt{\mathrm{Var}(\widehat{\boldsymbol{b}})}} \leq c_2$;
- **Large bias** regime (EDO estimator is expected to be optimal): $|\boldsymbol{b}| > c_2\sqrt{\mathrm{Var}(\widehat{\boldsymbol{b}})}$.

According to our theory, we set $c_1 = 1$ and $c_2 = \sqrt{\log(n)}$. This ensures:

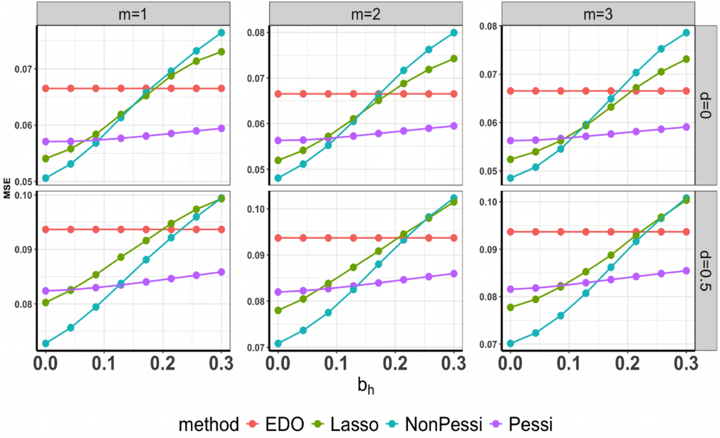
- Scenarios where variance dominates the bias are categorized within the small bias region.
- When the bias exceeds the established high confidence bound, it is classified under the large bias regime.

Simulation Study: Bandit Simulation

- **NonPessi**: the proposed non-pessimistic estimator.
- **Pessi**: the proposed pessimistic estimator.
- **EDO**: the doubly robust estimator $\widehat{\tau}_e$ constructed based on the experimental data only (see (1)).
- **Lasso**: a weighted estimator $\widehat{\tau}_{Lasso} = w\widehat{\tau}_e + (1-w)\widehat{\tau}_h$ that linearly combines the ATE estimator $\widehat{\tau}_e$ based on experimental data and $\widehat{\tau}_h$ based on historical data, where the weight w is chosen to minimize the estimated variance of the final ATE estimator with the Lasso penalty (Cheng & Cai, 2021),
- **SPE**: the semi-parametrically efficient estimator proposed by Li et al. (2023) developed under the assumption of no reward shift between the experimental and historical data, i.e., $r_e(0, s) = r_h(s)$ for any s.

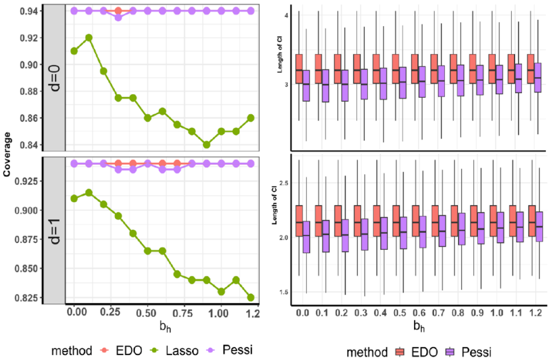


Ridesharing Data-based Sequential Simulation



Pessimistic estimator shows robustness in dealing with distributional shift

Simulation Study: Confidence Intervals



- While maintaining nominal coverage, the **pessimistic estimator** yields narrower confidence intervals compared to the EDO estimator
- Improvement in efficiency by incorporating historical data.

Summary

- Policy evaluation using both **experimental** and **historical** datasets, allowing distributional shifts between the two datasets.
- Two weighted estimators that leverage both data sources.
- The proposed **non-pessimistic estimator** chooses the weight by minimizing an estimated MSE.
- The proposed **pessimistic estimator** further employs the pessimistic principle to boost its robustness.
- Our theoretical and empirical analyses identify the most effective estimator within each regime.

References I

Elynn Y Chen, Rui Song, and Michael I Jordan. Reinforcement learning in latent heterogeneous environments. *Journal of the American Statistical Association*, just-accepted, 2024.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *arXiv preprint arXiv:2112.09313*, 2021.

Larry Han, Zhu Shen, and Jose Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36:70453–70482, 2023.

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.

References II

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 31, 2018.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Xinyu Li, Wang Miao, Fang Lu, and Xiao-Hua Zhou. Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403, 2023.

Chengchun Shi, Rui Song, Wenbin Lu, and Bo Fu. Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):681–702, 2018.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

References III

Runzhe Wan, Sheng Zhang, Chengchun Shi, Shikai Luo, and Rui Song. Pattern transfer learning for reinforcement learning in order dispatching. *arXiv preprint arXiv:2105.13218*, 2021.

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2020.