Safe Exploration for Efficient Policy Evaluation and Comparison

Runzhe Wan[1], Branislav Kveton[2], Rui Song[1]

[1]North Carolina State University

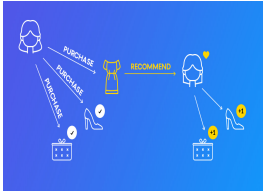
[2]Amazon

ICML 2022

Policy Evaluation and Comparison

(a) Economics

(b) Health Care



(c) E-commerce Platforms

(d) Ridesharing

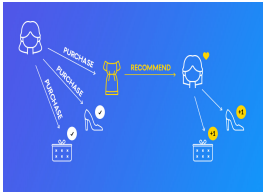
- Bandit policy/Optimal decision rule:

$$A \sim \pi(a|\mathbf{x})$$

Policy Evaluation and Comparison

(a) Economics

(b) Health Care



(c) E-commerce Platforms

(d) Ridesharing

- Bandit policy/Optimal decision rule:

$$A \sim \pi(a|\mathbf{x})$$

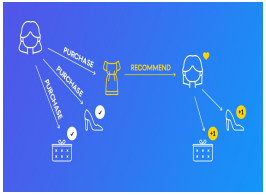
- **Policy Evaluation**:

$$V(\pi) = \mathbb{E}_{\mathbf{x},a\sim\pi(a|\mathbf{x})} r(a,x)$$

Policy Evaluation and Comparison

(a) Economics

(b) Health Care



(c) E-commerce Platforms

(d) Ridesharing

- Bandit policy/Optimal decision rule:

$$A \sim \pi(a|\mathbf{x})$$

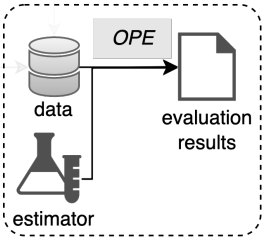
- **Policy Evaluation**:

$$V(\pi) = \mathbb{E}_{\mathbf{x},a\sim\pi(a|\mathbf{x})} r(a,x)$$

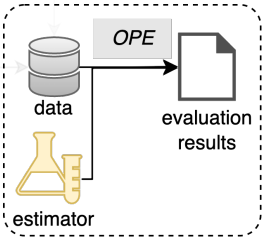
- **Policy Comparison**:

$$V(\pi) > V(\pi_0)?$$

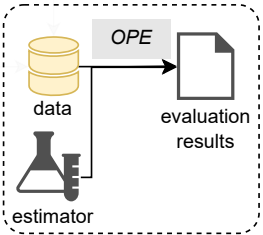
Classic OPE v.s. SEPEC



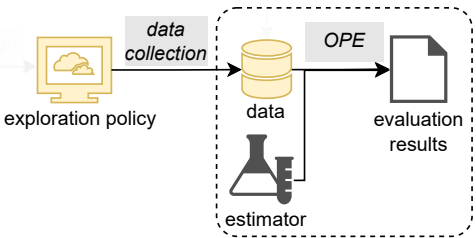
Classic OPE v.s. SEPEC



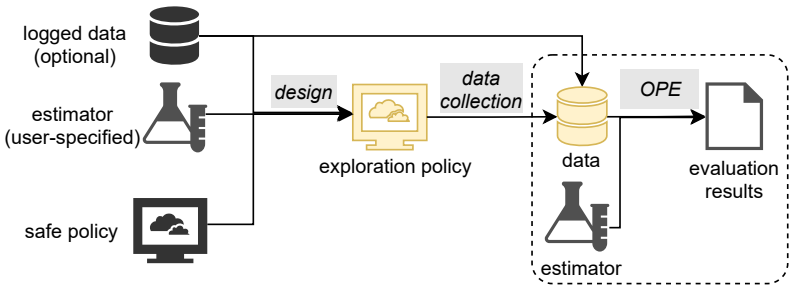
Classic OPE v.s. SEPEC



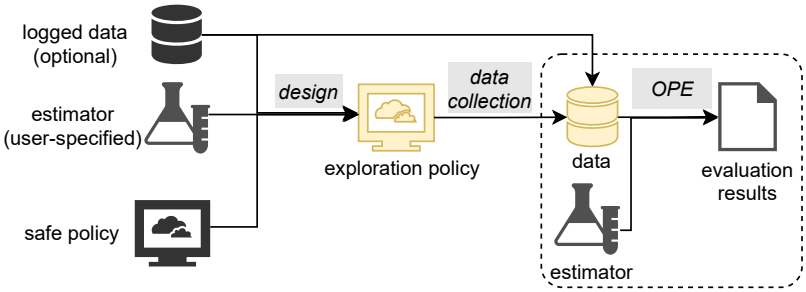
Classic OPE v.s. SEPEC



Classic OPE v.s. SEPEC

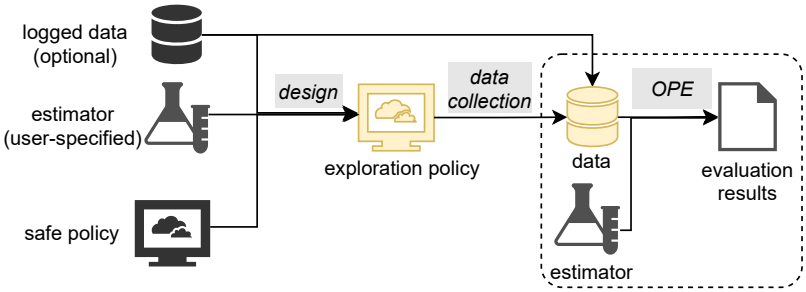


Classic OPE v.s. SEPEC



SEPEC: Safe Exploration for efficient Policy Evaluation and Comparison.

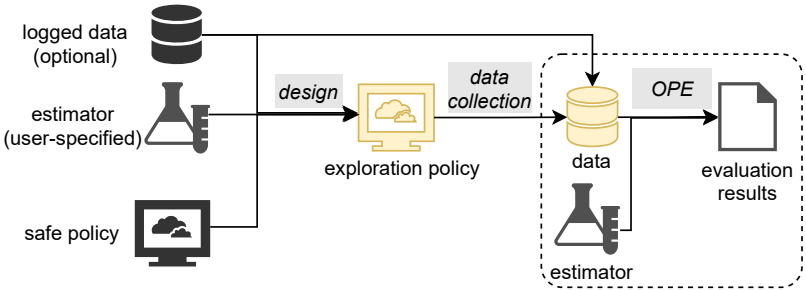
Classic OPE v.s. SEPEC



SEPEC: Safe Exploration for efficient Policy Evaluation and Comparison.

- Efficient: Minimize $\mathrm{var}(\hat{V}(\pi))$ from following the exploration policy to collect data

Classic OPE v.s. SEPEC



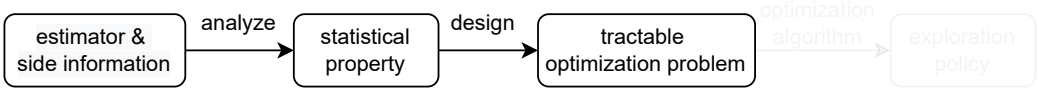
SEPEC: Safe Exploration for efficient Policy Evaluation and Comparison.

- Efficient: Minimize $\mathrm{var}(\hat{V}(\pi))$ from following the exploration policy to collect data
- Safe: $V(\text{exploration policy}) \geq (1-\epsilon) \times V(\text{safe policy})$

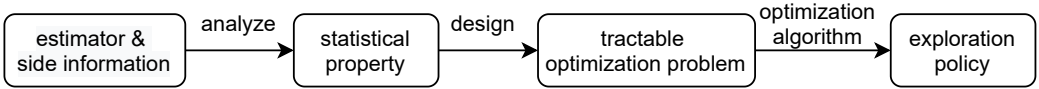
Contributions

Contributions

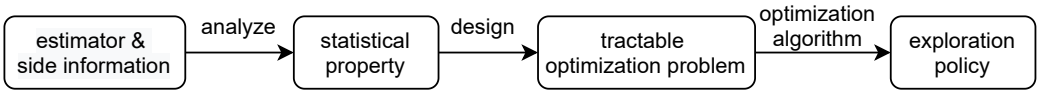
Contributions



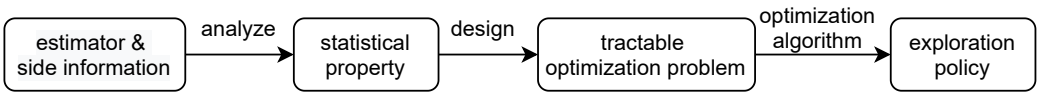
Contributions



Contributions

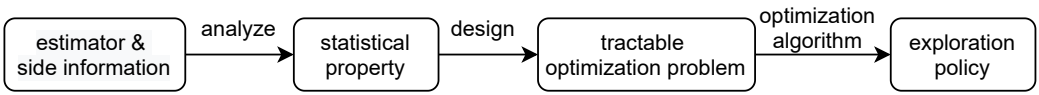


Contributions



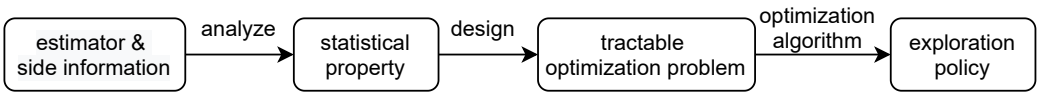
- Study **three representative variants**: MAB with inverse probability weighting (IPW), contextual MAB with IPW, and linear bandit with direct methods

Contributions



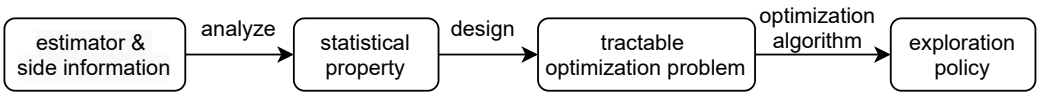
- Study **three representative variants**: MAB with inverse probability weighting (IPW), contextual MAB with IPW, and linear bandit with direct methods
- Investigate **differences** due to bandit setups, evaluation tasks, value estimators, and side information availability

Contributions



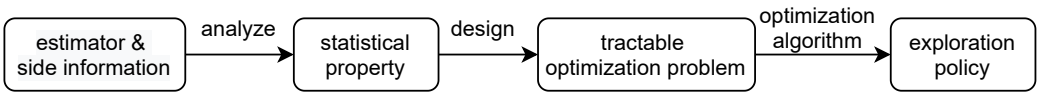
- Study **three representative variants**: MAB with inverse probability weighting (IPW), contextual MAB with IPW, and linear bandit with direct methods
- Investigate **differences** due to bandit setups, evaluation tasks, value estimators, and side information availability
- Present **extensions** including doubly robust (DR) estimators, pseudo-inverse estimator, contextual linear bandits, etc.

Contributions



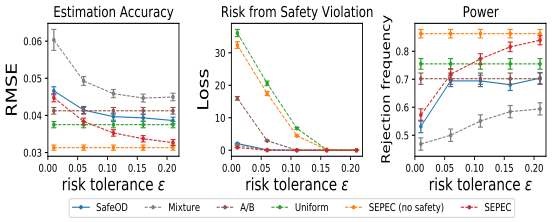
- Study **three representative variants**: MAB with inverse probability weighting (IPW), contextual MAB with IPW, and linear bandit with direct methods
- Investigate **differences** due to bandit setups, evaluation tasks, value estimators, and side information availability
- Present **extensions** including doubly robust (DR) estimators, pseudo-inverse estimator, contextual linear bandits, etc.
- Formulate as **constrained convex optimization** problems and solve with cutting-plane method / Frank–Wolfe algorithm

Contributions

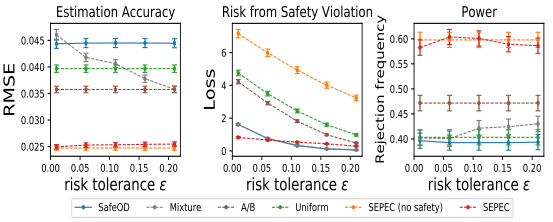


- Study **three representative variants**: MAB with inverse probability weighting (IPW), contextual MAB with IPW, and linear bandit with direct methods
- Investigate **differences** due to bandit setups, evaluation tasks, value estimators, and side information availability
- Present **extensions** including doubly robust (DR) estimators, pseudo-inverse estimator, contextual linear bandits, etc.
- Formulate as **constrained convex optimization** problems and solve with cutting-plane method / Frank–Wolfe algorithm
- Prove both **optimality** and **safety**

Experiments

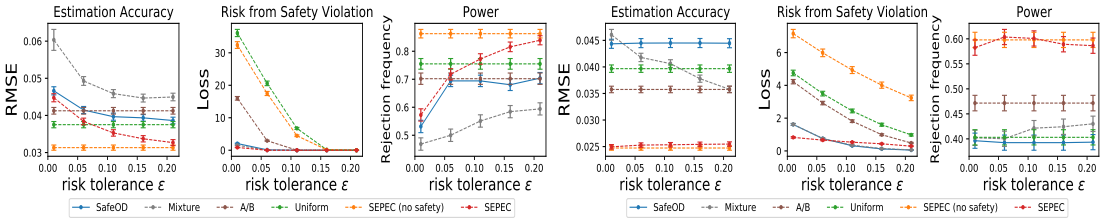


(a) CMAB with IPW.



(b) Linear bandits with DM.

Experiments

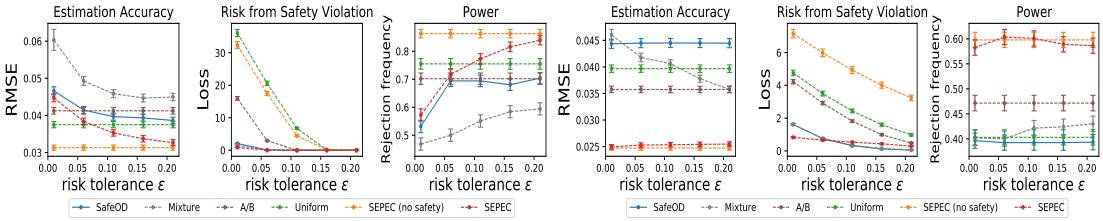


(a) CMAB with IPW.

(b) Linear bandits with DM.

SEPEC: Safe Exploration for efficient Policy Evaluation and Comparison.

Experiments



(a) CMAB with IPW.

(b) Linear bandits with DM.

SEPEC: Safe Exploration for efficient Policy Evaluation and Comparison.

Poster: #**638**