OFF POLICY EVALUATION AND LEARNING FOR INTERACTIVE SYSTEMS

Yi Su
Cornell University
July 15th, 2021

News Feed

Search Engine

Food Recommendation

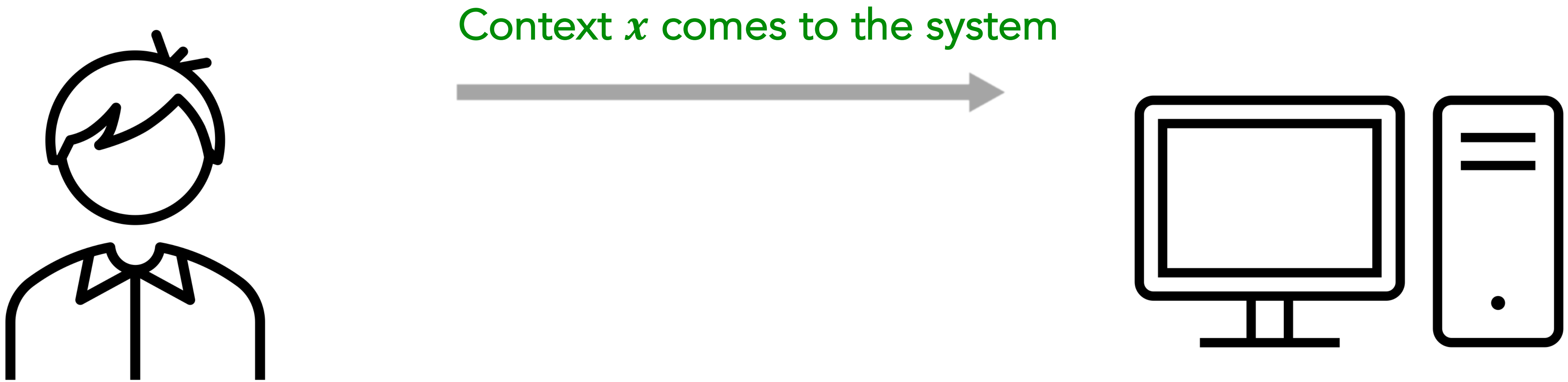
Entertainment

Autonomous Driving

Online Shopping

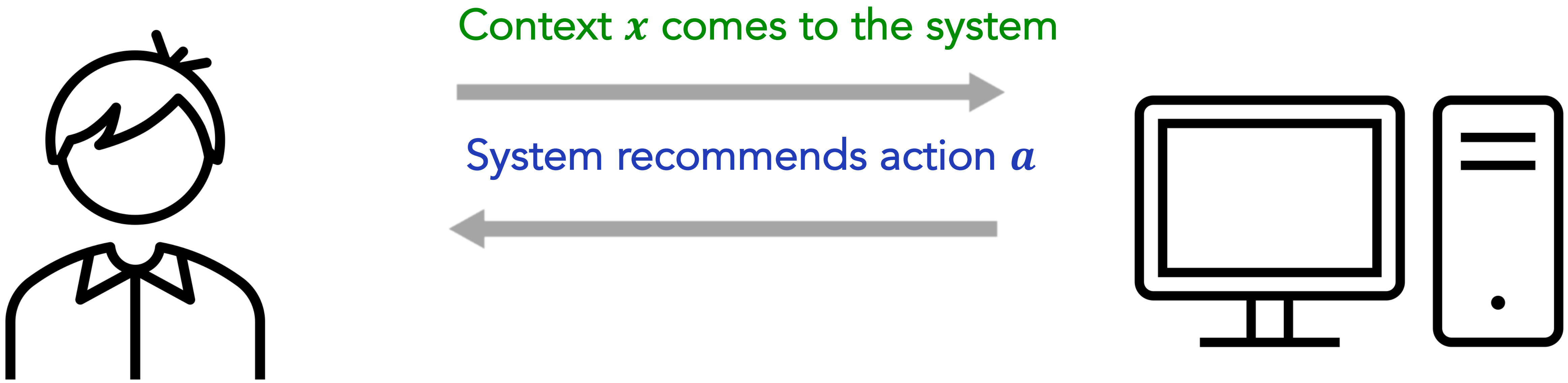
Interactive systems are everywhere

INTERACTIVE SYSTEMS SCHEMATIC



x: user information, query information, etc.

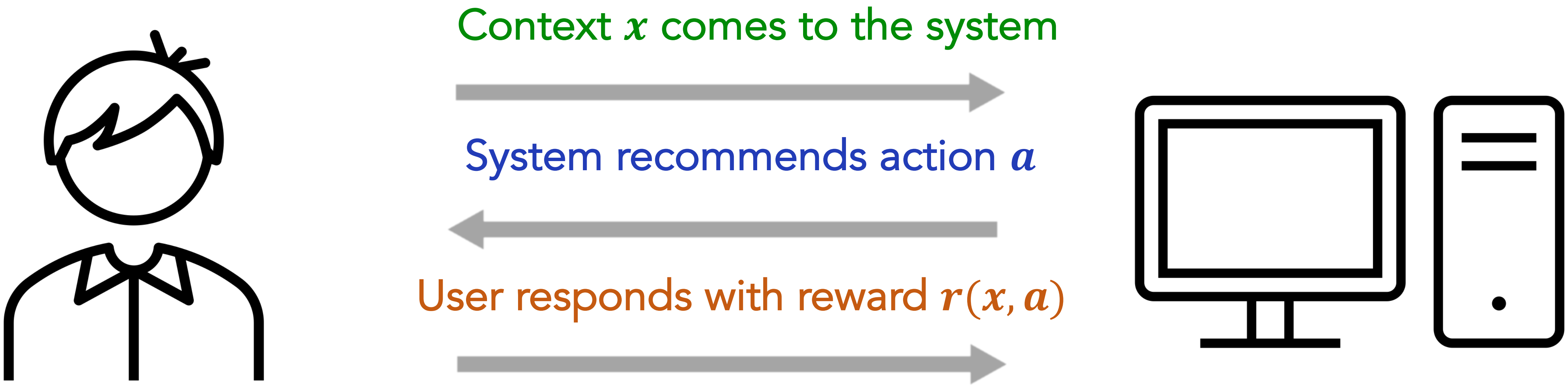
INTERACTIVE SYSTEMS SCHEMATIC



x: user information, query information, etc.

a: ranking, recommended music/news, etc.

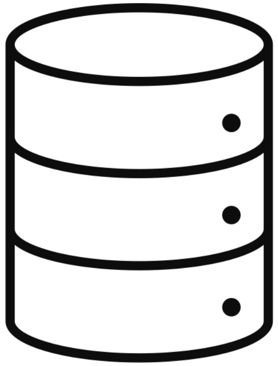
INTERACTIVE SYSTEMS SCHEMATIC



x: user information, query information, etc.
a: ranking, recommended music/news, etc.
r: click, dwell time, transactions, etc.

INTERACTIVE SYSTEMS SCHEMATIC

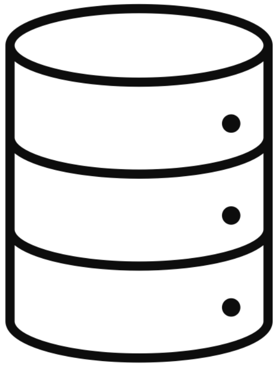
x: user information, query information, etc.
a: ranking, recommended music/news, etc.
r: click, dwell time, transactions, etc.



$$\mathcal{D} = \{x_i, a_i, r_i\}_{i=1}^{n}$$

Logged Dataset

INTERACTIVE SYSTEMS SCHEMATIC



We collect user interactions for:

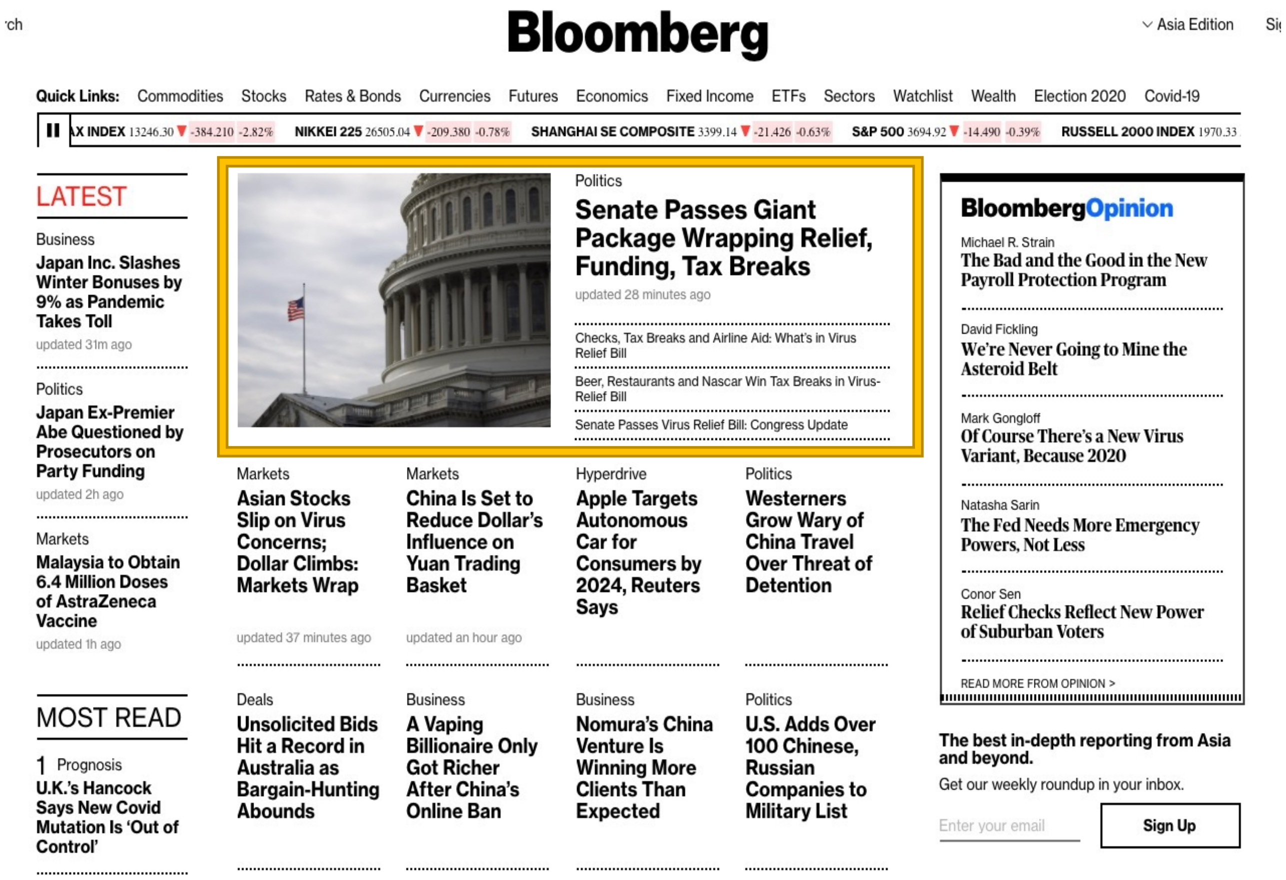
- Evaluating the system performance
- Learning an improved system

EXAMPLE: NEWS RECOMMENDER

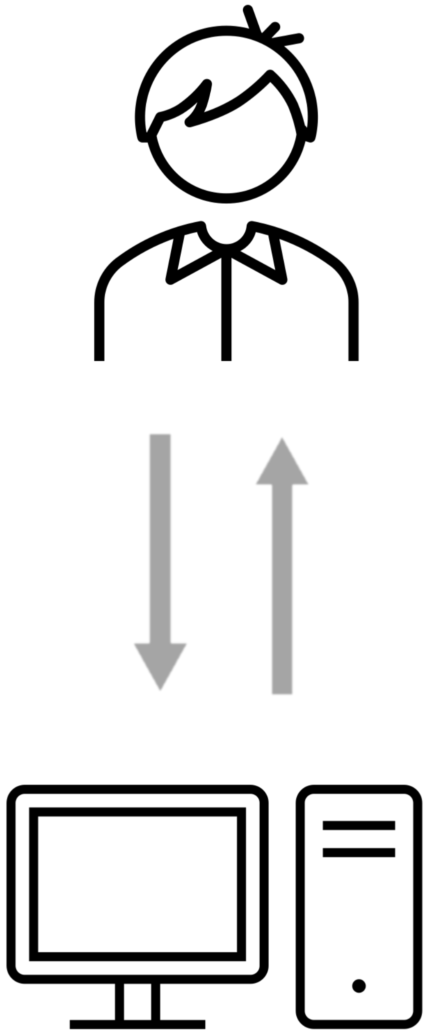
Context x:
- User information/ Visiting history

Action a:
- News article featured in the main panel.

Reward $r(x,a)$:
- Reading time



CONTEXTUAL BANDIT PROTOCOL



Repeated Interaction:

Context $\boldsymbol{x}$ i.i.d follows some distribution $P(x)$.
(user information, visiting history etc.)

System chooses **action $\boldsymbol{a}$** according to some policy $\pi(a|x)$.
(recommended music/news, ranking, etc.)

The user provides **feedback $\boldsymbol{r(x,a)}$** to the presented action.
(click, dwell time, likes/shares, etc.)

Given a new system, how is the performance of it?

Policy Evaluation

How do we improve and learn new systems?

Policy Learning

POLICY EVALUATION

➤ Definition [Utility of Policy]:

The expected reward/utility of a policy π is:

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

ONLINE EVALUATION: A/B TESTING

- Evaluation of Policy $\boldsymbol{\pi}$:
  - Deploy system π online.
  - For user $x \sim P(x)$, draws action $a \sim \pi(\cdot\,|x)$, receives feedback $r(x, a)$.
  - Collect dataset in the format $\mathcal{D} = \{x_i, a_i, r_i\}_{i=1}^{n}$.
  - Construct estimate of the policy utility:

$$\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} r_i$$

ONLINE EVALUATION: A/B TESTING

Draw $\mathcal{D}_1$ from π_1 $\longrightarrow$ Evaluate $\hat{V}(\pi_1)$ — $\hat{V}(\pi_1)$

Draw $\mathcal{D}_2$ from π_2 $\longrightarrow$ Evaluate $\hat{V}(\pi_2)$ — $\hat{V}(\pi_2)$

Draw $\mathcal{D}_3$ from π_3 $\longrightarrow$ Evaluate $\hat{V}(\pi_3)$ — $\hat{V}(\pi_3)$

$\cdots$

Draw $\mathcal{D}_{|\mathcal{H}|}$ from $\pi_{|\mathcal{H}|}$ $\longrightarrow$ Evaluate $\hat{V}(\pi_{|\mathcal{H}|})$ — $\hat{V}(\pi_{|\mathcal{H}|})$

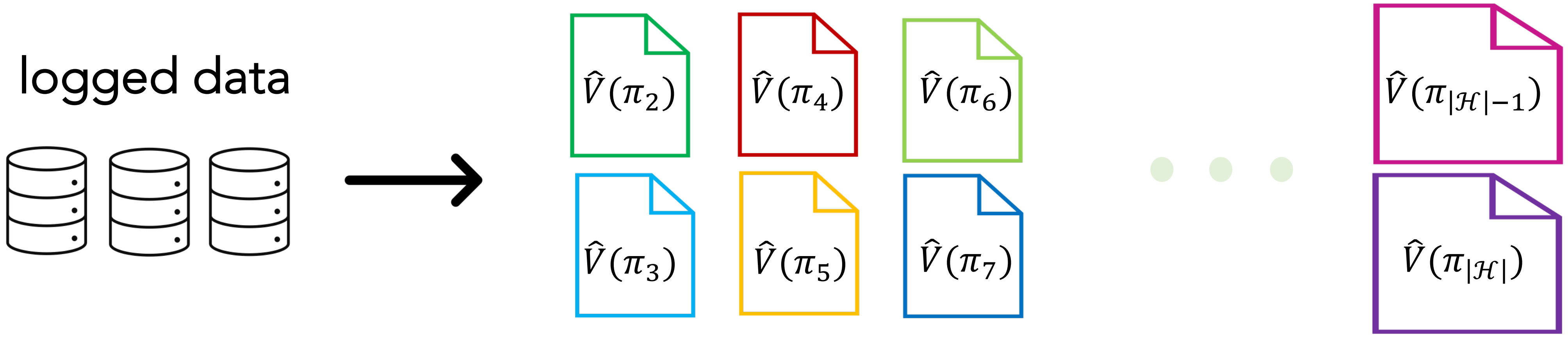
MOVE ONLINE EVALUATION TO OFFLINE

- Problems with online A/B Testing:
  - Long turnaround time.
  - High engineering cost.
  - Limited number of policies being evaluated.
  - High risk of deploying bad policy.

MOVE ONLINE EVALUATION TO OFFLINE

- Problems with online A/B Testing:
  - Long turnaround time.
  - High engineering cost.
  - Limited number of policies being evaluated.
  - High risk of deploying bad policy.

- Idea: Move online to offline:



GOALS

Provide statistically and computationally efficient way to evaluate and optimize interactive systems by exploiting logs of past user interactions.

GOALS

Provide statistically and computationally efficient way to evaluate and optimize interactive systems by exploiting logs of past user interactions. Specifically:

1. Off-policy Evaluation

GOALS

Provide statistically and computationally efficient way to evaluate and optimize interactive systems by exploiting logs of past user interactions. Specifically:

1. Off-policy Evaluation
2. Off-policy Model Selection

GOALS

Provide statistically and computationally efficient way to evaluate and optimize interactive systems by exploiting logs of past user interactions. Specifically:

1. Off-policy Evaluation
2. Off-policy Model Selection
3. Off-policy Learning

TALK OUTLINE

Off-policy Evaluation

Introduction and Background.

Counterfactual family of estimators.
[ICML, 2019]

Optimization-based framework for estimator design.
[ICML, 2020]

Off-policy Model Selection

SLOPE: A model selection procedure in OPE.
[ICML, 2020]

Off-policy Learning

Multiple logging policies.
[CausalML, 2018]

Deficient support data.
[KDD, 2020]

TALK OUTLINE

Off-policy Evaluation

Introduction and Background.

Counterfactual family of estimators.
[ICML, 2019]

Optimization-based framework for estimator design.
[ICML, 2020]

Off-policy Model Selection

SLOPE: A model selection procedure in OPE.
[ICML, 2020]

Off-policy Learning

Multiple logging policies
[CausalML, 2018]

Deficient support data
[KDD, 2020]

OFF-POLICY EVALUATION

- **Goal:**

Find an estimate $\hat{V}(\pi)$ to measure the expected reward of a new policy π

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

Using the **logged data** from a different known logging policy μ

$$\mathcal{D} = \{x_i, a_i, \mu(a_i|x_i), r_i\}_{i=1}^{n}$$

- **Quality of the estimate $\hat{V}(\pi)$:**

$$\boxed{MSE\left(\hat{V}(\pi)\right) = \mathbb{E}\left(\hat{V}(\pi) - V(\pi)\right)^2 = Bias\left(\hat{V}(\pi)\right)^2 + Var\left(\hat{V}(\pi)\right)}$$

Challenges

Bias data: selection-bias due to the logging policy.

Partial information data: only observe the reward for recommended action.

OFF-POLICY EVALUATION: EXISTING APPROACHES

- Model the bias: Inverse propensity scores (IPS).
  - A weighted average of the data according to importance sampling weights.

$$\hat{V}_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)\, r_i$$

[Horitz & Thompson, 1952; Strehl et al., 2011]

OFF-POLICY EVALUATION: EXISTING APPROACHES

- **Model the bias:** Inverse propensity scores (IPS).
  - A weighted average of the data according to importance sampling weights.

$$\hat{V}_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \underline{w(x_i, a_i)}\, r_i$$

$$\boxed{w(x,a) = \frac{\pi(a|x)}{\mu(a|x)}}$$

[Horitz & Thompson, 1952; Strehl et al., 2011]

OFF-POLICY EVALUATION: EXISTING APPROACHES

- **Model the bias: Inverse propensity scores (IPS).**
  - A weighted average of the data according to importance sampling weights.

$$\hat{V}_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \underline{w(x_i, a_i)}\, r_i$$

$$w(x, a) = \frac{\pi(a|x)}{\mu(a|x)}$$

Unbiased estimator under full support.

High variance when logging policy and target policy differ a lot.

[Horitz & Thompson, 1952; Strehl et al., 2011]

OFF-POLICY EVALUATION: EXISTING APPROACHES

- Model the world: Direct Model (DM).
  - Use logged data $\mathcal{D} = \{x_i, a_i, r_i\}_{i=1}^{n}$ to estimate reward predictor $\hat{\delta}(x, a)$, then using this estimate to do the imputation.

$$\hat{V}_{DM}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a}\pi(a|x_i)\,\hat{\delta}(x_i, a)$$

[Dudik et al., 2011]

OFF-POLICY EVALUATION: EXISTING APPROACHES

- Model the world: Direct Model (DM).
  - Use logged data $\mathcal{D} = \{x_i, a_i, r_i\}_{i=1}^{n}$ to estimate reward predictor $\hat{\delta}(x, a)$, then using this estimate to do the imputation.

$$\hat{V}_{DM}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a}\pi(a|x_i)\,\hat{\delta}(x_i, a)$$

Low variance.

Typically has **high bias** due to model misspecification.

[Dudik et al., 2011]

OFF-POLICY EVALUATION: EXISTING APPROACHES

- Doubly Robust Estimator
  - Use Direct Model as a baseline, also leverages IPS weighting to measure the departure from the baseline.

$$\hat{V}_{DR}(\pi) = \hat{V}_{DM}(\pi) + \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)\,(r_i - \hat{\delta}(x_i, a_i))$$

[Robins & Rotnitzky, 1995; Dudik et al., 2011]

OFF-POLICY EVALUATION: EXISTING APPROACHES

- Doubly Robust Estimator
  - Use Direct Model as a baseline, also leverages IPS weighting to measure the departure from the baseline.

$$\hat{V}_{DR}(\pi) = \hat{V}_{DM}(\pi) + \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)\,(r_i - \hat{\delta}(x_i, a_i))$$

[Robins & Rotnitzky, 1995; Dudik et al., 2011]

OFF-POLICY EVALUATION: EXISTING APPROACHES

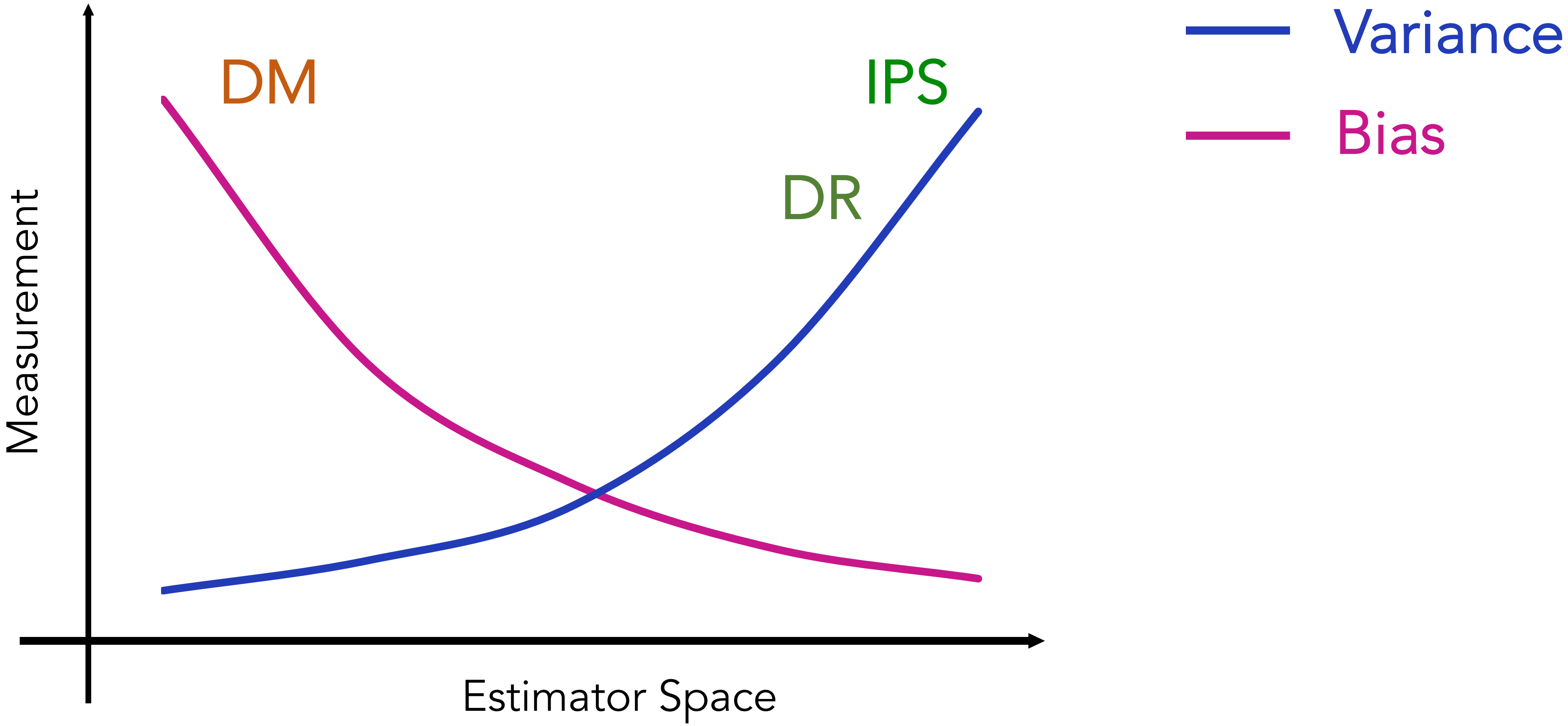
- **Doubly Robust Estimator**
  - Use Direct Model as a baseline, also leverages IPS weighting to measure the departure from the baseline.

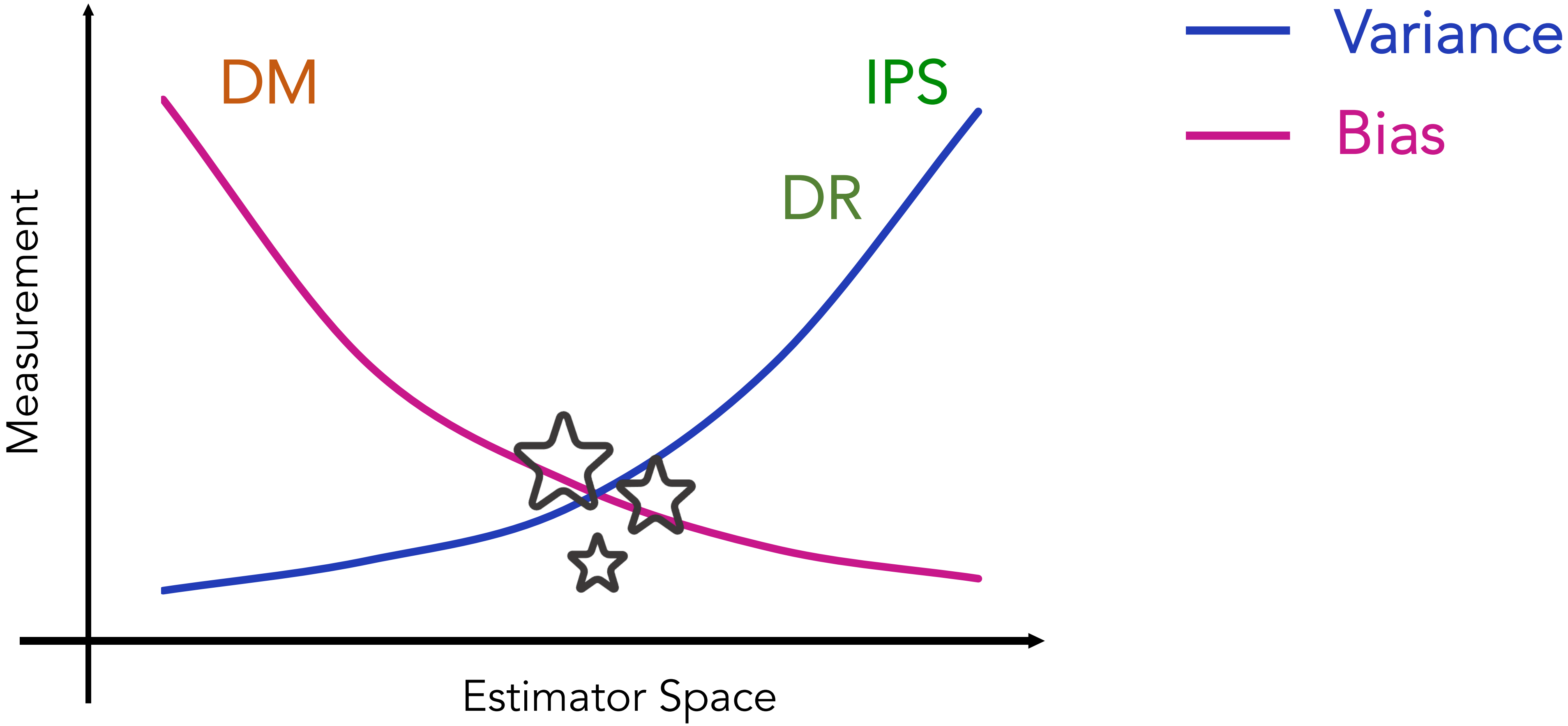
$$\hat{V}_{DR}(\pi) = \hat{V}_{DM}(\pi) + \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)\,(r_i - \hat{\delta}(x_i, a_i))$$

Unbiased estimator, **asymptotically optimal** under mild conditions.

Variance improvement over IPS, but still suffer from high variance.

[Robins & Rotnitzky, 1995; Dudik et al., 2011]


DM
IPS
DR
Measurement
Estimator Space
Variance
Bias



1. How do we **quantify** estimators in between?
2. What is the estimator in the **sweet spot**?

TALK OUTLINE

Off-policy Evaluation

Introduction and Background.

Counterfactual family of estimators.
[ICML, 2019]

Optimization-based framework for estimator design.
[ICML, 2020]

Off-policy Model Selection

SLOPE: A model selection procedure in OPE.
[ICML, 2020]

Off-policy Learning

Multiple logging policies
[CausalML, 2018]

Deficient support data
[KDD, 2020]

INTERPOLATED COUNTERFACTUAL ESTIMATOR FAMILY

Given a triplet $\boldsymbol{\mathcal{W}} = (\boldsymbol{w}^{\alpha}, \boldsymbol{w}^{\beta}, \boldsymbol{w}^{\gamma})$ of weighting functions:

$$\hat{V}^{w}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\pi(a|x_i)\,\boldsymbol{w}_{\boldsymbol{ia}}^{\boldsymbol{\alpha}}\alpha_{ia} + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{\beta}}\beta_i + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{\gamma}}\gamma_i$$

Su,Y.*, Wang.L,*, Santacatterina, M., and Joachims,T. CAB: Continuous adaptive blending estimator for policy evaluation and learning. ICML 2019.

INTERPOLATED COUNTERFACTUAL ESTIMATOR FAMILY

Given a triplet $\boldsymbol{\mathcal{W}} = (\boldsymbol{w^{\alpha}}, \boldsymbol{w^{\beta}}, \boldsymbol{w^{\gamma}})$ of weighting functions:

$$\hat{V}^{w}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\pi(a|x_i)\,\boldsymbol{w_{ia}^{\alpha}}\alpha_{ia} + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w_i^{\beta}}\beta_i + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w_i^{\gamma}}\gamma_i$$

- First Component (Model part): $\boldsymbol{\alpha_{ia} = \hat{\delta}(x_i, a)}$.
  - "Model the world" by having a reward estimator for all (x, a) pairs.
  - The estimator that purely relies on this is DM, which has weights $w = (1,0,0)$.
  - Induce high bias, but typically low variance.

INTERPOLATED COUNTERFACTUAL ESTIMATOR FAMILY

Given a triplet $\boldsymbol{\mathcal{W}} = (\boldsymbol{w^{\alpha}}, \boldsymbol{w^{\beta}}, \boldsymbol{w^{\gamma}})$ of weighting functions:

$$\hat{V}^{w}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\pi(a|x_i)\, w_{ia}^{\alpha}\alpha_{ia} + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w_i^{\beta}}\beta_i + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)w_i^{\gamma}\gamma_i$$

- Second Component (Weighting part): $\boldsymbol{\beta_i := \beta(x_i, a_i) = \frac{r(x_i,a_i)}{\mu(a_i|x_i)}}$
  - "Model the bias" by correcting the probability mismatch.
  - The estimator that purely relies on this is IPS, which put weights $w = (0,1,0)$
  - Induce high variance, but unbiased under mild conditions.

INTERPOLATED COUNTERFACTUAL ESTIMATOR FAMILY

Given a triplet $\mathcal{W} = (w^{\alpha}, w^{\beta}, w^{\gamma})$ of weighting functions:

$$\hat{V}^{w}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\pi(a|x_i)\, w_{ia}^{\alpha}\alpha_{ia} + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)w_i^{\beta}\beta_i + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w_i^{\gamma}}\gamma_i$$

➤ Third Component (Control Variate): $\boldsymbol{\gamma_i \coloneqq \gamma(x_i, a_i) = \frac{\hat{\delta}\,(x_i, a_i)}{\mu(a_i|x_i)}}$

➤ Used as control variate for variance reduction, example: DR.

➤ This part could not be used in some partial information setting, such as Learning to Rank.

INTERPOLATED COUNTERFACTUAL ESTIMATOR FAMILY

Given a triplet $\mathcal{W} = (\boldsymbol{w}^{\alpha}, \boldsymbol{w}^{\beta}, \boldsymbol{w}^{\gamma})$ of weighting functions:

$$\hat{V}^{w}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\pi(a|x_i)\,\boldsymbol{w}_{ia}^{\alpha}\alpha_{ia} + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w}_{i}^{\beta}\beta_i + \frac{1}{n}\sum_{i=1}^{n}\pi(a_i|x_i)\boldsymbol{w}_{i}^{\gamma}\gamma_i$$

$\hat{V}^{w}(\pi) = \boldsymbol{w}_{ia}^{\alpha}$ Model Part + $\boldsymbol{w}_{i}^{\beta}$ Weighting Part + $\boldsymbol{w}_{i}^{\gamma}$ Control Variate

INTERPOLATED COUNTERFACTUAL ESTIMATOR FAMILY

Estimator	w_{ia}^{α} (Model)	w_i^{β} (Weighting)	w_i^{γ} (Control Variate)
DM	1	0	0
IPS	0	1	0
DR	1	1	-1
cIPS	0	$\min\{\frac{M\mu(a_i\|x_i)}{\pi(a_i\|x_i)}, 1\}$	0
MAGIC/SB	$1-\tau$	τ	0
SWITCH	$\mathbb{I}\{\frac{\pi(a\|x_i)}{\mu(a\|x_i)} > M\}$	$\mathbb{I}\{\frac{\pi(a_i\|x_i)}{\mu(a_i\|x_i)} \leq M\}$	0

INTERPOLATED COUNTERFACTUAL ESTIMATOR FAMILY

Estimator	w_{ia}^{α} (Model)	w_i^{β} (Weighting)	w_i^{γ} (Control Variate)
DM	1	0	0
IPS	0	1	0
DR	1	1	-1
cIPS	0	$\min\{\frac{M\mu(a_i\|x_i)}{\pi(a_i\|x_i)}, 1\}$	0
MAGIC/SB	$1-\tau$	τ	0
SWITCH	$\mathbb{I}\{\frac{\pi(a\|x_i)}{\mu(a\|x_i)} > M\}$	$\mathbb{I}\{\frac{\pi(a_i\|x_i)}{\mu(a_i\|x_i)} \leq M\}$	0

SB(Static Blending)

[Thomas & Brunskill, 2016]

Static weighting and does not depend on importance weights.

SWITCH

[Wang, et.al., 2017]

Hard switching makes it not differentiable w.r.t. parameter of policy and could not be used in gradient-based learning algorithms.

DESIRABLE PROPERTIES

- Applicable for a wide range of settings, like LTR, need to make control variate term to be 0.

DESIRABLE PROPERTIES

- Applicable for a wide range of settings, like LTR, need to make control variate term to be 0.

- Low MSE: data dependent weights that allow an instance dependent trade-off between bias and variance.

DESIRABLE PROPERTIES

- Applicable for a wide range of settings, like LTR, need to make control variate term to be 0.

- Low MSE: data dependent weights that allow an instance dependent trade-off between bias and variance.

- Sub-differentiable for gradient based learning.

CONTINUOUS ADAPTIVE BLENDING (CAB)

CAB is a specific estimator in the interpolated counterfactual estimator family with:

$$\hat{V}_{CAB}(\pi) = \hat{V}^{w}(\pi) \quad with \quad \begin{cases} w_{ia}^{\alpha} = 1 - \min\{M \dfrac{\mu(a|x_i)}{\pi(a|x_i)}, 1\} \\ \boldsymbol{w_i^{\beta}} = \min\{M \dfrac{\mu(a_i|x_i)}{\pi(a_i|x_i)}, 1\} \\ \boldsymbol{w_i^{\gamma}} = 0 \end{cases}$$

$$\hat{V}_{CAB}(\pi) = \left(1 - \min\left\{M \frac{\mu(a|x_i)}{\pi(a|x_i)}, 1\right\}\right) \times \text{Model Part} + \min\left\{M \frac{\mu(a_i|x_i)}{\pi(a_i|x_i)}, 1\right\} \times \text{Weighting Part}$$

PROPERTIES OF CAB

- Can be substantially less biased than clipped IPS and DM.
- While having **low variance** compared to IPS and DR.
- Subdifferentiable and capable of **gradient based learning**: POEM (Swaminathan & Joachims, 2015a), BanditNet (Joachims et.al., 2018)
- Unlike DR, can be used in **off-policy Learning to Rank** (LTR) algorithms. (Joachims et.al., 2017)

Estimator	w_{ia}^{α} (Model)	w_i^{β} (Weighting)	w_i^{γ}
DM	1	0	0
cIPS	0	$\min\{\frac{M\mu(a_i\|x_i)}{\pi(a_i\|x_i)}, 1\}$	0
CAB	$1 - \min\left\{M\frac{\mu(a\|x_i)}{\pi(a\|x_i)}, 1\right\}$	$\min\left\{M\frac{\mu(a_i\|x_i)}{\pi(a_i\|x_i)}, 1\right\}$	0

PROPERTIES OF CAB

- Can be substantially **less biased** than clipped IPS and DM.

- While having low variance compared to IPS and DR.

- Subdifferentiable and capable of **gradient based learning**: POEM (Swaminathan & Joachims, 2015a), BanditNet (Joachims et.al., 2018)

- Unlike DR, can be used in **off-policy Learning to Rank** (LTR) algorithms. (Joachims et.al., 2017)

Estimator	w_{ia}^{α} (Model)	w_i^{β} (Weighting)	w_i^{γ}
IPS	0	1	0
DR	1	1	-1
CAB	$1-\min\left\{M\frac{\mu(a\|x_i)}{\pi(a\|x_i)},1\right\}$	$\min\left\{M\frac{\mu(a_i\|x_i)}{\pi(a_i\|x_i)},1\right\}$	0

PROPERTIES OF CAB

- Can be substantially **less biased** than clipped IPS and DM.
- While having **low variance** compared to IPS and DR.
- Subdifferentiable and capable of **gradient based learning**: POEM (Swaminathan & Joachims, 2015a), BanditNet (Joachims et.al., 2018)
- Unlike DR, can be used in **off-policy Learning to Rank** (LTR) algorithms. (Joachims et.al., 2017)

Estimator	w_{ia}^{α} (Model)	w_i^{β} (Weighting)	w_i^{γ}
SWITCH	$\mathbb{I}\{\frac{\pi(a\|x_i)}{\mu(a\|x_i)} > M\}$	$\mathbb{I}\{\frac{\pi(a_i\|x_i)}{\mu(a_i\|x_i)} \leq M\}$	0
CAB	$1 - \min\left\{M\frac{\mu(a\|x_i)}{\pi(a\|x_i)}, 1\right\}$	$\min\left\{M\frac{\mu(a_i\|x_i)}{\pi(a_i\|x_i)}, 1\right\}$	0

PROPERTIES OF CAB

- Can be substantially **less biased** than clipped IPS and DM.
- While having **low variance** compared to IPS and DR.
- Subdifferentiable and capable of **gradient based learning**: POEM (Swaminathan & Joachims, 2015a), BanditNet (Joachims et.al., 2018)
- Unlike DR, can be used in off-policy Learning to Rank (LTR) algorithms. (Joachims et.al., 2017)

Estimator	w_{ia}^{α} (Model)	w_i^{β} (Weighting)	w_i^{γ}
DR	1	1	-1
CAB	$1 - \min\left\{M\frac{\mu(a\|x_i)}{\pi(a\|x_i)}, 1\right\}$	$\min\left\{M\frac{\mu(a_i\|x_i)}{\pi(a_i\|x_i)}, 1\right\}$	0

EXPERIMENTS: SETTINGS

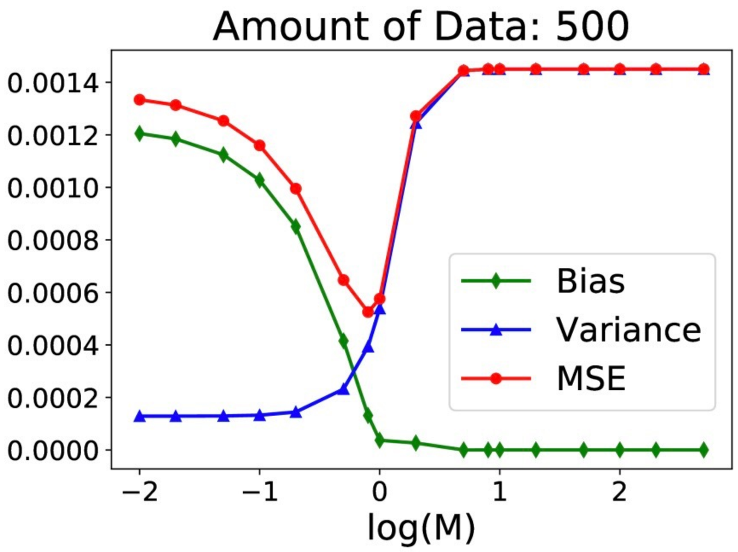
- Batch Learning from Bandit Feedback.
  - Datasets: UCI multi-class classification, bandit conversion.
  - Model: Logistic Regression
  - Policy: Softmax Policy
- Learning to Rank.
  - Datasets: Yahoo LTR!
  - Model: Gradient Boosted Decision Tree
  - Policy: SVM-Rank

EXPERIMENTS: UCI DATASET

➤ **Question 1:** Can CAB achieve improved estimation by trading bias-variance through M?

$$\hat{V}_{CAB}(\pi) = \left(1 - \min\left\{\boldsymbol{M}\frac{\mu(a|x_i)}{\pi(a|x_i)}, 1\right\}\right) \times \text{Model Part} + \min\left\{\boldsymbol{M}\frac{\mu(a_i|x_i)}{\pi(a_i|x_i)}, 1\right\} \times \text{Weighting Part}$$

Performance of CAB: satImage

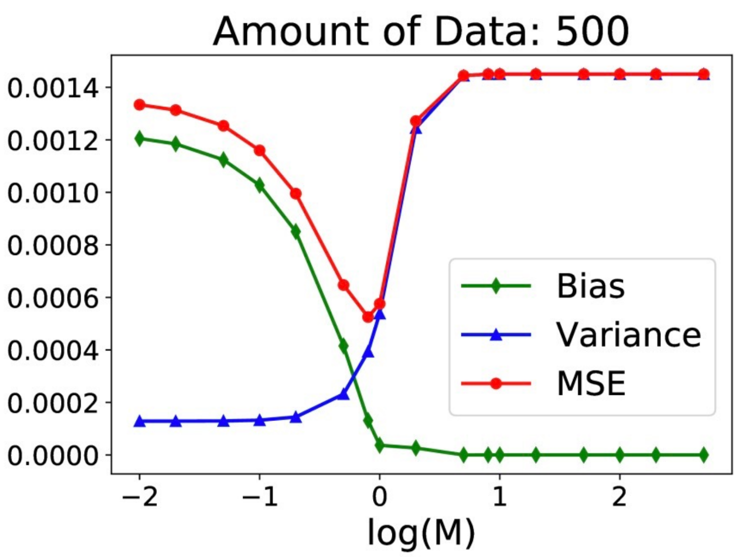


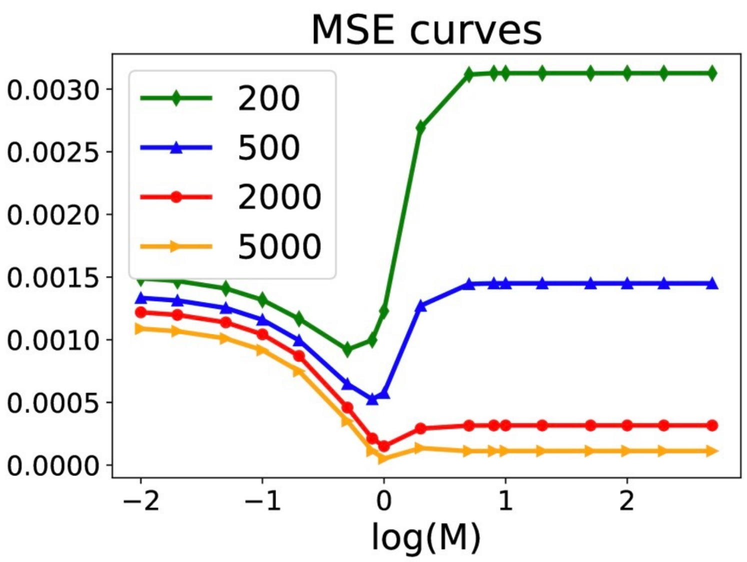
EXPERIMENTS: UCI DATASET

➤ **Question 1:** Can CAB achieve improved estimation by trading bias-variance through M?

$$\hat{V}_{CAB}(\pi) = \left(1 - \min\left\{\boldsymbol{M}\frac{\mu(a|x_i)}{\pi(a|x_i)}, 1\right\}\right) \times \text{Model Part} + \min\left\{\boldsymbol{M}\frac{\mu(a_i|x_i)}{\pi(a_i|x_i)}, 1\right\} \times \text{Weighting Part}$$

Performance of CAB: satImage



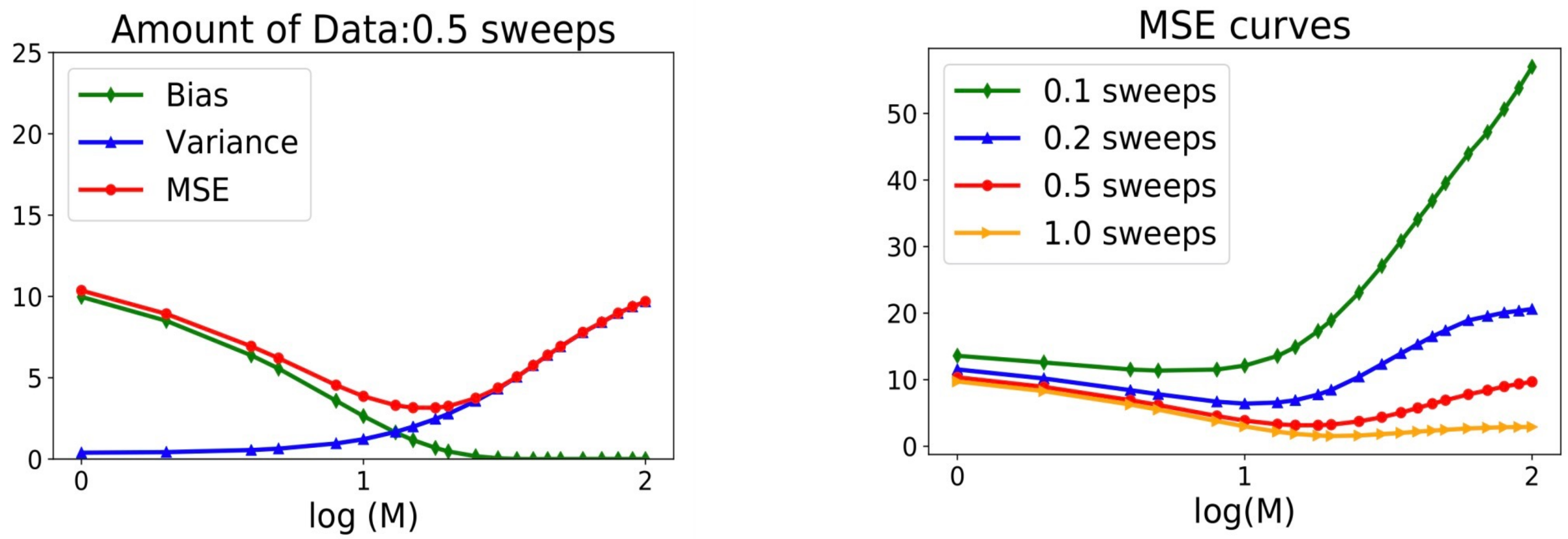


EXPERIMENTS: YAHOO LTR!

➤ **Question 1:** Can CAB achieve improved estimation by trading bias-variance through M?

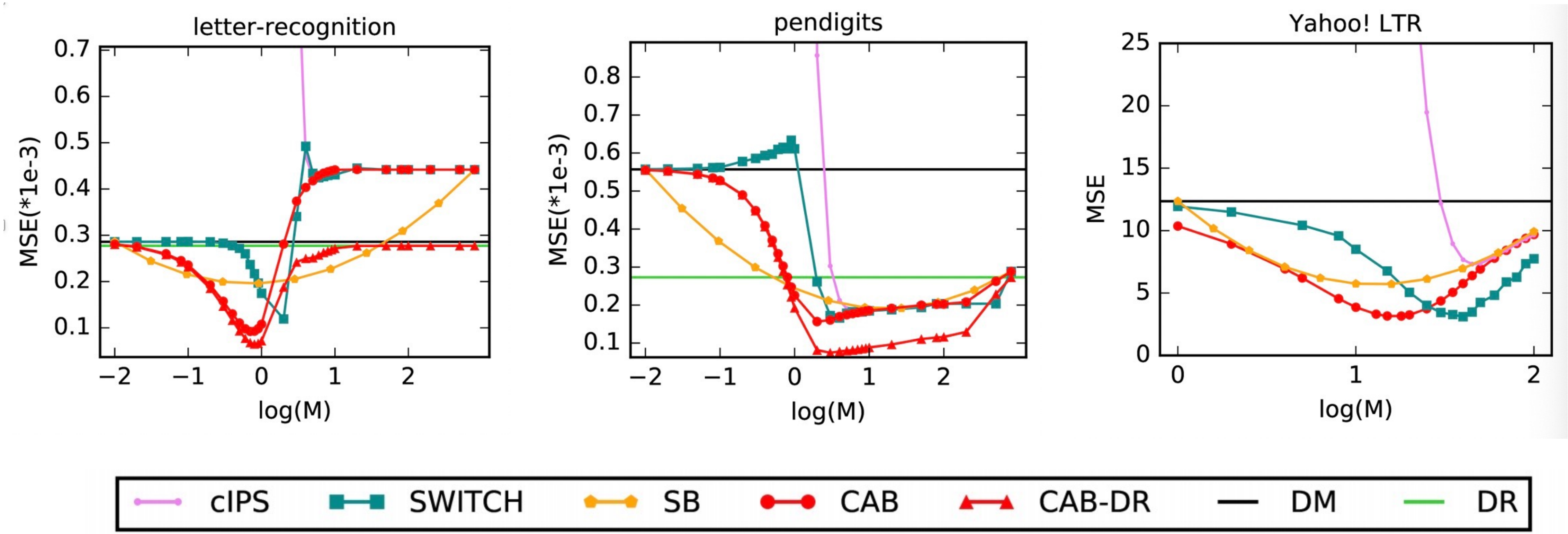
$$\hat{V}_{CAB}(\pi) = \left(1 - \min\left\{\boldsymbol{M}\frac{\mu(a|x_i)}{\pi(a|x_i)}, 1\right\}\right) \times \text{Model Part} + \min\left\{\boldsymbol{M}\frac{\mu(a_i|x_i)}{\pi(a_i|x_i)}, 1\right\} \times \text{Weighting Part}$$

Performance of CAB: Yahoo LTR!

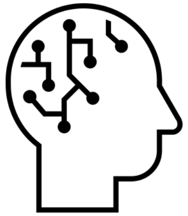


EXPERIMENTS

➤ **Question 2:** How does CAB compared with other estimators?



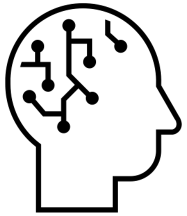
LESSONS LEARNT



A family of estimators

↓

Flexible bias variance tradeoff



CAB (smooth weight clipping)

↓

slightly higher bias

+

substantially lower variance

A specific weight design → CAB

Is there any ***systematic way*** to **design the weights** for better bias-variance tradeoff?

TALK OUTLINE

Off-policy Evaluation

Introduction and Background.

Counterfactual family of estimators.
[ICML, 2019]

Optimization-based framework for estimator design.
[ICML, 2020]

Off-policy Model Selection

SLOPE: A model selection procedure in OPE.
[ICML, 2020]

Off-policy Learning

Multiple logging policies
[CausalML, 2018]

Deficient support data
[KDD, 2020]

DOUBLY ROBUST ESTIMATOR WITH SHRINKAGE (DRS)

$$\hat{V}_{DR}(\pi) = \hat{V}_{DM}(\pi) + \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)\,(r_i - \hat{\delta}(x_i, a_i))$$

DR is asymptotically optimal.

However, it still suffers from the large variance due to utilizing the importance sampling weight.

Su.Y, Dimakopoulou.M, Krishnamurthy.A, and Dudík.M. Doubly robust off-policy evaluation with shrinkage. ICML, 2020.

DOUBLY ROBUST ESTIMATOR WITH SHRINKAGE (DRS)

Replace the original weight $\boldsymbol{w}(\boldsymbol{x}, \boldsymbol{a})$ by a shrinkage version $\widehat{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{a})$.

$$\hat{V}_{DR}(\pi, \widehat{w}, \hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)\,(r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

$$\hat{V}_{DRS}(\pi, \widehat{w}, \hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(x_i, a_i)\,(r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

$$0 \leq \widehat{w}(x_i, a_i) \leq w(x_i, a_i)$$

DOUBLY ROBUST ESTIMATOR WITH SHRINKAGE

Replace the original weight $\boldsymbol{w}(\boldsymbol{x}, \boldsymbol{a})$ by a shrinkage version $\widehat{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{a})$.

$$\hat{V}_{DRS}(\pi, \widehat{w}, \hat{\delta}) = \frac{1}{n} \sum_{i=1}^{n} \widehat{w}(x_i, a_i)\, (r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

Which form of shrinkage should we use?

Which one should we use for our specific reward predictor?

DOUBLY ROBUST ESTIMATOR WITH SHRINKAGE

Our approach:

Directly finding the optimal weights by minimizing an upper bound of the MSE

APPROACH I: BEING PESSIMISTIC

Assume $\sup_{x,a} \left| r - \hat{\delta}(x,a) \right| \le 1$

- Bias: $Bias(\widehat{w}) \le UB(Bias) = \mathbb{E}_{\mu}[|\widehat{w}(x,a) - w(x,a)|]$
- Variance: $Var(\widehat{w}) \lesssim UB(Var) = \frac{1}{n}\mathbb{E}_{\mu}[\widehat{w}(x,a)^2]$

APPROACH I: BEING PESSIMISTIC

Assume $\sup_{x,a} \left| r - \hat{\delta}(x,a) \right| \leq 1$

- **Bias**: $Bias(\widehat{w}) \leq UB(Bias) = \mathbb{E}_{\mu}[|\widehat{w}(x,a) - w(x,a)|]$
- **Variance**: $Var(\widehat{w}) \lesssim UB(Var) = \frac{1}{n}\mathbb{E}_{\mu}[\widehat{w}(x,a)^2]$
- The optimal weights can be obtained by minimizing:

$$UB(Bias) + \lambda \cdot UB(Var)$$

APPROACH I: BEING PESSIMISTIC

Assume $\sup_{x,a} \left| r - \hat{\delta}(x,a) \right| \leq 1$

- Bias: $Bias(\widehat{w}) \leq UB(Bias) = \mathbb{E}_{\mu}[|\widehat{w}(x,a) - w(x,a)|]$
- Variance: $Var(\widehat{w}) \lesssim UB(Var) = \frac{1}{n}\mathbb{E}_{\mu}[\widehat{w}(x,a)^2]$
- The optimal weights can be obtained by minimizing:

$$UB(Bias) + \lambda \cdot UB(Var)$$

- Solution: $\widehat{w}(x,a) = \min\{\lambda, w(x,a)\}$ $\longrightarrow$ Clipping Estimator

APPROACH II: BEING OPTIMISTIC

Typically, the reward estimator $\hat{\delta}(x, a)$ is trained to minimize the weighted square loss based on some weighting function $z(x, a)$:

$$L(\hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} z(x_i, a_i)\left(r_i - \hat{\delta}(x_i, a_i)\right)^2$$

Popular choices include $z = 1,\ z = w(x, a),\ z = w(x, a)^2$

APPROACH II: BEING **OPTIMISTIC**

- **Bias:** $Bias^2(\hat{w}) \leq \mathbb{E}_\mu[\frac{1}{z(x,a)}(\hat{w}(x,a) - w(x,a))^2]L(\hat{\delta})$
- **Variance:** $Var(\hat{w}) \lesssim \sqrt{\mathbb{E}_\mu[\frac{w(x,a)^2}{z(x,a)}\ \hat{w}(x,a)^2]}\sqrt{L(\hat{\delta})}$
- Using similar trick to minimize an upper bound of MSE.

- **Solution:** $\hat{\boldsymbol{w}}(\boldsymbol{x},\boldsymbol{a}) = \frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda}+\boldsymbol{w}(\boldsymbol{x},\boldsymbol{a})^2}\boldsymbol{w}(\boldsymbol{x},\boldsymbol{a})$ $\longrightarrow$ **Shrinkage Estimator**

DOUBLY ROBUST ESTIMATOR WITH SHRINKAGE

$$\hat{V}_{DRS-p}(\pi, \hat{w}, \hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} \min\{\lambda, w(x,a)\}(r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

$$\hat{V}_{DRS-o}(\pi, \hat{w}, \hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda} + \boldsymbol{w}(\boldsymbol{x},\boldsymbol{a})^2} \boldsymbol{w}(\boldsymbol{x}, \boldsymbol{a})\ (r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

➤ Interpolating between DM and DR:

- $\lambda = 0 \ \rightarrow\ \hat{V}_{DM}(\pi)$, small variance, large bias
- $\lambda = \infty \rightarrow \hat{V}_{DR}(\pi)$, large variance, small bias

EMPIRICAL EVALUATION

For non-combinatorial bandit, we perform 108 settings:

- 9 UCI multi-class classification datasets
- 6 different logging policies
- 2 reward conditions: deterministic reward and stochastic reward

EMPIRICAL EVALUATION

➤ Ablation Studies for DR with shrinkage.

$$\hat{V}_{DRS}(\pi, \hat{w}, \hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}(x_i, a_i)\, (r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

- evaluating different reward predictors: $z = 1, w(x,a), w(x,a)^2$.

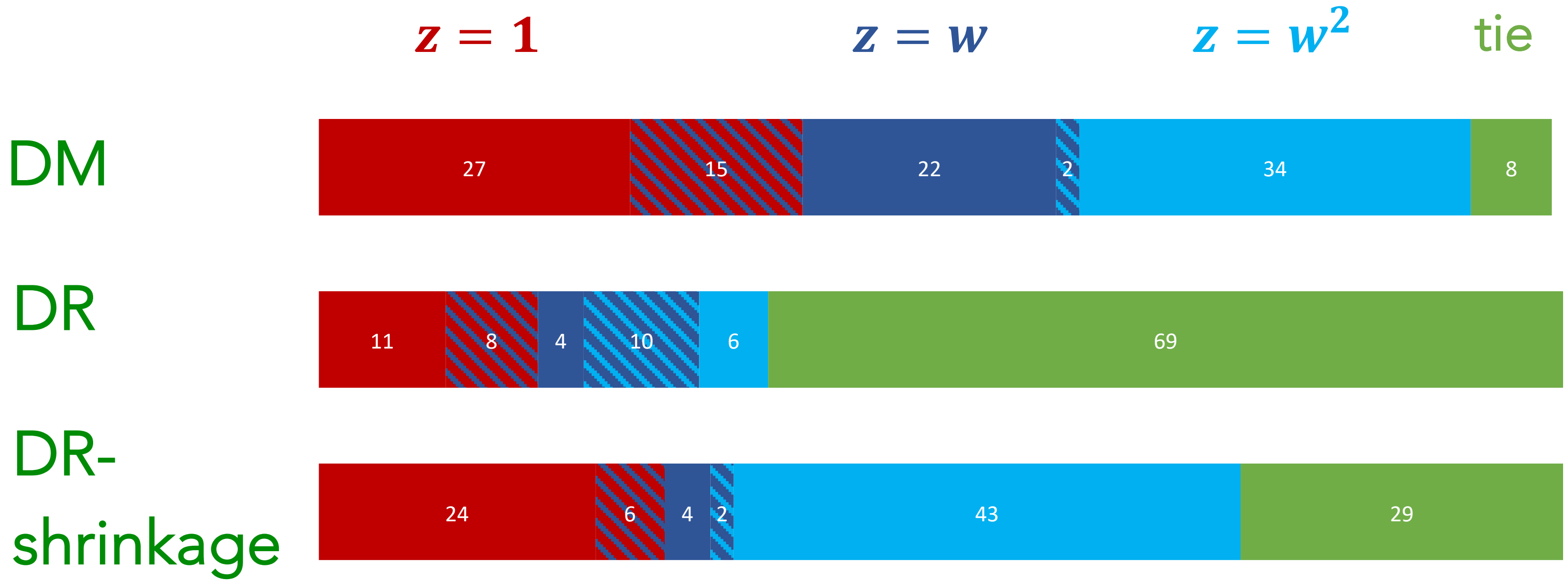
$$L(\hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} z(x_i, a_i)\left(r_i - \hat{\delta}(x_i, a_i)\right)^2$$

- evaluating the optimistic and pessimistic shrinkage types.

EMPIRICAL EVALUATION

Do we need all different reward predictors?

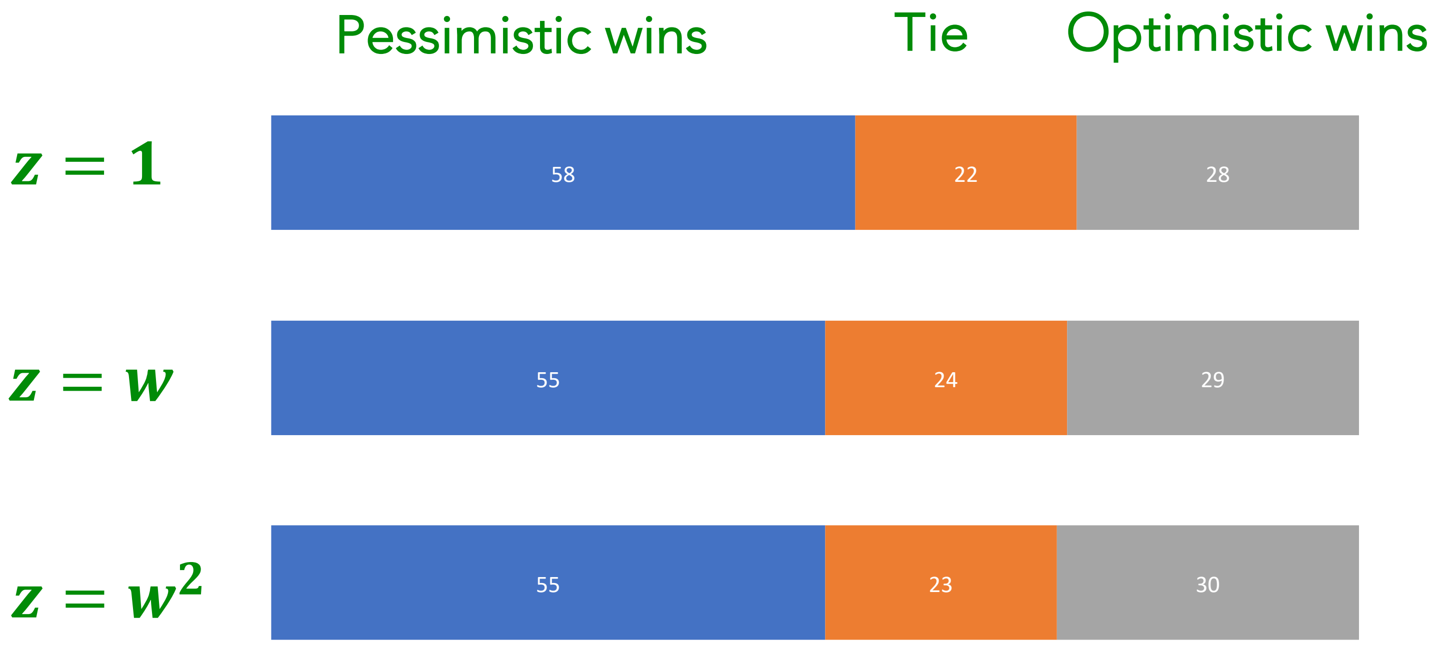
How often across 108 conditions is each of the reward predictor the best?



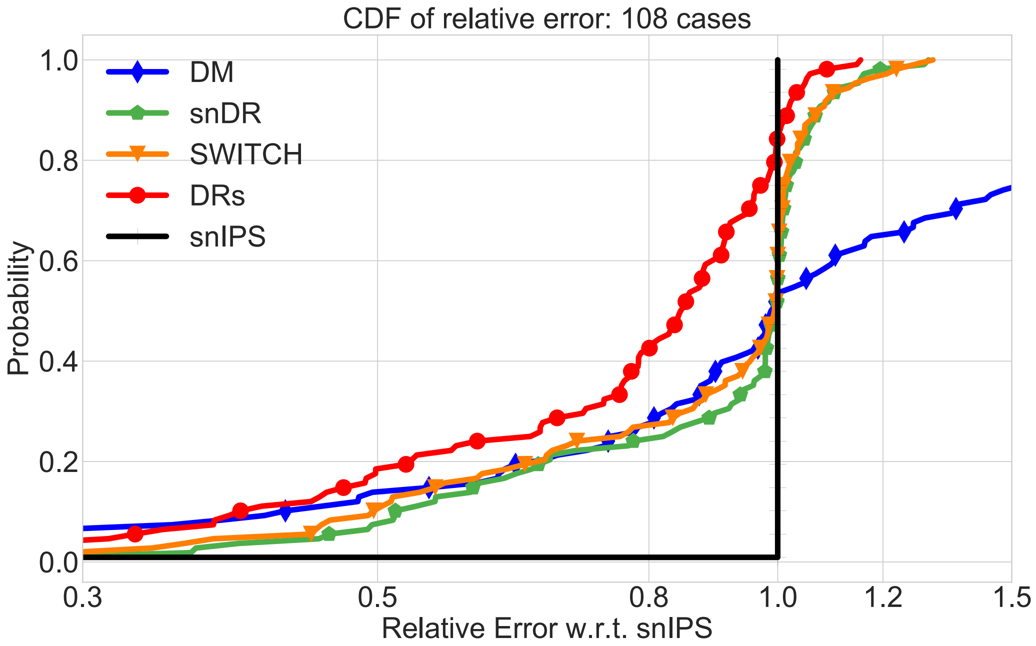
EMPIRICAL EVALUATION

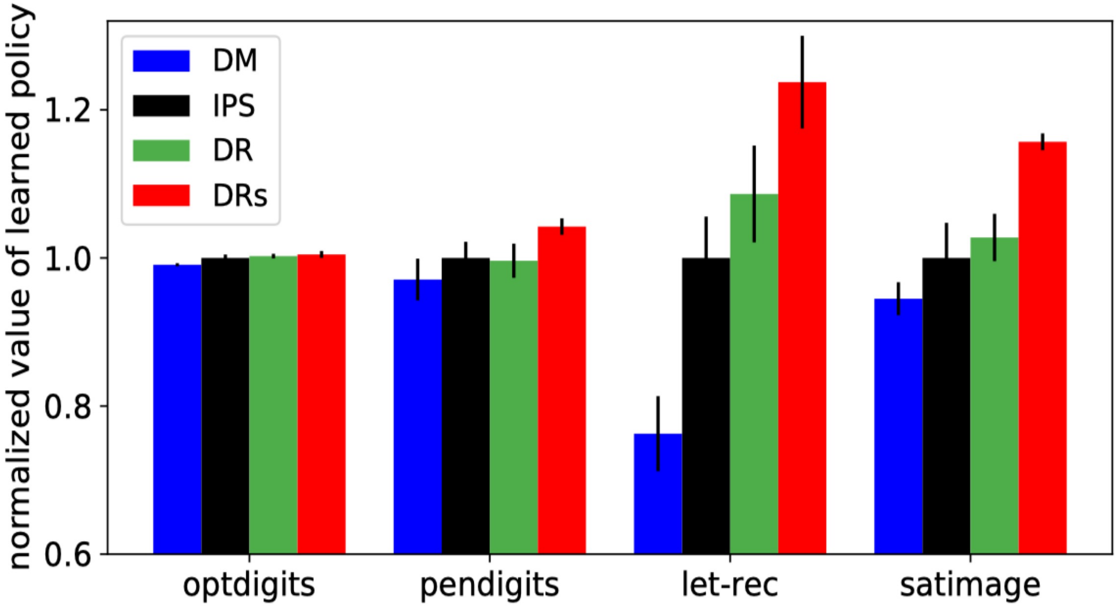
Do we need both pessimistic shrinkage and optimistic shrinkage?

How often across 108 conditions is each of them better in DR with shrinkage?



EMPIRICAL EVALUATION

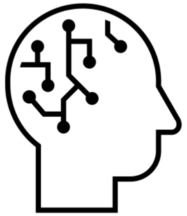




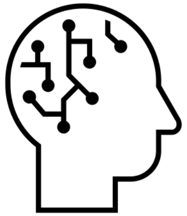
Evaluation Performance

Learning Performance

LESSONS LEARNT



Instead of manually constructing estimators, there is an optimization-based framework to design estimators.



Different reward predictors and weight shrinkage types perform well in different settings.

$$\hat{V}_{CAB}(\pi) = \left(1 - \min\left\{M \frac{\mu(a|x_i)}{\pi(a|x_i)}, 1\right\}\right) \times \text{Model Part} + \min\left\{M \frac{\mu(a_i|x_i)}{\pi(a_i|x_i)}, 1\right\} \times \text{Weighting Part}$$

$$\hat{V}_{DRS-p}(\pi, \hat{w}, \hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} \min\{\lambda, w(x,a)\}(r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

$$\hat{V}_{DRS-o}(\pi, \hat{w}, \hat{\delta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda}{\lambda + \boldsymbol{w}(\boldsymbol{x},\boldsymbol{a})^2} \boldsymbol{w}(\boldsymbol{x},\boldsymbol{a}) \, (r_i - \hat{\delta}(x_i, a_i)) + \hat{V}_{DM}(\pi)$$

How do we select the ***hyper-parameters*** in OPE?

TALK OUTLINE

Off-policy Evaluation

Introduction and Background.

Counterfactual family of estimators.

[ICML, 2019]

Optimization-based framework for estimator design.

[ICML, 2020]

Off-policy Model Selection

SLOPE: A model selection procedure in OPE.

[ICML, 2020]

Off-policy Learning

Multiple logging policies

[CausalML, 2018]

Deficient support data

[KDD, 2020]

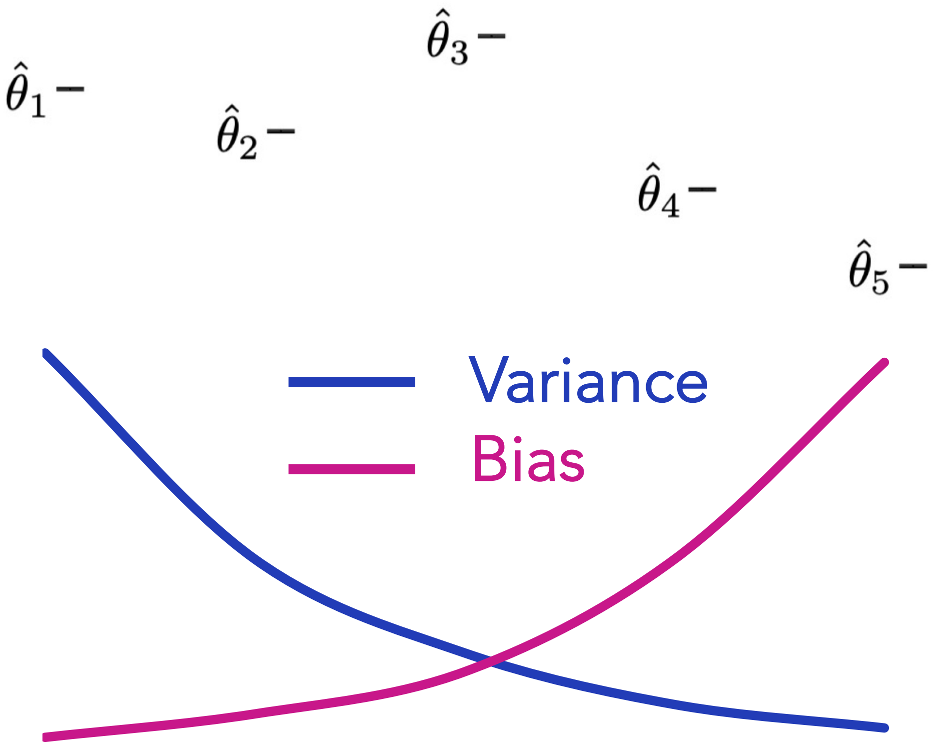
Off-policy Model Selection:

Among a family of off-policy estimates $\hat{V}(\pi)$,

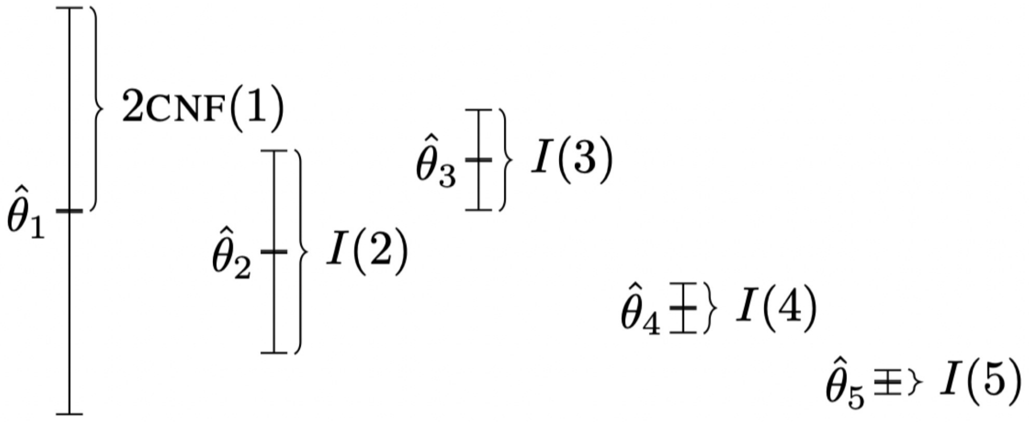
selects the one with highest evaluation accuracy.

OFF POLICY MODEL SELECTION: SLOPE

1 Ordering



2 Building CIs

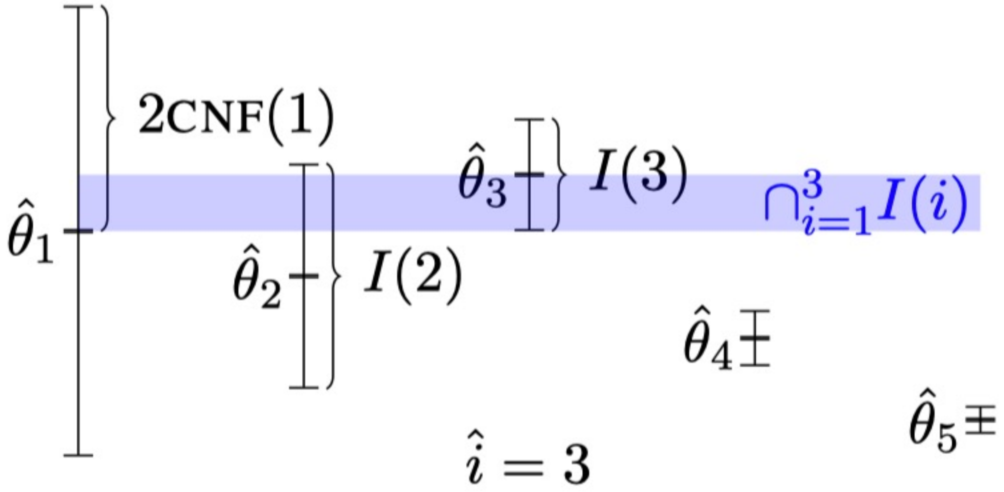


$$I_j = [\hat{\theta}_i - 2CNF(i), \hat{\theta}_i + 2CNF(i)]$$

Su.Y, Srinath.P, Krishnamurthy.A, Adaptive Estimator Selection for Off-Policy Evaluation, ICML 2020

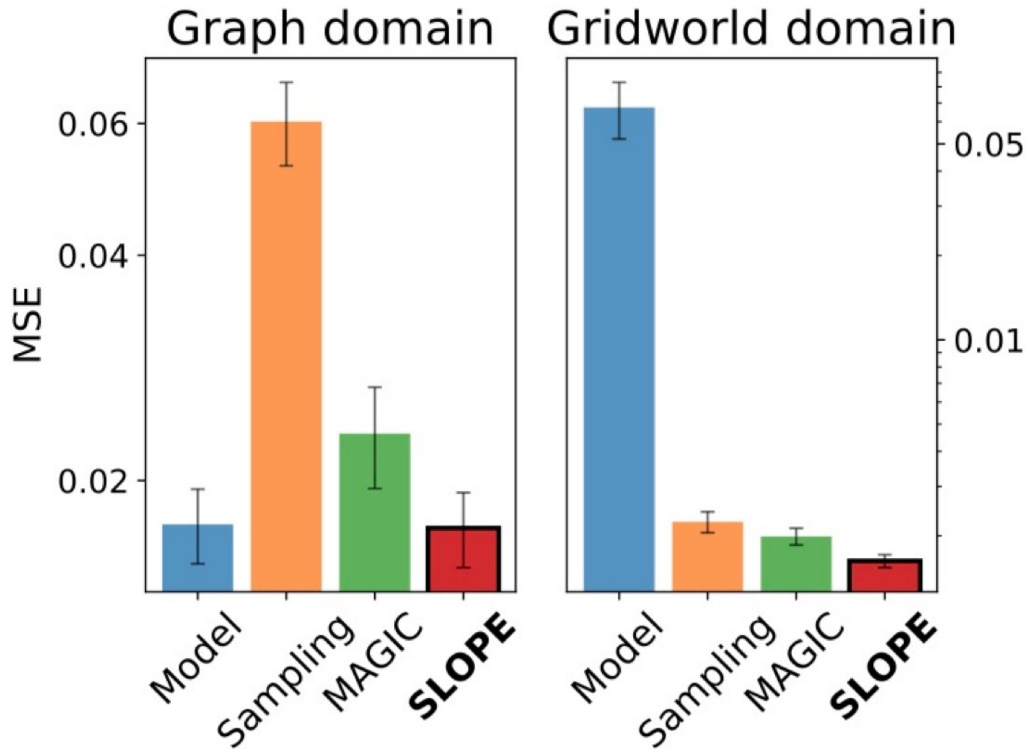
OFF POLICY MODEL SELECTION: SLOPE

③ Index Selection



$$\hat{\imath} := \max\{i \in [M]: \cap_{j=1}^{M} I_j \neq \emptyset\}$$

Performance



Su.Y, Srinath.P, Krishnamurthy.A, Adaptive Estimator Selection for Off-Policy Evaluation, ICML 2020

Off-policy Learning:

Learn an optimal policy $\boldsymbol{\pi}^*$ in some hypothesis space Π

$$\pi^* = argmax_{\pi \in \Pi} V(\pi)$$

Tool: ERM based on an OPE estimate

$$\hat{\pi}^* = argmax_{\pi \in \Pi} \boldsymbol{\hat{V}(\pi)}$$

TALK OUTLINE

Off-policy Evaluation

Introduction and Background.

Counterfactual family of estimators.
[ICML, 2019]

Optimization-based framework for estimator design.
[ICML, 2020]

Off-policy Model Selection

SLOPE: A model selection procedure in OPE.
[ICML, 2020]

Off-policy Learning

Multiple logging policies
[CausalML, 2018]

Deficient support data
[KDD, 2020]

OFF POLICY LEARNING: MULTIPLE POLICIES

logged data
$\mathcal{D}_1$ from $\boldsymbol{\pi}_1$

logged data
$\mathcal{D}_2$ from $\boldsymbol{\pi}_2$

logged data
$\mathcal{D}_k$ from $\boldsymbol{\pi}_k$

Training logs are collected under **multiple** policies.

Naively using IPS in learning will give sub-optimal results.

Utilize a **weighted** estimator, to track the divergence between the learned policy and various logging policies.

Su.Y, Agarwal.A, Joachims.T, Learning from Logged Bandit Feedback of Multiple Loggers, CausalML 2018

TALK OUTLINE

Off-policy Evaluation

Introduction and Background.

Counterfactual family of estimators.
[ICML, 2019]

Optimization-based framework for estimator design.
[ICML, 2020]

Off-policy Model Selection

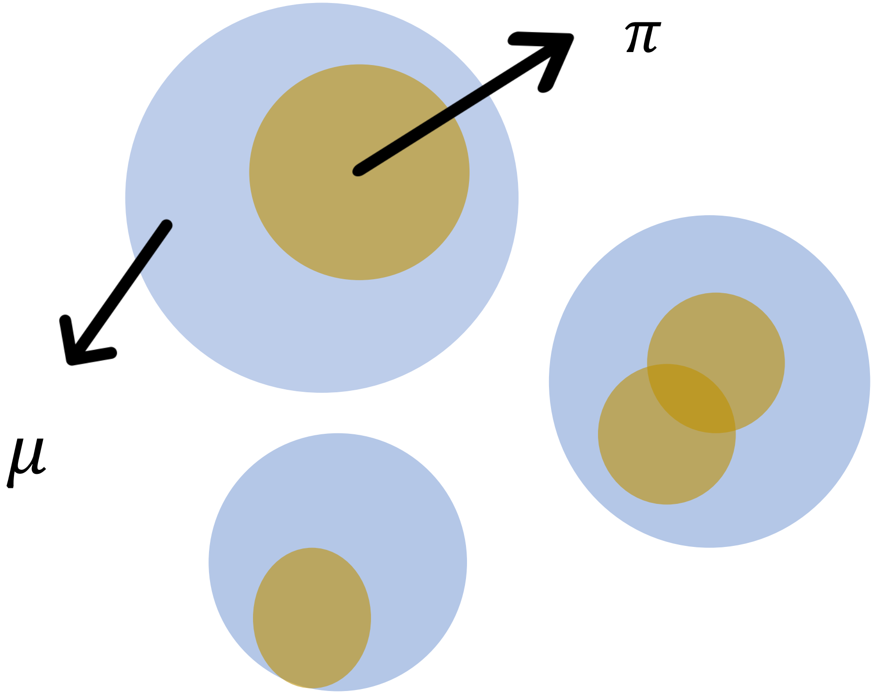
SLOPE: A model selection procedure in OPE.
[ICML, 2020]

Off-policy Learning

Multiple logging policies
[CausalML, 2018]

Deficient support data
[KDD, 2020]

OFF POLICY LEARNING: DEFICIENT SUPPORT DATA

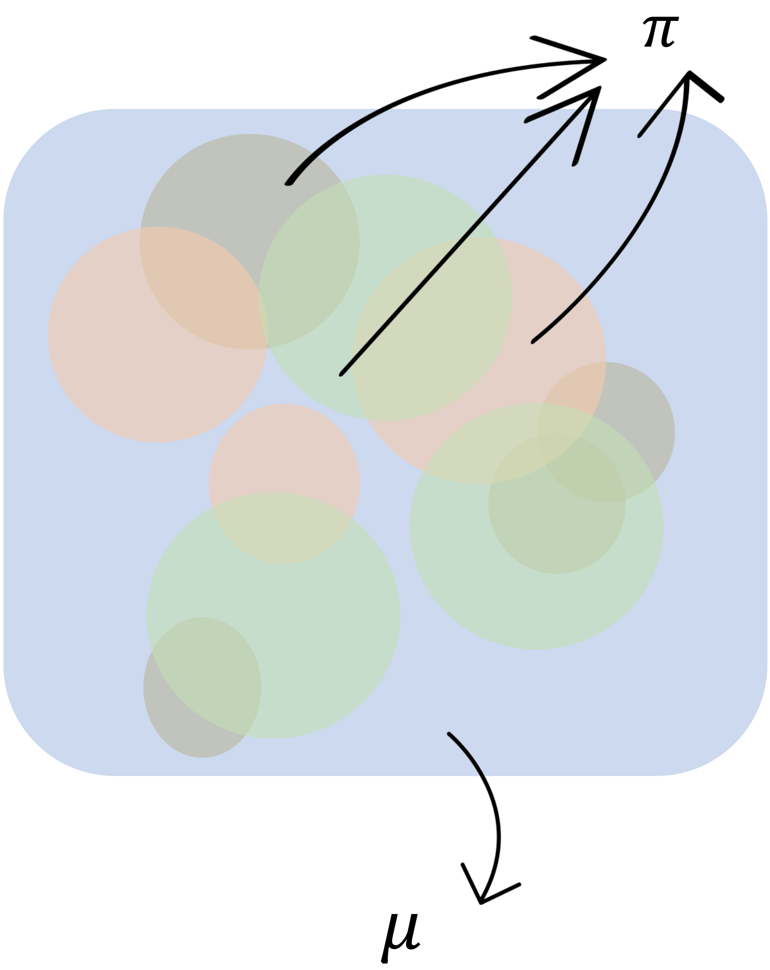


Effectiveness of IPS relies on the crucial **full support** assumption

The logging policy μ is said to have full support for π:
$\mu(a|x) > 0$ whenever $\pi(a|x) > 0$

Sachdeva.N*, Su.Y*, Joachims.T, Off-policy Bandits with Deficient Support. KDD, 2020

OFF POLICY LEARNING: DEFICIENT SUPPORT DATA

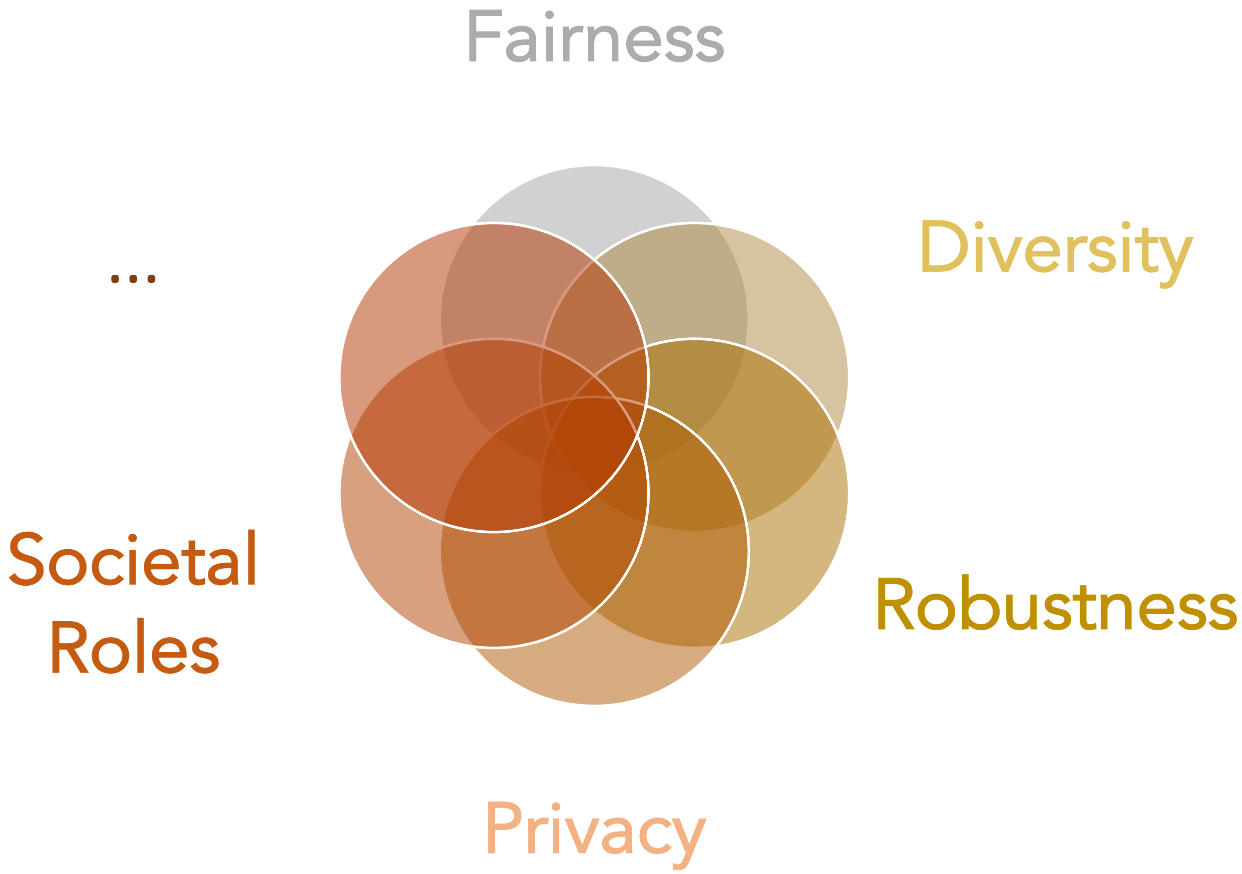


The logging policy needs to assign
non-zero probability to every action a
for every context x !

We propose three efficient approaches
to overcome the support deficient issue
by restricting action space,
policy space
and reward extrapolation.

Sachdeva.N*, Su.Y*, Joachims.T, Off-policy Bandits with Deficient Support. KDD, 2020

Beyond off-policy evaluation and learning …



MULTI-SIDED MARKET PLATFORM

Traditional Recommender Systems

NETFLIX Spotify The New York Times

☆ Only users have preference.

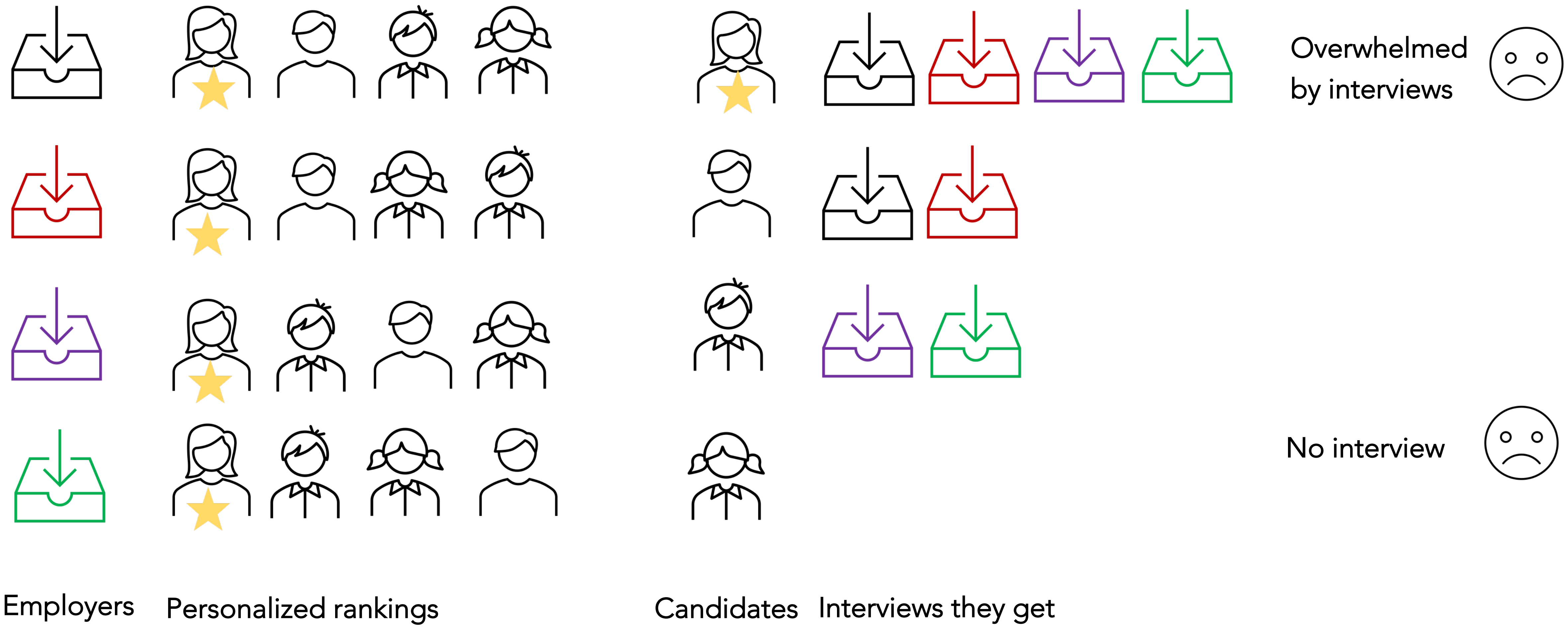
Multi-sided Market Platforms

LinkedIn airbnb tinder

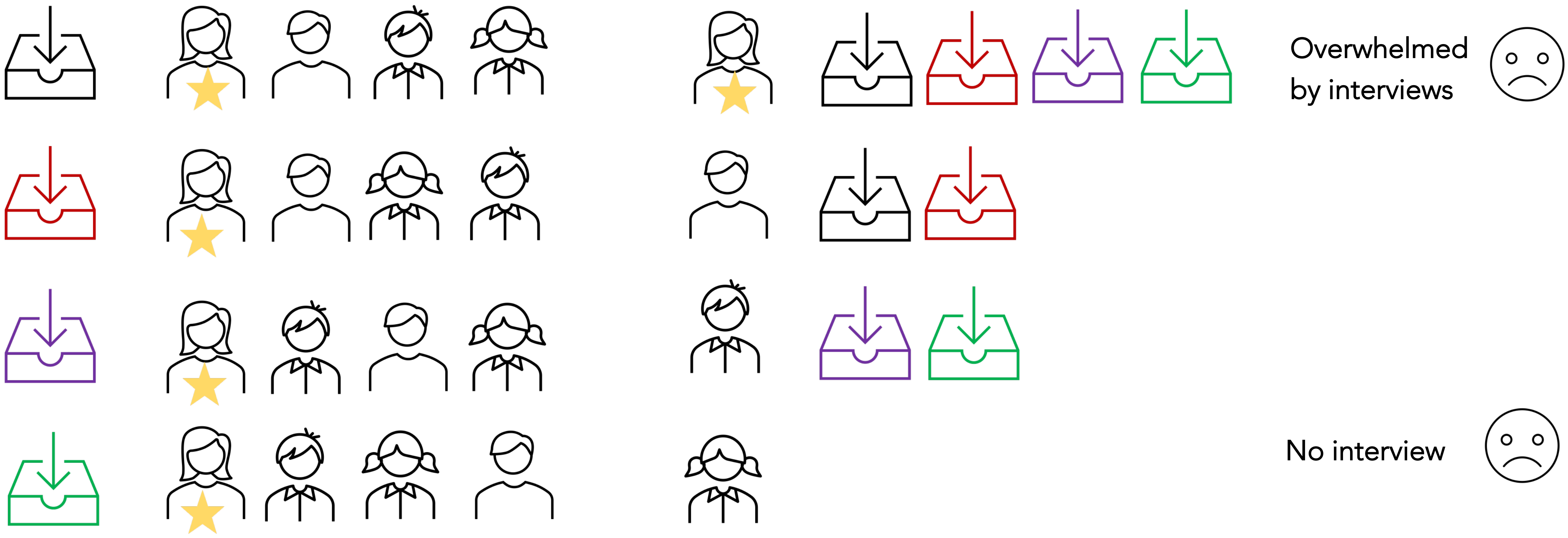
☆ Preference from both sides.

☆ Scarcity in the supply side.

MULTI-SIDED MARKET PLATFORM



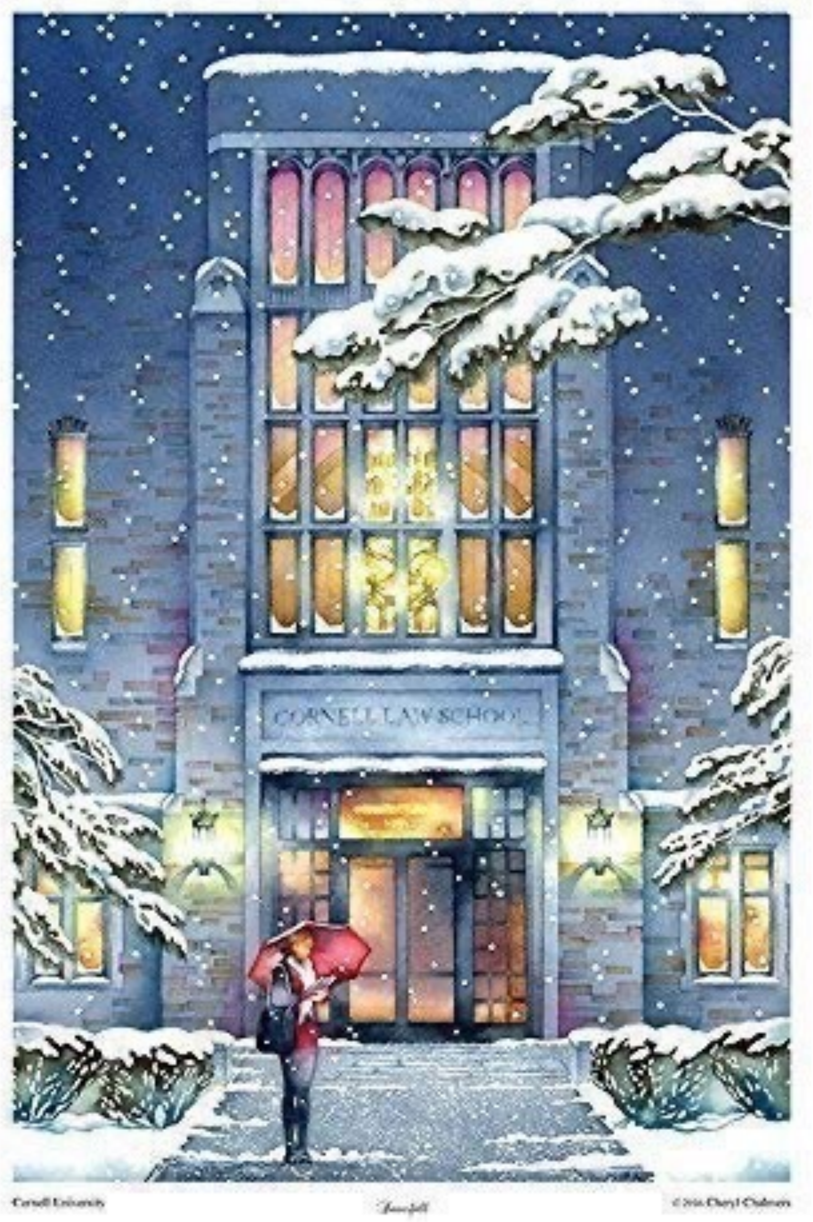
MULTI-SIDED MARKET PLATFORM



Societal Roles of Recommender Systems

Thorsten Joachims (Cornell)
Miro Dudik (Microsoft Research, NYC)
Akshay Krishnamurthy (Microsoft Research, NYC)
Pavithra Srinath (Microsoft Research, NYC)
Maria Dimakopoulou (Netflix)
Michele Santacatterina (Cornell)
Luke Wang (Cornell)
Noveen Sachdeva (UCSB)

Thank you!



REFERENCES

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In International Conference on Machine Learning, 2011.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In International Conference on Machine Learning, 2015.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In International Conference on Machine Learning, 2017.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In International Conference on Machine Learning, 2016.

Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In International Conference on Machine Learning, 2019.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In International Conference on Machine Learning, 2020.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy, Adaptive Estimator Selection for Off-Policy Evaluation. In International Conference on Machine Learning, 2020.

REFERENCES

Noveen Sachdeva, Yi Su, Thorsten Joachims, Off-policy bandits with deficient support. International Conference on Knowledge Discovery & Data Mining, 2020.

Yi Su, Aman Agarwal, Thorsten Joachims, Learning from logged bandit feedback of multiple loggers. In CausalML, 2018.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. The Journal of Machine Learning Research, 2013.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 90(429):122– 129, 1995.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. Biometrics, 2005.

Swaminathan, Adith, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. "Off-policy evaluation for slate recommendation." In Advances in Neural Information Processing Systems, pp. 3632-3642. 2017.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In International Conference on Machine Learning, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.