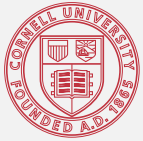
Identification and Estimation for Algorithmic Frontiers with Selective Labels

Yiqi Liu (Harvard CMSA), Francesca Molinari (Johns Hopkins), Amilcar Velez



LACEA–LAMES 2026 Annual Meeting, Lima, Peru

November 13, 2026

Motivation



- Algorithms play an increasingly prevalent role in everyday life.

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
  - Medical treatment, bail decisions, college admissions, lending decisions, hiring...

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
- Regulatory and oversight processes have emerged in response, as *AI governance*.
  - Design, deployment, documentation and monitoring of AI systems.

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
- Regulatory and oversight processes have emerged in response, as *AI governance*.
- Fairness and accuracy evaluations play a central role in AI governance.

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
- Regulatory and oversight processes have emerged in response, as *AI governance*.
- Fairness and accuracy evaluations play a central role in AI governance.
- Regulators may ask:

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
- Regulatory and oversight processes have emerged in response, as *AI governance*.
- Fairness and accuracy evaluations play a central role in AI governance.
- Regulators may ask:
  - For a class of algorithms: When does the **tradeoff** between fairness and accuracy arise?

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
- Regulatory and oversight processes have emerged in response, as *AI governance*.
- Fairness and accuracy evaluations play a central role in AI governance.
- Regulators may ask:
  - For a class of algorithms: When does the **tradeoff** between fairness and accuracy arise?
  - For a specific algorithm: Are there **less discriminatory alternatives** (LDA)?

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
- Regulatory and oversight processes have emerged in response, as *AI governance*.
- Fairness and accuracy evaluations play a central role in AI governance.
- Regulators may ask:
  - For a class of algorithms: When does the **tradeoff** between fairness and accuracy arise?
  - For a specific algorithm: Are there **less discriminatory alternatives** (LDA)?
  - What can we learn if the outcome is **selectively observed**?

Motivation

- Algorithms play an increasingly prevalent role in everyday life.
- Regulatory and oversight processes have emerged in response, as *AI governance*.
- Fairness and accuracy evaluations play a central role in AI governance.
- Regulators may ask:
  - For a class of algorithms: When does the **tradeoff** between fairness and accuracy arise?
  - For a specific algorithm: Are there **less discriminatory alternatives** (LDA)?
  - What can we learn if the outcome is **selectively observed**?
- How can we answer these questions with only **finite samples**?

What We Do

- We extend the theoretical framework of Liang, Lu, Mu & Okumura (2026) [LLMO] by:
  - Characterizing a **fairness-accuracy (FA) frontier** w/different loss functions for fairness and accuracy.
  - Providing IFF condition for when improving fairness does **not** come at the cost of reducing accuracy.

What We Do

- We extend the theoretical framework of Liang, Lu, Mu & Okumura (2026) [LLMO] by:
  - Characterizing a **fairness-accuracy (FA) frontier** w/different loss functions for fairness and accuracy.
  - Providing IFF condition for when improving fairness does **not** come at the cost of reducing accuracy.
- We extend the inference methods of Liu & Molinari (2025) and allow selectively observed Y^* by:
  - Deriving **sharp identification region** for FA frontier with binary Y^* and specific loss functions.
  - **Point-identifying** FA frontier under missing-at-random assumption for generic Y^* and losses.
    - ⋆ Providing DML estimator for FA frontier using inverse propensity score weighting.
    - ⋆ Testing whether an LDA exists for a given algorithm.
    - ⋆ Constructing a confidence set for the FA frontier.

What We Do

- We extend the theoretical framework of Liang, Lu, Mu & Okumura (2026) [LLMO] by:
  - Characterizing a **fairness-accuracy (FA) frontier** w/different loss functions for fairness and accuracy.
  - Providing IFF condition for when improving fairness does **not** come at the cost of reducing accuracy.
- We extend the inference methods of Liu & Molinari (2025) and allow selectively observed Y^* by:
  - Deriving **sharp identification region** for FA frontier with binary Y^* and specific loss functions.
  - **Point-identifying** FA frontier under missing-at-random assumption for generic Y^* and losses.
    - ⋆ Providing DML estimator for FA frontier using inverse propensity score weighting.
    - ⋆ Testing whether an LDA exists for a given algorithm.
    - ⋆ Constructing a confidence set for the FA frontier. Not today.

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by:

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by: $(Y^*$ $)$
  - Outcome $Y^* \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ (number of active chronic conditions and medical costs)

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by: $(Y^*, G \quad)$
  - Group $G \in \{r, b\}$ (race)

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by: (Y^*, G, X)
  - Covariate inputs $X \in \mathcal{X}$ (age, gender, comorbidity and medication variables, biomarkers, ...)

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by: (Y^*, G, X)
- Each individual receives a binary **decision** $d \in \{0, 1\}$.
  (whether automatically enrolled in high-risk care management program)

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by: (Y^*, G, X)
- Each individual receives a binary **decision** $d \in \{0, 1\}$.
- **Algorithm** $a : \mathcal{X} \to [0, 1]$ predicts the probability of an event.
  (whether benefit from treatment is large enough)

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by: (Y^*, G, X)
- Each individual receives a binary **decision** $d \in \{0, 1\}$.
- **Algorithm** $a : \mathcal{X} \to [0, 1]$ predicts the probability of an event.
  - This prediction determines the probability with which the decision equals 1.

Setup and Notation

Ideal case: suppose no selective label problem

We extend the framework proposed by LLMO.

- A population of individuals described by: (Y^*, G, X)
- Each individual receives a binary **decision** $d \in \{0, 1\}$.
- **Algorithm** $a : \mathcal{X} \to [0, 1]$ predicts the probability of an event.
  - This prediction determines the probability with which the decision equals 1.
- Evaluate the **accuracy** (**fairness**) of a through $\ell^A : \{0, 1\} \times \mathcal{Y} \to \mathbb{R}$ ($\ell^F : \{0, 1\} \times \mathcal{Y} \to \mathbb{R}$).
  - For example, $\ell^A(d, Y^*) = \mathbb{1}\{Y_1^* \neq d\}$ ($\ell^F(d, Y^*) = \mathbb{1}\{d = 1\}$).

Feasible Set

- The **group accuracy** of algorithm a for group g is defined as

$$e_g^A(a) := \mathbb{E}\left[a(X) \cdot \ell^A(1, Y^*) + (1 - a(X)) \cdot \ell^A(0, Y^*) | G = g\right]$$

- The **fairness** of algorithm a is measured by $\left|e_r^F(a) - e_b^F(a)\right|$, where

$$e_g^F(a) := \mathbb{E}\left[a(X) \cdot \ell^F(1, Y^*) + (1 - a(X)) \cdot \ell^F(0, Y^*) | G = g\right]$$

- The algorithm designer chooses an algorithm, yielding a **fairness-accuracy** allocation

$$\varepsilon(a) := \left(e_r^A(a), e_b^A(a), e_r^F(a) - e_b^F(a)\right).$$

Feasible Set

- The **group accuracy** of algorithm a for group g is defined as

$$e_g^A(a) := \mathbb{E}\left[a(X) \cdot \ell^A(1, Y^*) + (1 - a(X)) \cdot \ell^A(0, Y^*) | G = g\right]$$

- The **fairness** of algorithm a is measured by $\left|e_r^F(a) - e_b^F(a)\right|$, where

$$e_g^F(a) := \mathbb{E}\left[a(X) \cdot \ell^F(1, Y^*) + (1 - a(X)) \cdot \ell^F(0, Y^*) | G = g\right]$$

- The algorithm designer chooses an algorithm, yielding a **fairness-accuracy** allocation

$$\varepsilon(a) := \left(e_r^A(a), e_b^A(a), e_r^F(a) - e_b^F(a)\right).$$

- Given $\mathcal{A}(\mathcal{X}) := \{a : \mathcal{X} \to [0, 1]\}$, the **feasible set** of fairness-accuracy allocations is defined as

$$\mathcal{E} := \{\varepsilon(a) : a \in \mathcal{A}(\mathcal{X})\}$$

- $\mathcal{E}$ is a **closed and convex** set in $\mathbb{R}^3$.

Preferences & FA-Frontier

The designer has preferences over fairness-accuracy allocations.

- $\varepsilon = \left(e_r^A, e_b^A, e_r^F - e_b^F\right)$ and $\tilde{\varepsilon} = \left(\tilde{e}_r^A, \tilde{e}_b^A, \tilde{e}_r^F - \tilde{e}_b^F\right)$

Preferences & FA-Frontier

The designer has preferences over fairness-accuracy allocations.

- $\varepsilon = \left(e_r^A, e_b^A, e_r^F - e_b^F\right)$ and $\tilde{\varepsilon} = \left(\tilde{e}_r^A, \tilde{e}_b^A, \tilde{e}_r^F - \tilde{e}_b^F\right)$

- ε is FA-preferred over $\tilde{\varepsilon}$ if

$$\underbrace{e_r^A \leq \tilde{e}_r^A,\ e_b^A \leq \tilde{e}_b^A}_{\varepsilon \text{ is more accurate}} \quad \text{and} \quad \underbrace{\left|e_r^F - e_b^F\right| \leq \left|\tilde{e}_r^F - \tilde{e}_b^F\right|}_{\varepsilon \text{ is fairer}}.$$

  with at least one inequality strict.

Preferences & FA-Frontier

The designer has preferences over fairness-accuracy allocations.

- $\varepsilon = \left(e_r^A, e_b^A, e_r^F - e_b^F\right)$ and $\tilde{\varepsilon} = \left(\tilde{e}_r^A, \tilde{e}_b^A, \tilde{e}_r^F - \tilde{e}_b^F\right)$
- ε is FA-preferred over $\tilde{\varepsilon}$ if

$$\underbrace{e_r^A \leq \tilde{e}_r^A,\ e_b^A \leq \tilde{e}_b^A}_{\varepsilon \text{ is more accurate}} \quad \text{and} \quad \underbrace{\left|e_r^F - e_b^F\right| \leq \left|\tilde{e}_r^F - \tilde{e}_b^F\right|}_{\varepsilon \text{ is fairer}}.$$

  with at least one inequality strict.

- Goal: Characterize the **Fairness-Accuracy Frontier**

$$\mathcal{F} := \{\varepsilon \in \mathcal{E} : \nexists \tilde{\varepsilon} \in \mathcal{E} \ s.t.\ \tilde{\varepsilon} \succ_{FA} \varepsilon\}$$

Preferences & Pareto Frontier

The designer has preferences over accuracy allocations.

- $\varepsilon = (e_r^A, e_b^A, e_r^F - e_b^F)$ and $\tilde{\varepsilon} = (\tilde{e}_r^A, \tilde{e}_b^A, \tilde{e}_r^F - \tilde{e}_b^F)$

- ε is Pareto preferred over $\tilde{\varepsilon}$ if

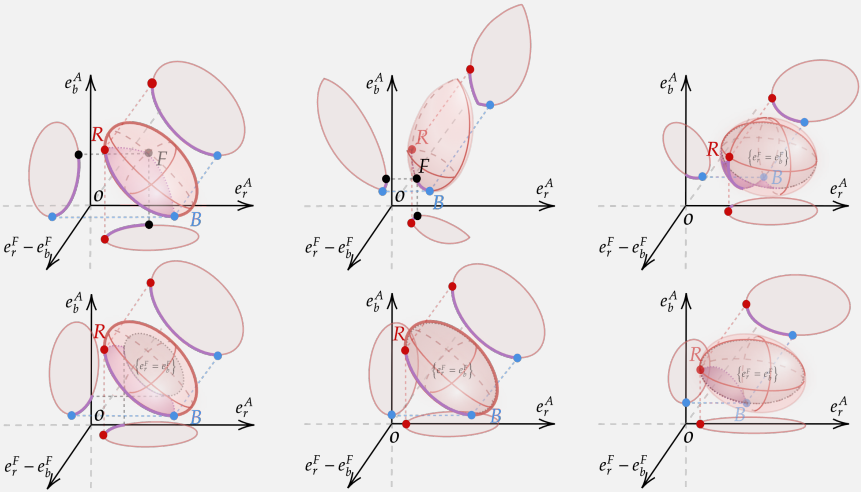
$$\underbrace{e_r^A \leq \tilde{e}_r^A,\ e_b^A \leq \tilde{e}_b^A}_{\varepsilon \text{ is more accurate}} \quad \text{and} \quad \underbrace{\left|e_r^F - e_b^F\right| \leq \left|\tilde{e}_r^F - \tilde{e}_b^F\right|}_{\varepsilon \text{ is fairer}}.$$

  with at least one inequality strict.

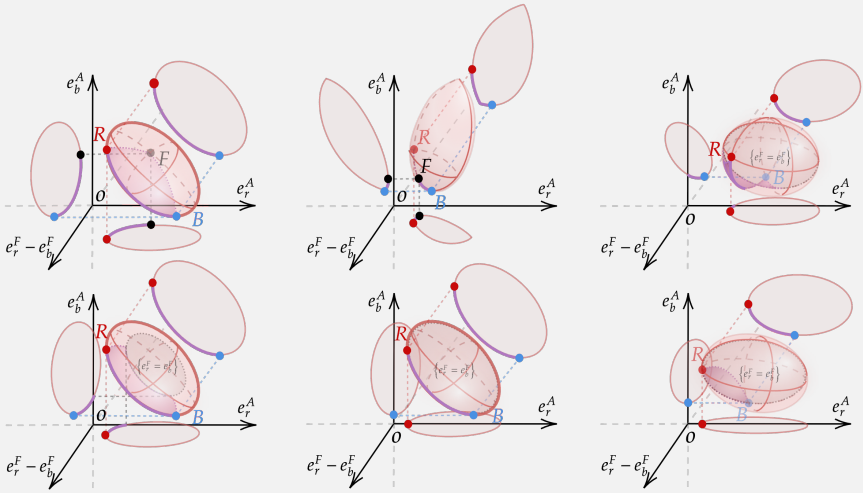
- Goal: Characterize the **Pareto Frontier**

$$\mathcal{PF} := \{\varepsilon \in \mathcal{E} : \nexists \tilde{\varepsilon} \in \mathcal{E} \ s.t.\ \tilde{\varepsilon} \succ_{PD} \varepsilon\}$$

Geometry of $\mathcal{F}$ in $\mathbb{R}^3$ Is...Complicated



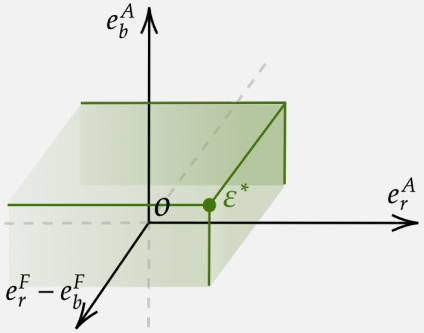
Geometry of $\mathcal{F}$ in $\mathbb{R}^3$ Is...Complicated



We provide a simple characterization of $\mathcal{F}$ **without** pre-testing which case we are in!

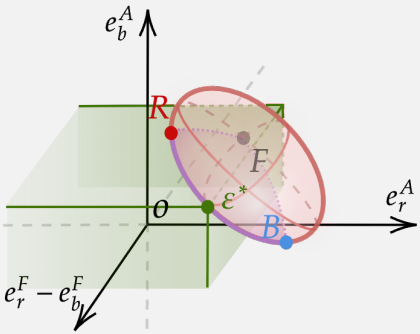
The FA Frontier via the Separating Hyperplane Theorem

- Define $\mathcal{C}(\varepsilon^*) := \{\varepsilon \in \mathbb{R}^3 : \varepsilon_1 \leq \varepsilon_1^*, \varepsilon_2 \leq \varepsilon_2^*, |\varepsilon_3| \leq |\varepsilon_3^*|\}$.

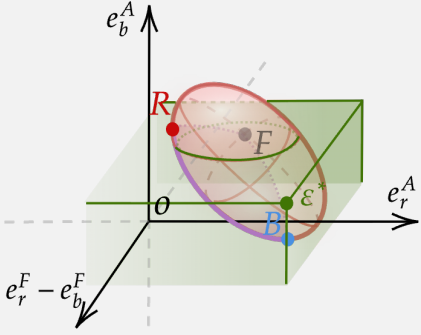


The FA Frontier via the Separating Hyperplane Theorem

- Define $\mathcal{C}(\varepsilon^*) := \{\varepsilon \in \mathbb{R}^3 : \varepsilon_1 \leq \varepsilon_1^*, \varepsilon_2 \leq \varepsilon_2^*, |\varepsilon_3| \leq |\varepsilon_3^*|\}$.



(a) ε^* is on $\mathcal{F}$



(b) ε^* is off $\mathcal{F}$

We characterize the set of points on $\mathcal{F}$ by judicious use of the **separating hyperplane theorem**.

The FA Frontier via the Support Function

- Let $h_{\mathcal{C}(\varepsilon^*)}(q) = \max_{\varepsilon \in \mathcal{C}(\varepsilon^*)} q^\intercal \varepsilon$ be the support function of $\mathcal{C}(\varepsilon^*)$
- Let $h_{\mathcal{E}}(q) = \max_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$ be the support function of $\mathcal{E}$

Theorem

When $\mathcal{E}$ is strictly convex, $\varepsilon^* \in \mathcal{F}$ if and only if there exists a hyperplane that properly separates $\mathcal{C}(\varepsilon^*)$ and $\mathcal{E}$, i.e., there exists $q \in \mathbb{S}^2$ such that

$$h_{\mathcal{C}(\varepsilon^*)}(q) = -h_{\mathcal{E}}(-q).$$

The FA Frontier via the Support Function

- Let $h_{\mathcal{C}(\varepsilon^*)}(q) = \max_{\varepsilon \in \mathcal{C}(\varepsilon^*)} q^\intercal \varepsilon$ be the support function of $\mathcal{C}(\varepsilon^*)$
- Let $h_{\mathcal{E}}(q) = \max_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$ be the support function of $\mathcal{E}$

Theorem

When $\mathcal{E}$ is strictly convex, $\varepsilon^* \in \mathcal{F}$ if and only if there exists a hyperplane that properly separates $\mathcal{C}(\varepsilon^*)$ and $\mathcal{E}$, i.e., there exists $q \in \mathbb{S}^2$ such that

$$h_{\mathcal{C}(\varepsilon^*)}(q) = -h_{\mathcal{E}}(-q).$$

We then have that

$$\mathcal{F} = \left\{ \varepsilon^* \in \mathcal{E} : \ \min_{q \in \mathbb{S}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \right) = 0 \right\}.$$

No need to distinguish between cases or pre-test!

Algorithms on the Frontier

Assume $\sup_{q \in \mathbb{S}^2} \mathbb{P}_X(q^\intercal(\boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X)) = 0) = 0$.

An application of the Neyman-Pearson lemma then gives,

$$\varepsilon(a) \in \partial\mathcal{E} \iff \exists q \in \mathbb{S}^2 : \; a(X) = \mathbb{1}\{q^\intercal(\boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X)) > 0\}, \; \mathbb{P}_X\text{-a.s.}$$

Algorithms on the Frontier

Assume $\sup_{q \in \mathbb{S}^2} \mathbb{P}_X(q^\intercal(\boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X)) = 0) = 0$.

An application of the Neyman-Pearson lemma then gives,

$$\varepsilon(a) \in \partial\mathcal{E} \iff \exists q \in \mathbb{S}^2 : \; a(X) = \mathbb{1}\{q^\intercal(\boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X)) > 0\}, \; \mathbb{P}_X\text{-a.s.}$$

Moving forward, we ask whether binary algorithms yield expected losses on $\mathcal{F}$.

Our sharpness claims account for the assumed smoothness in the distribution of $\boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X)$.

Selectively Observed Y^*

Observe $(Y, G, X, Z) \sim \mathbb{P}$ with $Y \equiv ZY^*$, $Z \in \{0, 1\}$

Selectively Observed Y^*

Observe $(Y, G, X, Z) \sim \mathbb{P}$ with $Y \equiv ZY^*$, $Z \in \{0, 1\}$

For simplicity, let:

$$Y^* \in \{0, 1\}$$

$$\ell^A(d, Y^*) = \mathbb{1}\{d \neq Y^*\} \quad \text{(classification loss)} \quad \rightarrowtail \quad \text{selectively observed!}$$

$$\ell^F(d, Y^*) = \mathbb{1}\{1 = d\} \quad \text{(statistical disparity loss)}$$

Selectively Observed Y^*

Observe $(Y, G, X, Z) \sim \mathbb{P}$ with $Y \equiv ZY^*$, $Z \in \{0, 1\}$

For simplicity, let:

$$Y^* \in \{0, 1\}$$

$$\ell^A(d, Y^*) = \mathbb{1}\{d \neq Y^*\} \quad \text{(classification loss)} \quad \rightarrowtail \quad \text{selectively observed!}$$

$$\ell^F(d, Y^*) = \mathbb{1}\{1 = d\} \quad \text{(statistical disparity loss)}$$

The following objects are *not* point identified:

Selectively Observed Y^*

Observe $(Y, G, X, Z) \sim \mathbb{P}$ with $Y \equiv ZY^*$, $Z \in \{0, 1\}$

For simplicity, let:

$$
\begin{aligned}
Y^* &\in \{0, 1\} \\
\ell^A(d, Y^*) &= \mathbb{1}\{d \neq Y^*\} \quad \text{(classification loss)} \quad \rightarrowtail \quad \text{selectively observed!} \\
\ell^F(d, Y^*) &= \mathbb{1}\{1 = d\} \quad \text{(statistical disparity loss)}
\end{aligned}
$$

The following objects are *not* point identified:

$\Rightarrow$ $\lambda_g(X) := \mathbb{P}[Y^* = 1 | G = g, Z = 0, X] \in [0, 1]$.

$\Rightarrow$ $h_{\mathcal{C}(\varepsilon^*;\lambda)}(q) = f_1(q, \lambda)$, where $f_1(\cdot)$ is linear in λ.

$\Rightarrow$ $h_{\mathcal{E}}(q; \lambda) = \max\{f_2(q, \lambda), f_3(q, \lambda)\}$, where $f_2(\cdot), f_3(\cdot)$ are linear in λ.

$\Rightarrow$ $\mathcal{F}(\lambda)$ – FA frontier associated with λ

Selective Labels: The Fundamental Challenge

Observe $(Y, G, X, Z) \sim \mathbb{P}$ with $Y \equiv ZY^*$, $Z \in \{0, 1\}$

The goal of the analysis is to ascertain properties of *algorithms*.

For given $a^* \in \mathcal{A}(\mathcal{X})$, the empirical evidence should come from $\varepsilon(a^*)$ and $\mathcal{E}$.

Selective Labels: The Fundamental Challenge

Observe $(Y, G, X, Z) \sim \mathbb{P}$ with $Y \equiv ZY^*$, $Z \in \{0, 1\}$

The goal of the analysis is to ascertain properties of *algorithms*.

For given $a^* \in \mathcal{A}(\mathcal{X})$, the empirical evidence should come from $\varepsilon(a^*)$ and $\mathcal{E}$.

Neither is point identified, **but** they are linked by the distribution of the missing data, λ.

Selective Labels: The Fundamental Challenge

Observe $(Y, G, X, Z) \sim \mathbb{P}$ with $Y \equiv ZY^*$, $Z \in \{0, 1\}$

The goal of the analysis is to ascertain properties of *algorithms*.

For given $a^* \in \mathcal{A}(\mathcal{X})$, the empirical evidence should come from $\varepsilon(a^*)$ and $\mathcal{E}$.

Neither is point identified, **but** they are linked by the distribution of the missing data, λ.

To determine if a^* yields expected losses on the FA frontier,

- Must use the *same* selection mechanism/completion,

$$\lambda_g(X) := \mathbb{P}[Y^* = 1 | G = g, Z = 0, X] \in [0, 1], \quad g \in \{r, b\}$$

  to obtain a candidate $\varepsilon(a^*; \lambda)$ and a candidate $\mathcal{E}(\lambda)$.

- Must loop over all possible $\lambda_g : \mathcal{X} \mapsto [0, 1]$, $g \in \{r, b\}$.

Algorithms on the FA Frontier



We reduce the task to a finite dimensional optimization problem

Theorem

For a given $a^* \in \{0, 1\}$,

Algorithms on the FA Frontier



We reduce the task to a finite dimensional optimization problem

Theorem

For a given $a^* \in \{0,1\}$, $\varepsilon(a^*) \in \mathcal{F}$ *for some admissible distribution of* Y^* *IFF* $\exists\, \lambda : \mathcal{X} \to [0,1]^2$ *s.t.*

$$\min_{q \in S^2} \left(h_{\mathcal{C}(\varepsilon(a^*))}(q;\lambda) + h_{\mathcal{E}}(-q;\lambda) \right) = 0.$$

Algorithms on the FA Frontier

We reduce the task to a finite dimensional optimization problem

Theorem

For a given $a^* \in \{0,1\}$, $\varepsilon(a^*) \in \mathcal{F}$ *for some admissible distribution of* Y^* *IFF* $\exists\, \lambda : \mathcal{X} \to [0,1]^2$ *s.t.*

$$\min_{q \in S^2} \left(h_{\mathcal{C}(\varepsilon(a^*))}(q;\lambda) + h_{\mathcal{E}}(-q;\lambda) \right) = 0.$$

This occurs if and only if, for $\lambda_g(x;a^*) = a^*(x),\ g \in \{r,b\}$,

Key insights:

- $\lambda(\cdot;a^*)$ is *least favorable* to finding an improvement over a^*.

Algorithms on the FA Frontier

We reduce the task to a finite dimensional optimization problem

Theorem

For a given $a^* \in \{0,1\}$, $\varepsilon(a^*) \in \mathcal{F}$ *for some admissible distribution of* Y^* *IFF* $\exists\, \lambda : \mathcal{X} \to [0,1]^2$ *s.t.*

$$\min_{q \in \mathbb{S}^2} \Big(h_{\mathcal{C}(\varepsilon(a^*))}(q;\lambda) + h_{\mathcal{E}}(-q;\lambda) \Big) = 0.$$

This occurs if and only if, for $\lambda_g(x;a^*) = a^*(x),\ g \in \{r,b\}$,

$$\min_{q \in \mathbb{S}^2} \Big(h_{\mathcal{C}(\varepsilon(a^*))}(q;\ \lambda(\cdot;a^*)\ ) + h_{\mathcal{E}}(-q;\ \lambda(\cdot;a^*)\ ) \Big) = 0.$$

Key insights:

- $\lambda(\cdot;a^*)$ is *least favorable* to finding an improvement over a^*.

Uniformly Improving Algorithms

When $\min_{q \in \mathbb{S}^2} \underbrace{\left(h_{\mathcal{C}(\varepsilon(a^*))}(q; \lambda(\cdot; a^*)) + h_{\mathcal{E}}(-q; \lambda(\cdot; a^*)) \right)}_{J(q; a^*)} > 0 \ldots$

Uniformly Improving Algorithms

When $\min_{q \in \mathbb{S}^2} \underbrace{\left(h_{\mathcal{C}(\varepsilon(a^*))}(q; \lambda(\cdot; a^*)) + h_{\mathcal{E}}(-q; \lambda(\cdot; a^*)) \right)}_{J(q; a^*)} > 0$...

...it means that for all $\lambda \in [0,1]^2$, $\exists\ a(\cdot; \lambda)$ that FA-improves over a^*.

Uniformly Improving Algorithms

When $\min_{q \in S^2} \underbrace{\left(h_{\mathcal{C}(\varepsilon(a^*))}(q; \lambda(\cdot; a^*)) + h_{\mathcal{E}}(-q; \lambda(\cdot; a^*)) \right)}_{J(q; a^*)} > 0$...

...it means that for all $\lambda \in [0,1]^2$, $\exists\ a(\cdot; \lambda)$ that FA-improves over a^*.

We show a much stronger result. Let $\widetilde{q} \in \arg\min_{q \in R^3:\ q_1, q_2 \geq 0,\ q_1 + q_2 = 1} J(q; a^*)$. Then,

$$\widetilde{a}(x) \equiv \mathbb{1}\{\widetilde{q}^\intercal \Delta\boldsymbol{\theta}(x; \lambda(\cdot; a^*)) < 0\},$$

strictly reduces both accuracy losses for every $\lambda \in [0,1]^2$ while weakly reducing statistical disparity.

Conclusion

- We extend the theoretical framework of Liang, Lu, Mu & Okumura (2026) [LLMO] by:
  - ✓ Characterizing **fairness-accuracy (FA) frontier** with different loss functions for fairness and accuracy.
  - ▶ Providing **IFF** condition for when improving fairness does **not** come at the cost of reducing accuracy
- We extend the inference methods of Liu & Molinari (2025) and allow selectively observed Y^* by:
  - ✓ Deriving **sharp characterization** for the FA frontier for binary Y^* and specific loss functions
  - ▶ **Point-identifying** the FA frontier using a missing-at-random assumption
    - ⋆ Provide a method to test whether an LDA exists for a given algorithm

Conclusion

- We extend the theoretical framework of Liang, Lu, Mu & Okumura (2026) [LLMO] by:
  - ✓ Characterizing **fairness-accuracy (FA) frontier** with different loss functions for fairness and accuracy.
  - ▶ Providing **IFF** condition for when improving fairness does **not** come at the cost of reducing accuracy
- We extend the inference methods of Liu & Molinari (2025) and allow selectively observed Y^* by:
  - ✓ Deriving **sharp characterization** for the FA frontier for binary Y^* and specific loss functions
  - ▶ **Point-identifying** the FA frontier using a missing-at-random assumption
    - ⋆ Provide a method to test whether an LDA exists for a given algorithm
- In Progress:
  - ▶ Test for existence of LDA under partial identification
  - ▶ Test for $\mathcal{PF} = \mathcal{F}$ under partial identification or MAR.
  - ▶ Empirical application + package.

Conclusion

- We extend the theoretical framework of Liang, Lu, Mu & Okumura (2026) [LLMO] by:
  - ✓ Characterizing **fairness-accuracy (FA) frontier** with different loss functions for fairness and accuracy.
  - ▶ Providing **IFF** condition for when improving fairness does **not** come at the cost of reducing accuracy
- We extend the inference methods of Liu & Molinari (2025) and allow selectively observed Y^* by:
  - ✓ Deriving **sharp characterization** for the FA frontier for binary Y^* and specific loss functions
  - ▶ **Point-identifying** the FA frontier using a missing-at-random assumption
    - ⋆ Provide a method to test whether an LDA exists for a given algorithm
- In Progress:
  - ▶ Test for existence of LDA under partial identification
  - ▶ Test for $\mathcal{PF} = \mathcal{F}$ under partial identification or MAR.
  - ▶ Empirical application + package.

Thanks for listening!

Paper & slides here

Comments/questions are welcome!

Appendix

When Fairness and Accuracy Trade Off



- Intuitively, optimizing fairness aligns with optimizing accuracy if
  - every point on $\mathcal{PF}$ is already the fairest; or
  - every point on $\mathcal{F}$ is already Pareto-optimal.

When Fairness and Accuracy Trade Off

- Intuitively, optimizing fairness aligns with optimizing accuracy if
  - every point on $\mathcal{PF}$ is already the fairest; or
  - every point on $\mathcal{F}$ is already Pareto-optimal.

Theorem

If $\mathcal{E}$ is strictly convex, then $\mathcal{PF} \subseteq \mathcal{F}$, and $\mathcal{PF} = \mathcal{F}$ if and only if for all $\varepsilon^* \in \mathcal{F}$, there exists $q \in \mathbb{S}^2$ such that $q_1 \leq 0, q_2 \leq 0, q_3 = 0$, and

$$\varepsilon^* = \mathcal{S}_{\mathcal{E}}(q) = \arg\max_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon.$$

Support Function Characterization

By definition, $h_{\mathcal{E}}(q) = \max_{e \in \mathcal{E}} q^{\mathsf{T}} e$, $q \in \mathbb{S}^2 = \{q \in \mathbb{R}^3 : ||q|| = 1\}$.
For $g \in \{r, b\}$ and $d \in \{0, 1\}$, let:

$$\begin{aligned}
\mu_g &\equiv \mathbb{P}(G = g) \\
L_d^F &\equiv \ell^F(d, Y^*) \left(\frac{\mathbb{1}\{G=r\}}{\mu_r} - \frac{\mathbb{1}\{G=b\}}{\mu_b} \right) \\
L_d^{g,A} &\equiv \ell^A(d, Y^*) \frac{\mathbb{1}\{G=g\}}{\mu_g} \\
\theta_d^F(X) &\equiv \mathbb{E}[L_d^F | X] \\
\theta_d^{g,A}(X) &\equiv \mathbb{E}[L_d^{g,A} | X]
\end{aligned}$$

Denoting $\mathsf{L}_d := \left[L_d^{r,A},\ L_d^{b,A},\ L_d^F \right]^{\mathsf{T}}$, algebraic manipulations yield

$$h_{\mathcal{E}}(q) = \mathbb{E}\left[q^{\mathsf{T}} \mathsf{L}_0 + q^{\mathsf{T}} (\mathsf{L}_1 - \mathsf{L}_0) \underbrace{\mathbb{1}\{q^{\mathsf{T}} (\boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X)) > 0\}}_{a_q^{\mathrm{opt}}(X)} \right]$$

Missing at Random (MAR)

For **generic** $Y^* \in \mathcal{Y}$ and loss functions, if one is willing to assume

Assumption (MAR)

$(Y^*, G) \perp Z \mid X$ *and* $\pi(X) := \mathbb{E}[Z|X] \in (0, 1)$.

Missing at Random (MAR)

For **generic** $Y^* \in \mathcal{Y}$ and loss functions, if one is willing to assume

Assumption (MAR)

$(Y^*, G) \perp Z \mid X$ *and* $\pi(X) := \mathbb{E}[Z|X] \in (0,1)$.

Then $\boldsymbol{\theta}_d$, and hence $h_{\mathcal{E}}(q)$, are point-identified after adjusting for $\pi(X)$:

$$\boldsymbol{\theta}_d(X) = \mathbb{E}\left[\frac{Z}{\pi(X)}\mathsf{L}_d \,\middle|\, X\right], \quad h_{\mathcal{E}}(q) = \mathbb{E}\left[q^\intercal \mathsf{L}_0 \frac{Z}{\pi(X)} + q^\intercal \Delta\mathsf{L} \frac{Z}{\pi(X)} \mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\right]$$

Missing at Random (MAR)

For **generic** $Y^* \in \mathcal{Y}$ and loss functions, if one is willing to assume

Assumption (MAR)

$(Y^*, G) \perp Z \mid X$ *and* $\pi(X) := \mathbb{E}[Z|X] \in (0,1)$.

Then $\boldsymbol{\theta}_d$, and hence $h_{\mathcal{E}}(q)$, are point-identified after adjusting for $\pi(X)$:

$$\boldsymbol{\theta}_d(X) = \mathbb{E}\left[\frac{Z}{\pi(X)}\mathsf{L}_d \,\middle|\, X\right], \quad h_{\mathcal{E}}(q) = \mathbb{E}\left[q^\intercal \mathsf{L}_0 \frac{Z}{\pi(X)} + q^\intercal \Delta\mathsf{L} \frac{Z}{\pi(X)} \mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\right]$$

Theorem

Under regularity conditions, with cross-fitting and $n^{1/4}$*-consistent estimator of* $\boldsymbol{\theta}$ *and* π*, the plug-in estimator of* $h_{\mathcal{E}}(\cdot)$ *with appropriate influence-function adjustment term converges to a Gaussian process,*

$$\sqrt{n}\left(\widehat{h}_{\mathcal{E}}(q; \widehat{\boldsymbol{\theta}}, \widehat{\pi}) - h_{\mathcal{E}}(q; \boldsymbol{\theta})\right) =_d \mathbb{G}[\zeta_i(q; \boldsymbol{\theta}, \pi)] + o_p(1) \quad \text{in } \ell^\infty(\mathbb{S}^2).$$

Testing for Less Discriminatory Alternatives

- Given an existing algorithm with $\varepsilon^* \in \mathcal{E}$, **we would like to test**:

$$H_0 : \text{There } \textbf{does not exist} \text{ an LDA to } a^*.$$
$$H_1 : \text{There } \textbf{exists} \text{ an LDA to } a^*.$$

Testing for Less Discriminatory Alternatives

- Given an existing algorithm with $\varepsilon^* \in \mathcal{E}$, **we would like to test**:

$$H_0 : \varepsilon^* \in \mathcal{F}$$
$$H_1 : \varepsilon^* \notin \mathcal{F}$$

Testing for Less Discriminatory Alternatives

- Given an existing algorithm with $\varepsilon^* \in \mathcal{E}$, **we would like to test**:

$$\mathsf{H}_0 : \varepsilon^* \in \mathcal{F}$$
$$\mathsf{H}_1 : \varepsilon^* \notin \mathcal{F}$$

- Based on previous derivation, we can express this hypothesis as

$$\mathsf{H}_0 : \min_{q \in \mathsf{S}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \right) = 0$$
$$\mathsf{H}_1 : \min_{q \in \mathsf{S}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \right) > 0$$

Testing for Less Discriminatory Alternatives

- Given an existing algorithm with $\varepsilon^* \in \mathcal{E}$, **we would like to test**:

$$\begin{aligned} \mathsf{H}_0 &: \varepsilon^* \in \mathcal{F} \\ \mathsf{H}_1 &: \varepsilon^* \notin \mathcal{F} \end{aligned}$$

- Based on previous derivation, we can express this hypothesis as

$$\begin{aligned} \mathsf{H}_0 &: \min_{q \in \mathsf{S}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \right) = 0 \\ \mathsf{H}_1 &: \min_{q \in \mathsf{S}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \right) > 0 \end{aligned}$$

- We build on
  - Liu & Molinari (2025) to derive a statistic for testing H_0.
  - Fang & Santos (2019) to prove size control and propose a consistent bootstrap critical value.

Asymptotic Theory: Main Assumptions

Assumption (Margin Condition)

$\exists\ 0 < m \leq 1$ *such that* $\forall\ \delta > 0$, *we have* $\sup_{q \in \mathbb{S}^2} \mathbb{P}\big(|q^\intercal(\boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X))| < \delta\big) \lesssim \delta^m$.

Assumption (Behavior of Nuisance Parameters)

We can partition $X = (X_1, X_2)$ *such that (1)* $\forall \delta > 0$, $\sup_{q \in \mathbb{S}^2} \mathbb{P}(|q^\intercal \Delta\boldsymbol{\theta}(X)| < \delta \,|\, X_2) \lesssim \delta$; *(2)* $(\pi(X), \Delta\boldsymbol{\theta}(X)) = F(\nu(X_1), \gamma(X_2))$ *for unknown functions* ν *and* γ *and known link function* F, *where the derivatives of* F *with respect to* (ν, γ) *are uniformly bounded.*

Assumption (Rate Requirement)

The k*-th fold estimators satisfy* $\|\widehat{\nu}_k(X_1) - \nu(X_1)\|_\infty = o_p(n^{-1/4})$, $\|\widehat{\gamma}_k(X_2) - \gamma(X_2)\|_{L^2} = o_p(n^{-1/4})$.