Learning Optimal Fair Policies

Razieh Nabi

Computer Science Department

Johns Hopkins University

rnabi@jhu.edu

Joint work with **Daniel Malinsky** and **Ilya Shpitser**

ICML-19: Thirty-sixth International Conference on Machine Learning

Fairness in automated decision-making

- Algorithms are increasingly prevalent in socially-impactful settings
  - Criminal justice, welfare policy, hiring, personal finance

- Data includes potentially sensitive attributes (and/or proxies)
  - Risk of perpetuating injustice:
    naively maximizing utility may maintain, reinforce, or even introduce unfair dependence between sensitive features, decisions, and outcomes

- Automated decisions should respect principles of fairness

Fairness in automated decision-making

- Algorithms are increasingly prevalent in socially-impactful settings
  - Criminal justice, welfare policy, hiring, personal finance
- Data includes potentially sensitive attributes (and/or proxies)
  - Risk of perpetuating injustice:
    naively maximizing utility may maintain, reinforce, or even introduce unfair dependence between sensitive features, decisions, and outcomes
- Automated decisions should respect principles of fairness

Fairness in automated decision-making

- Algorithms are increasingly prevalent in socially-impactful settings
  - Criminal justice, welfare policy, hiring, personal finance
- Data includes potentially sensitive attributes (and/or proxies)
  - Risk of perpetuating injustice:
    naively maximizing utility may maintain, reinforce, or even introduce unfair dependence between sensitive features, decisions, and outcomes
- Automated decisions should respect principles of fairness

Fairness in automated decision-making

- Algorithms are increasingly prevalent in socially-impactful settings
  - Criminal justice, welfare policy, hiring, personal finance

- Data includes potentially sensitive attributes (and/or proxies)
  - Risk of perpetuating injustice:
    naively maximizing utility may maintain, reinforce, or even introduce unfair dependence between sensitive features, decisions, and outcomes

- Automated decisions should respect principles of fairness

Objective

Want to make **optimal but fair** decisions, which “break the cycle of injustice" by correcting for the unfair dependence of both decisions and outcomes on sensitive features.

Child welfare example

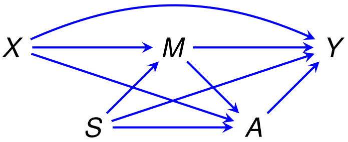
X: baseline factors

S: sensitive features

M: mediators

A: action/decision

Y: outcome/utility



- Child welfare hotline: decision A to dispatch case-worker may depend on all available information, and optimal decision would minimize negative outcomes (e.g. child separation and/or hospitalization).
- Concern that unconstrained optimal decision-making may lead to unacceptable racial disparities.
  - Ignoring race information is insufficient: dependence due to proxies

Child welfare example

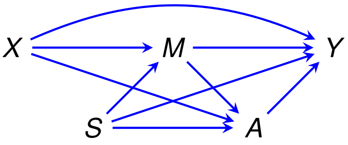
X: baseline factors

S: sensitive features

M: mediators

A: action/decision

Y: outcome/utility



- Child welfare hotline: decision A to dispatch case-worker may depend on all available information, and optimal decision would minimize negative outcomes (e.g. child separation and/or hospitalization).
- Concern that unconstrained optimal decision-making may lead to unacceptable racial disparities.
  - Ignoring race information is insufficient: dependence due to proxies

Impermissible causal paths

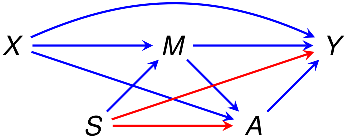
X: baseline factors

S: sensitive features

M: mediators

A: action/decision

Y: outcome/utility



Our perspective:

- In a “fairer world,” certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects; generalizing a view in (Nabi and Shpitser, 2018).
- Try to approximate the “nearest fair world” and learn optimal policies there.
- Must sacrifice some optimality to make decisions fairly.

Impermissible causal paths

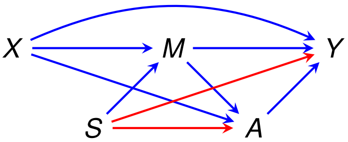
X: baseline factors

S: sensitive features

M: mediators

A: action/decision

Y: outcome/utility



Our perspective:

- In a "fairer world," certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects; generalizing a view in (Nabi and Shpitser, 2018).
- Try to approximate the "nearest fair world" and learn optimal policies there.
- Must sacrifice some optimality to make decisions fairly.

Impermissible causal paths

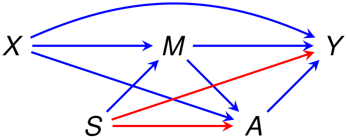
X: baseline factors

S: sensitive features

M: mediators

A: action/decision

Y: outcome/utility



Our perspective:

- In a "fairer world," certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects; generalizing a view in (Nabi and Shpitser, 2018).
- Try to approximate the "nearest fair world" and learn optimal policies there.
- Must sacrifice some optimality to make decisions fairly.

Impermissible causal paths

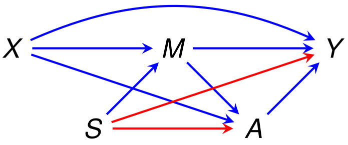
X: baseline factors

S: sensitive features

M: mediators

A: action/decision

Y: outcome/utility



Our perspective:

- In a “fairer world,” certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects; generalizing a view in (Nabi and Shpitser, 2018).
- Try to approximate the “nearest fair world” and learn optimal policies there.
- Must sacrifice some optimality to make decisions fairly.

Outline

Causal inference: preliminaries

- Data $\mathcal{D} \sim p(X, A, Y)$
- $Y(a)$: outcome Y had A been assigned to a
- Average causal effect: $\text{ACE} = \mathbb{E}[Y(A=1)] - \mathbb{E}[Y(A=0)]$
  - Randomized experiments: compare cases, $(A=1)$ and controls $(A=0)$
  - Observational data: people choose to smoke
- Identifiability under standard assumptions:
  - Consistency: $Y(A) = Y$,
  - Ignorability: $Y(a) \perp\!\!\!\perp A \mid X, \forall a$,
  - Positivity: $p(a|X) > 0, \forall a$.

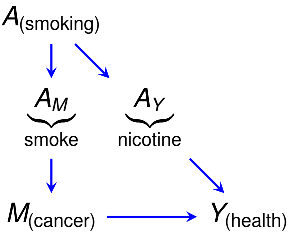
$$\text{ACE} = \sum_X \Big\{ \mathbb{E}[Y \mid A=1, X] - \mathbb{E}[Y \mid A=0, X] \Big\}\, p(X)$$

Mediation analysis: preliminaries

- Causal mechanisms: how does A cause Y?
- ACE = Direct effect ($A \rightarrow Y$) + Indirect effect ($A \rightarrow M \rightarrow Y$)
  - $\mathcal{D} = \{\mathbf{X}, A, M, Y\}$. M mediates the effect of A on Y
- Nested counterfactuals $Y(a, M(a'))$
  - Outcome Y had A been assigned to a and M been assigned to whatever value it would have had under a'

Mediation analysis: preliminaries

- Causal mechanisms: how does A cause Y?
- ACE = Direct effect ($A \rightarrow Y$) + Indirect effect ($A \rightarrow M \rightarrow Y$)
  - $\mathcal{D} = \{\mathbf{X}, A, M, Y\}$. M mediates the effect of A on Y
- Nested counterfactuals $Y(a, M(a'))$
  - Outcome Y had A been assigned to a and M been assigned to whatever value it would have had under a'



	a_Y **nicotine**	a_M **smoke**	potential outcome
$Y(1, M(0))$	1	0	nicotine patch

- **Direct Effect** $= \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0)]$ $(A \rightarrow Y)$
- **Indirect Effect** $= \mathbb{E}[Y(1)] - \mathbb{E}[Y(1, M(0))]$ $(A \rightarrow M \rightarrow Y)$

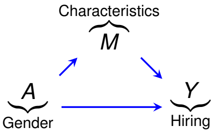
Unfair path-specific effects

Example: name-swapping experiments to evaluate gender bias in hiring

$Y(a')$: $Y(A = male, M(A = male)) = Y(A = male)$,

$Y(a, M(a'))$: $Y(A = female, M(A = male))$.

Direct effect: $\mathbb{E}[Y(a, M(a')) - \mathbb{E}[Y(a')]]$



- **Path-specific effects (PSEs)**
  - Along a set of paths, all nodes behave as if $A = a$,
  - Along all other paths, nodes behave as if $A = a'$.
- Some PSEs should be considered impermissible, depending on context.
  - Policymakers, (bio)ethicists, general public should determine which mechanisms are problematic in applied settings.

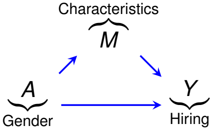
Unfair path-specific effects

Example: name-swapping experiments to evaluate gender bias in hiring

$Y(a')$: $Y(A = male, M(A = male)) = Y(A = male)$,

$Y(a, M(a'))$: $Y(A = female, M(A = male))$.

Direct effect: $\mathbb{E}[Y(a, M(a')) - \mathbb{E}[Y(a')]]$



- **Path-specific effects (PSEs)**
  - Along a set of paths, all nodes behave as if $A = a$,
  - Along all other paths, nodes behave as if $A = a'$.
- Some PSEs should be considered impermissible, depending on context.
  - Policymakers, (bio)ethicists, general public should determine which mechanisms are problematic in applied settings.

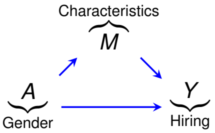
Unfair path-specific effects

Example: name-swapping experiments to evaluate gender bias in hiring

$Y(a')$: $Y(A = male, M(A = male)) = Y(A = male)$,

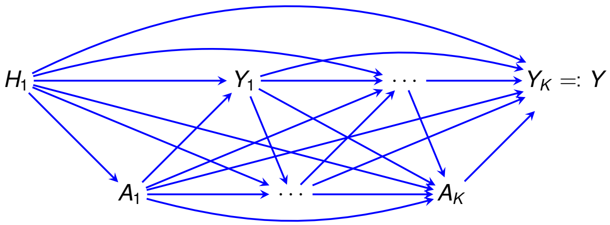
$Y(a, M(a'))$: $Y(A = female, M(A = male))$.

Direct effect: $\mathbb{E}[Y(a, M(a')) - \mathbb{E}[Y(a')]]$



- **Path-specific effects (PSEs)**
  - Along a set of paths, all nodes behave as if $A = a$,
  - Along all other paths, nodes behave as if $A = a'$.
- Some PSEs should be considered impermissible, depending on context.
  - Policymakers, (bio)ethicists, general public should determine which mechanisms are problematic in applied settings.

Sequential decision-making



- K decision points.
  - H_1: available information prior to the first decision A_1.
  - Y_k: intermediate outcome between k^{th} and $(k+1)^{th}$ decision points.
  - H_k: history prior to kth decision point.
- Policy $f_A = \{f_{A_1}, \ldots, f_{A_K}\}$ where $f_{A_k} : \mathcal{H}_k \mapsto \mathcal{A}_k, \forall k = 1, \ldots, K$.

Optimal policy in causal inference

- Counterfactual response under f_A is denoted by $Y(f_A(H))$
- **Optimal policy**:

$$f_A^* := \arg\max_{f_A} \mathbb{E}[Y(f_A)]$$

- Under certain **identification** assumptions (Shpitser, 2013):
- Fairness concerns arise since $H_1 = \{X, S, M\}$

Optimal policy in causal inference

- Counterfactual response under f_A is denoted by $Y(f_A(H))$
- **Optimal policy**:

$$f_A^* := \arg\max_{f_A} \mathbb{E}[Y(f_A)]$$

- Under certain **identification** assumptions (Shpitser, 2013):

$$\mathbb{E}[Y(f_A)] = \sum_{h_1, y_1} p\big(Y|h_1, a_1 = f_{A_1}, y_1, a_2 = f_{A_2}\big)\ p(y_1|a_1 = f_{A_1}, h_1)\ p(h_1).$$

- Fairness concerns arise since $H_1 = \{X, S, M\}$

Optimal policy in causal inference

- Counterfactual response under f_A is denoted by $Y(f_A(H))$
- **Optimal policy**:

$$f_A^* := \arg\max_{f_A} \mathbb{E}[Y(f_A)]$$

- Under certain **identification** assumptions (Shpitser, 2013):

$$\mathbb{E}[Y(f_A)] = \sum_{h_1, y_1} p\big(Y | h_1, a_1 = f_{A_1}, y_1, a_2 = f_{A_2}\big)\ p(y_1 | a_1 = f_{A_1}, h_1)\ p(h_1).$$

- Fairness concerns arise since $H_1 = \{X, S, M\}$

Sources of bias

Consider the following effects impermissible:

- The effect of S on Y *not through M*

$$\begin{aligned} \text{PSE}^{sy} &= \mathbb{E}[Y(s, M(s')) - Y(s', M(s'))] \\ &\equiv g_1(Z) \end{aligned}$$

- The effect of S on A_k *not through M*

$$\begin{aligned} \text{PSE}^{sa_k} &= \mathbb{E}[A_k(s, M(s')) - A_k(s', M(s'))] \\ &\equiv g_k(Z) \end{aligned}$$

$$Z = \{X, S, M, A_1, ..., A_K, Y_1, ..., Y_K\}$$

The closest "fair world"

Given *a set of identified PSEs* $g_j(Z)\ \forall j \in \{1, ..., J\}$ and *lower/upper tolerance bounds* $\epsilon_j^-, \epsilon_j^+$, the fair distribution $p^*(Z)$ is defined:

$$p^*(Z) \equiv \arg\min_q \ D_{KL}(p||q)$$
$$\text{subject to} \ \ \epsilon_j^- \leq g_j(Z) \leq \epsilon_j^+, \ \ \forall j \in \{1, ..., J\},$$

Approximating the “fair world” in finite samples

Approximate $p^*(Z)$ by solving:

$$\widehat{\alpha} = \arg\max_{\alpha} \quad \mathcal{L}(Z;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \leq \widehat{g}_j(Z) \leq \epsilon_j^+, j = 1, \ldots, J.$$

Consistent estimators of PSE^{sy} and PSE^{sa_k}:

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} Y_n,$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} A_{kn}, \quad k = 1, 2.$$

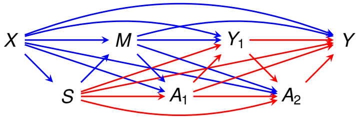
Constraints involve $p(S|X;\alpha_s)$ and $p(M|S,X;\alpha_m)$ models.

Approximating the "fair world" in finite samples

Approximate $p^*(Z)$ by solving:

$$\widehat{\alpha} = \arg\max_{\alpha} \quad \mathcal{L}(Z;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \leq \widehat{g}_j(Z) \leq \epsilon_j^+, j = 1, \ldots, J.$$



Consistent estimators of PSE^{sy} and PSE^{sa_k}:

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} Y_n,$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} A_{kn}, \quad k = 1, 2.$$

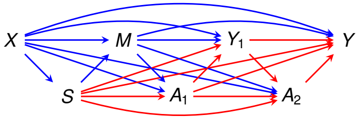
Constraints involve $p(S|X;\alpha_s)$ and $p(M|S,X;\alpha_m)$ models.

Approximating the "fair world" in finite samples

Approximate $p^*(Z)$ by solving:

$$\widehat{\alpha} = \arg\max_{\alpha} \; \mathcal{L}(Z;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \leq \widehat{g}_j(Z) \leq \epsilon_j^+, j = 1, \ldots, J.$$



Consistent estimators of PSE^{sy} and PSE^{sa_k}:

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} Y_n,$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} A_{kn}, \quad k = 1, 2.$$

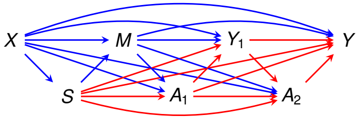
Constraints involve $p(S|X;\alpha_s)$ and $p(M|S,X;\alpha_m)$ models.

Approximating the “fair world” in finite samples

Approximate $p^*(Z)$ by solving:

$$\widehat{\alpha} = \arg\max_{\alpha} \ \mathcal{L}(Z;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \le \widehat{g}_j(Z) \le \epsilon_j^+, j = 1, \ldots, J.$$



Consistent estimators of PSE^{sy} and PSE^{sa_k}:

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} Y_n,$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} A_{kn}, \ \ k = 1, 2.$$

Constraints involve $p(S|X;\alpha_s)$ and $p(M|S, X;\alpha_m)$ models.

Breaking the cycle of injustice

- Let $p^*(M|S,X;\alpha_m)$ and $p^*(S|X;\alpha_s)$ be the constrained models chosen to satisfy $\text{PSE}^{sy} = 0$ and $\text{PSE}^{sa_k} = 0$,
- Let $\widetilde{p}(Z)$ be the joint distribution induced by $p^*(M|S,X;\alpha_m)$ and $p^*(S|X;\alpha_s)$:

$$\widetilde{p}(Z) \equiv p(X) p^*(S|X;\alpha_s)\, p^*(M|S,X;\alpha_m) \prod_{k=1}^{K} p(A_k|H_k) p(Y_k|A_k,H_k).$$

Breaking the cycle of injustice

- Let $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$ be the constrained models chosen to satisfy $\text{PSE}^{sy} = 0$ and $\text{PSE}^{sa_k} = 0$,
- Let $\widetilde{p}(Z)$ be the joint distribution induced by $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$:

$$\widetilde{p}(Z) \equiv p(X) p^*(S|X; \alpha_s)\, p^*(M|S, X; \alpha_m) \prod_{k=1}^{K} p(A_k|H_k) p(Y_k|A_k, H_k).$$

- Then PSE^{sy} and PSE^{sa_k} taken wrt $\widetilde{p}(Z)$ are also zero.

  $\implies$ constraining the S and M models induces a “fair distribution” no matter how A_k or Y_k are determined.

Three strategies for policy estimation

In the paper, we consider three strategies for estimating the optimal policy:

- Q-learning
- Value search
- G-estimation

In each case, we must modify these procedures to operate wrt the fair distribution.

We focus on Q-learning in this talk.

Optimal fair policy: Q-learning

- **Unfair world**: expectations wrt to $p(Z)$

$\mathrm{H_1, A_1}$ $\longleftarrow$ $\mathrm{H_k, A_k}$ $\longleftarrow$ $\mathrm{H_K, A_K}$

$$Q_1(H_1, A_1) = E[V_2(H_2, a_1) | H_1] \quad \ldots \quad Q_k(H_k, A_k) = E[V_{k+1}(H_{k+1}, a_k) | H_k] \quad \ldots \quad Q_K(H_K, A_K) = E[Y(a_K) | H_K]$$

$$V_1(H_1) = \max_{a_1} Q_1(H_1, A_1) \qquad V_k(H_k) = \max_{a_k} Q_k(H_k, A_k) \qquad V(H_K) = \max_{a_K} Q_K(H_K, A_K)$$

- **Optimal policy**: $f^*_{A_k} = \arg\max_{a_k} Q_k(H_k, a_k; \beta_k)$

- **Fair world** expectations wrt to $p^*(Z)$

Optimal fair policy: Q-learning

- **Unfair world**: expectations wrt to $p(Z)$

$\mathrm{H}_1, \mathrm{A}_1 \longleftarrow \mathrm{H}_k, \mathrm{A}_k \longleftarrow \mathrm{H}_K, \mathrm{A}_K$

$$Q_1(H_1, A_1) = E[V_2(H_2, a_1) | H_1] \quad \ldots \quad Q_k(H_k, A_k) = E[V_{k+1}(H_{k+1}, a_k) | H_k] \quad \ldots \quad Q_K(H_K, A_K) = E[Y(a_K) | H_K]$$

$$V_1(H_1) = \max_{a_1} Q_1(H_1, A_1) \qquad V_k(H_k) = \max_{a_k} Q_k(H_k, A_k) \qquad V(H_K) = \max_{a_K} Q_K(H_K, A_K)$$

- **Optimal policy**: $f^*_{A_k} = \arg\max_{a_k} Q_k(H_k, a_k; \beta_k)$

- **Fair world** expectations wrt to $p^*(Z)$

Optimal fair policy: Q-learning

- **Unfair world**: expectations wrt to $p(Z)$

$\text{H}_1, \text{A}_1 \longleftarrow \quad \text{H}_k, \text{A}_k \longleftarrow \quad \text{H}_K, \text{A}_K$

$Q_1(H_1, A_1) = E[V_2(H_2, a_1) | H_1]$ ... $Q_k(H_k, A_k) = E[V_{k+1}(H_{k+1}, a_k) | H_k]$ ... $Q_K(H_K, A_K) = E[Y(a_K) | H_K]$

$V_1(H_1) = \max_{a_1} Q_1(H_1, A_1)$ $\quad V_k(H_k) = \max_{a_k} Q_k(H_k, A_k)$ $\quad V(H_K) = \max_{a_K} Q_K(H_K, A_K)$

- **Optimal policy**: $f^*_{A_k} = \arg\max_{a_k} Q_k(H_k, a_k; \beta_k)$

- **Fair world** expectations wrt to $p^*(Z)$

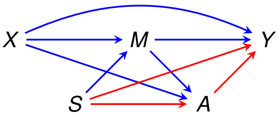
$$\widetilde{Q}_k(H_k \setminus \{M, S\}, A_k) = \frac{1}{Z_k} \sum_{m,s} Q^*_k(H_k, A_k)\, p^*(m | X, s; \alpha_m)\, p^*(s | X; \alpha_s),$$

for $k = 1, \ldots, K$.

COMPAS data

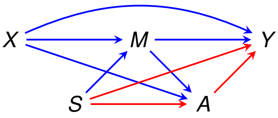
- COMPAS is a criminal justice risk assessment tool created by the company Northpointe (now Equivant)
- Used across the US to determine whether to release or detain a defendant before their trial
- COMPAS scores: recidivism score

COMPAS application



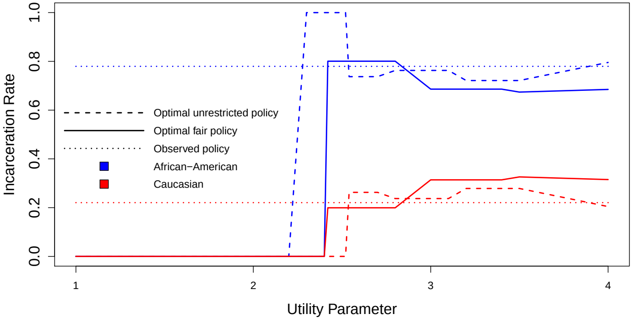
- S: race, X: other demographics, M: prior convictions
- A: incarceration (based on risk of recidivism)
- Heuristic utility: $Y \equiv (1 - A) \times \{\theta R + (1 - R)\} - A$
  - R: whether or not recidivism occurred in a span of two years
  - Negative utility (social, economical costs) associated with incarceration $A = 1$.
  - Some cost to releasing individuals who go on to reoffend (i.e., for whom $A = 0$ and $R = 1$) controlled by θ
  - Positive utility associated with releasing individuals who do not go on to recidivate (i.e., for whom $A = 0$ and $R = 0$)

COMPAS application



- S: race, X: other demographics, M: prior convictions
- A: incarceration (based on risk of recidivism)
- Heuristic utility: $Y \equiv (1 - A) \times \{\theta R + (1 - R)\} - A$
  - R: whether or not recidivism occurred in a span of two years
  - Negative utility (social, economical costs) associated with incarceration $A = 1$.
  - Some cost to releasing individuals who go on to reoffend (i.e., for whom $A = 0$ and $R = 1$) controlled by θ
  - Positive utility associated with releasing individuals who do not go on to recidivate (i.e., for whom $A = 0$ and $R = 0$)

COMPAS application



Question: What would be the resulting difference in pre-trial incarceration rate under a “fair” vs. unconstrained optimal policy?

Result: “fair” vs. unconstrained policies differ, and incarceration rates depend crucially on the utility function.

Summary

- We extended a formalization of algorithmic fairness from Nabi and Shpitser (2018) to the setting of learning optimal policies under fairness constraints.
- We showed how to constrain a set of statistical models and learn a policy that induces high-quality outcomes while satisfying the specified fairness constraints in the induced joint distribution.
- We investigated the performance of our proposals on synthetic and real data.
- We hope to develop and implement more sophisticated constrained optimization methods, to use information as efficiently as possible while satisfying the desired theoretical guarantees.

References

R. Nabi and I. Shpitser
Fair Inference on Outcomes.
Proceedings of AAAI, 2018.

R. Nabi, D. Malinsky, and I. Shpitser
Learning Optimal Fair Policies.
Proceedings of ICML, 2019.

J. Pearl
Direct and indirect effects.
Proceedings of UAI, 2001.

I. Shpitser and J. Pearl
Complete identification methods for the causal hierarchy.
JMLR 9(Sep):1941-1979, 2008.

I. Shpitser
Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding.
Cognitive Science (Rumelhart special issue) 37:1011-1035, 2013.

Thank you for listening.

Razieh Nabi, rnabi@jhu.edu
Daniel Malinsky, malinsky@jhu.edu
Ilya Shpitser, ilyas@cs.jhu.edu

Poster ID: 126

Appendix A: mapping instances to $p^*(Z)$

- A new instance Z is not drawn from the "fair world" $p^*(Z)$, but from "unfair world" $p(Z)$.
  - Map new instances from p to p^*
- Assume $Z = \{Z_1, Z_2\}$ such that $p^*(Y, Z) = p^*(Y, Z_1 | Z_2) p(Z_2)$.
  - Z_2 is shared between p and p^*, i.e. $p^*(Z_2) = p(Z_2)$ but $p^*(Z_1|Z_2) \neq p(Z_1|Z_2)$.
- There is no obvious principled way of knowing exactly what values of Z_1 the "fair version" of the new instance would attain.
  - All such possible values are averaged out, weighted appropriately by how likely they are according to the estimated p^*.
  - This entails predicting Y as the expected value $E^*[Y|Z_2]$ (with respect to the distribution $\sum_{Z_1} p^*(Y, Z_1|Z_2)$).

Appendix B: value search

- **Optimal policy**: $f_A^* = \arg\max_{f_A} \mathbb{E}[Y(f_A)]$
- **Unfair world**: expectations wrt to $p(Z)$

$$\mathbb{E}[Y(f_A)] = \mathbb{E}\left[\frac{\mathbb{I}(A_1 = f_{A_1}(H_1))}{p(A_1 | H_1; \psi)} \times \frac{\mathbb{I}(A_2 = f_{A_2}(H_2))}{p(A_2 | H_2; \psi)} \times Y\right],$$

- **Fair world**: expectations wrt to $p^*(Z)$

$$\widetilde{\mathbb{E}}[Y(f_A)] = \frac{1}{Z} \sum_{m,s} \mathbb{E}[Y(f_A)] \, p^*(m | X, s; \alpha_m) \, p^*(s | X; \alpha_s)$$

Appendix C: synthetic data

	Unfair Policy	**Fair Policy**
Q-learning	7.219 ± 0.005	6.104 ± 0.006

Table: Comparison of population outcomes $\mathbb{E}[Y]$

- For observed policy, the value is 4.82 ± 0.007
- $(X, Y) \sim$ Normal, $(M, A) \sim$ Logistic, $S \sim$ Bernoulli.

Appendix D: additional COMPAS results

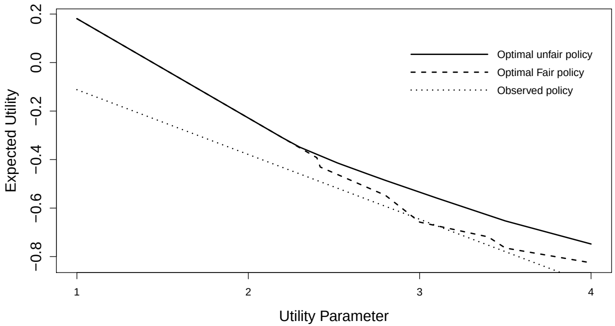
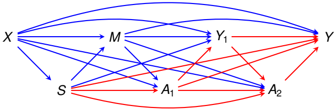
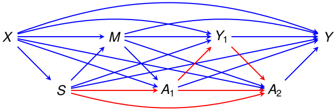


Figure: Relative expected utilities for policies as function of θ

Appendix E: sources of bias





- **Retrospective bias**: bias in historical data used as input to learning procedure.

  Example: unfair paths from S to Y:
  $\{S \to Y,$
  $S \to A_1 \to \ldots \to Y,$
  $S \to A_2 \to \ldots \to Y\}.$

  $$\text{PSE}^{sy} = g_1(Z)$$

- **Prospective bias**: functional form of policy depends on sensitive features.

  Example: unfair paths from S to A_1, A_2:
  $\{S \to A_1\}$,
  and
  $\{S \to A_2, \quad S \to A_1 \to \ldots \to A_2\}$

  $$\text{PSE}^{sa_k} = g_k(Z)$$

- **Confounding bias**: spurious associations between sensitive features, decisions, and outcomes.

- **Statistical bias**: reliance on misspecified statistical models.