Algorithm Design: A Fairness-Accuracy Frontier

Annie Liang[1] Jay Lu[2] Xiaosheng Mu[3]

[1]Northwestern

[2]UCLA [3]Princeton

background

algorithms are used to guide many high-stakes decisions

- which patients should be treated? which borrowers should receive a loan? which defendants should receive bail?

background

algorithms are used to guide many high-stakes decisions

- which patients should be treated? which borrowers should receive a loan? which defendants should receive bail?

these algorithms often have errors that vary systematically across subgroups of the population

- false positive rate of algorithm used to predict criminal reoffense twice as high for Black defendants (Angwin and Larson, 2016)
- patients assigned to same risk score have substantially different actual health risks depending on race (Obermeyer et al., 2019)
- accuracy of facial-recognition technologies varies substantially across racial and gender groups (Klare et al., 2012)

background

algorithms are used to guide many high-stakes decisions

- which patients should be treated? which borrowers should receive a loan? which defendants should receive bail?

these algorithms often have errors that vary systematically across subgroups of the population

- false positive rate of algorithm used to predict criminal reoffense twice as high for Black defendants (Angwin and Larson, 2016)
- patients assigned to same risk score have substantially different actual health risks depending on race (Obermeyer et al., 2019)
- accuracy of facial-recognition technologies varies substantially across racial and gender groups (Klare et al., 2012)

algorithm designers increasingly optimize not only for accuracy but also "fairness" (maintain comparable error rates across groups)

fairness vs. accuracy

what is the tradeoff between fairness and accuracy, and how does it depend on the information available for prediction?

fairness vs. accuracy

what is the tradeoff between fairness and accuracy, and how does it depend on the information available for prediction?

1. the designer chooses the algorithm
   - define a **fairness-accuracy frontier** that ranges across a broad class of preferences/optimization criteria
   - identify simple properties of the inputs that govern the shape of this frontier

fairness vs. accuracy

what is the tradeoff between fairness and accuracy, and how does it depend on the information available for prediction?

1. the designer chooses the algorithm
   - define a **fairness-accuracy frontier** that ranges across a broad class of preferences/optimization criteria
   - identify simple properties of the inputs that govern the shape of this frontier
2. the designer flexibly regulates the inputs to the algorithm (info design), another agent chooses the algorithm

fairness vs. accuracy

what is the tradeoff between fairness and accuracy, and how does it depend on the information available for prediction?

1. the designer chooses the algorithm
   - define a **fairness-accuracy frontier** that ranges across a broad class of preferences/optimization criteria
   - identify simple properties of the inputs that govern the shape of this frontier
2. the designer flexibly regulates the inputs to the algorithm (info design), another agent chooses the algorithm
   - characterize what part of this frontier can be achieved through appropriate garbling of inputs

fairness vs. accuracy

what is the tradeoff between fairness and accuracy, and how does it depend on the information available for prediction?

1. the designer chooses the algorithm
   - define a **fairness-accuracy frontier** that ranges across a broad class of preferences/optimization criteria
   - identify simple properties of the inputs that govern the shape of this frontier

2. the designer flexibly regulates the inputs to the algorithm (info design), another agent chooses the algorithm
   - characterize what part of this frontier can be achieved through appropriate garbling of inputs
   - ask whether the optimal garbling might involve excluding a covariate (group identity, test scores) entirely

part i:

designer chooses algorithm

setup

- single designer and population of (non-strategic) subjects

setup

- single designer and population of (non-strategic) subjects
- each subject is described by three variables:
  - **type** Y taking values in $\mathcal{Y}$
    (e.g. need for medical procedure)
  - **group** $G \in \mathcal{G} = \{r, b\}$
    (e.g. race)
  - **covariate** vector X taking values in $\mathcal{X}$
    (e.g. image scans, # past hospital visits, blood tests)

setup

- single designer and population of (non-strategic) subjects
- each subject is described by three variables:
  - **type** Y taking values in $\mathcal{Y}$
    (e.g. need for medical procedure)
  - **group** $G \in \mathcal{G} = \{r, b\}$
    (e.g. race)
  - **covariate** vector X taking values in $\mathcal{X}$
    (e.g. image scans, # past hospital visits, blood tests)
- X is observed by the designer, Y and G are not directly observed (but may be revealed by X)

how covariates, group identity, and type are related

in the population, $(Y, G, X) \sim \mathbb{P}$

how covariates, group identity, and type are related

in the population, $(Y, G, X) \sim \mathbb{P}$

don't impose any assumptions on $\mathbb{P}$, could be that:

- X reveals or closely proxies for G
  - e.g., consumption patterns predict gender and correlate highly with other group identities (Bertrand and Kamenica, 2020)

how covariates, group identity, and type are related

in the population, $(Y, G, X) \sim \mathbb{P}$

don't impose any assumptions on $\mathbb{P}$, could be that:

- X reveals or closely proxies for G
  - e.g., consumption patterns predict gender and correlate highly with other group identities (Bertrand and Kamenica, 2020)
- X is systematically biased up or down for one group
  - e.g., test scores may be shifted up for a high-income group

how covariates, group identity, and type are related

in the population, $(Y, G, X) \sim \mathbb{P}$

don't impose any assumptions on $\mathbb{P}$, could be that:

- X reveals or closely proxies for G
  - e.g., consumption patterns predict gender and correlate highly with other group identities (Bertrand and Kamenica, 2020)
- X is systematically biased up or down for one group
  - e.g., test scores may be shifted up for a high-income group
- X is more informative about Y for one group than the other
  - e.g., the covariate is selectively reported or more accurately measured for one group

algorithm

each subject receives a **decision** $d \in \mathcal{D} = \{0, 1\}$

(e.g. whether the procedure is recommended)

algorithm

each subject receives a **decision** $d \in \mathcal{D} = \{0, 1\}$

(e.g. whether the procedure is recommended)

the designer chooses an **algorithm**

$$a : \mathcal{X} \to \Delta(\mathcal{D})$$

for determining (distributions over) decisions based on the observed covariate vector

group errors

fix a **loss function** $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$

- measures inaccuracy or harm for a given subject
- e.g., a convex combination of Type I and Type II errors

group errors

fix a **loss function** $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$

- measures inaccuracy or harm for a given subject
- e.g., a convex combination of Type I and Type II errors

Definition

the **error** for group $g \in \mathcal{G}$ given algorithm a is

$$e_g\left(a\right) := \mathbb{E}_{D \sim a(X)}\left[\ell\left(D, Y\right) \mid G = g\right]$$

i.e., the average/expected loss for subjects in group g

group errors

fix a **loss function** $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$

- measures inaccuracy or harm for a given subject
- e.g., a convex combination of Type I and Type II errors

Definition

the **error** for group $g \in \mathcal{G}$ given algorithm a is

$$e_g(a) := \mathbb{E}_{D \sim a(X)} \left[\ell(D, Y) \mid G = g \right]$$

i.e., the average/expected loss for subjects in group g

- improving **accuracy**: lowering e_r and e_b
- improving **fairness**: lowering $|e_r - e_b|$

how to trade off fairness and accuracy?

- there is a large literature on social preferences
- this literature documents substantial heterogeneity in how individuals trade off equity and efficiency
  - Fehr and Schmidt (1999), Andreoni and Miller (2002), Charness and Rabin (2002), Sullivan (2022)
- moreover, no evidence of consensus on how to make this tradeoff for real applications of algorithmic prediction rules

preferences

we consider a broad class of designer preferences:

Definition (fairness-accuracy (FA) dominance)

let $>_{FA}$ be the partial order on $\mathbb{R}^2$ satisfying $(e_r, e_b) >_{FA} (e'_r, e'_b)$ if

$$\underbrace{e_r \leq e'_r, \quad e_b \leq e'_b,}_{\text{higher accuracy}} \quad \text{and} \quad \underbrace{|e_r - e_b| \leq |e'_r - e'_b|}_{\text{higher fairness}}$$

with at least one of these inequalities strict

preferences

we consider a broad class of designer preferences:

Definition (fairness-accuracy (FA) dominance)

let $>_{FA}$ be the partial order on $\mathbb{R}^2$ satisfying $(e_r, e_b) >_{FA} (e'_r, e'_b)$ if

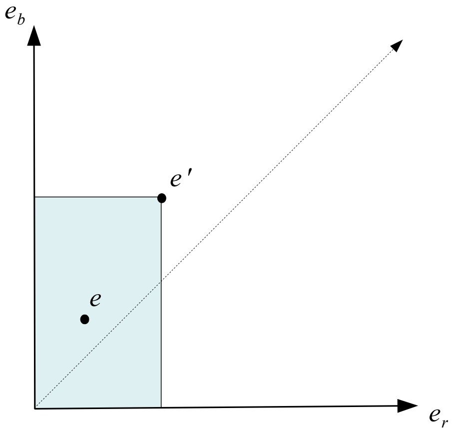
$$\underbrace{e_r \leq e'_r, \quad e_b \leq e'_b,}_{\text{higher accuracy}} \quad \text{and} \quad \underbrace{|e_r - e_b| \leq |e'_r - e'_b|}_{\text{higher fairness}}$$

with at least one of these inequalities strict

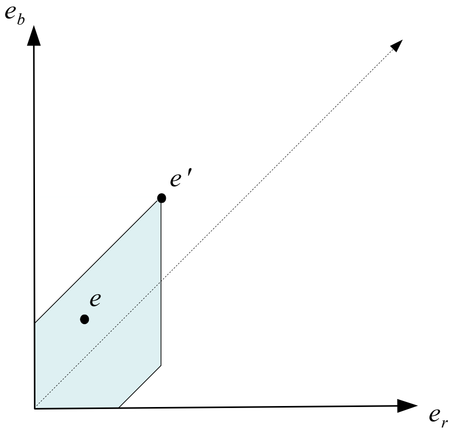
Definition

a **fairness-accuracy preference** $\succeq$ is any total order on $\mathbb{R}^2$ such that $e \succ e'$ whenever $e >_{FA} e'$

fairness-accuracy dominance



fairness-accuracy dominance



set of error pairs that **all** designers agree improve upon e'

example FA preferences

1. **utilitarian:** $w_u(e_r, e_b) = -p_r e_r - p_b e_b$ where p_r and p_b are the proportions of either group
   - generalizations of this rule put other weights on the two groups (Charness and Rabin, 2002; Dworczak et al., 2021)

example FA preferences

1. **utilitarian:** $w_u(e_r, e_b) = -p_r e_r - p_b e_b$ where p_r and p_b are the proportions of either group
   - generalizations of this rule put other weights on the two groups (Charness and Rabin, 2002; Dworczak et al., 2021)

2. **egalitarian:** order errors by $-|e_r - e_b|$, break ties using w_u
   - related formulation used in "difference aversion" models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000)

example FA preferences

1. **utilitarian:** $w_u(e_r, e_b) = -p_r e_r - p_b e_b$ where p_r and p_b are the proportions of either group
   - generalizations of this rule put other weights on the two groups (Charness and Rabin, 2002; Dworczak et al., 2021)

2. **egalitarian:** order errors by $-|e_r - e_b|$, break ties using w_u
   - related formulation used in "difference aversion" models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000)

3. **rawlsian**: order errors by $-\max\{e_r, e_b\}$, break ties using w_u

example FA preferences

1. **utilitarian:** $w_u(e_r, e_b) = -p_r e_r - p_b e_b$ where p_r and p_b are the proportions of either group
   - generalizations of this rule put other weights on the two groups (Charness and Rabin, 2002; Dworczak et al., 2021)

2. **egalitarian:** order errors by $-|e_r - e_b|$, break ties using w_u
   - related formulation used in "difference aversion" models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000)

3. **rawlsian**: order errors by $-\max\{e_r, e_b\}$, break ties using w_u

4. **constrained optimization** (e.g., Hardt et al., 2016):

$$\min_{a:\mathcal{X}\to\Delta(\mathcal{D})} \quad p_r e_r(a) + p_b e_b(a) \quad \text{s.t. } |e_r(a) - e_b(a)| \leq \varepsilon$$

fairness-accuracy frontier

Definition

the **feasible set** given X is

$$\mathcal{E}(X) := \{(e_r(a), e_b(a)) : a \in \mathcal{A}_X\}$$

where $\mathcal{A}_X$ is the set of all algorithms $a : \mathcal{X} \to \Delta(\mathcal{D})$

fairness-accuracy frontier

Definition

the **feasible set** given X is

$$\mathcal{E}(X) := \{(e_r(a), e_b(a)) : a \in \mathcal{A}_X\}$$

where $\mathcal{A}_X$ is the set of all algorithms $a : \mathcal{X} \to \Delta(\mathcal{D})$

Definition

the **fairness-accuracy frontier** given X is

$$\mathcal{F}(X) := \left\{e \in \mathcal{E}(X) : \nexists\ e' \in \mathcal{E}(X) \text{ s.t. } e' \succ_{FA} e\right\}$$

fairness-accuracy frontier

Definition

the **feasible set** given X is

$$\mathcal{E}(X) := \{(e_r(a), e_b(a)) : a \in \mathcal{A}_X\}$$

where $\mathcal{A}_X$ is the set of all algorithms $a : \mathcal{X} \to \Delta(\mathcal{D})$

Definition

the **fairness-accuracy frontier** given X is

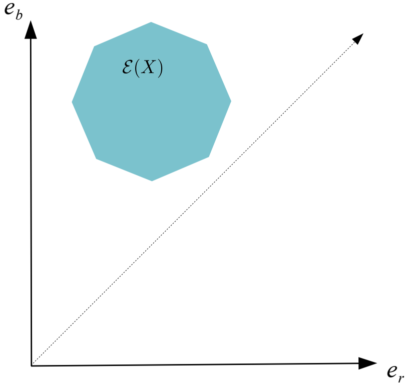
$$\mathcal{F}(X) := \left\{e \in \mathcal{E}(X) : \nexists\ e' \in \mathcal{E}(X) \text{ s.t. } e' \succ_{FA} e\right\}$$

- describes optimal points across the broad range of preferences consistent with FA-dominance

feasible set of group error pairs

lemma: for any X, the feasible set $\mathcal{E}(X)$ is compact and convex

(if $\mathcal{X}$ is finite, it is a convex polygon)



important points

group-optimal points:

$$R_X := \arg\min_{e \in \mathcal{E}(X)} e_r \qquad B_X := \arg\min_{e \in \mathcal{E}(X)} e_b$$

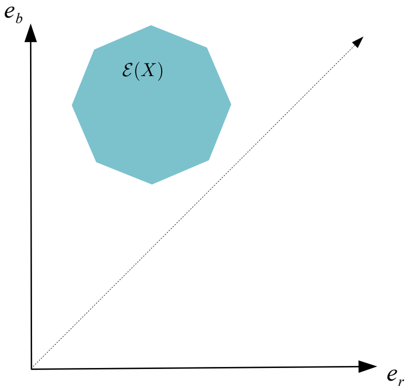
fairness-maximizing point:

$$F_X := \arg\min_{e \in \mathcal{E}(X)} |e_r - e_b|$$

(break all ties in favor of aggregate accuracy)

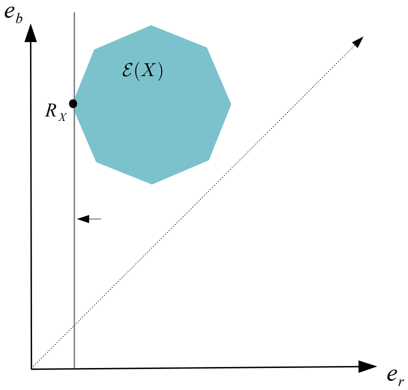
important points

easy to locate geometrically:



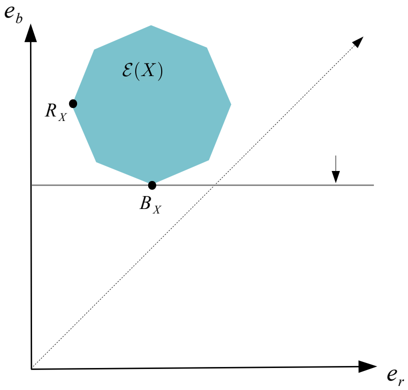
important points

easy to locate geometrically:



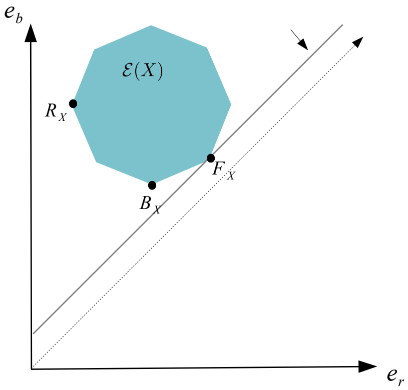
important points

easy to locate geometrically:



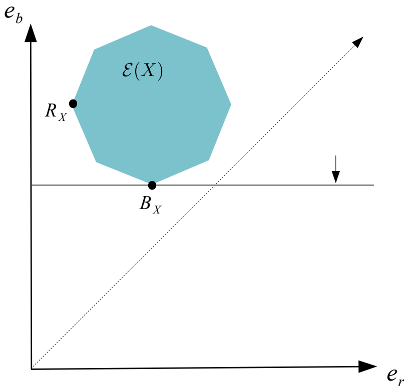
important points

easy to locate geometrically:



important points

easy to locate geometrically:



group-skewed vs group-balanced

Definition

covariate vector X is

- r-**skewed** if $e_r < e_b$ at R_X and $e_r \leq e_b$ at B_X

  "group r's error is lower both at group r's favorite point and also at group b's favorite point"

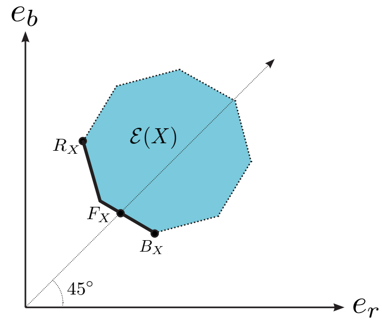
- b-**skewed** if $e_b < e_r$ at B_X and $e_b \leq e_r$ at R_X
- **group-balanced** otherwise

characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced*

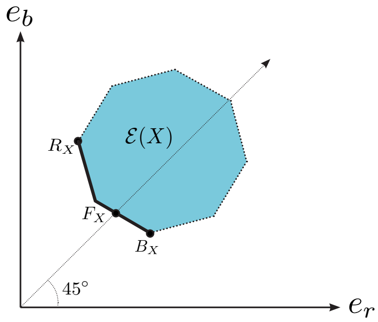


characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced* (= usual Pareto frontier!)

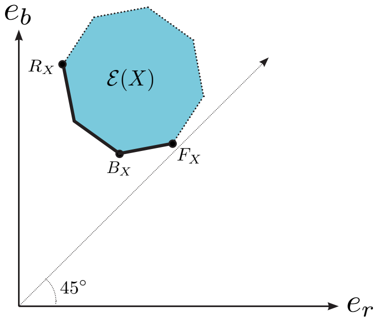


characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced*
- *G_X and F_X if X is g-skewed*

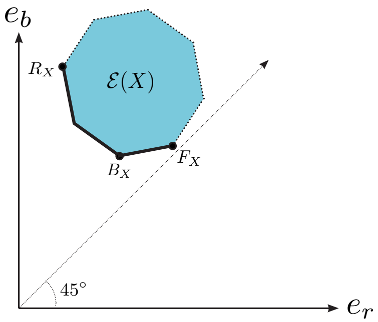


characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced*
- *G_X and F_X if X is g-skewed* (usual Pareto frontier + more)



strong fairness-accuracy conflict

- compare the error pairs $e = (1/2, 1/2)$ and $e' = (1/3, 1/4)$
  - e is Pareto-dominated but more equal

strong fairness-accuracy conflict

- compare the error pairs $e = (1/2, 1/2)$ and $e' = (1/3, 1/4)$
  - e is Pareto-dominated but more equal
- when given choices between allocations like e and e', some experimental subjects choose e
  - 31% of subjects in an experiment in Fisman et al. (2007)

strong fairness-accuracy conflict

- compare the error pairs $e = (1/2, 1/2)$ and $e' = (1/3, 1/4)$
  - e is Pareto-dominated but more equal
- when given choices between allocations like e and e', some experimental subjects choose e
  - 31% of subjects in an experiment in Fisman et al. (2007)
- can points like e and e' both be on the FA frontier?

strong fairness-accuracy conflict

- compare the error pairs $e = (1/2, 1/2)$ and $e' = (1/3, 1/4)$
  - e is Pareto-dominated but more equal
- when given choices between allocations like e and e', some experimental subjects choose e
  - 31% of subjects in an experiment in Fisman et al. (2007)
- can points like e and e' both be on the FA frontier?

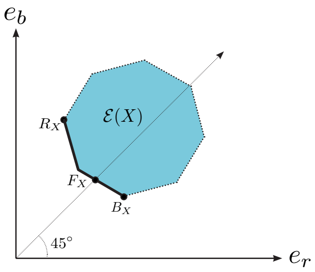
Definition

e, e' are a **strong accuracy-fairness conflict** if

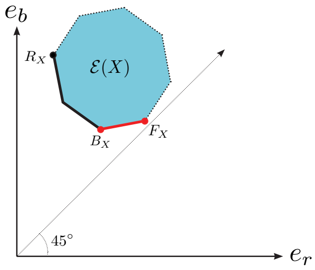
- $e_r \leq e'_r$ and $e_b \leq e'_b$ (with at least one inequality strict)
- $|e_r - e_b| > |e'_r - e'_b|$

strong fairness-accuracy conflict

corollary: suppose $F_X \notin \{R_X, B_X\}$; then X is group-skewed $\iff$ some $e, e' \in \mathcal{F}(X)$ represent a strong accuracy-fairness conflict



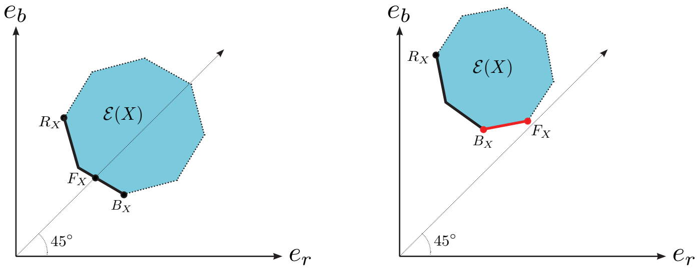
(a) X is group-balanced



(b) X is r-skewed

strong fairness-accuracy conflict

corollary: suppose $F_X \notin \{R_X, B_X\}$; then X is group-skewed $\iff$ some $e, e' \in \mathcal{F}(X)$ represent a strong accuracy-fairness conflict



(a) X is group-balanced (b) X is r-skewed

in practice, moving up that red line could correspond to choosing not to condition on certain available information

POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"**Yes**, they are **group-balanced**."

POLICYMAKER

“I have a policy proposal, which would decrease accuracy for both groups, but increase fairness.”

ACADEMIC

“Are the inputs to your algorithm group-balanced?”

POLICYMAKER

“**Yes**, they are **group-balanced**.”

ACADEMIC

“Your proposal is not optimal for you by your own preferences, **regardless** of how you tradeoff fairness and accuracy.”

POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"**No**, they are **group-skewed**."

POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"**No**, they are **group-skewed**."

ACADEMIC

"If you care sufficiently about fairness relative to accuracy, then your proposal **may be optimal** for your goals."

which of group balance and group skew is more common?

difficult to anticipate without an empirical analysis

which of group balance and group skew is more common?

difficult to anticipate without an empirical analysis

why might X be **group-balanced**?

- X has a group-dependent meanings
  - high X implies high Y for group r, but low Y for group b
- different inputs in X are informative for either group
  - $X = (X_1, X_2)$ where X_1 is uninformative about Y for group r and X_2 is uninformative about Y for group b

which of group balance and group skew is more common?

difficult to anticipate without an empirical analysis

why might X be **group-balanced**?

- X has a group-dependent meanings
  - high X implies high Y for group r, but low Y for group b
- different inputs in X are informative for either group
  - $X = (X_1, X_2)$ where X_1 is uninformative about Y for group r and X_2 is uninformative about Y for group b

why might X be **group-skewed**?

- X is asymmetrically informative
  - $Y \mid X, G = r$ more dispersed than $Y \mid X, G = b$
- e.g., medical data is recorded more accurately for high-income patients than low-income patients

generalizations

beyond absolute difference

- results extend when unfairness is measured as $|\phi(e_r) - \phi(e_b)|$ where ϕ is some continuous strictly increasing function
- if ϕ is log, then this corresponds to evaluating fairness using the ratio of errors rather than their difference

generalizations

beyond absolute difference

- results extend when unfairness is measured as $|\phi(e_r) - \phi(e_b)|$ where ϕ is some continuous strictly increasing function
- if ϕ is log, then this corresponds to evaluating fairness using the ratio of errors rather than their difference

different loss functions for evaluating fairness and accuracy

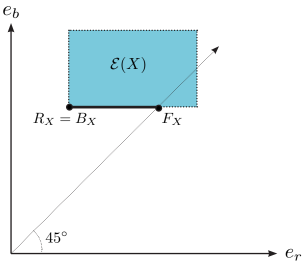
- qualitative result extends whenever the two loss functions aren't "directly opposed"
- group-balance generalizes to whether F_X belongs to usual Pareto frontier
  - X is group-balanced $\Longrightarrow$ FA frontier is usual Pareto frontier
  - X fails group-balance $\Longrightarrow$ FA frontier is union of the Pareto frontier and a positively-sloped sequence of lines

special case: X reveals G

in special cases, the frontier simplifies further.

Proposition

suppose $G \mid X$ is degenerate; then, $\mathcal{E}(X)$ is a rectangle with sides parallel to axes and $\mathcal{F}(X)$ is the line segment from $R_X = B_X$ to F_X

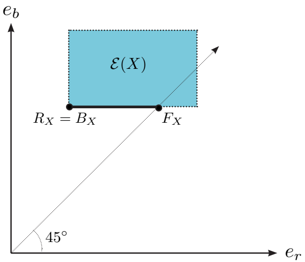


special case: X reveals G

in special cases, the frontier simplifies further.

Proposition

suppose $G \mid X$ is degenerate; then, $\mathcal{E}(X)$ is a rectangle with sides parallel to axes and $\mathcal{F}(X)$ is the line segment from $R_X = B_X$ to F_X

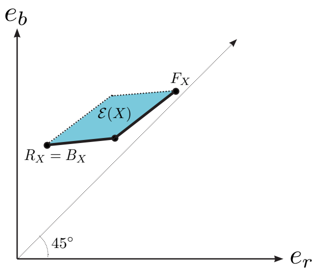


- frontier is **rawlsian**: worse-off group gets best feasible error

special case: conditional independence

Proposition

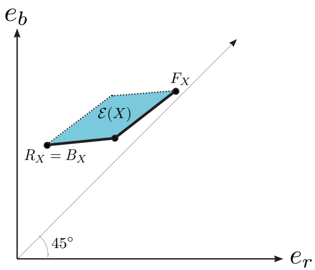
Suppose $G \perp\!\!\!\perp Y \mid X$; then, $\mathscr{F}(X)$ is that part of the lower boundary of the feasible set from the point $B_X = R_X$ to the point F_X.



special case: conditional independence

Proposition

Suppose $G \perp\!\!\!\perp Y \mid X$; then, $\mathscr{F}(X)$ is that part of the lower boundary of the feasible set from the point $B_X = R_X$ to the point F_X.



- the only difference across designers that matters is how they choose to resolve strong fairness-accuracy conflicts

part ii:

designer regulates inputs

regulating algorithmic inputs

- so far we've given the designer control of the algorithm

regulating algorithmic inputs

- so far we've given the designer control of the algorithm
- in practice sometimes
  - the algorithm is set by an agent who does not care about fairness across groups
  - the inputs used by the algorithm are constrained by a designer who does

regulating algorithmic inputs

- so far we've given the designer control of the algorithm
- in practice sometimes
  - the algorithm is set by an agent who does not care about fairness across groups
  - the inputs used by the algorithm are constrained by a designer who does
- e.g., in 1997, Berkeley law school administrators reported to their admissions committee only a coarsened LSAT score
  (Chan and Eyster, 2003)

regulating algorithmic inputs

- so far we've given the designer control of the algorithm
- in practice sometimes
  - the algorithm is set by an agent who does not care about fairness across groups
  - the inputs used by the algorithm are constrained by a designer who does
- e.g., in 1997, Berkeley law school administrators reported to their admissions committee only a coarsened LSAT score
  (Chan and Eyster, 2003)
- we'll model this as an information design problem

input design model

there is a primitive covariate vector X

input design model

there is a primitive covariate vector X

- $t = 1$: the designer chooses a garbling of X, i.e., a stochastic map $T : \mathcal{X} \to \Delta(\mathcal{T})$

input design model

there is a primitive covariate vector X

- $t = 1$: the designer chooses a garbling of X, i.e., a stochastic map $T : \mathcal{X} \to \Delta(\mathcal{T})$
- $t = 2$: the agent chooses the algorithm $a : \mathcal{T} \to \Delta(\mathcal{D})$ that maximizes the utilitarian criterion

input design model

there is a primitive covariate vector X

- $t = 1$: the designer chooses a garbling of X, i.e., a stochastic map $T : \mathcal{X} \to \Delta(\mathcal{T})$
- $t = 2$: the agent chooses the algorithm $a : \mathcal{T} \to \Delta(\mathcal{D})$ that maximizes the utilitarian criterion

Definition

the **input design feasible set** given X is

$$\mathcal{E}^* (X) := \{e (a_T) \ : \ T \text{ is a garbling of } X\}$$

where a_T denotes the utilitarian-optimal algorithm given T

input design model

there is a primitive covariate vector X

- $t = 1$: the designer chooses a garbling of X, i.e., a stochastic map $T : \mathcal{X} \to \Delta(\mathcal{T})$
- $t = 2$: the agent chooses the algorithm $a : \mathcal{T} \to \Delta(\mathcal{D})$ that maximizes the utilitarian criterion

Definition

the **input design feasible set** given X is

$$\mathcal{E}^*(X) := \{e(a_T) \ : \ T \text{ is a garbling of } X\}$$

where a_T denotes the utilitarian-optimal algorithm given T

Definition

the **input design fairness-accuracy frontier** given X, denoted $\mathcal{F}^*(X)$, is the set of all FA-undominated points in $\mathcal{E}^*(X)$

example garblings

real examples of such garblings are abundant

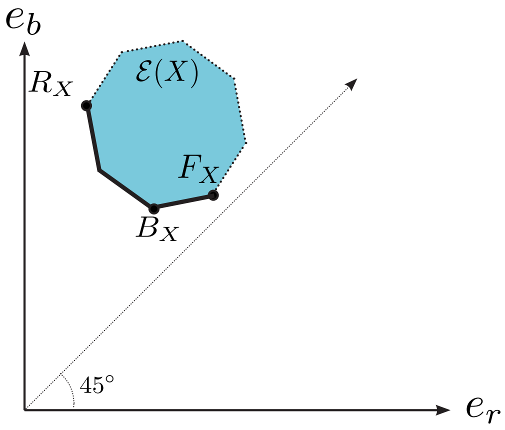
- **drop an input:**
  - "Ban the Box" campaign prohibited employers from asking about criminal history (Agan and Starr, 2018)
  - some researchers advocate for race-blind algorithms in the context of health predictions (Manski, 2022)
- **coarsen an input:**
  - essentially any test score
- **add noise:**
  - differential privacy initiatives adopted by the US Census Bureau, Apple, and Google

input design versus control of the algorithm

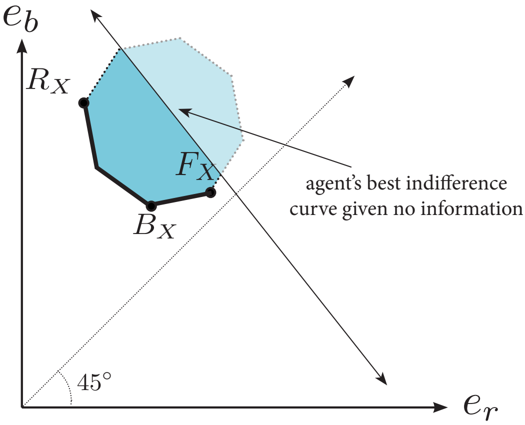
we'll ask two questions:

- how powerful is input design relative to control of the algorithm?
- could it be optimal for the designer to exclude an input altogether?

how powerful is input design?

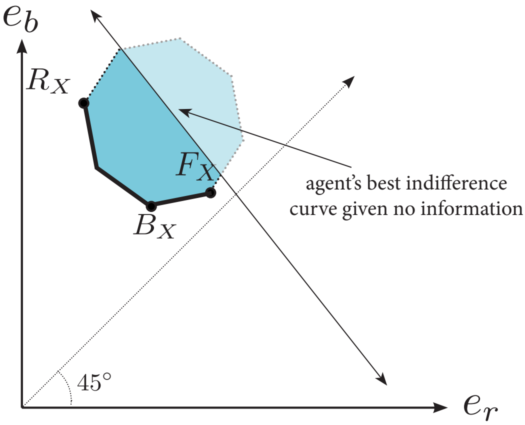


how powerful is input design?



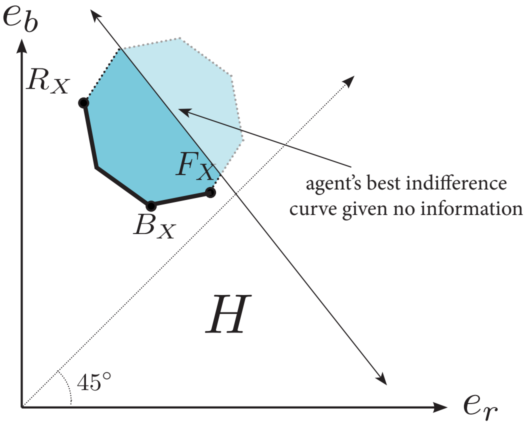
let e_0 be the minimal achievable aggregate error given no information

how powerful is input design?



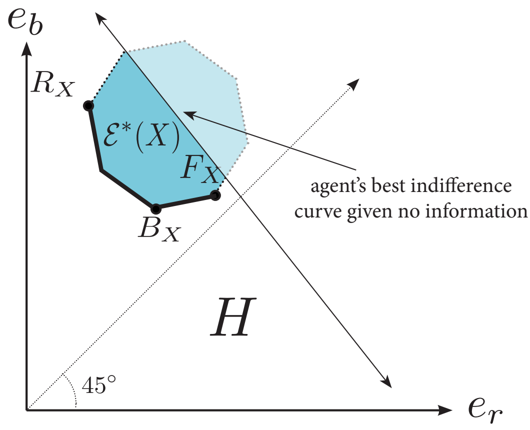
cannot force the agent to implement an error pair (e_r, e_b) satisfying $p_r e_r + p_b e_b > e_0$

how powerful is input design?



define $H = \{(e_r, e_b) : p_r e_r + p_b e_b \leq e_0\}$

how powerful is input design?



lemma: $\mathcal{E}^*(X) = \mathcal{E}(X) \cap H$

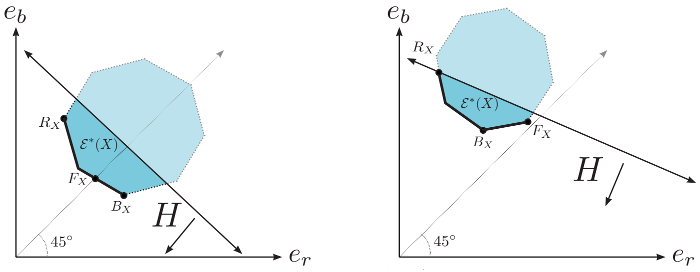
(see also Alonso and Camâra, 2016)

how powerful are informational constraints?

Proposition

(a) *If X is group-balanced, then $\mathcal{F}(X) = \mathcal{F}^*(X)$ iff $R_X, B_X \in H$*

(b) *If X is r-skewed, then $\mathcal{F}(X) = \mathcal{F}^*(X)$ iff $R_X, F_X \in H$*



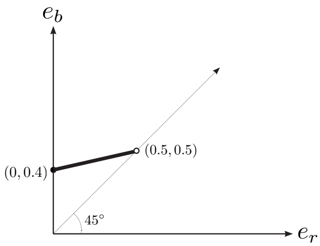
takeaway: under weak conditions, designer can implement favorite (unconstrained) outcome by designing the algorithmic inputs

add/ban covariates?

- constraints on algorithmic inputs sometimes take the form of a ban on use of a specific covariate
  - e.g., banning use of race in medical predictions, or banning test scores in college admissions
- because of misaligned preferences between the designer and agent, banning a covariate **can be optimal**

simple example where banning an input is optimal

- $Y \in \{0, 1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0, 1\}$ is a binary signal
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the designer is Egalitarian (payoff is $-|e_r - e_b|$)
- sending no information leads to a payoff of $|0.5 - 0.5| = 0$.
- sending **any** information about X leads to a negative payoff



uniform worsening of the frontier

at the other extreme:

Definition

excluding X' given X **uniformly worsens the frontier** if every point in $\mathcal{F}^*(X)$ is FA-dominated by a point in $\mathcal{F}^*(X, X')$

uniform worsening of the frontier

at the other extreme:

Definition

excluding X' given X **uniformly worsens the frontier** if every point in $\mathcal{F}^*(X)$ is FA-dominated by a point in $\mathcal{F}^*(X, X')$

- any point that belongs to $\mathcal{F}^*(X, X')$ but not to $\mathcal{F}^*(X)$ can only be implemented by sending information about X'

uniform worsening of the frontier

at the other extreme:

Definition

excluding X' given X **uniformly worsens the frontier** if every point in $\mathcal{F}^*(X)$ is FA-dominated by a point in $\mathcal{F}^*(X, X')$

- any point that belongs to $\mathcal{F}^*(X, X')$ but not to $\mathcal{F}^*(X)$ can only be implemented by sending information about X'
- condition guarantees that **no** designer's optimal garbling excludes X'

uniform worsening of the frontier

at the other extreme:

Definition

excluding X' given X **uniformly worsens the frontier** if every point in $\mathcal{F}^*(X)$ is FA-dominated by a point in $\mathcal{F}^*(X, X')$

- any point that belongs to $\mathcal{F}^*(X, X')$ but not to $\mathcal{F}^*(X)$ can only be implemented by sending information about X'
- condition guarantees that **no** designer's optimal garbling excludes X'

remark: this is different from comparing the information policies of completely revealing X versus completely revealing (X, X')

two comparisons

excluding group identity:

garblings of X vs. garblings of (X, G)

excluding an arbitrary covariate when G is present:

garblings of (X, G) vs. garblings of (X, G, X')

compare X to (X, G)

Proposition

suppose $R_X, B_X \in H$. excluding G uniformly worsens the frontier if and only if X is group-balanced

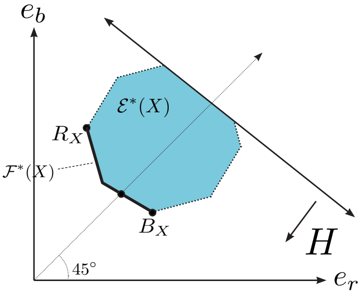


Figure: X is group-balanced

compare X to (X, G)

Proposition

suppose $R_X, B_X \in H$. excluding G uniformly worsens the frontier if and only if X is group-balanced

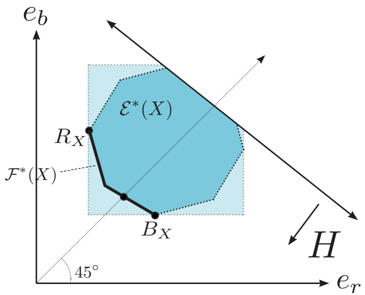


Figure: X is group-balanced

compare X to (X, G)

Proposition

suppose $R_X, B_X \in H$. excluding G uniformly worsens the frontier if and only if X is group-balanced

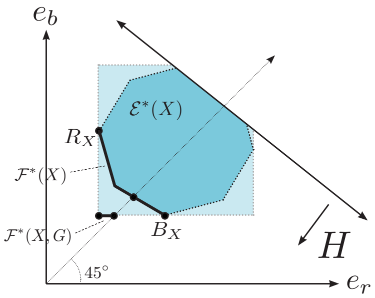


Figure: X is group-balanced

compare X to (X, G)

Proposition

suppose $R_X, B_X \in H$. excluding G uniformly worsens the frontier if and only if X is group-balanced

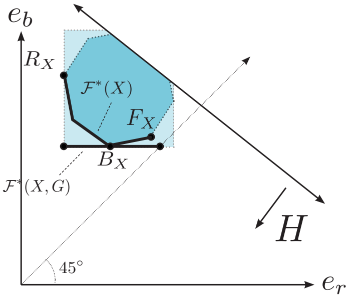


Figure: X is group-skewed

takeaways

- so long as X is group-balanced, then every designer can find a way of combining the information in G and X that is superior to sending information about X alone
- echoes previous findings that **disparate treatment** may be necessary to preclude **disparate impact**
  - Lundberg (1991), Chan and Eyster (2003), Ellison and Pathak (2022)
- here disparate treatment is via an asymmetric information policy rather than through the algorithm itself

compare (X, G) to (X, G, X')

definition: say that X' is **decision-relevant over** X **for group** g if there are realizations (x, x') and $(x, \tilde{x}')$ of (X, X') such that

$$\{1\} = \operatorname*{argmin}_{d \in \mathcal{D}} \mathbb{E}[\ell(d, Y) \mid X = x, X' = x', G = g]$$

while

$$\{0\} = \operatorname*{argmin}_{d \in \mathcal{D}} \mathbb{E}[\ell(d, Y) \mid X = x, X' = \tilde{x}', G = g]$$

i.e., the additional information in X' changes the optimal assignment for some individual in group g relative to X alone

compare (X, G) to (X, G, X')

Proposition

(a) *suppose* (X, G) *is* g*-skewed. then:*

excluding X' *given* (X, G) *uniformly worsens the frontier* $\iff$ X' *is decision-relevant over* X *for group* $g' \neq g$.

(b) *suppose* (X, G) *is group-balanced. then:*

excluding X' *given* (X, G) *uniformly worsens the frontier* $\iff$ X' *is decision-relevant over* X *for both groups.*

takeaways

consider the question of whether to ban test scores in admissions decisions

takeaways

consider the question of whether to ban test scores in admissions decisions

test scores are likely to be decision-relevant for both groups, so our result suggests that:

- if G is available, then excluding test scores is welfare-reducing for all designers with the ability to garble available covariates
- if G is not available, then it may be better for a sufficiently fairness-minded designer to completely exclude test scores

takeaways

consider the question of whether to ban test scores in admissions decisions

test scores are likely to be decision-relevant for both groups, so our result suggests that:

- if G is available, then excluding test scores is welfare-reducing for all designers with the ability to garble available covariates
- if G is not available, then it may be better for a sufficiently fairness-minded designer to completely exclude test scores

if affirmative action is banned nationwide, then universities with certain preferences may have reason to ban use of test scores

takeaways

- our framework abstracts away from many important features of the college admissions process
- but the link between the availability of G and the value of additional information holds more generally
- access to group identity has a positive spillover effect for the value of other covariates
- there is always some group-dependent garbling of the other information that aligns the agent and designer's incentives.

related literature

equity-efficiency tradeoffs: taxation (Saez and Stantcheva, 2016; Dworczak et al., 2021), policing (Persico, 2002; Jung et al., 2020), admissions (Chan and Eyster, 2003; Ellison and Pathak, 2021)

related literature

equity-efficiency tradeoffs: taxation (Saez and Stantcheva, 2016; Dworczak et al., 2021), policing (Persico, 2002; Jung et al., 2020), admissions (Chan and Eyster, 2003; Ellison and Pathak, 2021)

$\longrightarrow$ we focus on a special equity-efficiency tradeoff motivated by...

algorithmic bias:

- empirical documentation of the disparate impact of algorithms (Obermeyer et al., 2019; Arnold et al., 2021; Fuster et al., 2021)
- much of the theoretical literature posits a particular objective criterion (Roth and Kearns, 2019), engagement with fairness accuracy tradeoffs in special cases (Menon and Williamson, 2018)

related literature

equity-efficiency tradeoffs: taxation (Saez and Stantcheva, 2016; Dworczak et al., 2021), policing (Persico, 2002; Jung et al., 2020), admissions (Chan and Eyster, 2003; Ellison and Pathak, 2021)

⟶ we focus on a special equity-efficiency tradeoff motivated by...

algorithmic bias:

- empirical documentation of the disparate impact of algorithms (Obermeyer et al., 2019; Arnold et al., 2021; Fuster et al., 2021)
- much of the theoretical literature posits a particular objective criterion (Roth and Kearns, 2019), engagement with fairness accuracy tradeoffs in special cases (Menon and Williamson, 2018)

⟶ we provide general results for how this tradeoff is moderated by the inputs to the algorithm, and also...

info design: model the design of algorithm inputs as information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019)

conclusion

- framework for evaluating the accuracy/fairness tradeoffs of algorithms
- characterized the fairness-accuracy frontier over different designer preferences for how to trade off these criteria
- explained how certain statistical properties of the algorithm's inputs impact the shape of this frontier
- in some cases (e.g., when the inputs are group-balanced), there are conclusions/policy recommmendations that hold **for all** designer preferences in a broad class

thank you

simple example where banning an input is optimal

- $Y \in \{0, 1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0, 1\}$ is a binary signal
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the designer is Egalitarian (payoff is $-|e_r - e_b|$)

simple example where banning an input is optimal

- $Y \in \{0, 1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0, 1\}$ is a binary signal
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the designer is Egalitarian (payoff is $-|e_r - e_b|$)
- sending no information leads to a payoff of $|0.5 - 0.5| = 0$.