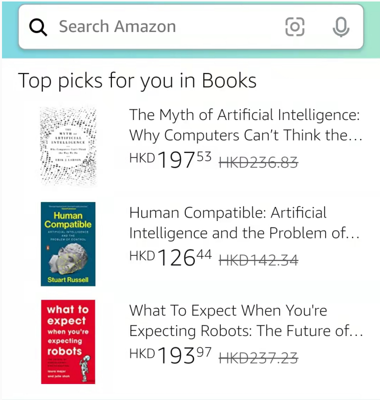
Self-fulfilling Bandits

Dynamic Selection in Algorithmic Decision-making

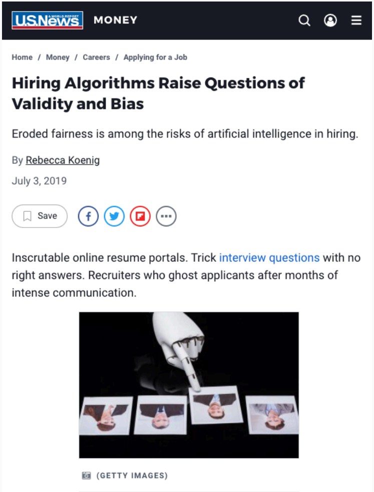
Jin Li Ye Luo Xiaowei Zhang

University of Hong Kong

Algorithmic Decision-making



Recommendation


U.S.News MONEY

Home / Money / Careers / Applying for a Job

Hiring Algorithms Raise Questions of Validity and Bias

Eroded fairness is among the risks of artificial intelligence in hiring.

By Rebecca Koenig

July 3, 2019

Save

Inscrutable online resume portals. Trick interview questions with no right answers. Recruiters who ghost applicants after months of intense communication.

(GETTY IMAGES)

Hiring

Features

- Personalization: Decisions depend on individual characteristics
- Online: Decision rules are constantly adjusted in response to the new information from the behaviors of the targeted users

$$\{\text{Decision-makers}\} \subset \{\text{Data-generators}\}$$

Example: Recommendation on E-Commerce Platforms

1. Observe visitor characteristics (e.g., age, gender, browsing history)
2. Decide which product to recommend
3. Collect a random reward (e.g., clicks, purchases)
4. Adjust the estimate of the reward function and the decision rule

Contextual Bandit Problem

At each time $t = 1, \ldots, T$, an agent

- Observes a vector of covariates $\boldsymbol{v}_t \in \mathbb{R}^p$
- Takes an action (i.e., pulls an arm) $a_t \in \mathcal{A} = \{1, \ldots, M\}$
- Receives a *random* reward R_t:

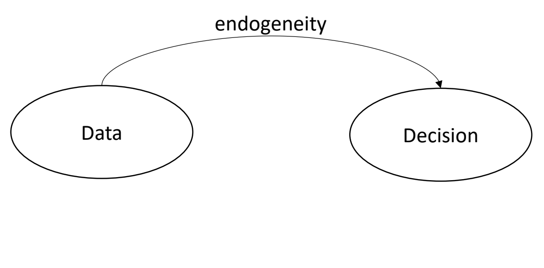
$$R_t = \sum_{i=1}^{M} \mu_i(\boldsymbol{v}_t)\, \mathbb{I}(a_t = i) + \epsilon_t,$$

where $\mu_i(\boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{\alpha}_i$ is a linear reward function with unknown parameters $\boldsymbol{\alpha}_i \in \mathbb{R}^p$

Goal: A policy π that maps $(\text{past data}, \boldsymbol{v}_t)$ to a_t to minimize the cumulative regret

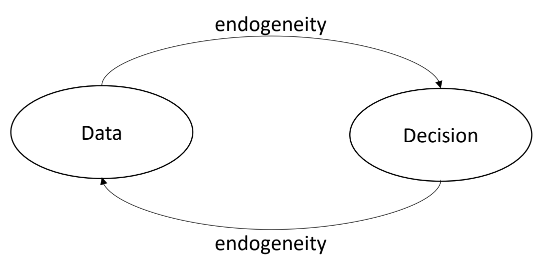
Endogeneity

- Most studies in bandit literature assume $\boldsymbol{v}_t$ and ϵ_t are independent
- Endogeneity problems ($\mathbb{E}[\epsilon_t|\boldsymbol{v}_t] \neq 0$) are key to analyzing human behaviors
- Measurement error, model misspecification, sample selection, or omitted variables



Endogeneity in Online Learning Algorithms

- Endogeneity problem is exacerbated by online learning environments
- Dynamic selection problem: Endogeneity affects the outcomes of data analysis and, therefore, influences the actions taken and the data generated



Contributions

- Identify a novel type of bias—self-fulfilling bias—in online learning environments
- Propose algorithms that not only correct for the bias but also generate actions that attain low levels of regret
- Develop a technique that facilitates theoretical analysis of online learning algorithms with endogeneity problems

An Simple Example

- Arm 1 (safe arm): known reward c independent of the covariates
- Arm 2 (risky arm): linear expected reward with unknown coefficient $\alpha > 0$

$$\mu_1(v) \equiv c \quad \text{and} \quad \mu_2(v) = \alpha v, \quad \forall v \in \mathbb{R}$$

- Optimal policy: pull arm 2 if and only if $\alpha v > c$ (if α is known)

Dynamic Selection

- The agent makes his decisions according to

$$\text{Select} \begin{cases} \text{arm } 2, & \text{if } \widehat{\alpha}_t v_t > c \\ \text{arm } 1, & \text{otherwise} \end{cases}$$

- He uses OLS to update the estimate of α over time

$$\widehat{\alpha}_{t+1} = \begin{cases} \text{run OLS with the addition of } (v_t, R_t), & \text{if } \widehat{\alpha}_t v_t > c \\ \widehat{\alpha}_t, & \text{otherwise} \end{cases}$$

- The data used for estimating α is only available when arm 2 is pulled

Self-fulfilling Bias

- Suppose $\widehat{\alpha}_t \to \widehat{\alpha}$
- Then, in the long run,

$$\widehat{\alpha} = \alpha + \frac{\mathrm{Cov}[v_t, \epsilon_t \mid \widehat{\alpha} v_t > c]}{\mathrm{Var}[v_t \mid \widehat{\alpha} v_t > c]}$$

- $\widehat{\alpha}$ is a fixed point
  - The limit policy of the agent is induced by his limit belief (the limit estimate $\widehat{\alpha}$)
  - The limit belief is confirmed by the data generated from the limit policy
- Self-fulfilling bias:

$$\frac{\mathrm{Cov}[v_t, \epsilon_t | v_t > c/\widehat{\alpha}]}{\mathrm{Var}[v_t | v_t > c/\widehat{\alpha}]} - \underbrace{\frac{\mathrm{Cov}[v_t, \epsilon_t]}{\mathrm{Var}[v_t]}}_{\text{OLS bias}}$$

- $\widehat{\alpha}$ may have multiple values
- May also exist for non-greedy policies

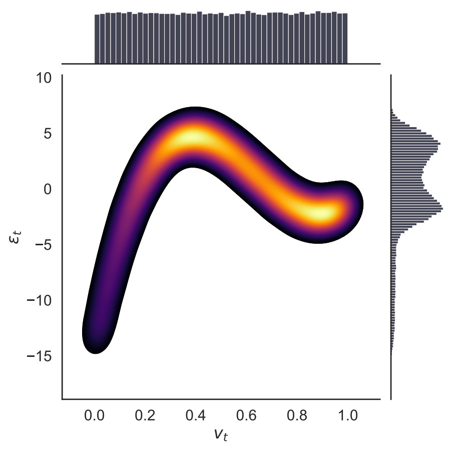


Figure 1: Joint distribution of (v_t, ϵ_t)

Multiplicity of Self-fulfilling Bias with Greedy Policy

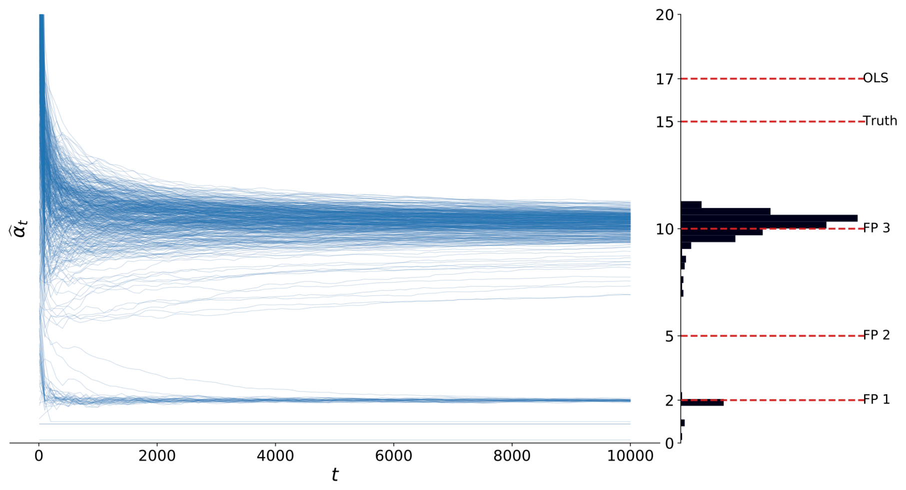


Figure 2: Multiplicity of Self-fulfilling Bias

Endogeneity Spillover

- v_t exogenous

$$R_t = c\mathbb{I}(a_t = 1) + \alpha v_t \mathbb{I}(a_t = 2) + \epsilon_t$$

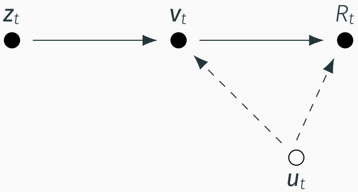
- v_t endogenous + Greedy/UCB

$$R_t = c\mathbb{I}(a_t = 1) + \alpha v_t \mathbb{I}(a_t = 2) + \epsilon_t$$

- v_t endogenous + random arm choice (ex-post randomization)

$$R_t = c\mathbb{I}(a_t = 1) + \alpha v_t \mathbb{I}(a_t = 2) + \epsilon_t$$

Instrumental Variables as Ex-ante Randomization



- $\boldsymbol{v}_t$: visitor characteristics
- ϵ_t and $\boldsymbol{v}_t$ are both positively affected by unobserved consumer sentiment ($\boldsymbol{u}_t$)
- $\boldsymbol{z}_t$ is correlated with $\boldsymbol{v}_t$, but $\boldsymbol{z}_t \perp \epsilon_t$ given $(\boldsymbol{v}_t, a_t)$
  - $\boldsymbol{z}_t$ should affect website traffic without affecting consumer sentiment
  - Google search ads ranking or promotional activities
- IVs + ex-post randomization

$$R_t = c\mathbb{I}(a_t = 1) + \alpha v_t \mathbb{I}(a_t = 2) + \epsilon_t$$

IV-Greedy Algorithm

- Phase 1 ($t = 1, \ldots, T_1$)
  - Take a random action $a_t = 1, 2, \ldots, M$ with equal probability
  - At $t = T_1$, run arm-specific-2SLS to obtain an estimate of $\boldsymbol{\alpha}$
- Phase 2 ($t = T_1 + 1, \ldots, T_2$)
  - Take a greedy action without parameter updates: $a_t = \mathbf{argmax}_i \left\{ \mathbf{v}_t^\mathsf{T} \widehat{\boldsymbol{\alpha}}_{i,T_1} \right\}$
- Phase 3 ($t = T_2 + 1, \ldots, T$)
  - Take a greedy action with parameter updates: $a_t = \mathbf{argmax}_i \left\{ \mathbf{v}_t^\mathsf{T} \widehat{\boldsymbol{\alpha}}_{i,t-1} \right\}$
  - At each $t = T_2 + 1, \ldots, T$, run joint-2SLS on data collected from T_1 to t to update estimates of $\boldsymbol{\Omega}^*$ and $\boldsymbol{\alpha}$:

$$R_s = \sum_{i=1}^{M} \mathbb{I}(a_s = i) \mathbf{v}_s^\mathsf{T} \boldsymbol{\alpha}_i + \epsilon_s, \quad s = T_1 + 1, \ldots, t$$

Theorem

Let $T_1 = C_1 \log(T)$ and $T_2 = (C_1 + C_2) \log(T)$ for some sufficiently large constants C_1 and C_2. Then, the IV-Greedy Algorithm satisfies

- $\sqrt{T - T_1}(\widehat{\boldsymbol{\alpha}}_T - \boldsymbol{\alpha}) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{\Omega}^*\right)$
- *Regret* = $\mathcal{O}(\log(T))$

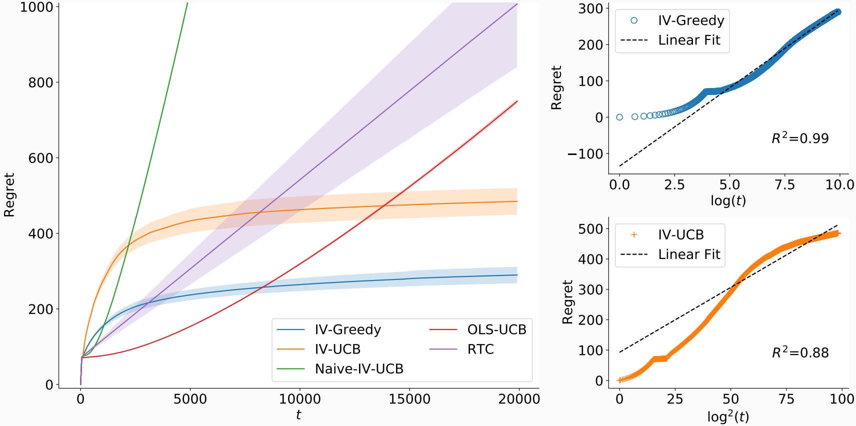


Figure 3: Regret

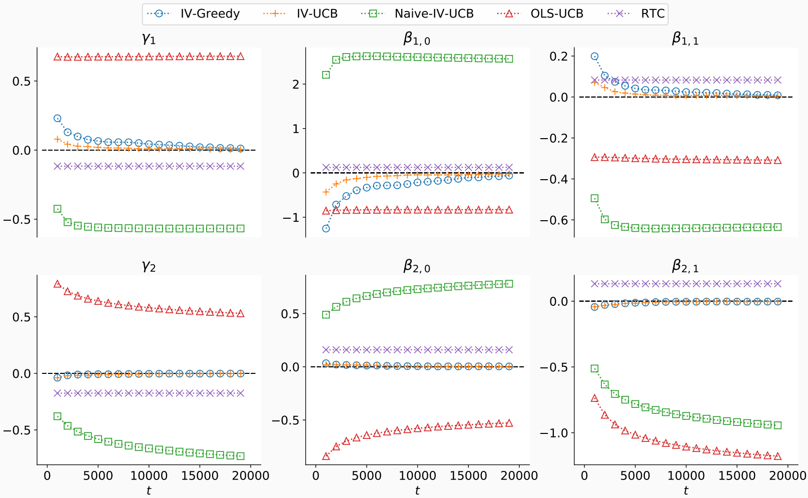


Figure 4: Bias

Conclusions

- In online algorithms, endogeneity in data spills over to actions, resulting in self-fulfilling bias
- IV-based algorithms
- Ex-ante randomization (IV) v.s. ex-post randomization ("exploration")
- Preprint available at *`https://ssrn.com/abstract=3912989`*