Fair Inference on Outcomes

Razieh Nabi
rnabiab1@jhu.edu

Ilya Shpitser
ilyas@cs.jhu.edu

Computer Science Department

JOHNS HOPKINS UNIVERSITY

AAAI-18: Thirty-Second Conference on Artificial Intelligence

Bias and Discrimination

- ML algorithms are making making influential decisions in people's lives
  - Insurance approval, hiring decision, recidivism prediction
  - Based on complicated regression or classification algorithms $\mathbb{E}[Y \mid \mathbf{X}; \alpha]$ or $p(Y \mid \mathbf{X}; \alpha)$, ($Y$: outcome, $\mathbf{X}$: features, α: model parameters)
- Algorithms can reinforce human prejudices
  - Data is collected from the "unfair" world
    - Example: racial profiling (police officers vs African-Americans)
  - No (default) correction for discriminatory biases in statistical models
    - Selection bias is not the same as statistical bias
- How to define and measure discrimination/fairness?
- How to make statistical inference "fair"?

Bias and Discrimination

- ML algorithms are making making influential decisions in people's lives
  - Insurance approval, hiring decision, recidivism prediction
  - Based on complicated regression or classification algorithms $\mathbb{E}[Y \mid \mathbf{X}; \alpha]$ or $p(Y \mid \mathbf{X}; \alpha)$, ($Y$: outcome, $\mathbf{X}$: features, α: model parameters)
- Algorithms can reinforce human prejudices
  - Data is collected from the "unfair" world
    - Example: racial profiling (police officers vs African-Americans)
  - No (default) correction for discriminatory biases in statistical models
    - Selection bias is not the same as statistical bias
- How to define and measure discrimination/fairness?
- How to make statistical inference "fair"?

Bias and Discrimination

- ML algorithms are making making influential decisions in people's lives
  - Insurance approval, hiring decision, recidivism prediction
  - Based on complicated regression or classification algorithms $\mathbb{E}[Y \mid \mathbf{X}; \alpha]$ or $p(Y \mid \mathbf{X}; \alpha)$, ($Y$: outcome, $\mathbf{X}$: features, α: model parameters)
- Algorithms can reinforce human prejudices
  - Data is collected from the "unfair" world
    - Example: racial profiling (police officers vs African-Americans)
  - No (default) correction for discriminatory biases in statistical models
    - Selection bias is not the same as statistical bias
- How to define and measure discrimination/fairness?
- How to make statistical inference "fair"?

Bias and Discrimination

- ML algorithms are making making influential decisions in people's lives
  - Insurance approval, hiring decision, recidivism prediction
  - Based on complicated regression or classification algorithms
    $\mathbb{E}[Y \mid \mathbf{X}; \alpha]$ or $p(Y \mid \mathbf{X}; \alpha)$, ($Y$: outcome, $\mathbf{X}$: features, α: model parameters)
- Algorithms can reinforce human prejudices
  - Data is collected from the "unfair" world
    - Example: racial profiling (police officers vs African-Americans)
  - No (default) correction for discriminatory biases in statistical models
    - Selection bias is not the same as statistical bias
- How to define and measure discrimination/fairness?
- How to make statistical inference "fair"?

Fairness in Machine Learning

- Problem setup:
  - **X**: a set of covariates, A: sensitive variable, Y: outcome variable
  - Given data on $(\mathbf{X}, A, Y)$, are predictions of Y from **X** and A discriminatory (with respect to A)?
- A mathematical definition **+** "analytic philosophy" argument
  - Define "discrimination" as X
  - Why is X a good definition?
- Fairness is something rooted in **human intuition**
- Our approach is inspired by causal inference
  - Causal inference: move from a *factual* to a *counterfactual* world
  - Fair inference: move from an *"unfair"* to a *"fair world"*
- Discrimination wrt A for Y is the presence of an effect of A on Y along "unfair causal paths."

Fairness in Machine Learning

- Problem setup:
  - $\mathbf{X}$: a set of covariates, A: sensitive variable, Y: outcome variable
  - Given data on $(\mathbf{X}, A, Y)$, are predictions of Y from $\mathbf{X}$ and A discriminatory (with respect to A)?
- A mathematical definition + "analytic philosophy" argument
  - Define "discrimination" as X
  - Why is X a good definition?
- Fairness is something rooted in **human intuition**
- Our approach is inspired by causal inference
  - Causal inference: move from a *factual* to a *counterfactual* world
  - Fair inference: move from an *"unfair"* to a *"fair world"*
- Discrimination wrt A for Y is the presence of an effect of A on Y along "unfair causal paths."

Fairness in Machine Learning

- Problem setup:
  - $\mathbf{X}$: a set of covariates, A: sensitive variable, Y: outcome variable
  - Given data on $(\mathbf{X}, A, Y)$, are predictions of Y from $\mathbf{X}$ and A discriminatory (with respect to A)?
- A mathematical definition **+** "analytic philosophy" argument
  - Define "discrimination" as X
  - Why is X a good definition?
- Fairness is something rooted in **human intuition**
- Our approach is inspired by causal inference
  - Causal inference: move from a *factual* to a *counterfactual* world
  - Fair inference: move from an *"unfair"* to a *"fair world"*
- Discrimination wrt A for Y is the presence of an effect of A on Y along "unfair causal paths."

Fairness in Machine Learning

- Problem setup:
  - **X**: a set of covariates, A: sensitive variable, Y: outcome variable
  - Given data on $(\mathbf{X}, A, Y)$, are predictions of Y from **X** and A discriminatory (with respect to A)?
- A mathematical definition **+** "analytic philosophy" argument
  - Define "discrimination" as X
  - Why is X a good definition?
- Fairness is something rooted in **human intuition**
- Our approach is inspired by causal inference
  - Causal inference: move from a *factual* to a *counterfactual* world
  - Fair inference: move from an *"unfair"* to a *"fair world"*
- Discrimination wrt A for Y is the presence of an effect of A on Y along "unfair causal paths."

Fairness in Machine Learning

- Problem setup:
  - $\mathbf{X}$: a set of covariates, A: sensitive variable, Y: outcome variable
  - Given data on $(\mathbf{X}, A, Y)$, are predictions of Y from $\mathbf{X}$ and A discriminatory (with respect to A)?
- A mathematical definition **+** "analytic philosophy" argument
  - Define "discrimination" as X
  - Why is X a good definition?
- Fairness is something rooted in **human intuition**
- Our approach is inspired by causal inference
  - Causal inference: move from a *factual* to a *counterfactual* world
  - Fair inference: move from an *"unfair"* to a *"fair world"*
- Discrimination wrt A for Y is the presence of an effect of A on Y along "unfair causal paths."

Fairness in Machine Learning

- Problem setup:
  - $\mathbf{X}$: a set of covariates, A: sensitive variable, Y: outcome variable
  - Given data on ($\mathbf{X}$, A, Y), are predictions of Y from $\mathbf{X}$ and A discriminatory (with respect to A)?
- A mathematical definition **+** "analytic philosophy" argument
  - Define "discrimination" as X
  - Why is X a good definition?
- Fairness is something rooted in **human intuition**
- Our approach is inspired by causal inference
  - Causal inference: move from a *factual* to a *counterfactual* world
  - Fair inference: move from an *"unfair"* to a *"fair world"*
- Discrimination wrt A for Y is the presence of an effect of A on Y along "unfair causal paths."

Intuitions for Defining Discrimination

- Gender discrimination and hiring:
  - Data: features $\mathbf{X}$ (collected from resumes), gender A, hiring decision Y
  - Title VII of the Civil Rights Act of 1964 forbids employment discrimination on the basis of gender, race, national origin, etc
  - Is there hiring discrimination wrt A?
- Intuition: hypothetical experiments

Intuitions for Defining Discrimination

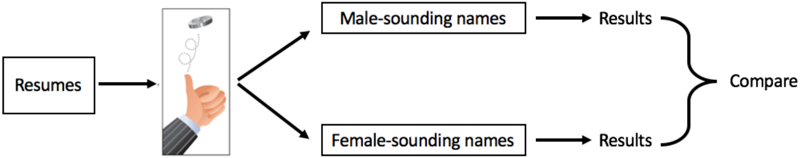
- Gender discrimination and hiring:
  - Data: features $\mathbf{X}$ (collected from resumes), gender A, hiring decision Y
  - Title VII of the Civil Rights Act of 1964 forbids employment discrimination on the basis of gender, race, national origin, etc
  - Is there hiring discrimination wrt A?
- Intuition: hypothetical experiments

Intuitions for Defining Discrimination

- Gender discrimination and hiring:
  - Data: features $\mathbf{X}$ (collected from resumes), gender A, hiring decision Y
  - Title VII of the Civil Rights Act of 1964 forbids employment discrimination on the basis of gender, race, national origin, etc
  - Is there hiring discrimination wrt A?
- Intuition: hypothetical experiments

Intuitions for Defining Discrimination

- Gender discrimination and hiring:
  - Data: features **X** (collected from resumes), gender A, hiring decision Y
  - Title VII of the Civil Rights Act of 1964 forbids employment discrimination on the basis of gender, race, national origin, etc
  - Is there hiring discrimination wrt A?
- Intuition: hypothetical experiments



Intuitions in Defining Discrimination

- 7th circuit court case (Carson versus Bethlehem Steel Corp, 1996):

  “The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different gender (age, race, religion, national origin etc.) and everything else had been the same"

- Intuitive definitions of fairness are **counterfactual**
  - Causal inference: study of hypothetical experiments and counterfactuals
  - “Fairness” is a causal inference problem

- Mediation analysis: study of causal mechanisms
  - Our approach to fairness uses tools from mediation analysis

Intuitions in Defining Discrimination

- 7th circuit court case (Carson versus Bethlehem Steel Corp, 1996):

  "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different gender (age, race, religion, national origin etc.) and everything else had been the same"

- Intuitive definitions of fairness are **counterfactual**
  - Causal inference: study of hypothetical experiments and counterfactuals
  - "Fairness" is a causal inference problem

- Mediation analysis: study of causal mechanisms
  - Our approach to fairness uses tools from mediation analysis

Intuitions in Defining Discrimination

- 7th circuit court case (Carson versus Bethlehem Steel Corp, 1996):

  "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different gender (age, race, religion, national origin etc.) and everything else had been the same"

- Intuitive definitions of fairness are **counterfactual**
  - Causal inference: study of hypothetical experiments and counterfactuals
  - "Fairness" is a causal inference problem

- Mediation analysis: study of causal mechanisms
  - Our approach to fairness uses tools from mediation analysis

Causal Inference: preliminaries

- Data $\mathcal{D} \sim p(\mathbf{X}, A, Y)$, $\mathbf{X}$ baselines, A treatment, Y outcome
- $Y(a)$: outcome Y had A been assigned to a
- Average causal effect: $\text{ACE} = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$
  - Randomized experiments: compare cases $(A = a)$ and controls $(A = a')$
  - Observational data: people choose to smoke
- Consistency $(Y(A) = Y)$ and ignorability $(Y(a) \perp\!\!\!\perp A \mid \mathbf{X}, \forall a)$

Causal Inference: preliminaries

- Data $\mathcal{D} \sim p(\mathbf{X}, A, Y)$, $\mathbf{X}$ baselines, A treatment, Y outcome
- $Y(a)$: outcome Y had A been assigned to a
- Average causal effect: $\text{ACE} = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$
  - Randomized experiments: compare cases $(A = a)$ and controls $(A = a')$
  - Observational data: people choose to smoke
- Consistency $(Y(A) = Y)$ and ignorability $(Y(a) \perp\!\!\!\perp A \mid \mathbf{X}, \forall a)$

Causal Inference: preliminaries

- Data $\mathcal{D} \sim p(\mathbf{X}, A, Y)$, $\mathbf{X}$ baselines, A treatment, Y outcome
- $Y(a)$: outcome Y had A been assigned to a
- Average causal effect: $\text{ACE} = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$
  - Randomized experiments: compare cases $(A = a)$ and controls $(A = a')$
  - Observational data: people choose to smoke
- Consistency $(Y(A) = Y)$ and ignorability $(Y(a) \perp\!\!\!\perp A \mid \mathbf{X}, \forall a)$

Causal Inference: preliminaries

- Data $\mathcal{D} \sim p(\mathbf{X}, A, Y)$, $\mathbf{X}$ baselines, A treatment, Y outcome
- $Y(a)$: outcome Y had A been assigned to a
- Average causal effect: $\text{ACE} = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$
  - Randomized experiments: compare cases ($A = a$) and controls ($A = a'$)
  - Observational data: people choose to smoke
- Consistency ($Y(A) = Y$) and ignorability ($Y(a) \perp\!\!\!\perp A \mid \mathbf{X}, \forall a$)

Causal Inference: preliminaries

- Data $\mathcal{D} \sim p(\mathbf{X}, A, Y)$, $\mathbf{X}$ baselines, A treatment, Y outcome
- $Y(a)$: outcome Y had A been assigned to a
- Average causal effect: $\text{ACE} = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$
  - Randomized experiments: compare cases $(A = a)$ and controls $(A = a')$
  - Observational data: people choose to smoke
- Consistency $(Y(A) = Y)$ and ignorability $(Y(a) \perp\!\!\!\perp A \mid \mathbf{X}, \forall a)$

Causal Inference: preliminaries

- Data $\mathcal{D} \sim p(\mathbf{X}, A, Y)$, $\mathbf{X}$ baselines, A treatment, Y outcome
- $Y(a)$: outcome Y had A been assigned to a
- Average causal effect: ACE $= \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')]$
  - Randomized experiments: compare cases $(A = a)$ and controls $(A = a')$
  - Observational data: people choose to smoke
- Consistency $(Y(A) = Y)$ and ignorability $(Y(a) \perp\!\!\!\perp A \mid \mathbf{X}, \forall a)$

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

Mediation Analysis: preliminaries

- Causal mechanisms: how A causes 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

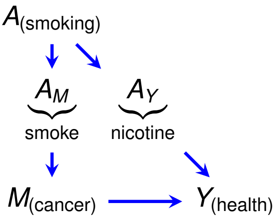
- Causal mechanisms: how A causes Y?
- ACE = Direct effect ($A \rightarrow Y$) + Indirect effect ($A \rightarrow M \rightarrow Y$)
  - $\mathcal{D} = \{\mathbf{X}, A, M, Y\}$. M mediates the effect of A on Y
- Nested counterfactuals $Y(a, M(a'))$
  - Outcome Y had A been assigned to a and M been assigned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

Mediation Analysis: preliminaries

- Causal mechanisms: how A causes 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'

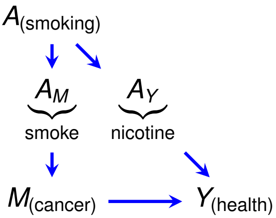
Mediation Analysis: preliminaries

- Causal mechanisms: how A causes 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

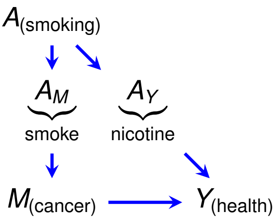
- Causal mechanisms: how A causes Y?
- ACE = Direct effect ($A \rightarrow Y$) + Indirect effect ($A \rightarrow M \rightarrow Y$)
  - $\mathcal{D} = \{\mathbf{X}, A, M, Y\}$. M mediates the effect of A on Y
- Nested counterfactuals $Y(a, M(a'))$
  - Outcome Y had A been assigned to a and M been assigned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



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

Mediation Analysis: preliminaries

- Causal mechanisms: how A causes 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 in:
$Y(1, M(0))$	1	0	nicotine patch

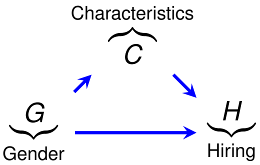
- **Direct Effect** $= \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0)]$ $(A \rightarrow Y)$
- **Indirect Effect** $= \mathbb{E}[Y(1)] - \mathbb{E}[Y(1, M(0))]$ $(A \rightarrow M \rightarrow Y)$

Direct Effect and Path-Specific Effect

- Compare resumes of men and **same** resumes with names switched to female ones:

$$Y(a') : H(G = male, C(G = male)) = H(G = male)$$

$$Y(a, M(a')) : H(G = female, C(G = male))$$



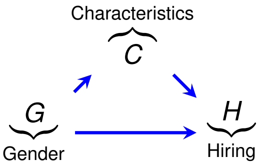
1. **Path-specific effect (PSE)**
   - Along a path, all nodes behave as if $A = a$,
   - Along all other paths, nodes behave as if $A = a'$
2. Discrimination as the presence of effect along unfair causal pathways
3. Fairness is a domain specific issue!

Direct Effect and Path-Specific Effect

- Compare resumes of men and **same** resumes with names switched to female ones:

$$Y(a') : H(G = male, C(G = male)) = H(G = male)$$

$$Y(a, M(a')) : H(G = female, C(G = male))$$



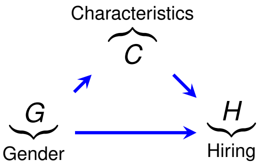
1. **Path-specific effect (PSE)**
   - Along a path, all nodes behave as if $A = a$,
   - Along all other paths, nodes behave as if $A = a'$
2. Discrimination as the presence of effect along unfair causal pathways
3. Fairness is a domain specific issue!

Direct Effect and Path-Specific Effect

- Compare resumes of men and **same** resumes with names switched to female ones:

$$Y(a') : H(G = male, C(G = male)) = H(G = male)$$

$$Y(a, M(a')) : H(G = female, C(G = male))$$



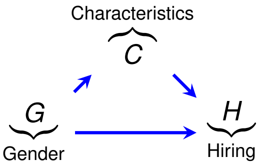
1. **Path-specific effect (PSE)**
   - Along a path, all nodes behave as if $A = a$,
   - Along all other paths, nodes behave as if $A = a'$
2. Discrimination as the presence of effect along unfair causal pathways
3. Fairness is a domain specific issue!

Direct Effect and Path-Specific Effect

- Compare resumes of men and **same** resumes with names switched to female ones:

$$Y(a') : H(G = male, C(G = male)) = H(G = male)$$

$$Y(a, M(a')) : H(G = female, C(G = male))$$



1. **Path-specific effect (PSE)**
   - Along a path, all nodes behave as if $A = a$,
   - Along all other paths, nodes behave as if $A = a'$
2. Discrimination as the presence of effect along unfair causal pathways
3. Fairness is a domain specific issue!

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark "unfair" causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is "unfair"
  - Find a "fair world" p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance
- Approximate "Fair World" p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \alpha)$

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark "unfair" causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is "unfair"
  - Find a "fair world" p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance
- Approximate "Fair World" p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \alpha)$

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark "unfair" causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is "unfair"
  - Find a "fair world" p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance
- Approximate "Fair World" p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \alpha)$

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark “unfair” causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is “unfair”
  - Find a “fair world” p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance
- Approximate “Fair World” p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \alpha)$

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark "unfair" causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is "unfair"
  - Find a "fair world" p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance
- Approximate "Fair World" p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \alpha)$

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark "unfair" causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is "unfair"
  - Find a "fair world" p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense
      (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance
- Approximate "Fair World" p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \alpha)$

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark "unfair" causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is "unfair"
  - Find a "fair world" p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance
- Approximate "Fair World" p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \alpha)$

Our Approach

- Predict Y from $\mathbf{X}, A, \mathbf{M}$ in a fair way:
  - Consider all causal paths from A to Y
  - Mark "unfair" causal paths and
  - Compute PSE along those paths: $g(\mathcal{D})$
  - If there is PSE, then $p(\mathbf{X}, A, \mathbf{M}, Y)$ is "unfair"
  - Find a "fair world" p^* close to p where there is no PSE
    - Close in Kullback-Leibler divergence sense
      (use data as well as possible while remaining fair)
    - $(\epsilon_l; \epsilon_u)$: discrimination tolerance

- Approximate "Fair World" p^*:
  - Likelihood function: $\mathcal{L}(\mathcal{D}; \boldsymbol{\alpha})$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \quad \mathcal{L}(\mathcal{D}; \boldsymbol{\alpha})$$
$$\text{subject to} \quad \epsilon_l \leq g(\mathcal{D}) \leq \epsilon_u.$$

Inference on New Instances

- Inference on new instances $(\mathbf{x}, a, \mathbf{m})$:
  - New instances are drawn from unfair p
  - Cannot classify/regress new instances using $p^*(Y \mid \mathbf{x}, a, \mathbf{m}, \hat{\alpha})$
  - Use only shared information between p and p^*
    - If $p^*(\mathbf{X}, A, \mathbf{M}, Y) = p(\mathbf{X})p^*(A, \mathbf{M}, Y \mid \mathbf{X})$, use $\mathbb{E}[Y \mid \mathbf{X}; \hat{\alpha}]$.
    - Depends on $g(\mathcal{D})$: inverse weighting, g-formula, semi-parametric

Inference on New Instances

- Inference on new instances $(\mathbf{x}, a, \mathbf{m})$:
  - New instances are drawn from unfair p
  - Cannot classify/regress new instances using $p^*(Y \mid \mathbf{x}, a, \mathbf{m}, \hat{\alpha})$
  - Use only shared information between p and p^*
    - If $p^*(\mathbf{X}, A, \mathbf{M}, Y) = p(\mathbf{X})p^*(A, \mathbf{M}, Y \mid \mathbf{X})$, use $\mathbb{E}[Y \mid \mathbf{X}; \hat{\alpha}]$.
    - Depends on $g(\mathcal{D})$: inverse weighting, g-formula, semi-parametric

Inference on New Instances

- Inference on new instances $(\mathbf{x}, a, \mathbf{m})$:
  - New instances are drawn from unfair p
  - Cannot classify/regress new instances using $p^*(Y \mid \mathbf{x}, a, \mathbf{m}, \hat{\alpha})$
  - Use only shared information between p and p^*
    - If $p^*(\mathbf{X}, A, \mathbf{M}, Y) = p(\mathbf{X})p^*(A, \mathbf{M}, Y \mid \mathbf{X})$, use $\mathbb{E}[Y \mid \mathbf{X}; \hat{\alpha}]$.
    - Depends on $g(\mathcal{D})$: inverse weighting, g-formula, semi-parametric

Inference on New Instances

- Inference on new instances $(\mathbf{x}, a, \mathbf{m})$:
  - New instances are drawn from unfair p
  - Cannot classify/regress new instances using $p^*(Y \mid \mathbf{x}, a, \mathbf{m}, \hat{\alpha})$
  - Use only shared information between p and p^*
    - If $p^*(\mathbf{X}, A, \mathbf{M}, Y) = p(\mathbf{X})p^*(A, \mathbf{M}, Y \mid \mathbf{X})$, use $\mathbb{E}[Y \mid \mathbf{X}; \hat{\alpha}]$.
    - Depends on $g(\mathcal{D})$: inverse weighting, g-formula, semi-parametric

Application: COMPAS

Machine Bias (ProPublica)

COMPAS: risk assessments in criminal sentencing (developed by Northpointe)

Experiment 1: Bias in Data

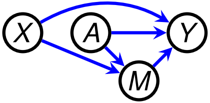
- Is there any **bias in the data** wrt race in predicting recidivism?

  Y: recidivism

  A: race

  $\mathbf{X}$: demographics

  M: criminal record



- Discriminatory path: $A \to Y$
- Constrained MCMC and Bayesian random forests to obtain "fair world"
- No hope to beat the MLE, by definition. Do as well as possible while remaining fair.

Experiment 1: Bias in Data

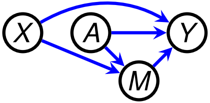
- Is there any **bias in the data** wrt race in predicting recidivism?

  Y: recidivism

  A: race

  $\mathbf{X}$: demographics

  M: criminal record



- Discriminatory path: $A \rightarrow Y$
- Constrained MCMC and Bayesian random forests to obtain "fair world"
- No hope to beat the MLE, by definition. Do as well as possible while remaining fair.

Experiment 1: Bias in Data

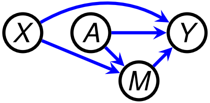
- Is there any **bias in the data** wrt race in predicting recidivism?

  Y: recidivism

  A: race

  $\mathbf{X}$: demographics

  M: criminal record



- Discriminatory path: $A \rightarrow Y$
- Constrained MCMC and Bayesian random forests to obtain "fair world"
- No hope to beat the MLE, by definition. Do as well as possible while remaining fair.

Experiment 1: Bias in Data

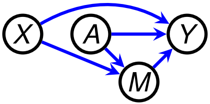
- Is there any **bias in the data** wrt race in predicting recidivism?

  Y: recidivism

  A: race

  $\mathbf{X}$: demographics

  M: criminal record



- Discriminatory path: $A \to Y$
- Constrained MCMC and Bayesian random forests to obtain “fair world”

	Direct Effect (odds ratio scale, null = 1)
$\mathbb{E}[Y \mid A, M, \mathbf{X}]$	1.3

- No hope to beat the MLE, by definition. Do as well as possible while remaining fair.

Experiment 1: Bias in Data

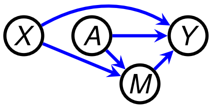
- Is there any **bias in the data** wrt race in predicting recidivism?

  Y: recidivism

  A: race

  $\mathbf{X}$: demographics

  M: criminal record



- Discriminatory path: $A \to Y$

- Constrained MCMC and Bayesian random forests to obtain "fair world"

	Direct Effect (odds ratio scale, null = 1)
$\mathbb{E}[Y \mid A, M, \mathbf{X}]$	1.3
(our method) $\mathbb{E}^*[Y \mid \mathbf{X}]$	$0.95 \leq \text{PSE} \leq 1.05$

- No hope to beat the MLE, by definition. Do as well as possible while remaining fair.

Experiment 1: Bias in Data

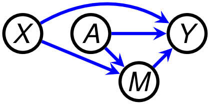
- Is there any **bias in the data** wrt race in predicting recidivism?

  Y: recidivism

  A: race

  $\mathbf{X}$: demographics

  M: criminal record



- Discriminatory path: $A \to Y$

- Constrained MCMC and Bayesian random forests to obtain "fair world"

	Direct Effect (odds ratio scale, null = 1)	Accuracy %
$\mathbb{E}[Y \mid A, M, \mathbf{X}]$	1.3	67.8
(our method) $\mathbb{E}^*[Y \mid \mathbf{X}]$	$0.95 \leq \text{PSE} \leq 1.05$	66.4

- No hope to beat the MLE, by definition. Do as well as possible while remaining fair.

Experiment 1: Bias in Data

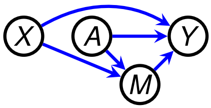
- Is there any **bias in the data** wrt race in predicting recidivism?

  Y: recidivism

  A: race

  $\mathbf{X}$: demographics

  M: criminal record



- Discriminatory path: $A \rightarrow Y$
- Constrained MCMC and Bayesian random forests to obtain "fair world"

	Direct Effect (odds ratio scale, null = 1)	Accuracy %
$\mathbb{E}[Y \mid A, M, \mathbf{X}]$	1.3	67.8
(our method) $\mathbb{E}^*[Y \mid \mathbf{X}]$	$0.95 \leq \text{PSE} \leq 1.05$	66.4

- No hope to beat the MLE, by definition. Do as well as possible while remaining fair.

Experiment 2: Bias in Algorithm

- Is there any **bias in the algorithm** that generates COMPAS scores?
- Northpointe claims they do not use race in generating COMPAS scores. Therefore, they are fair.
- Do not have access to Northpointe's model
- Best we can do:
  - Try to learn $\tilde{\mathbb{E}}[Y \mid M, \mathbf{X}]$ with what we have
  - Check for PSE of A on Y in the model for $p(Y, A, M, \mathbf{X})$ where $\mathbb{E}[Y \mid M, \mathbf{X}]$ is constrained to be $\tilde{\mathbb{E}}$.
- PSE (direct effect) is 2.1

  There's something wrong with Northpinte's claim.

Experiment 2: Bias in Algorithm

- Is there any **bias in the algorithm** that generates COMPAS scores?
- Northpointe claims they do not use race in generating COMPAS scores. Therefore, they are fair.

- Do not have access to Northpointe's model
- Best we can do:
  - Try to learn $\tilde{\mathbb{E}}[Y \mid M, \mathbf{X}]$ with what we have
  - Check for PSE of A on Y in the model for $p(Y, A, M, \mathbf{X})$ where $\mathbb{E}[Y \mid M, \mathbf{X}]$ is constrained to be $\tilde{\mathbb{E}}$.
- PSE (direct effect) is 2.1

  There's something wrong with Northpinte's claim.

Experiment 2: Bias in Algorithm

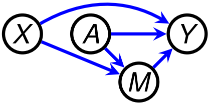
- Is there any **bias in the algorithm** that generates COMPAS scores?
- Northpointe claims they do not use race in generating COMPAS scores. Therefore, they are fair.

Y: COMPAS score

A: race

$\mathbf{X}$: demographics

M: criminal record



- Do not have access to Northpointe's model
- Best we can do:
  - Try to learn $\tilde{\mathbb{E}}[Y \mid M, \mathbf{X}]$ with what we have
  - Check for PSE of A on Y in the model for $p(Y, A, M, \mathbf{X})$ where $\mathbb{E}[Y \mid M, \mathbf{X}]$ is constrained to be $\tilde{\mathbb{E}}$.
- PSE (direct effect) is 2.1

  There's something wrong with Northpinte's claim.

Experiment 2: Bias in Algorithm

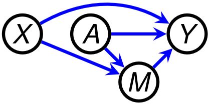
- Is there any **bias in the algorithm** that generates COMPAS scores?
- Northpointe claims they do not use race in generating COMPAS scores. Therefore, they are fair.

Y: COMPAS score

A: race

$\mathbf{X}$: demographics

M: criminal record



- Do not have access to Northpointe's model
- Best we can do:
  - Try to learn $\tilde{\mathbb{E}}[Y \mid M, \mathbf{X}]$ with what we have
  - Check for PSE of A on Y in the model for $p(Y, A, M, \mathbf{X})$ where $\mathbb{E}[Y \mid M, \mathbf{X}]$ is constrained to be $\tilde{\mathbb{E}}$.
- PSE (direct effect) is 2.1

  There's something wrong with Northpinte's claim.

Experiment 2: Bias in Algorithm

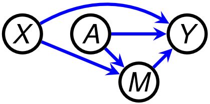
- Is there any **bias in the algorithm** that generates COMPAS scores?
- Northpointe claims they do not use race in generating COMPAS scores. Therefore, they are fair.

Y: COMPAS score

A: race

$\mathbf{X}$: demographics

M: criminal record



- Do not have access to Northpointe's model
- Best we can do:
  - Try to learn $\tilde{\mathbb{E}}[Y \mid M, \mathbf{X}]$ with what we have
  - Check for PSE of A on Y in the model for $p(Y, A, M, \mathbf{X})$ where $\mathbb{E}[Y \mid M, \mathbf{X}]$ is constrained to be $\tilde{\mathbb{E}}$.
- PSE (direct effect) is 2.1

  There's something wrong with Northpinte's claim.

Summary

- An approach to fair inference based on mediation analysis.
- Argued approach isn't arbitrary, but rooted in human intuition on what is fair in practice.
- Fairness may be characterized as the absence (or dampening) of a path-specific effect (PSE).
- Restriction of a PSE is expressed as a likelihood maximization problem that features constraining the magnitude of the undesirable PSE.
- Extensions (not covered):
  - What if the path-specific effect is not identified?
  - Easy ways to do constrained MLE.
- Evidence existing prediction models may be quite discriminatory.

Summary

- An approach to fair inference based on mediation analysis.
- Argued approach isn't arbitrary, but rooted in human intuition on what is fair in practice.
- Fairness may be characterized as the absence (or dampening) of a path-specific effect (PSE).
- Restriction of a PSE is expressed as a likelihood maximization problem that features constraining the magnitude of the undesirable PSE.
- Extensions (not covered):
  - What if the path-specific effect is not identified?
  - Easy ways to do constrained MLE.
- Evidence existing prediction models may be quite discriminatory.

Summary

- An approach to fair inference based on mediation analysis.
- Argued approach isn't arbitrary, but rooted in human intuition on what is fair in practice.
- Fairness may be characterized as the absence (or dampening) of a path-specific effect (PSE).
- Restriction of a PSE is expressed as a likelihood maximization problem that features constraining the magnitude of the undesirable PSE.
- Extensions (not covered):
  - What if the path-specific effect is not identified?
  - Easy ways to do constrained MLE.
- Evidence existing prediction models may be quite discriminatory.

Summary

- An approach to fair inference based on mediation analysis.
- Argued approach isn't arbitrary, but rooted in human intuition on what is fair in practice.
- Fairness may be characterized as the absence (or dampening) of a path-specific effect (PSE).
- Restriction of a PSE is expressed as a likelihood maximization problem that features constraining the magnitude of the undesirable PSE.
- Extensions (not covered):
  - What if the path-specific effect is not identified?
  - Easy ways to do constrained MLE.
- Evidence existing prediction models may be quite discriminatory.

Summary

- An approach to fair inference based on mediation analysis.
- Argued approach isn't arbitrary, but rooted in human intuition on what is fair in practice.
- Fairness may be characterized as the absence (or dampening) of a path-specific effect (PSE).
- Restriction of a PSE is expressed as a likelihood maximization problem that features constraining the magnitude of the undesirable PSE.
- Extensions (not covered):
  - What if the path-specific effect is not identified?
  - Easy ways to do constrained MLE.
- Evidence existing prediction models may be quite discriminatory.

Summary

- An approach to fair inference based on mediation analysis.
- Argued approach isn't arbitrary, but rooted in human intuition on what is fair in practice.
- Fairness may be characterized as the absence (or dampening) of a path-specific effect (PSE).
- Restriction of a PSE is expressed as a likelihood maximization problem that features constraining the magnitude of the undesirable PSE.
- Extensions (not covered):
  - What if the path-specific effect is not identified?
  - Easy ways to do constrained MLE.
- Evidence existing prediction models may be quite discriminatory.

References

R. Nabi and I. Shpitser
Fair Inference on Outcomes.
n Proceedings of the Thirty-Second Conference on AAAI, 2018.

J. Pearl
Causality: Models, Reasoning, and Inference,
Cambridge University Press 2009.

J. Pearl
Direct and indirect effects.
In Proceedings of the Seventeenth Conference on UAI, 411-420, 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`
Ilya Shpitser, `ilyas@cs.jhu.edu`