Adaptive Correspondence Experiments

Hadar Avivi,[1] Patrick Kline,[1] Evan Rose[2] and Christopher Walters[1]

[1]UC Berkeley

[2]Microsoft Research

January 5, 2021

Motivation

- There is a growing interest in adopting algorithmic predictions to advise decision making
  - This talk - detection of discriminatory jobs
  - Potential tool for regulators such as the Equal Employment Opportunity Commission (EEOC) which are charged with preventing and remedying discrimination by individual employers
- Kline and Walters (forthcoming) show that correspondence experiments sending multiple applications to each job can be used to detect discrimination by individual employers
  - Correspondence experiments can be seen as ensembles of mini-experiments
  - Using these ensembles, we can learn the distribution of discrimination across jobs, and use Empirical Bayes (EB) methods to predict the probability a job is discriminating
  - Only few apps are required because discriminatory behavior is highly variable across jobs

Motivation

- There is a growing interest in adopting algorithmic predictions to advise decision making
  - This talk - detection of discriminatory jobs
  - Potential tool for regulators such as the Equal Employment Opportunity Commission (EEOC) which are charged with preventing and remedying discrimination by individual employers
- Kline and Walters (forthcoming) show that correspondence experiments sending multiple applications to each job can be used to detect discrimination by individual employers
  - Correspondence experiments can be seen as ensembles of mini-experiments
  - Using these ensembles, we can learn the distribution of discrimination across jobs, and use Empirical Bayes (EB) methods to predict the probability a job is discriminating
  - Only few apps are required because discriminatory behavior is highly variable across jobs

Motivation - cont

- Obstacle: these experiments are costly
  - Typically send a fixed number of apps per job
  - More apps increase likelihood of detection
  - Some jobs have a very low callback probability
- **Potential solution:** adaptive correspondence experiments
  - Similar to dynamic treatment regime to patients in the medical sciences Chakraborty and Murphy (2014)
  - Inspired by research in econometrics that update estimators, decision rules, and experimental designs in response to realized data Kasy and Sautmann (forthcoming); Tabord-Meehan (2020)
- Adaptive methods can be useful in other domains where discrimination is a concern, such as healthcare (Alsan et al., 2019; Obermeyer et al., 2019) and criminal justice (Arnold et al., 2020; Rose, forthcoming)

Motivation - cont

- Obstacle: these experiments are costly
  - Typically send a fixed number of apps per job
  - More apps increase likelihood of detection
  - Some jobs have a very low callback probability
- **Potential solution:** adaptive correspondence experiments
  - Similar to dynamic treatment regime to patients in the medical sciences Chakraborty and Murphy (2014)
  - Inspired by research in econometrics that update estimators, decision rules, and experimental designs in response to realized data Kasy and Sautmann (forthcoming); Tabord-Meehan (2020)
- Adaptive methods can be useful in other domains where discrimination is a concern, such as healthcare (Alsan et al., 2019; Obermeyer et al., 2019) and criminal justice (Arnold et al., 2020; Rose, forthcoming)

Motivation - cont

- Obstacle: these experiments are costly
  - Typically send a fixed number of apps per job
  - More apps increase likelihood of detection
  - Some jobs have a very low callback probability
- **Potential solution:** adaptive correspondence experiments
  - Similar to dynamic treatment regime to patients in the medical sciences Chakraborty and Murphy (2014)
  - Inspired by research in econometrics that update estimators, decision rules, and experimental designs in response to realized data Kasy and Sautmann (forthcoming); Tabord-Meehan (2020)
- Adaptive methods can be useful in other domains where discrimination is a concern, such as healthcare (Alsan et al., 2019; Obermeyer et al., 2019) and criminal justice (Arnold et al., 2020; Rose, forthcoming)

This paper

- Consider a hypothetical regulator seeking to detect discriminatory jobs (e.g. the EEOC who is charge of enforcing anti-discrimination law)
- The auditor draws new vacancies from a known distribution and sends fictitious applications in attempt to infer the job's type
- Unlike a static audit experiment, at each step the auditor can decide whether to keep sending applications, initiate an investigation, or give up
- **Key result:** # of apps are cut by more than half without reducing accuracy of detection
  - Giving up early on jobs with very low callback rates, or those that call black applicants
  - Choosing application characteristics optimally

This paper

- Consider a hypothetical regulator seeking to detect discriminatory jobs (e.g. the EEOC who is charge of enforcing anti-discrimination law)
- The auditor draws new vacancies from a known distribution and sends fictitious applications in attempt to infer the job's type
- Unlike a static audit experiment, at each step the auditor can decide whether to keep sending applications, initiate an investigation, or give up
- **Key result:** # of apps are cut by more than half without reducing accuracy of detection
  - Giving up early on jobs with very low callback rates, or those that call black applicants
  - Choosing application characteristics optimally

Model

A model for callbacks

Following Kline and Walters (forthcoming):

- Callbacks are modeled as *iid* Bernoulli trials
- Callback probability of job j to applications of race $r \in \{b, w\}$ with characteristics x:

$$p_{jr}(x) = \Lambda(\alpha_j - \beta_j \mathbb{1}\{r = b\} + x'\gamma),$$

  where $\Lambda(z) \equiv [1 + exp(-z)]^{-1}$.

- (α_j, β_j) are random coefficients: $\beta_j = max\{0, \tilde{\beta}_j\}$, with

$$\begin{pmatrix} \alpha_j \\ \tilde{\beta}_j \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \rho \\ \rho & \sigma_\beta^2 \end{bmatrix} \right)$$

- Model allows for continuous heterogeneity in callback rates and discrimination severity, and a mass point at $\beta_j = 0$

Fitting the model - Nunley et al. (2015) data

- We estimate the model using data from Nunley et al. (2015)'s (NPRS) audit experiment
- The NPRS experiment submitted fictitious applications with racially distinctive names to 2,305 entry-level jobs for college graduates in the US
- 4 applications per job, typically 2 white and 2 black
- View this as a pilot study, e.g. commissioned by the EEOC

Maximum Simulated Likelihood estimates

	(1)	(2)
α_0	-4.922	-4.918
	(0.234)	(0.234)
σ_α	4.968	4.963
	(0.240)	(0.240)
β_0	-5.035	-5.022
	(0.176)	(0.329)
σ_β	6.347	6.521
	(0.148)	(0.154)
ρ		-0.013
		(0.017)
Likelihood	-2788.3	-2788.3
Number of jobs	2305	2305

No correlation between white CB and discrimination severity

	(1)	(2)
α_0	-4.922	-4.918
	(0.234)	(0.234)
σ_α	4.968	4.963
	(0.240)	(0.240)
β_0	-5.035	-5.022
	(0.176)	(0.329)
σ_β	6.347	6.521
	(0.148)	(0.154)
ρ		-0.013
		(0.017)
Likelihood	-2788.3	-2788.3
Number of jobs	2305	2305

Most jobs don't call anyone

	(1)	(2)
α_0	-4.922	-4.918
	(0.234)	(0.234)
σ_α	4.968	4.963
	(0.240)	(0.240)
β_0	-5.035	-5.022
	(0.176)	(0.329)
σ_β	6.347	6.521
	(0.148)	(0.154)
ρ		-0.013
		(0.017)
Likelihood	-2788.3	-2788.3
Number of jobs	2305	2305

$Pr(p_{jw} < 0.01) \approx 0.53$

Severe discrimination among a minority of jobs

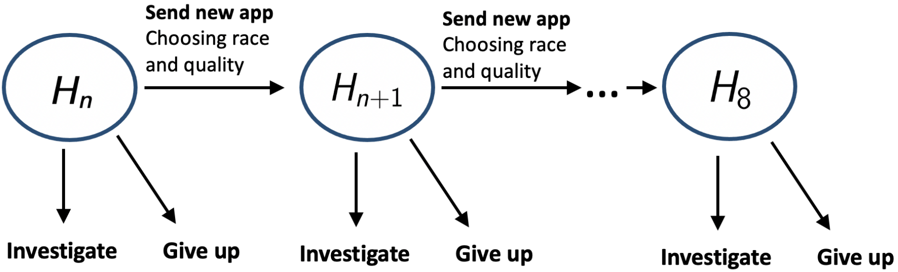
	(1)	(2)
α_0	-4.922	-4.918
	(0.234)	(0.234)
σ_α	4.968	4.963
	(0.240)	(0.240)
β_0	-5.035	-5.022
	(0.176)	(0.329)
σ_β	6.347	6.521
	(0.148)	(0.154)
ρ		-0.013
		(0.017)
Likelihood	-2788.3	-2788.3
Number of jobs	2305	2305

$Pr(\beta_j = 0) \approx 0.79, \quad E[\beta_j|\beta_j > 0] \approx 3.6$

The auditor's problem

- Consider an auditor that knows the parameters of the model
- The auditor's goal is to find discriminators by sending additional fictitious apps
- Can send up to 8 apps per job
- Simplify to two quality levels $q \in \{h, l\}$, corresponding to $x'\gamma$ one SD above and below its mean
- At every step, based on the observed callbacks, the auditor can decide to send another application, initiate an investigation, or give up

The auditor's problem



- H_n is the auditing history after sending n apps. Includes counts of apps and callbacks by race and quality
- For example: $H_4 = \begin{cases} \text{sent:} & (W_l, B_l, W_h, B_h) = (1,0,2,1) \\ \text{CB:} & (W_l, B_l, W_h, B_h) = (0,0,2,0) \end{cases}$

The auditor's payoff

- Once an investigation is initiated, the job's true type is revealed, yielding payoff:

$$\underbrace{\frac{1}{2} \sum_{q \in \{h,l\}} [p_{jw}(q) - p_{jb}(q)]}_{\equiv S_j} - \kappa,$$

  where S_j is the severity of discrimination, κ is the cost of investigation, and $q \in \{h, l\}$ indexes quality

- The auditor cares about the expected number of black callbacks lost relative to white applicants

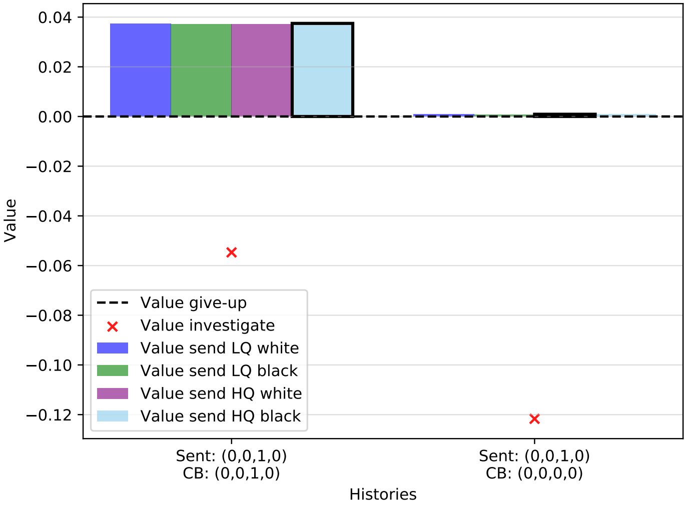
The auditor's value function

$$
V(H_n) = \begin{cases} \max \left\{ \underbrace{\max_{r,q} v_{rq}\left(H_n\right)}_{\text{send new app}}, \underbrace{v_I\left(H_n\right)}_{\text{investigate}}, 0 \right\} & \text{if } n < 8, \\ \max \left\{ \underbrace{v_I\left(H_n\right)}_{\text{investigate}}, 0 \right\} & \text{if } n = 8. \end{cases}
$$

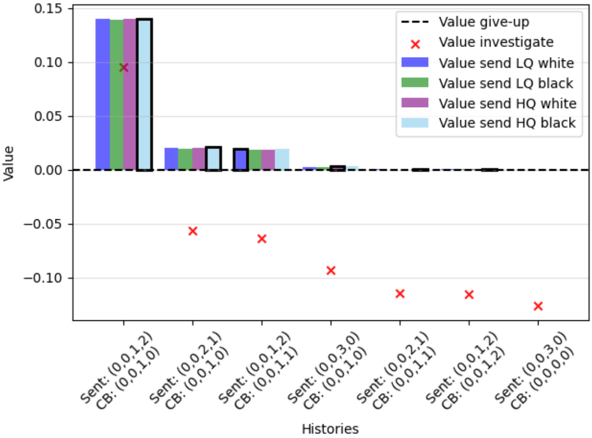
- Value of sending new app: $v_{rq}(H_n) = -c + \mathbb{E}[V(H_{n+1})|H_n]$
- Value of investigation: $v_I(H_n) = \mathbb{E}\left[S_j \middle| H_n\right] - \kappa$
- Expectations are evaluated via Bayes' rule starting with the population distribution as prior

Simulation Results

Expected value and optimal strategy after sending one application $(\kappa = .13, c = 10^{-4})$ more

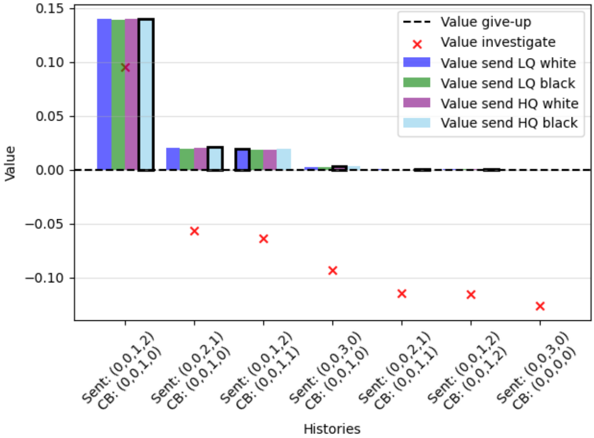


Expected value and optimal strategy after sending three applications $(\kappa = .13, c = 10^{-4})$ more



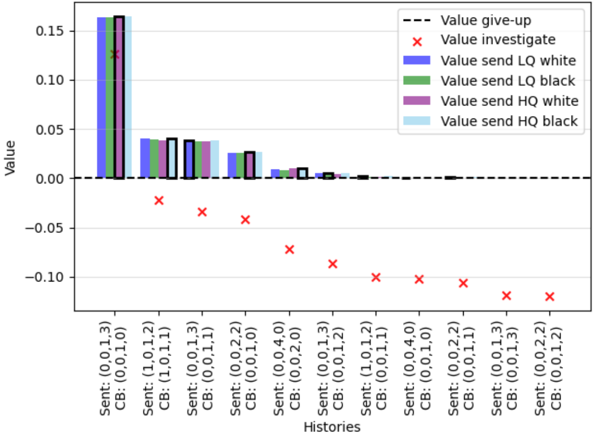
$\sim 72\%$ of jobs w/ history $(0, 0, 3, 0)$ and no CBs. If # of jobs $= 100$, then the auditor saves $0.72 \times 5 \times 100 = 360$ apps on average

Expected value and optimal strategy after sending three applications $(\kappa = .13, c = 10^{-4})$ more



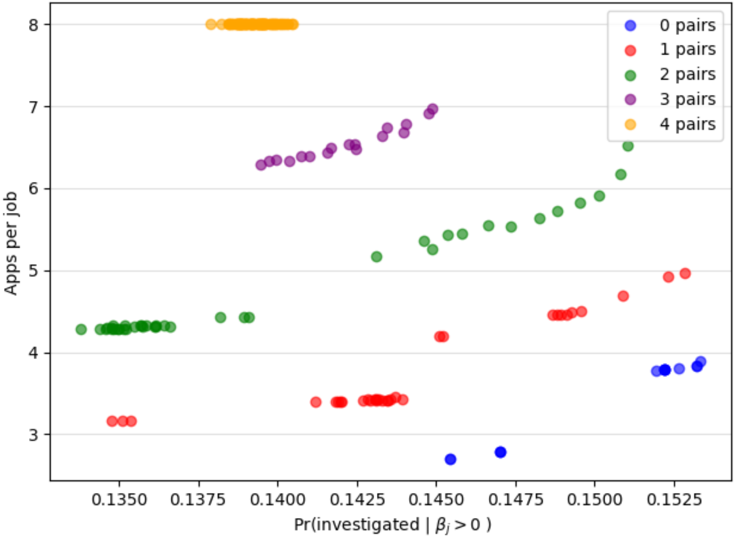
$\sim$ 72% of jobs w/ history $(0, 0, 3, 0)$ and no CBs. If # of jobs $= 100$, then the auditor saves $0.72 \times 5 \times 100 = 360$ apps on average

Expected value and optimal strategy after sending four applications $(\kappa = .13, c = 10^{-4})$ more

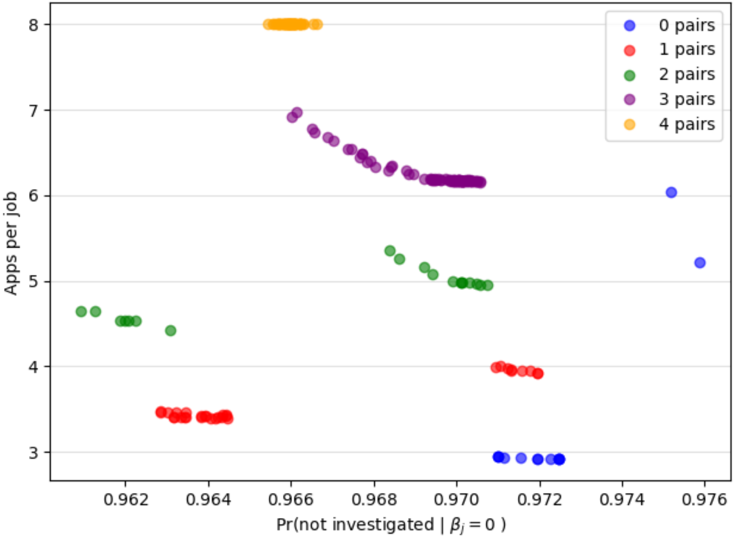


$\sim$ 12% of jobs w/ the two last histories. If # of jobs $= 100$, then the auditor saves $0.12 \times 4 \times 100 = 48$ apps on average

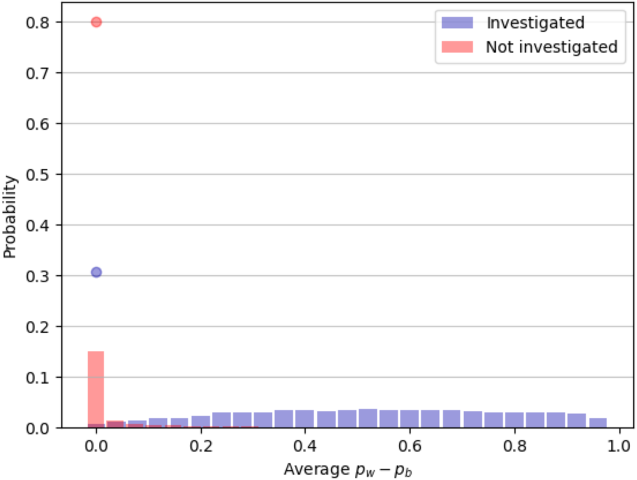
Apps sent vs. sensitivity Investigation probability fixed $\in [.055, 0.06]$



Apps sent vs. specificity sensitivity fixed $\in [.14, .145]$



Adaptive auditing catches the worst discriminators



$\kappa = .13, c = 10^{-4}$

Discussion

- Adaptive correspondence experiments have the potential to detect discrimination more efficiently than static experiments
  - Substantial reduction in the number of apps sent
  - Achieve the same levels of sensitivity and specificity
- These methods can contribute to other settings (e.g criminal justice, healthcare, policing and education) to detect discrimination efficiently
- Potential drawbacks:
  - Requires full knowledge of the distribution of callbacks (pilot study)
  - Assumes stable callback parameters
  - Dynamic programming is computationally expensive, especially as the dimension of the action space grows
- Potential extensions based on reinforcement learning e.g, Kasy and Sautmann (forthcoming)

Discussion

- Adaptive correspondence experiments have the potential to detect discrimination more efficiently than static experiments
  - Substantial reduction in the number of apps sent
  - Achieve the same levels of sensitivity and specificity
- These methods can contribute to other settings (e.g criminal justice, healthcare, policing and education) to detect discrimination efficiently
- Potential drawbacks:
  - Requires full knowledge of the distribution of callbacks (pilot study)
  - Assumes stable callback parameters
  - Dynamic programming is computationally expensive, especially as the dimension of the action space grows
- Potential extensions based on reinforcement learning e.g, Kasy and Sautmann (forthcoming)

Discussion

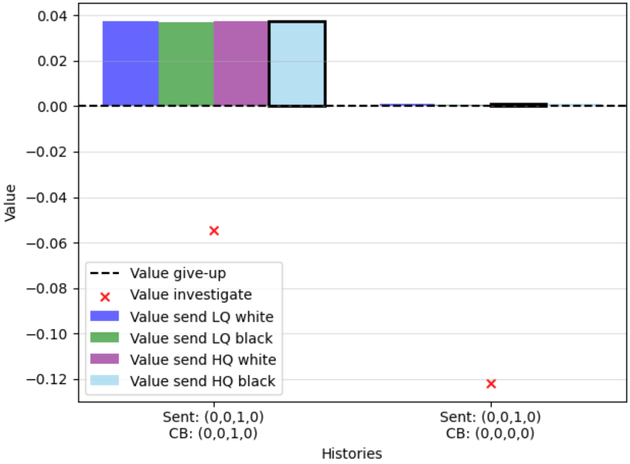
- Adaptive correspondence experiments have the potential to detect discrimination more efficiently than static experiments
  - Substantial reduction in the number of apps sent
  - Achieve the same levels of sensitivity and specificity
- These methods can contribute to other settings (e.g criminal justice, healthcare, policing and education) to detect discrimination efficiently
- Potential drawbacks:
  - Requires full knowledge of the distribution of callbacks (pilot study)
  - Assumes stable callback parameters
  - Dynamic programming is computationally expensive, especially as the dimension of the action space grows
- Potential extensions based on reinforcement learning e.g, Kasy and Sautmann (forthcoming)

Discussion

- Adaptive correspondence experiments have the potential to detect discrimination more efficiently than static experiments
  - Substantial reduction in the number of apps sent
  - Achieve the same levels of sensitivity and specificity
- These methods can contribute to other settings (e.g criminal justice, healthcare, policing and education) to detect discrimination efficiently
- Potential drawbacks:
  - Requires full knowledge of the distribution of callbacks (pilot study)
  - Assumes stable callback parameters
  - Dynamic programming is computationally expensive, especially as the dimension of the action space grows
- Potential extensions based on reinforcement learning e.g, Kasy and Sautmann (forthcoming)

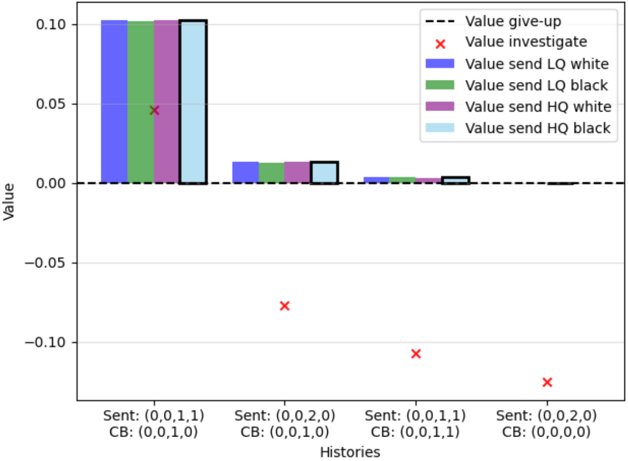
Thank You!

Expected value and optimal strategy after sending one app ($\kappa = .13, c = 10^{-4}$)



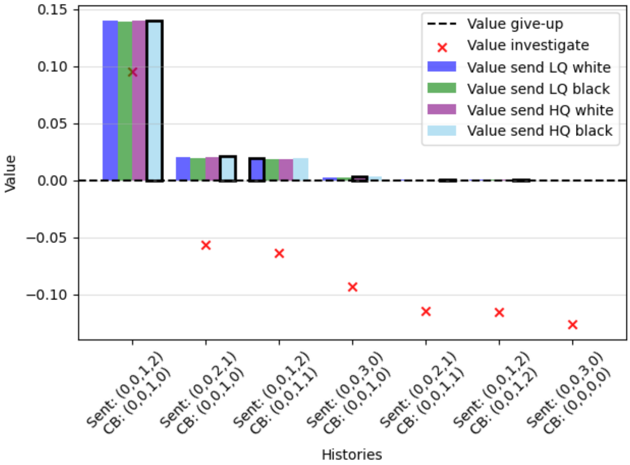
back

Expected value and optimal strategy after sending two apps ($\kappa = .13, c = 10^{-4}$)



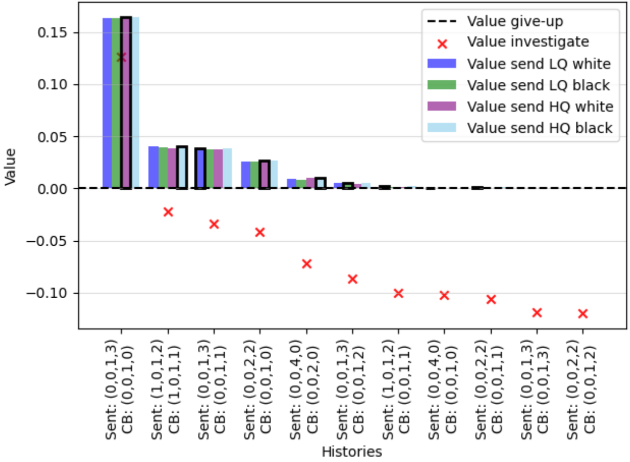
back

Expected value and optimal strategy after sending three apps ($\kappa = .13, c = 10^{-4}$)



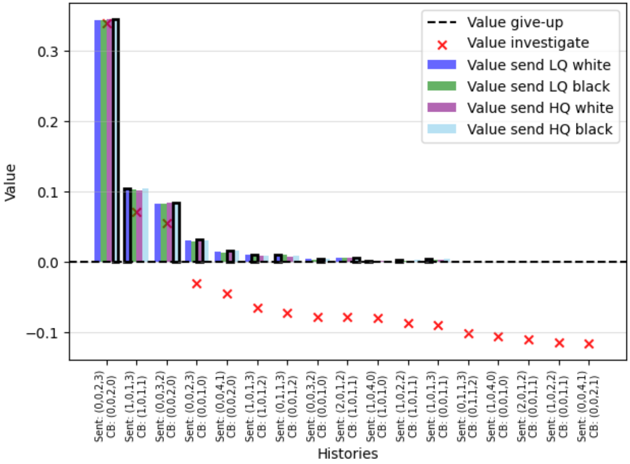
back

Expected value and optimal strategy after sending four apps ($\kappa = .13, c = 10^{-4}$)



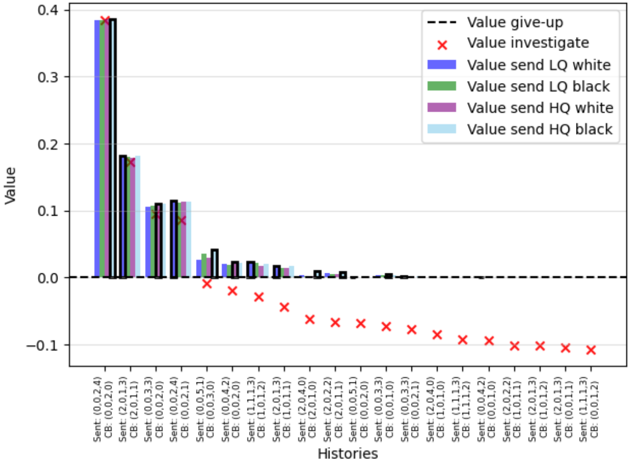
back

Expected value and optimal strategy after sending five apps ($\kappa = .13, c = 10^{-4}$)



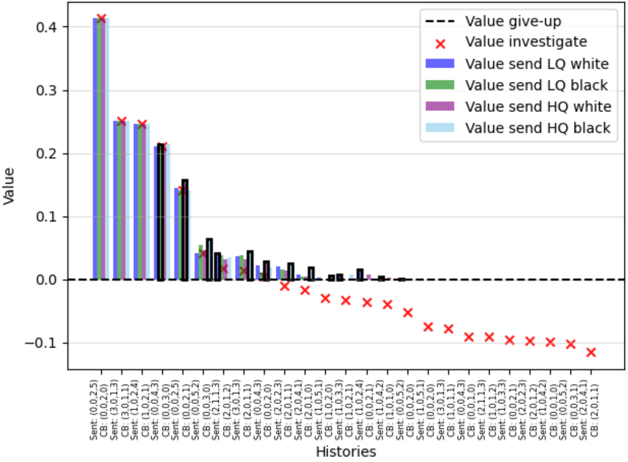
◂ back

Expected value and optimal strategy after sending six apps ($\kappa = .13, c = 10^{-4}$)



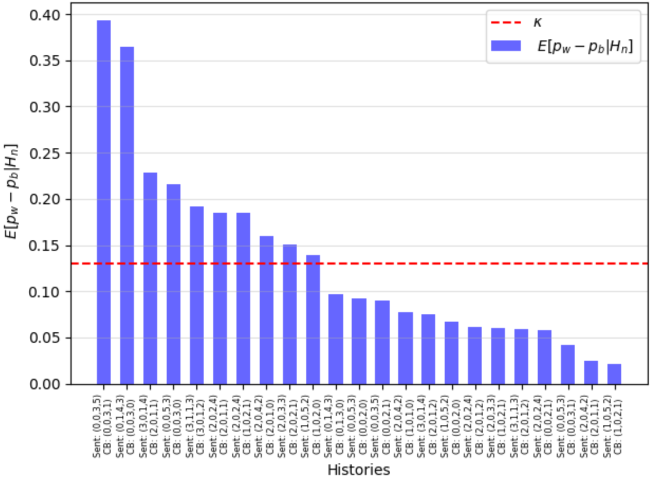
back

Expected value and optimal strategy after sending seven apps ($\kappa = .13, c = 10^{-4}$)



back

Expected value after sending eight apps ($\kappa = .13, c = 10^{-4}$)



◂ back

References I

Alsan, M., Garrick, O., and Graziani, G. (2019). Does diversity matter for health? experimental evidence from oakland. *American Economic Review*, 109(12):4071–4111.

Arnold, D., Dobbie, W. S., and Hull, P. (2020). Measuring racial discrimination in bail decisions. Technical report, National Bureau of Economic Research.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464.

Kasy, M. and Sautmann, A. (forthcoming). Adaptive treatment assignment in experiments for policy choice. *Econometrica*.

Kline, P. M. and Walters, C. R. (forthcoming). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica*.

Nunley, J. M., Pugh, A., Romero, N., and Seals, R. A. (2015). Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment. *The BE Journal of Economic Analysis & Policy*, 15(3):1093–1125.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

References II

Rose, E. K. (forthcoming). Who gets a second chance? effectiveness and equity in supervision of criminal offenders. Technical report.

Tabord-Meehan, M. (2020). Stratification trees for adaptive randomization in randomized controlled trials.