CSC2541 Guest Lecture

Economic Notions of Fairness in Machine Learning

Overview of This Lecture

- Background
  - Study of fairness in economics
  - Fairness in resource allocation (cake-cutting and indivisible goods)
- Adaptation to machine learning
  - Classification
  - Clustering
  - Future work

Study of Fairness in Economics

- Almost a century old
  - Started from the work of Steinhaus in 1948
  - Introduced fairness in the classic cake-cutting setting

- Notions of individual fairness
  - Proportionality (Prop) [Steinhaus, 1948]
  - Envy-freeness (EF) [Foley, 1967]
  - Equitability (EQ) [Pazner and Schmeidler, 1978]
    - More generally, "egalitarian-equivalence"
  - Maximin share (MMS) [Budish, 2011]

Study of Fairness in Economics

- Extended to groupwise notions of fairness
  - *Stronger* than individual fairness
  - The core [Varian, 1974]
    - Implies proportionality
  - Group envy-freeness (GEF) [Berliant, Thomson, Dunz, 1992]
    - Implies envy-freeness
  - Group fairness (GF) [Conitzer, Freeman, Shah, Wortman-Vaughan, 2019]
    - Implies both core and group envy-freeness

Study of Fairness in Economics

- Often, approximate versions are sought when exact versions cannot be guaranteed
  - Proportionality up to one (Prop1) [Conitzer, Freeman, Shah, 2017]
  - Envy-freeness up to one (EF1) [Budish 2011]
  - Core up to one (Core1) [Munagala, Fain, Shah, 2018]
  - Group fairness up to one (GF1) [Conitzer, Freeman, Shah, Wortman-Vaughan, 2019]

Fairness: Cake-Cutting & Indivisible Goods

Cake-Cutting

- A heterogeneous, divisible good
  - Heterogeneous: different parts valued differently by different individuals
  - Divisible: we can split it between individuals

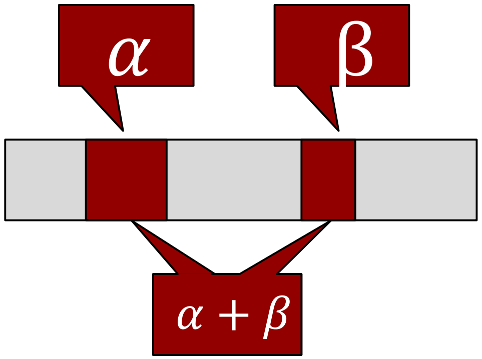
- Represented as $[0,1]$

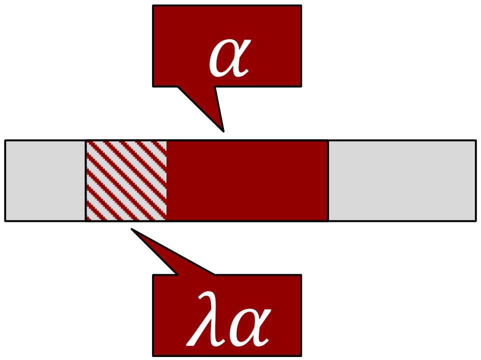
- How can we fairly divide the cake between n agents?

Agent Valuations

- Set of agents $N = \{1, \ldots, n\}$
- Agent i has utility function u_i
  - $u_i(X)$ = utility for getting $X \subseteq [0,1]$

- Additive: For $X \cap Y = \emptyset$, $u_i(X) + u_i(Y) = u_i(X \cup Y)$
- Normalized: $u_i([0,1]) = 1$
- Divisible: $\forall \lambda \in [0,1]$ and X, $\exists Y \subseteq X$ s.t. $u_i(Y) = \lambda u_i(X)$





Fairness Goals

- Allocation $A = (A_1, \ldots, A_n)$ is a partition of the cake into n disjoint bundles

- Proportionality (Prop):

$$\forall i \in N\colon\ u_i(A_i) \geq 1/n$$

- Envy-Freeness (EF):

$$\forall i, j \in N\colon u_i(A_i) \geq u_i(A_j)$$

- Equitability (EQ):

$$\forall i, j \in N\colon u_i(A_i) = u_j(A_j)$$

Fairness Goals

- Prop: $\forall i \in N: u_i(A_i) \geq 1/n$
- EF: $\forall i, j \in N: u_i(A_i) \geq u_i(A_j)$

- Question: What is the relation between Prop & EF?

1. Prop ⇒ EF
2. EF ⇒ Prop
3. Equivalent
4. Incomparable

Cut-and-Choose

- Algorithm for $n = 2$ agents

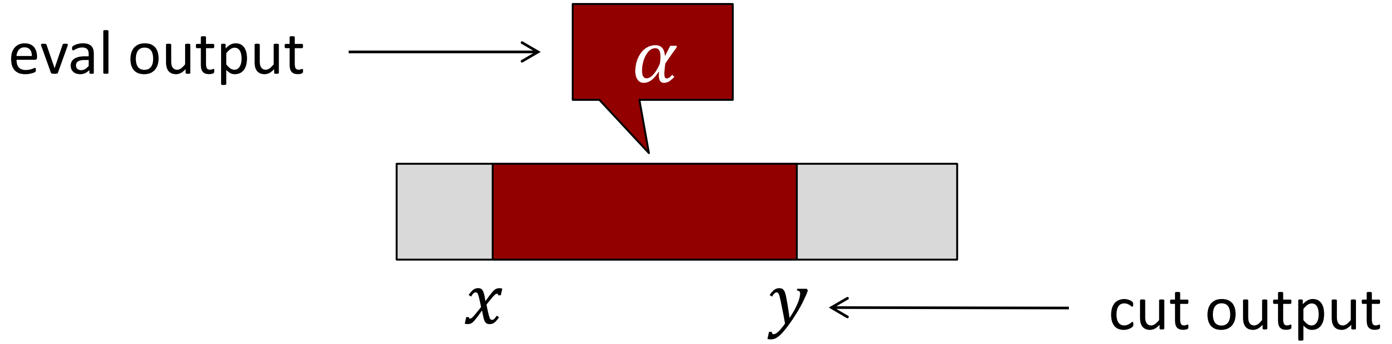
- Agent 1 divides the cake into two pieces X, Y s.t.
$$V_1(X) = V_1(Y) = 1/2$$

- Agent 2 chooses the piece she prefers.

- This is EF and therefore proportional.
  - Why?

Query Model

- To capture the complexity of computing various solution concepts, we need a model for accessing utilities
- Robertson-Webb model
  - $\text{Eval}_i(x, y)$ returns $u_i([x, y])$
  - $\text{Cut}_i(x, \alpha)$ returns y such that $u_i([x, y]) = \alpha$



Complexity of Proportionality

- Theorem [Even and Paz, 1984]
  - There exists a protocol for computing a proportional allocation using $O(n \log n)$ queries in the Robertson-Webb model.
  - Uses a simple divide-and-conquer idea

- Theorem [Edmonds and Pruhs, 2006]
  - Any protocol computing a proportional allocation needs $\Omega(n \log n)$ queries in the Robertson-Webb model.

Complexity of Envy-Freeness

- [Brams and Taylor, 1995]
  - First unbounded EF protocol
- [Procaccia 2009]
  - $\Omega(n^2)$ lower bound for EF
- Major open question: bounded EF protocol?
- [Aziz and Mackenzie, 2016]
  - Breakthrough $O(n^{n^{n^{n^{n^n}}}})$ protocol!
  - Not a typo!

Complexity of Equitability

- [Procaccia and Wang, 2017]
  - Any protocol for computing an equitable allocation requires an unbounded number of queries in the Robertson-Webb model.
  - An ϵ-equitable allocation can be computed in $O\left({}^{1}/_{\epsilon} \ln\left({}^{1}/_{\epsilon}\right)\right)$ queries
  - A corresponding lower bound is $\Omega\left(\ln\left({}^{1}/_{\epsilon}\right) \ln\ln\left({}^{1}/_{\epsilon}\right)\right)$

Other Desiderata

- Pareto optimality (PO)
  - Allocation A is PO if $\nexists B$ s.t. $u_i(B_i) \geq u_i(A_i)$ for all i, and at least one inequality is strict.
  - "There should be no unilaterally better allocation."
- Strategyproofness (SP)
  - If A and A' denote allocations obtained when agent i reports u_i and u_i' respectively, fixing the reports of the other agents, then $u_i(A_i) \geq u_i(A_i')$.
  - "Regardless of what the other agents do, there is no incentive for agent i to misreport."

PO and SP

- By themselves, PO and SP are easy to achieve

- Serial dictatorship
  - Agent 1 takes any part of the cake she likes
  - From what's left, agent 2 takes any part that she likes
  - …

- The goal is to achieve them along with fairness

PO + EF

- Theorem [Weller '85]
  - There always exists an allocation of the cake that is both envy-free and Pareto optimal.
  - One method: maximize Nash welfare
  $$\text{argmax}_A \Pi_i \, u_i(A_i)$$
  - Informal proof of EF on the board (if time permits)
  - Named after John Nash.

Special Case

- There are m "divisible" goods
  - E.g. a gold bar, a pile of money, …
  - Agents only care about the fraction of each good they get
- Notation
  - $u_{i,g}$ = utility to agent i for all of good g
  - $x_{i,g}$ = fraction of good g given to agent i
  - $u_i(A_i) = \sum_g x_{i,g} \cdot u_{i,g}$
  - Feasibility: $\sum_i x_{i,g} = 1$ for all g

Indivisible Goods

- Indivisible goods?
  - Allocation = partition of goods
  - Splitting not allowed

- If randomized allocations are permitted…
  - Any "divisible" allocation can be "implemented" [Birkhoff-von-Neumann theorem]

- What if only deterministic allocations are allowed?

Indivisible Goods

				
	8	7	20	5
	9	11	12	8
	9	10	18	3

Given such a matrix of numbers, assign each good to a agent.

We assume additive values. So, e.g., V$(\{$ , $\}) = 8 + 7 = 15$

Indivisible Goods

- Theorem [Caragiannis et al. 2016]
  - For indivisible goods, maximizing Nash welfare over integral allocations returns an allocation that is envy-free up to one good (EF1) and Pareto optimal (PO).
- EF1:
  - $\forall i, j, \exists g \in A_j$ s.t. $u_i(A_i) \geq u_i\left(A_j \setminus \{g\}\right)$
- EFX:
  - $\forall i, j, \forall g \in A_j$ s.t. $u_i(A_i) \geq u_i\left(A_j \setminus \{g\}\right)$
  - Open question: Does an EFX allocation always exist?

Enough about fair division!

How do I apply this
to machine learning?

Envy-Freeness for Classification

- Two key differences from resource allocation

- Q1: No resources being *partitioned* across people
  - Often, a single classifier is implemented
  - What does it mean for i to not envy j?

- Q2: Is it reasonable to require that no individual envies any other individual?
  - If not, what would be a good relaxation?

Envy-Freeness for Classification

- Q1: No resources being *partitioned* across people
  - Often, a single classifier is implemented
  - What does it mean for i to not envy j in this case?

- Idea 1:
  - Compare the classification outcomes
  - Let $\mathcal{Y}$ be the set of classes, $\mathcal{X}$ be the set of individuals represented by their feature vectors
  - Classifier $h : N \to C$ is EF if $\forall i, j \in \mathcal{X}, u_i\big(h(i)\big) \geq u_i\big(h(j)\big)$
    - "I prefer my label to the label assigned to anyone else"
  - [Balcan et al., 2019]

Envy-Freeness for Classification

- Q1: No resources being *partitioned* across people
  - Often, a single classifier is implemented
  - What does it mean for i to not envy j in this case?

- Idea 2:
  - Actually train two different classifiers h_1, h_2 for two different individuals/groups
  - Define their utility for a classifier
  - Ask that individual/group $i \in \{1,2\}$ prefer h_i to h_{3-i}
  - [Ustun et al., 2019]

Envy-Freeness for Classification

- Q2: Is it reasonable to require that no individual envies any other individual?
  - If not, what would be a good relaxation?

- Idea 1:
  - It may be reasonable if randomized (or soft) classification is allowed
  - This still imposes *many* constraints
  - How do we train for it? Does it generalize?
  - [Balcan et al., 2019]

Envy-Freeness for Classification

- Q2: Is it reasonable to require that no individual envies any other individual?
  - If not, what would be a good relaxation?

- Idea 2:
  - If deterministic classification is required, we can relax EF to require that no group, *on average,* envy another group
  - [Hossain et al., manuscript]

Envy-Free Classification

- $\mathcal{X}$ = space of individuals
  - Represented by feature vectors
- $\mathcal{Y}$ = space of possible labels
  - Sometimes there's a ground truth label $\hat{y}$ for each individual x, which can be treated as side information not available to the classifier but available during training
- Classifier $h : \mathcal{X} \to \mathcal{Y}$ or $h : \mathcal{X} \to \Delta(\mathcal{Y})$

Envy-Free Classification

- Two conflicting objectives

- Loss
  - $L(x, y) =$ loss when labeling individual x by y
  - For $c \in \Delta(\mathcal{Y})$, $L(x, c) = \mathbb{E}_{y \sim c}[L(x, y)]$

- Utilities
  - $u(x, y) =$ utility of individual x for receiving label y
  - For $c \in \Delta(\mathcal{Y})$, $u(x, c) = \mathbb{E}_{y \sim c}[u(x, y)]$
  - Assumed to be L-Lipschitz in x

Envy-Free Classification

- Envy-freeness:
  - Sample: $h : \mathcal{X} \to \Delta(\mathcal{Y})$ is EF on a set $S \subseteq \mathcal{X}$ if:

$$u\big(x, h(x)\big) \geq u\big(x, h(x')\big), \forall x, x' \in S$$

  - Distribution: h is (α, β)-EF w.r.t. a distribution P on $\mathcal{X}$ if:

$$\Pr_{x,x' \sim P}\left[u\big(x, h(x)\big) < u\big(x, h(x')\big) - \beta\right] \leq \alpha$$

  - Questions:
    - Is it reasonable to require h to be EF on training data?
    - If it is, does it generalize to the underlying distribution?

Envy-Free Classification

- Deterministic classifiers
  - Envy-freeness is very restrictive
  - Let $h(S)$ denote the set of all classes assigned to individuals in S
  - Then, clearly, h is EF on S iff each individual $x \in S$ is assigned her most preferred label in $h(S)$
- Randomized classifiers
  - Allow mixing a preferred label with a "low loss" label to achieve low empirical loss along with envy-freeness

Generalization

- "ERM subject to EF"
  - For arbitrary classifiers, we need an algorithm A to extend the classifier to unseen data (e.g., by nearest neighbor)

- Theorem:
  - There exists $\mathcal{X}$ and a distribution P over $\mathcal{X}$ s.t. for any A, w.p. $1 - \exp(-\exp(q))$, the following happens:
  - When training set S of size $\exp(q)$ is drawn from P and A is applied to derive a classifier, it violates (α, β)-EF w.r.t. P for $\alpha < 1/25$ and $\beta < L/8$.

Generalization

- Natarajan dimension
  - Generalizes VC dimension to multi-class classification
  - Low dimension: One-vs-all, multiclass SVM, tree-based classifiers, error-correcting code-based classifiers, …
- Theorem:
  - $\mathcal{G}$ = family of classifiers with Natarajan dimension d
  - $\mathcal{H}$ = mixtures of up to m classifiers from $\mathcal{G}$
  - (α, β)-EF on training set S implies $(\alpha + 7\gamma, \beta + 4\gamma)$-EF on the underlying distribution P w.p. $1 - \delta$ when

$$|S| \geq O\left(\frac{dm^2}{\gamma^2} \log \frac{dm|\mathcal{Y}|}{\gamma}\right)$$

Generalization

- Key lemma (informal):
  - If $\mathcal{H}$ is a mixture of up to m classifiers from a low dimension family $\mathcal{G}$, then a "small finite" subset of classifiers "cover" all of $\mathcal{H}$
  - Given any $h \in \mathcal{H}$, we can find some classifier in the small subset that matches h on almost all inputs

Training for EF Classification

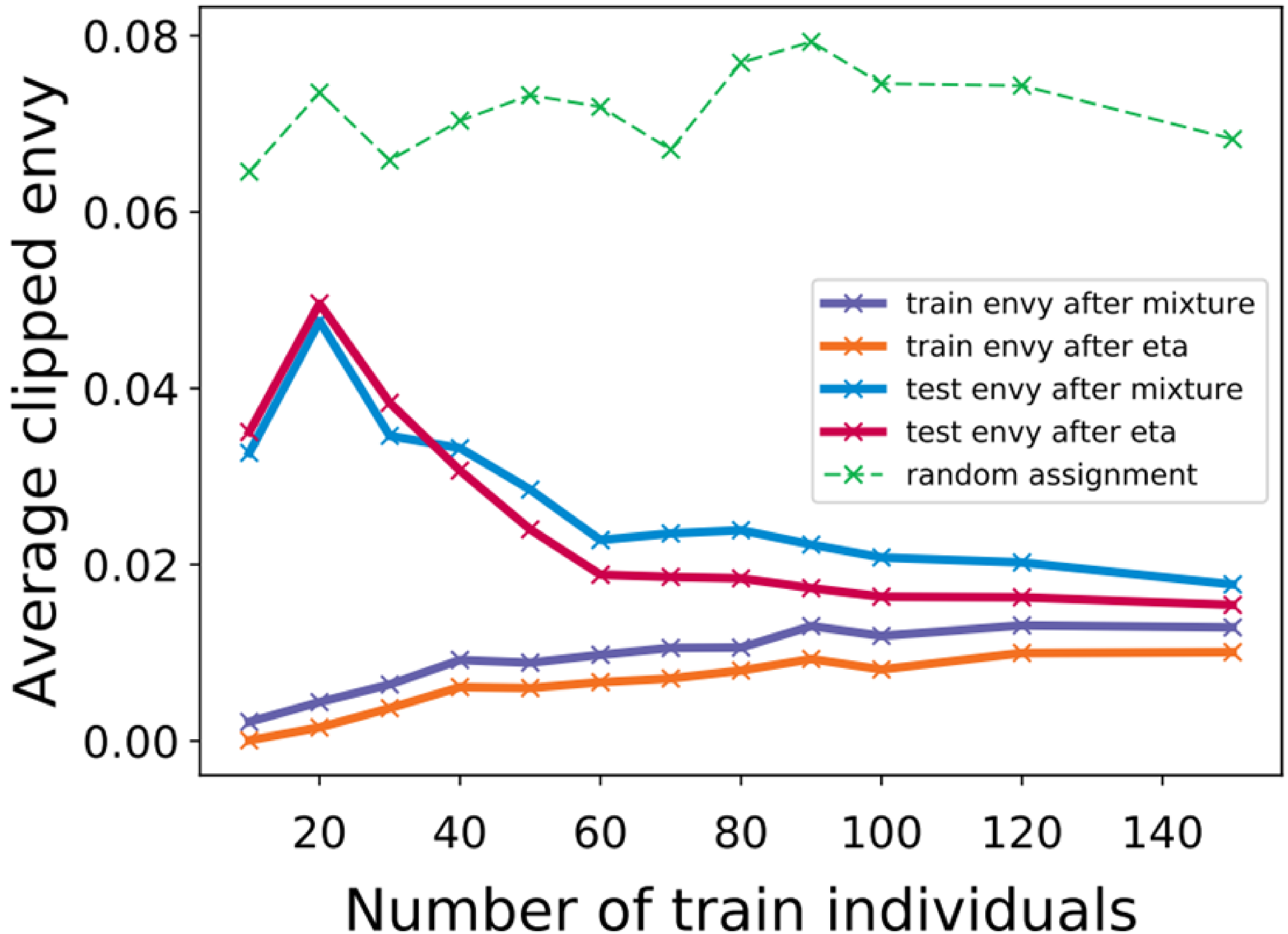
- Training a mixture through "ERM subject to EF" is not a convex program

$$\min_{\vec{g}\in\mathcal{G},\eta\in\Delta_m} \sum_{i=1}^{n}\sum_{k=1}^{m} \eta_k L\big(x_i, g_k(x_i)\big)$$

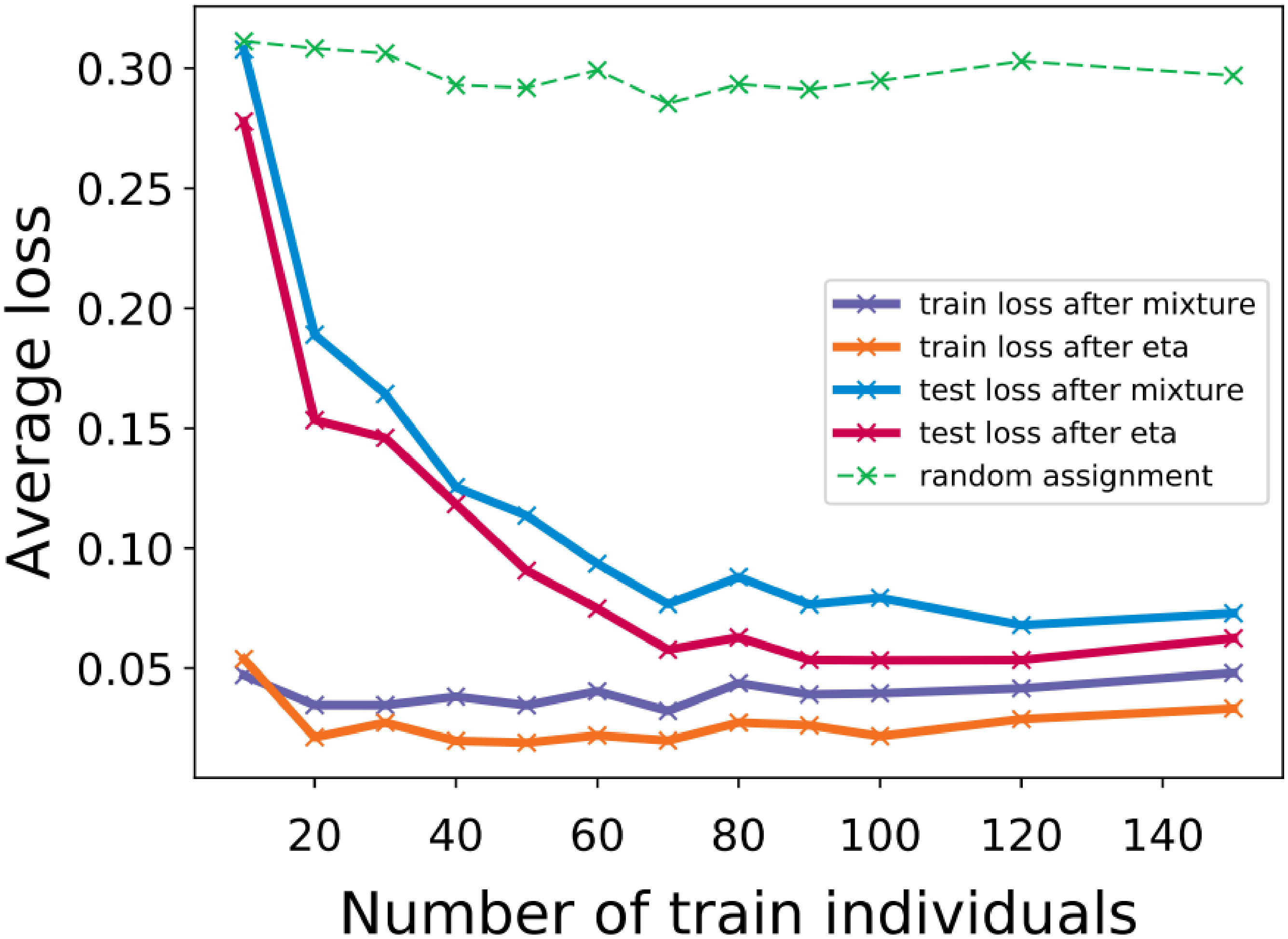
$$s.t. \sum_{k=1}^{m} \eta_k u\big(x_i, g_k(x_i)\big) \geq \sum_{k=1}^{m} \eta_k u\left(x_i, g_k\big(x_j\big)\right), \forall (i,j) \in [n]^2$$

- They introduce an SVM-style convex relaxation
  - Empirically results in low envy and low loss

Empirical Results



Empirical Results



Group EF & EQ

- Groups of individuals (G_1, G_2)
- GroupEF:
  - $\mathbb{E}_{x_1 \sim G_1, x_2 \sim G_2}[u\left(x_1, h(x_2)\right) - u(x_1, h(x_1))] \leq 0$
- GroupEQ:
  - $\left|\mathbb{E}_{x_1 \sim G_1} u\left(x_1, h(x_1)\right) - \mathbb{E}_{x_2 \sim 2} u\left(x_2, h(x_2)\right)\right| \leq 0$
- For both definitions…
  - Replace expectation with empirical average on finite S
  - ϵ-GroupEF / ϵ-GroupEQ if the LHS is at most ϵ

Group EF & EQ

- Applicable in a non-ground truth setting
  - E.g. targeted advertising context of Balcan et al. [2019]
  - Groups typically defined using sensitive attributes

- Also applicable in a ground truth setting
  - E.g. making loan/bail decisions
  - Groups defined using a combination of sensitive attributes and ground truth
  - E.g. $G_1 = \{\text{male applicants who can repay the loan}\}$, $G_2 = \{\text{female applicants who can repay the loan}\}$

Group EF & EQ

- Ground truth setting
  - Sensitive attribute A, ground truth $\hat{Y}$
- Generalizes demographic parity (DP)
  - $G_1 = \{A = a_1\}, G_2 = \{A = a_2\}$
- Generalizes equalized odds (EO)
  - $G_1^1 = \left\{A = a_1 \wedge \hat{Y} = 1\right\}, G_2^1 = \left\{A = a_2 \wedge \hat{Y} = 1\right\}$
  - $G_1^2 = \left\{A = a_1 \wedge \hat{Y} = 0\right\}, G_2^2 = \left\{A = a_2 \wedge \hat{Y} = 0\right\}$
- For group EF, also need to add reverse sets

Group EF & EQ

- Ground truth setting
  - Sensitive attribute A, ground truth $\hat{Y}$

- Generalizes demographic parity (DP) and equalized odds (EO)

  - Allows extending these definitions to multi-class classification

  - E.g. how should DP or EO be applied when there are k different types of loans available and applicants have different preferences over these loans?

Problems with Group EF/EQ

- Post-processing a given (unfair) classifier to achieve fairness by just "rebalancing" rates is not an option

- Theorem [Hossain et al., manuscript]
  - The only way to post-process a classifier to get group EF with respect to (G_1, G_2) without accessing utilities is to return h such that for each $x \in G_1$, $\Pr[h(x) = c]$ is the average of $\Pr[h(x_2) = c]$ over $x \in G_2$.
  - The only way to post-process a classifier to get group EQ with respect to (G_1, G_2) without accessing utilities is to assign a uniformly random label to each individual.

Generalization of Group EF/EQ

- Rademacher complexity approach
  - $Rad(A) = \frac{1}{m}\mathbb{E}[\sup_{a \in A} \sum_{i=1}^{m} \sigma_i \, a_i]$

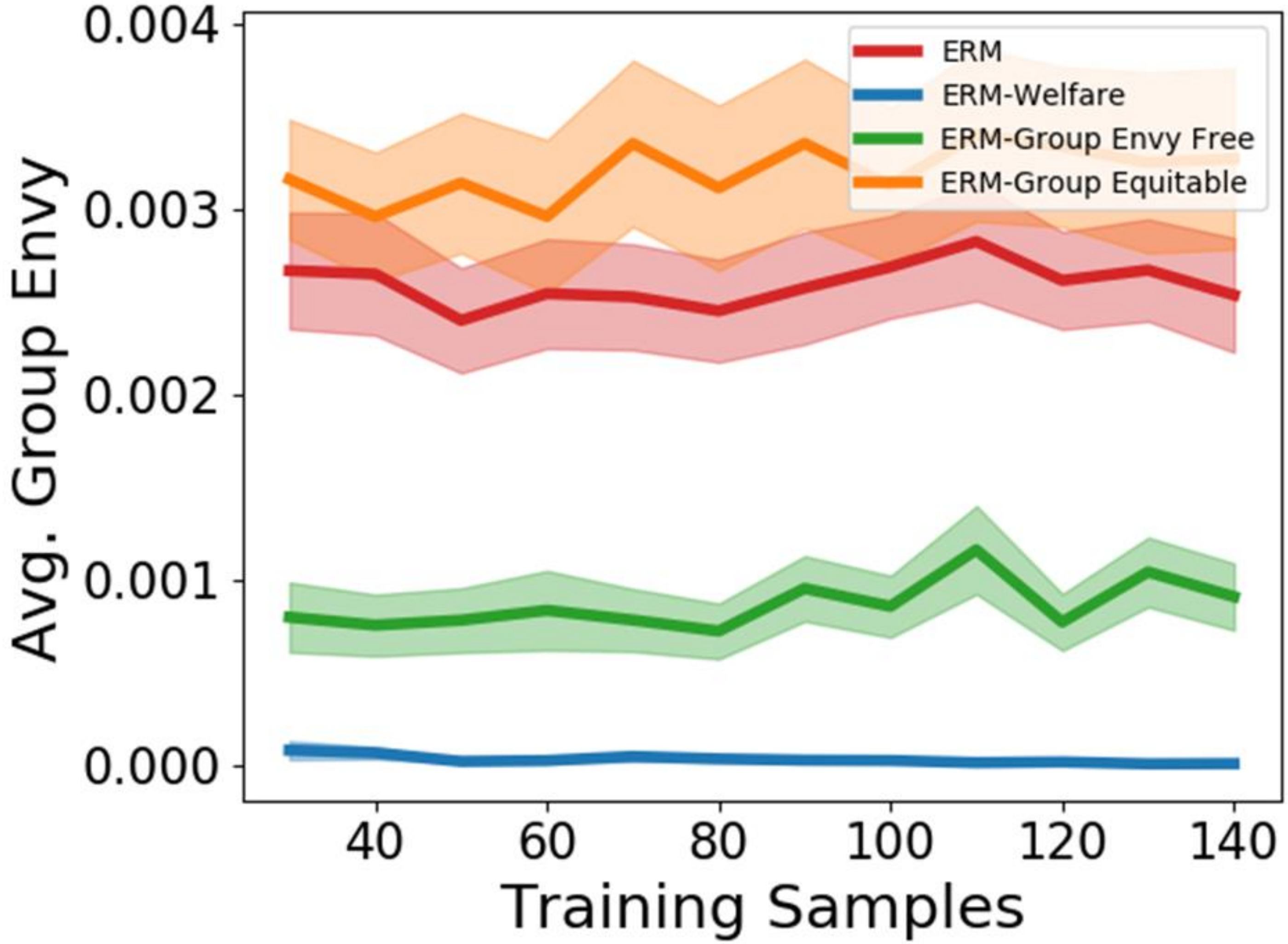
- Problems adapting to this framework
  - Usually defined for functions that map to $[0,1]$, not for multi-class classification
  - Writing group envy or equitability violation on population involves a product of utility and group membership indicators

Generalization of Group EF/EQ

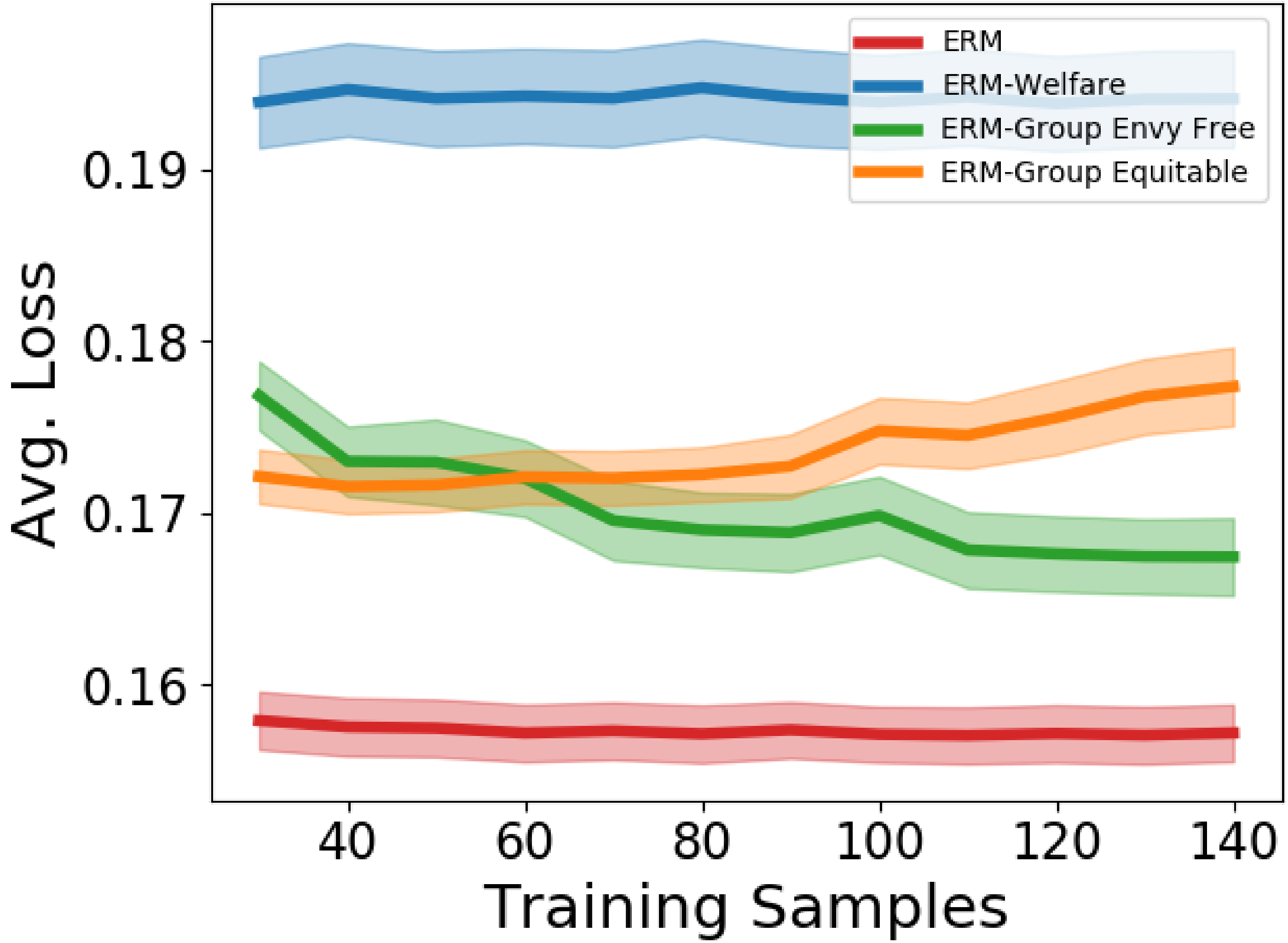
- Theorem (informal) [Hossain et al., manuscript]
  - $\mathcal{H}$ = family of classifiers
  - S = training set such that $\mathcal{R}(\mathcal{H} \circ S) \leq \epsilon/8$
  - If $|S| \geq O\left(\frac{1}{\epsilon^2} \ln\left(\frac{|\mathcal{G}|}{\delta}\right)\right)$, then w.p. $1 - \delta$, all constraints in $\mathcal{G}$ generalize up to ϵ additive error.
    - $\mathcal{G}$ = set of (G_1, G_2) pairs

- Theorem (informal)
  - For linear one-vs-all classifiers in d dimensions, $|S| = O\left(\frac{d^3 m}{\epsilon^2} \ln\left(\frac{dm}{\epsilon}\right)\right)$ is enough.

Empirical Results



Empirical Results



Other Approaches

- Decoupled Classifiers [Utsun et al., 2019]
  - Train a pair of classifiers: h_1 for group G_1 and h_2 for G_2
  - (h_1, h_2) is envy-free if

$$\mathbb{E}_{x \sim G_1}\left[u\left(x, h_1(x)\right)\right] \geq \mathbb{E}_{x \sim G_1}\left[u\left(x, h_2(x)\right)\right]$$

    and a similar inequality holds for group G_2.

  - One problem: Even when preferences are identical…
    - h_1 might assign bad labels to G_1
    - h_2 might assign great labels to G_2, but when applied on G_1, might apply even worse labels than h_1 by "detecting" certain features
    - Intuitively unfair but satisfies the fairness guarantee

Other Approaches

- Individual Fairness [Dwork et al., 2011]
  - "Similar individuals should be treated similarly"
  - Given a distance d, $\|h(x) - h(y)\| \leq d(x, y), \forall x, y$

- Preference-Informed Fairness [Kim et al., 2019]
  - What if the individuals have heterogeneous preferences?
  - y is similar to x, but doesn't like $h(x)$
  - $\forall x, y\ \exists c\ u\big(y, h(y)\big) \geq u(y, c) \wedge \|h(x) - c\| \leq d(x, y)$
    - "I could've given you c, which would have satisfied individual fairness. I'm only giving you something you like more."

Other Approaches

- Preference-Informed Fairness [Kim et al., 2019]
  - $\forall x, y \; \exists c \; u\big(y, h(y)\big) \geq u(y, c) \wedge \|h(x) - c\| \leq d(x, y)$
  - Almost a "justified envy-freeness" concept
  - When u is L-Lipschitz continuous, PIF implies

$$\left|u\big(y, h(x)\big) - u(y, c)\right| \leq L \cdot d(x, y)$$

$$\Rightarrow u\big(y, h(y)\big) \geq u\big(y, h(x)\big) - L \cdot d(x, y)$$

  - Every y envies x by at most $L \cdot d(x, y)$

Other Approaches

- Circumventing Harmful Fairness [Ben-Porat et al., 2019]
  - ERM subject to EO:
    - May harm the disadvantaged group in terms of welfare
  - ERM subject to group EQ:
    - Can never harm the disadvantaged group in terms of welfare
  - Characterize ERM subject to Group EQ outcomes, and give algorithms to compute them quickly

Other Approaches

- Fairness in clustering
  - n data points, k cluster centers
  - Sometimes clustering is used for facility location, where k facilities are located to serve n data points
  - Core
    - A clustering C is in the core if there exist no group S of n/k data points and a possible cluster center y such that $d(i, y) < d(i, C)$ for all $i \in S$, where $d(i, C) = \min_{c \in C} d(i, c)$
  - There exist instances with no core clustering, but $1 + \sqrt{2}$ approximation is possible [Munagala et al., 2019]

Other Approaches

- Incentives
  - How does fairness play with incentives?
  - Do fair algorithms provide greater incentives to individuals to lie about their sensitive attributes?
  - Ongoing research…


The New York Times

Rachel Dolezal, Who Pretended to Be Black, Is Charged With Welfare Fraud


BBC Sign in News Sport Reel Worklife Travel

NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Sto

Newsbeat

Blackfishing: The women accused of pretending to be black


Los Angeles Times

CALIFORNIA

Admissions scandal: Mom who rigged son's ACT, lied about his race gets 3 weeks in prison