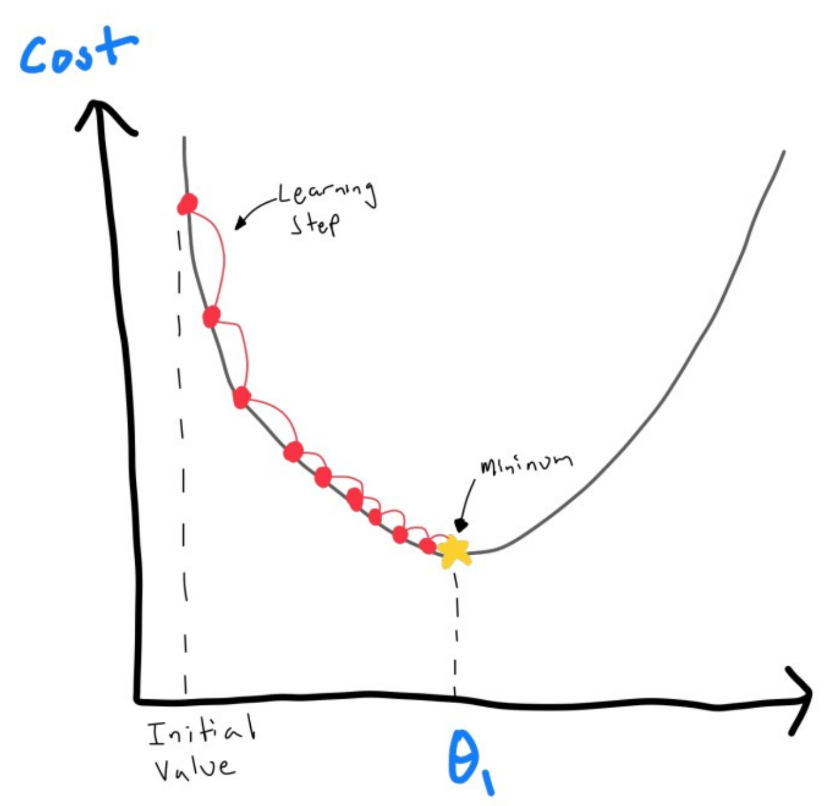
Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research

A. Feder Cooper and Ellen Abrams

afc78@cornell.edu, ema85@cornell.edu

Introduction

Machine learning (ML) typically trains models to optimize **accuracy**



Given: a dataset, model, algorithm

Goal: Find model parameters that minimize loss function

Where does **fairness** fit in?

Implicit normative assumptions bias every stage of the ML pipeline

Vox.com

Fairness attempts to correct for this bias

Models no longer just need to be **accurate**, they also need to be **fair**

The Fairness-Accuracy Trade-Off

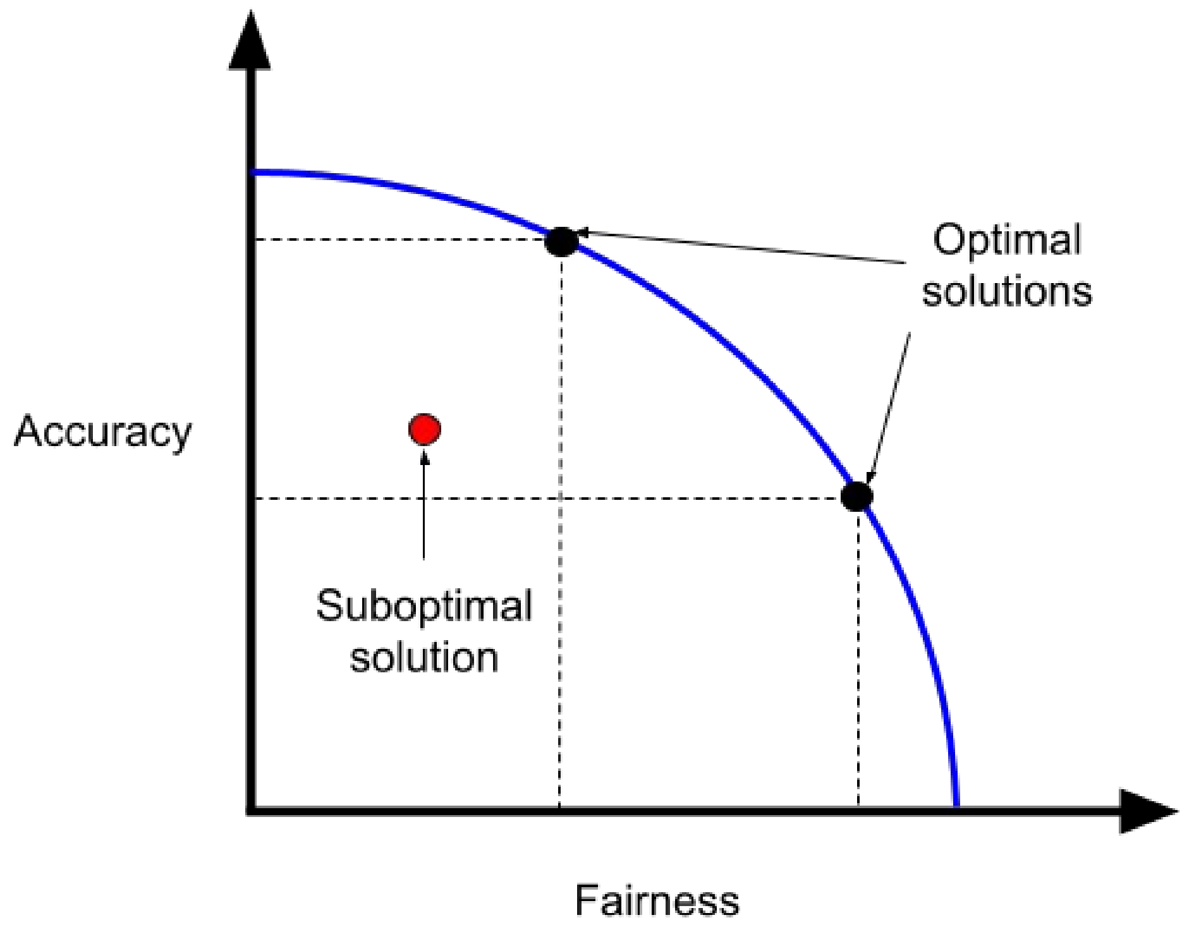
Researchers commonly pose the objectives of **fairness** and **accuracy** in a **trade-off** optimization problem

Researchers often call the trade-off "**inherent**" and "**unavoidable**"

The **blue curve** shows potential optimal trade-off solutions

Increases in **accuracy** require **decreases** in **fairness**

Increases in **fairness** require **decreases** in **accuracy**



Applying a Sociotechnical Lens

The **choice** to formulate an **optimization problem**

produces a particular kind of **knowledge** about fairness

that **cannot be detached** from broader **social context**

Public safety
? vs. ?
Fair policing

Hiring best candidates
? vs.?
not discriminating in hiring practices

Formulating a **trade-off** forecloses the possibility that **fairness** and **accuracy** could be **complementary**

Emergent Unfairness

Implicit normative assumptions in the trade-off formulation lead to **emergent unfairness**

1) Unfairness from assuming fairness = equality

What is **fair** and what is **strictly equal** are not always the same thing

Example: Affirmative action

2) Unfairness from assuming context can be ignored

Ignoring the **past**

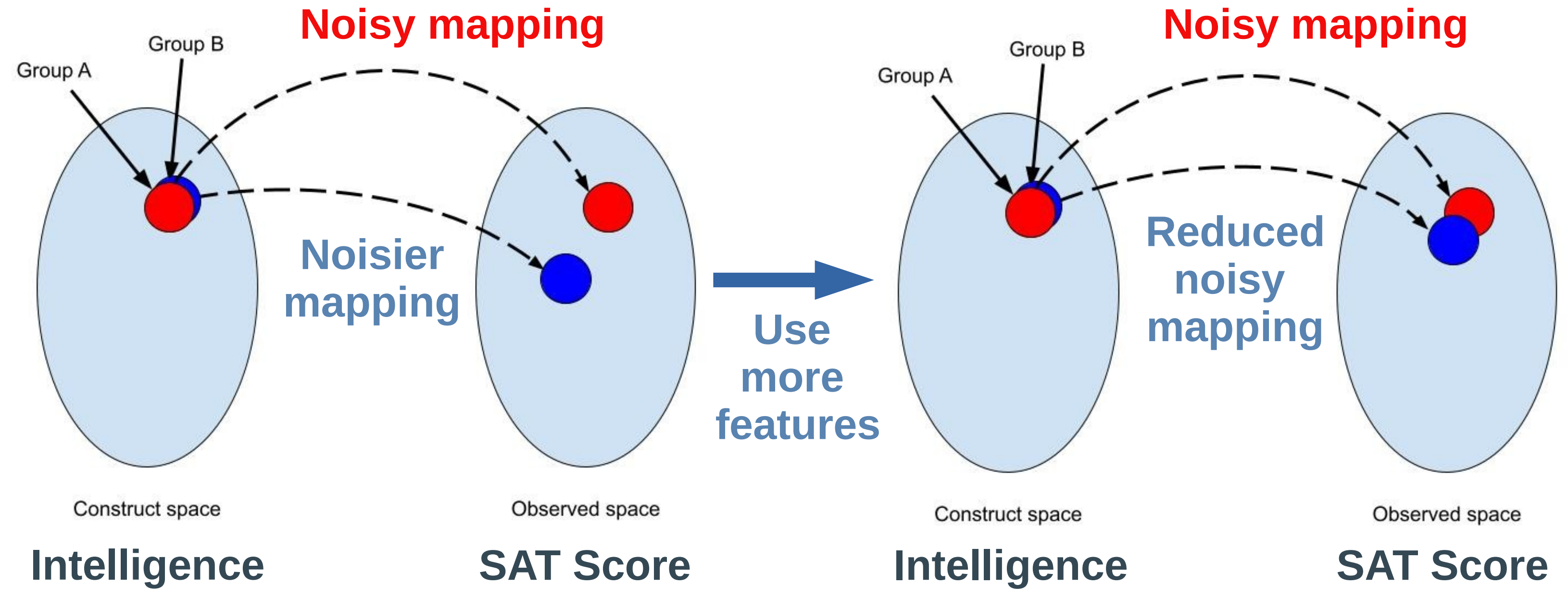
How we measure accuracy **implicates unfairness**, causes the trade-off to **break down**

Being blind to the **future**

Trade-off considers **local, immediate** decisions

Does not consider **long-term effects** that other stakeholders care about

3) Unfairness of "Active Fairness" remedies



Active fairness aims to decrease the disparity that comes from noisy mappings

EFF.org

By **collecting more data** (**increasing surveillance**) on the **less privileged** group

Takeaways

Make implicit normative assumptions **explicit** so that, just like mathematical ones, they can be **rigorously reviewed and tested**

Consider **revisiting** the fairness-accuracy trade-off problem formulation, **not using it anymore**

SCAN ME

Find our paper on arXiv.org