An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

Cyrus Cousins

Brown University
Department of Computer Science

December 2021

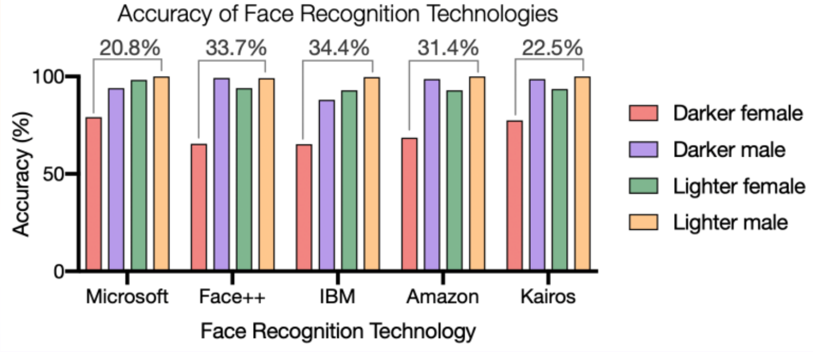
http://cs.brown.edu/people/ccousins/

BROWN
Computer Science

NEURAL INFORMATION PROCESSING SYSTEMS

Fairness in Machine Learning (or Lack Thereof)

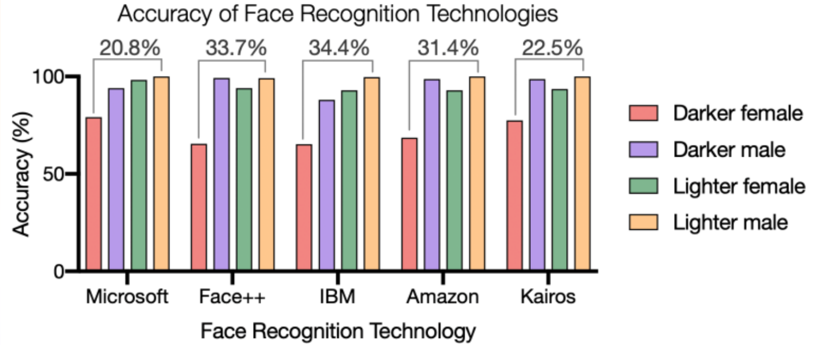
- ML systems often trained on *group A*, then applied to *group B*



[Gender Shades Project]

Fairness in Machine Learning (or Lack Thereof)

- ML systems often trained on *group A*, then applied to *group B*

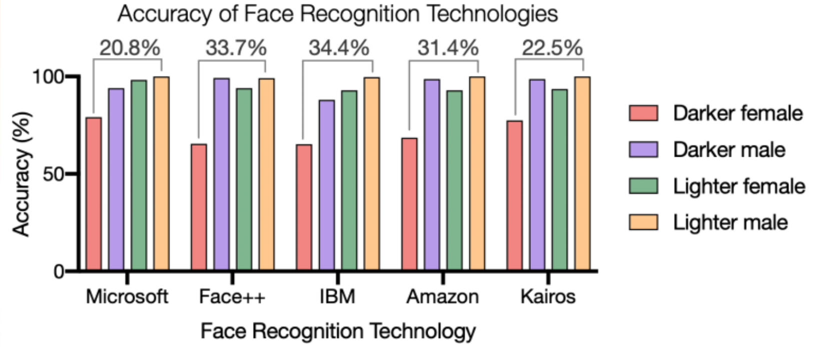


[Gender Shades Project]

- Differential performance $\implies$ algorithmic discrimination
  - Facial recognition and policing
  - Speech recognition and accessibility
  - Many more examples

Fairness in Machine Learning (or Lack Thereof)

- ML systems often trained on *group A*, then applied to *group B*



[Gender Shades Project]

- Differential performance $\implies$ algorithmic discrimination
  - Facial recognition and policing
  - Speech recognition and accessibility
  - Many more examples
- What has gone wrong? Is the problem:
  1. *that* a machine is learning;
  2. *from what* a machine is learning; or
  3. *how* a machine is learning?

Talk Outline

- Focus on *differential accuracy* between *protected groups*
  - Want *group level* fairness
  - Learn from sample of *many individuals* drawn from *each group*

Talk Outline

- Focus on *differential accuracy* between *protected groups*
  - Want *group level* fairness
  - Learn from sample of *many individuals* drawn from *each group*
- Claim: current ML systems are trained:
  - On the wrong data (well-known)
  - In the wrong way

Talk Outline

- Focus on *differential accuracy* between *protected groups*
  - Want *group level* fairness
  - Learn from sample of *many individuals* drawn from *each group*
- Claim: current ML systems are trained:
  - On the wrong data (well-known)
  - In the wrong way
- Optimize models sensitive to performance on *protected-groups*
  - Introduce malfare learning target
  - Consider all groups (possibly nonlinearly)

Talk Outline

- Focus on *differential accuracy* between *protected groups*
  - Want *group level* fairness
  - Learn from sample of *many individuals* drawn from *each group*
- Claim: current ML systems are trained:
  - On the wrong data (well-known)
  - In the wrong way
- Optimize models sensitive to performance on *protected-groups*
  - Introduce malfare learning target
  - Consider all groups (possibly nonlinearly)
- Theoretical treatment of learning and statistics
  - Overfitting and statistical estimation
  - Computational complexity issues in learning
  - Introduce *fair-PAC-learning* to theoretically treat these issues

First Canto: The Philosophy of Welfare and Malfare

Fair machine learning and the social planning problem

What is Welfare

- *Utility*: $\mathrm{U}(\cdot) : \mathcal{X} \to \mathbb{R}_{0+}$ Subjective measurement of *positive attribute*
  - Happiness, satisfaction, resource ownership

What is Welfare

- *Utility*: $\mathrm{U}(\cdot) : \mathcal{X} \to \mathbb{R}_{0+}$ Subjective measurement of *positive attribute*
  - Happiness, satisfaction, resource ownership

- *Welfare* summarizes *population-level* utility across $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_g)$
  "The subjective, measured objectively."

What is Welfare

- *Utility*: $\mathrm{U}(\cdot) : \mathcal{X} \to \mathbb{R}_{0+}$ Subjective measurement of *positive attribute*
  - Happiness, satisfaction, resource ownership
- *Welfare* summarizes *population-level* utility across $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_g)$
  "The subjective, measured objectively."
- *Utilitarian* welfare: *average utility* $\mathrm{U}(\cdot)$

$$\mathrm{W}_{\mathrm{Util}}\big(\mathrm{U}(\boldsymbol{x}_1), \ldots, \mathrm{U}(\boldsymbol{x}_g)\big) \doteq \frac{1}{g}\sum_{i=1}^{g}\mathrm{U}(\boldsymbol{x}_i)$$

What is Welfare

- *Utility*: $\mathrm{U}(\cdot) : \mathcal{X} \to \mathbb{R}_{0+}$ Subjective measurement of *positive attribute*
  - Happiness, satisfaction, resource ownership

- *Welfare* summarizes *population-level* utility across $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_g)$ *"The subjective, measured objectively."*
- *Utilitarian* welfare: *average utility* $\mathrm{U}(\cdot)$

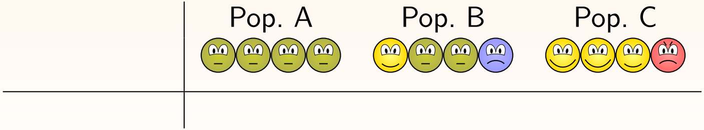
$$\mathrm{W}_{\mathrm{Util}}\big(\mathrm{U}(\boldsymbol{x}_1), \dots, \mathrm{U}(\boldsymbol{x}_g)\big) \doteq \frac{1}{g}\sum_{i=1}^{g} \mathrm{U}(\boldsymbol{x}_i)$$

- *Egalitarian* welfare: *worst-case utility* $\mathrm{U}(\cdot)$

$$\mathrm{W}_{\mathrm{Egal}}\big(\mathrm{U}(\boldsymbol{x}_1), \dots, \mathrm{U}(\boldsymbol{x}_g)\big) \doteq \min_{i \in 1, \dots, g} \mathrm{U}(\boldsymbol{x}_i)$$

  - A fair society should have *equality*
  - Incentivize aiding the most needy first

What is Welfare (cont.)



What is Welfare (cont.)

	Pop. A	Pop. B	Pop. C
	😐😐😐😐	🙂😐😐🙁	🙂🙂🙂😠
Utilitarian	😐	😐	🙂

What is Welfare (cont.)

	Pop. A 😐😐😐😐	Pop. B 🙂😐😐🙁	Pop. C 🙂🙂🙂😠
Utilitarian	😐	😐	🙂
Egalitarian	😐	🙁	😠

What is Welfare (cont.)

	Pop. A	Pop. B	Pop. C
	😐😐😐😐	🙂😐😐☹	🙂🙂🙂😠
Utilitarian	😐	😐	🙂
Egalitarian	😐	☹	😠

- Limitations
  - Nonnegativity
  - *Positive* directedness (utility is desirable)

What is Welfare (cont.)

	Pop. A 😐😐😐😐	Pop. B 🙂😐😐🙁	Pop. C 🙂🙂🙂😠
Utilitarian	😐	😐	🙂
Egalitarian	😐	🙁	😠

- Limitations
  - Nonnegativity
  - *Positive* directedness (utility is desirable)
- Which welfare function to use?
  - Analogy: *worst-case* vs *average case* bounds
  - Analogy: tail bounds vs expectation

The Power Mean

Suppose vector $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_g)$ representing *utility* or *loss* across a population

The Power Mean

Suppose vector $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_g)$ representing *utility* or *loss* across a population

$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}_i^p} \end{cases}$$

The Power Mean

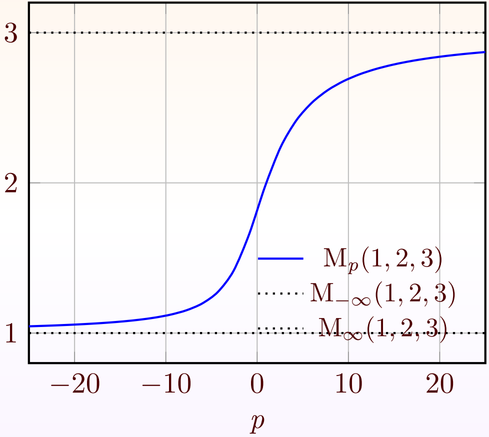
Suppose vector $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_g)$ representing *utility* or *loss* across a population

$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}_i^p} \\ p = -\infty & \min_{i \in 1,\ldots,g} \boldsymbol{\ell}_i \\ p = 0 & \sqrt[n]{\prod_{i=1}^{n} \boldsymbol{\ell}_i} \\ p = \infty & \max_{i \in 1,\ldots,g} \boldsymbol{\ell}_i \end{cases}$$

The Power Mean

Suppose vector $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_g)$ representing *utility* or *loss* across a population

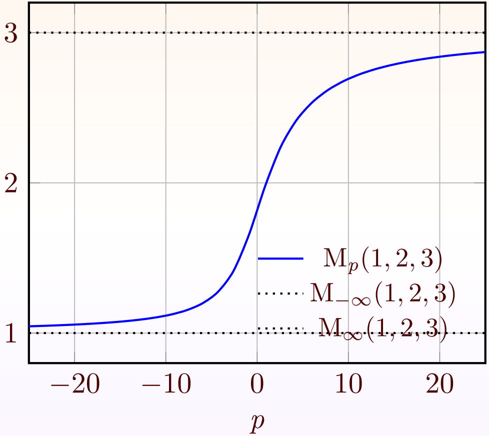
$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}_i^p} \\ p = -\infty & \min_{i \in 1,\ldots,g} \boldsymbol{\ell}_i \\ p = 0 & \sqrt[n]{\prod_{i=1}^{n} \boldsymbol{\ell}_i} \\ p = \infty & \max_{i \in 1,\ldots,g} \boldsymbol{\ell}_i \end{cases}$$



The Power Mean

Suppose vector $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_g)$ representing *utility* or *loss* across a population

$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}_i^p} \\ p = -\infty & \min_{i \in 1,\ldots,g} \boldsymbol{\ell}_i \\ p = 0 & \sqrt[n]{\prod_{i=1}^{n} \boldsymbol{\ell}_i} \\ p = \infty & \max_{i \in 1,\ldots,g} \boldsymbol{\ell}_i \end{cases}$$



- *Smooth interpolation* between *min*, *arithmetic mean*, and *max*
  - Other special cases: *geometric*, *harmonic*, and *quadratic* means
- Monotonic in p: *interpolate between* utilitarian and egalitarian

Welfare in Machine Learning

- Why do we care about cardinal welfare?
  - *Welfare* $\mathrm{W}(\cdot)$ encodes an *ideal notion* of *societal wellbeing* (fairness)
  - Utilitarian versus Egalitarian

Welfare in Machine Learning

- Why do we care about cardinal welfare?
  - *Welfare* $\mathrm{W}(\cdot)$ encodes an *ideal notion* of *societal wellbeing* (fairness)
  - Utilitarian versus Egalitarian
- The *social planning problem*
  - Select allocation of goods and services to *maximize welfare*
  - *Fair ML* is *learning* an *optimal allocation* from data?
  - Learn *policy* to maximize *welfare* of *per-group* utilities

Welfare in Machine Learning

- Why do we care about cardinal welfare?
  - *Welfare* $\mathrm{W}(\cdot)$ encodes an *ideal notion* of *societal wellbeing* (fairness)
  - Utilitarian versus Egalitarian
- The *social planning problem*
  - Select allocation of goods and services to *maximize welfare*
  - *Fair ML* is *learning* an *optimal allocation* from data?
  - Learn *policy* to maximize *welfare* of *per-group* utilities

Is it really that easy?

Welfare in Machine Learning

- Why do we care about cardinal welfare?
  - *Welfare* $\mathrm{W}(\cdot)$ encodes an *ideal notion* of *societal wellbeing* (fairness)
  - Utilitarian versus Egalitarian
- The *social planning problem*
  - Select allocation of goods and services to *maximize welfare*
  - *Fair ML* is *learning* an *optimal allocation* from data?
  - Learn *policy* to maximize *welfare* of *per-group* utilities

Is it really that easy?

What if we want to *minimize* a *loss function*?

Introducing Malfare

- Standard: *maximize* a *welfare measure* of *societal wellbeing* $\mathrm{W}(\cdot)$
- This work: *minimize* a *malfare measure* of *societal suffering* $\mathscr{M}(\cdot)$
  - Generically termed *aggregator functions* $\mathrm{M}(\cdot)$

Introducing Malfare

- Standard: *maximize* a *welfare measure* of *societal wellbeing* $\mathrm{W}(\cdot)$
- This work: *minimize* a *malfare measure* of *societal suffering* $\mathrm{M}(\cdot)$
  - Generically termed *aggregator functions* $\mathrm{M}(\cdot)$
- Is this really a new idea?
  - Everybody knows $\forall i: \operatorname*{argmax}_{h \in \mathcal{H}} -\boldsymbol{\ell}_i(h) = \operatorname*{argmin}_{h \in \mathcal{H}} \boldsymbol{\ell}_i(h)$
  - But we don't have $\operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}_p(-\boldsymbol{\ell}(h)) = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}_p(\boldsymbol{\ell}(h))$
  - Welfare is a *multivariate optimality concept*
  - Intuition from *univariate optimization* breaks down

Introducing Malfare

- Standard: *maximize* a *welfare measure* of *societal wellbeing* $\mathrm{W}(\cdot)$
- This work: *minimize* a *malfare measure* of *societal suffering* $\mathrm{M}(\cdot)$
  - Generically termed *aggregator functions* $\mathrm{M}(\cdot)$
- Is this really a new idea?
  - Everybody knows $\forall i: \operatorname*{argmax}_{h \in \mathcal{H}} -\boldsymbol{\ell}_i(h) = \operatorname*{argmin}_{h \in \mathcal{H}} \boldsymbol{\ell}_i(h)$
  - But we don't have $\operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}_p(-\boldsymbol{\ell}(h)) = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}_p(\boldsymbol{\ell}(h))$
  - Welfare is a *multivariate optimality concept*
  - Intuition from *univariate optimization* breaks down
- We shall see *equal axiomatic justification* for welfare and malfare

Introducing Malfare

- Standard: *maximize* a *welfare measure* of *societal wellbeing* $\mathrm{W}(\cdot)$
- This work: *minimize* a *malfare measure* of *societal suffering* $\mathcal{M}(\cdot)$
  - Generically termed *aggregator functions* $\mathrm{M}(\cdot)$
- Is this really a new idea?
  - Everybody knows $\forall i: \operatorname*{argmax}_{h \in \mathcal{H}} -\boldsymbol{\ell}_i(h) = \operatorname*{argmin}_{h \in \mathcal{H}} \boldsymbol{\ell}_i(h)$
  - But we don't have $\operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}_p(-\boldsymbol{\ell}(h)) = \operatorname*{argmin}_{h \in \mathcal{H}} \mathcal{M}_p(\boldsymbol{\ell}(h))$
  - Welfare is a *multivariate optimality concept*
  - Intuition from *univariate optimization* breaks down
- We shall see *equal axiomatic justification* for welfare and malfare

	😄	🙂	😐	🙁	😠
Utility	5	4	3	2	1
Loss	1	2	3	4	5

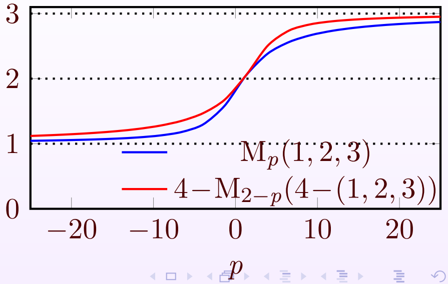
- *Malfare* extends the concept of *welfare* to *undesirable quantities* (disutility)
- Direct correspondence only for $p \in \{-\infty, 1, \infty\}$

Introducing Malfare

- Standard: *maximize* a *welfare measure* of *societal wellbeing* $\mathrm{W}(\cdot)$
- This work: *minimize* a *malfare measure* of *societal suffering* $\mathrm{M}(\cdot)$
  - Generically termed *aggregator functions* $\mathrm{M}(\cdot)$
- Is this really a new idea?
  - Everybody knows $\forall i: \operatorname*{argmax}_{h \in \mathcal{H}} -\boldsymbol{\ell}_i(h) = \operatorname*{argmin}_{h \in \mathcal{H}} \boldsymbol{\ell}_i(h)$
  - But we don't have $\operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}_p(-\boldsymbol{\ell}(h)) = \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}_p(\boldsymbol{\ell}(h))$
  - Welfare is a *multivariate optimality concept*
  - Intuition from *univariate optimization* breaks down
- We shall see *equal axiomatic justification* for welfare and malfare

	😀	🙂	😐	😟	😠
Utility	5	4	3	2	1
Loss	1	2	3	4	5

- *Malfare* extends the concept of *welfare* to *undesirable quantities* (disutility)
- Direct correspondence only for $p \in \{-\infty, 1, \infty\}$



Fair Machine Learning with Malfare Minimization

- Standard machine learning over *instance distribution* $\mathcal{D}$
  - Machine learning tasks often cast as *risk minimization* w.r.t. *loss function* ℓ

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} [\ell(h(x), y)]$$

Fair Machine Learning with Malfare Minimization

- Standard machine learning over *instance distribution* $\mathcal{D}$
  - Machine learning tasks often cast as *risk minimization* w.r.t. *loss function* ℓ

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\ell(h(x), y)]$$

- In fair machine learning, *different groups* have *different needs and preferences*
  - Need to consider *multiple distributions* $\mathcal{D}_1, \dots, \mathcal{D}_g$ over *multiple groups*
  - Analogously cast learning tasks as *malfare minimization*

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1} \left[\ell(h(x), y)\right], \dots, \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_g} \left[\ell(h(x), y)\right] \right)$$

Fair Machine Learning with Malfare Minimization

- Standard machine learning over *instance distribution* $\mathcal{D}$
  - Machine learning tasks often cast as *risk minimization* w.r.t. *loss function* ℓ

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$$

- In fair machine learning, *different groups* have *different needs and preferences*
  - Need to consider *multiple distributions* $\mathcal{D}_1, \ldots, \mathcal{D}_g$ over *multiple groups*
  - Analogously cast learning tasks as *malfare minimization*

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}[\ell(h(x), y)], \ldots, \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_g}[\ell(h(x), y)] \right)$$

- Contrast with welfare maximization:

  ⚠ Need to define a *utility function* $\mathrm{U}(\cdot)$

$$h^* \doteq \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}[\mathrm{U}(h(x), y)], \ldots, \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_g}[\mathrm{U}(h(x), y)] \right)$$

Fair Machine Learning with Malfare Minimization

- Standard machine learning over *instance distribution* $\mathcal{D}$
  - Machine learning tasks often cast as *risk minimization* w.r.t. *loss function* ℓ

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} [\ell(h(x), y)]$$

- In fair machine learning, *different groups* have *different needs and preferences*
  - Need to consider *multiple distributions* $\mathcal{D}_1, \dots, \mathcal{D}_g$ over *multiple groups*
  - Analogously cast learning tasks as *malfare minimization*

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\left(\mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}_1} [\ell(h(x), y)], \dots, \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}_g} [\ell(h(x), y)] \right)$$

- Contrast with welfare maximization:

  ⚠ Need to define a *utility function* $\mathrm{U}(\cdot)$

$$h^* \doteq \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}\left(\mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}_1} [\mathrm{U}(h(x), y)], \dots, \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}_g} [\mathrm{U}(h(x), y)] \right)$$

⚠ Don't know $\mathcal{D}_{1:g}$; have to work from *training samples*

- Select $\hat{h}$ to optimize empirical *risk* / *malfare* / *welfare* estimates

Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
   - Adding utility never harms welfare

Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
   - Adding utility never harms welfare
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
   - No *exceptionalism*; welfare is *identity blind*
   - Inherent tenet of *fairness* and *equality*

Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
   - Adding utility never harms welfare
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
   - No *exceptionalism*; welfare is *identity blind*
   - Inherent tenet of *fairness* and *equality*
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*

Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
   - Adding utility never harms welfare
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
   - No *exceptionalism*; welfare is *identity blind*
   - Inherent tenet of *fairness* and *equality*
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. Independence of Unconcerned Agents (IUA):
   $\forall a, b \in \mathbb{R}_+ :\ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
   - Compartmentalization and analysis of *subgroups*

Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
    - Adding utility never harms welfare
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
    - No *exceptionalism*; welfare is *identity blind*
    - Inherent tenet of *fairness* and *equality*
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. Independence of Unconcerned Agents (IUA):
   $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
    - Compartmentalization and analysis of *subgroups*
5. Independence of Common Scale (ICS):
   $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
    - Relative value *invariant* under (absolute) unit conversion: \$ versus ₽

Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
   - Adding utility never harms welfare
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
   - No *exceptionalism*; welfare is *identity blind*
   - Inherent tenet of *fairness* and *equality*
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. Independence of Unconcerned Agents (IUA):
   $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
   - Compartmentalization and analysis of *subgroups*
5. Independence of Common Scale (ICS):
   $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
   - Relative value *invariant* under (absolute) unit conversion: \$ versus ₽

Theorem (Debreu-Gorman (1959))

For some strictly increasing F, $p \in \mathbb{R}$, all welfare functions satisfying 1-5 take form

$$\mathrm{M}(\boldsymbol{\ell}) = F\left(\operatorname{sgn}(p)\sum_{i=1}^{g} \boldsymbol{\ell}_i^p\right) \quad \textit{or} \quad \mathrm{M}(\boldsymbol{\ell}) = F\left(\prod_{i=1}^{g} \boldsymbol{\ell}_i\right)$$

⚠ Can be *non-Lipschitz*, arbitrarily hard to *estimate* from *sample*

Extended Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$

Extended Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$;
   - Implies ICS
   - Units of $\mathrm{M}(\cdot)$ match units of $\boldsymbol{\ell}$

Extended Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$;
   - Implies ICS
   - Units of $\mathrm{M}(\cdot)$ match units of $\boldsymbol{\ell}$
7. Unit Scale: $\mathrm{M}(\mathbf{1}) = 1$
   - Scale of M matches units of $\boldsymbol{\ell}$
   - Absolute comparison: “My income is $x\%$ of average / maximum / minimum”

Extended Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$;
   - Implies ICS
   - Units of $\mathrm{M}(\cdot)$ match units of $\boldsymbol{\ell}$
7. Unit Scale: $\mathrm{M}(\mathbf{1}) = 1$
   - Scale of M matches units of $\boldsymbol{\ell}$
   - Absolute comparison: "My income is $x\%$ of average / maximum / minimum"

Theorem (Axiomatic Characterization of Welfare and Malfare)

For any aggregator function $\mathrm{M}(\cdot)$ *satisfying 1-7,* $\exists p \in \mathbb{R}$ *s.t.*

$$\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}_p(\boldsymbol{\ell}) = \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\ell}_i^p} \quad \textit{or} \quad \mathrm{M}(\boldsymbol{\ell}) = \sqrt[g]{\prod_{i=1}^{g} \boldsymbol{\ell}_i} \ ,$$

which are Lipschitz-continuous in $\boldsymbol{\ell}$ *for* $p \in (-\infty, 0) \cup [1, \infty)$.

Transfer Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$
7. Unit Scale: $\mathrm{M}(\mathbf{1}) = 1$

Transfer Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ :\ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ :\ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$
7. Unit Scale: $\mathrm{M}(\mathbf{1}) = 1$

All apply *equally well* to welfare and malfare

Transfer Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ : \ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$
7. Unit Scale: $\mathrm{M}(\mathbf{1}) = 1$
8. Pigou-Dalton Principle: Suppose $\boldsymbol{\ell}, \boldsymbol{\ell}'$ s.t. $\mathrm{M}_1(\boldsymbol{\ell}) = \mathrm{M}_1(\boldsymbol{\ell}') = \mu$. Then

$$\bigwedge_{i \in 1, \ldots, g} \left(|\boldsymbol{\ell}'_i - \mu| \geq |\boldsymbol{\ell}_i - \mu| \right) \implies \mathrm{W}(\boldsymbol{\ell}') \leq \mathrm{W}(\boldsymbol{\ell})$$

Transfer Axioms of Cardinal Welfare

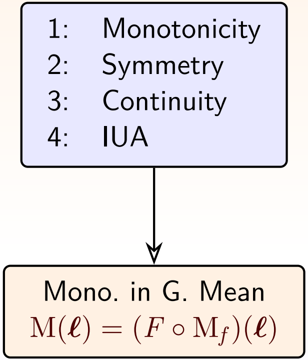
1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ :\ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ :\ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$
7. Unit Scale: $\mathrm{M}(\mathbf{1}) = 1$
8. Pigou-Dalton Principle: Suppose $\boldsymbol{\ell}, \boldsymbol{\ell}'$ s.t. $\mathrm{M}_1(\boldsymbol{\ell}) = \mathrm{M}_1(\boldsymbol{\ell}') = \mu$. Then
$$\bigwedge_{i \in 1,\dots,g} \left(|\boldsymbol{\ell}'_i - \mu| \geq |\boldsymbol{\ell}_i - \mu|\right) \implies \mathrm{W}(\boldsymbol{\ell}') \leq \mathrm{W}(\boldsymbol{\ell})$$
9. Anti-Pigou-Dalton Principle: Suppose as in (8). Then require *inverse*
$$\bigwedge_{i \in 1,\dots,g} \left(|\boldsymbol{\ell}'_i - \mu| \geq |\boldsymbol{\ell}_i - \mu|\right) \implies \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})$$

Transfer Axioms of Cardinal Welfare

1. Strict Monotonicity: $\forall \boldsymbol{\varepsilon} \succeq 0$ s.t. $\boldsymbol{\varepsilon} \neq \mathbf{0}$: $\mathrm{M}(\boldsymbol{\ell}) < \mathrm{M}(\boldsymbol{\ell} + \boldsymbol{\varepsilon})$
2. Symmetry: $\forall$ permutations π: $\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}(\pi(\boldsymbol{\ell}))$
3. Continuity: $\forall \boldsymbol{\ell} : \{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \leq \mathrm{M}(\boldsymbol{\ell})\}$ and $\{\boldsymbol{\ell}' \mid \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})\}$ are *closed sets*
4. IUA: $\forall a, b \in \mathbb{R}_+ :\ \mathrm{M}(\boldsymbol{\ell}, a) \leq \mathrm{M}(\boldsymbol{\ell}', a) \Leftrightarrow \mathrm{M}(\boldsymbol{\ell}, b) \leq \mathrm{M}(\boldsymbol{\ell}', b)$
5. ICS: $\forall \alpha \in \mathbb{R}_+ :\ \mathrm{M}(\boldsymbol{\ell}) \leq \mathrm{M}(\boldsymbol{\ell}') \implies \mathrm{M}(\alpha\boldsymbol{\ell}) \leq \mathrm{M}(\alpha\boldsymbol{\ell}')$
6. Multiplicative Linearity: $\mathrm{M}(\alpha\boldsymbol{\ell}) = \alpha\mathrm{M}(\boldsymbol{\ell})$
7. Unit Scale: $\mathrm{M}(\mathbf{1}) = 1$
8. Pigou-Dalton Principle: Suppose $\boldsymbol{\ell}, \boldsymbol{\ell}'$ s.t. $\mathrm{M}_1(\boldsymbol{\ell}) = \mathrm{M}_1(\boldsymbol{\ell}') = \mu$. Then
$$\bigwedge_{i \in 1,\dots,g} \left(\left|\boldsymbol{\ell}'_i - \mu\right| \geq \left|\boldsymbol{\ell}_i - \mu\right|\right) \implies \mathrm{W}(\boldsymbol{\ell}') \leq \mathrm{W}(\boldsymbol{\ell})$$
9. Anti-Pigou-Dalton Principle: Suppose as in (8). Then require *inverse*
$$\bigwedge_{i \in 1,\dots,g} \left(\left|\boldsymbol{\ell}'_i - \mu\right| \geq \left|\boldsymbol{\ell}_i - \mu\right|\right) \implies \mathrm{M}(\boldsymbol{\ell}') \geq \mathrm{M}(\boldsymbol{\ell})$$

	😐😐😐😐	😊😐😐☹
Welfare	w	$\leq w$
Malfare	m	$\geq m$

Properties of Welfare and Malfare

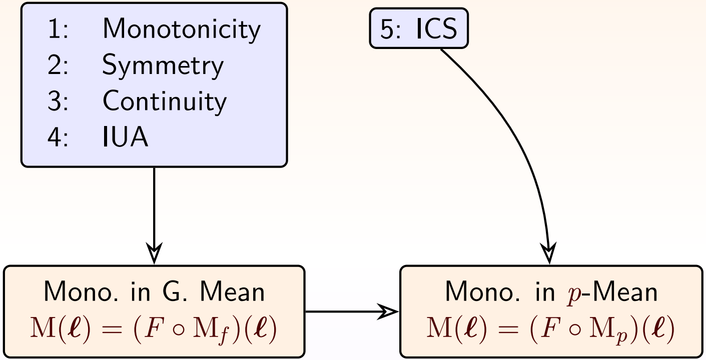


1: Monotonicity
2: Symmetry
3: Continuity
4: IUA

Mono. in G. Mean

$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_f)(\boldsymbol{\ell})$

Properties of Welfare and Malfare



1: Monotonicity
2: Symmetry
3: Continuity
4: IUA

5: ICS

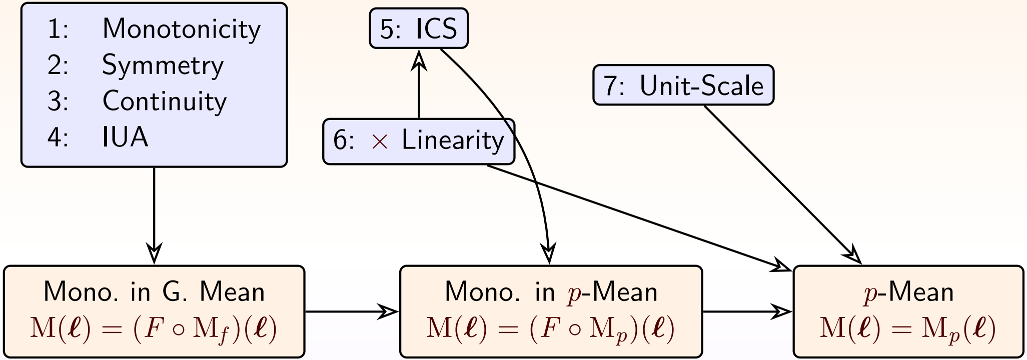
Mono. in G. Mean

$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_f)(\boldsymbol{\ell})$

Mono. in p-Mean

$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_p)(\boldsymbol{\ell})$

Properties of Welfare and Malfare



1: Monotonicity
2: Symmetry
3: Continuity
4: IUA

5: ICS

6: × Linearity

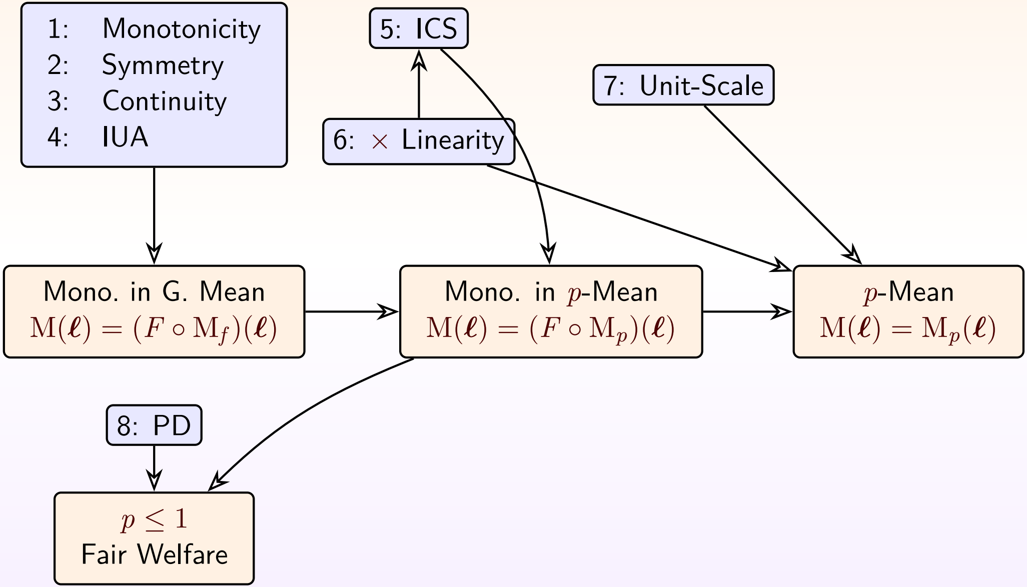
7: Unit-Scale

Mono. in G. Mean
$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_f)(\boldsymbol{\ell})$

Mono. in p-Mean
$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_p)(\boldsymbol{\ell})$

p-Mean
$\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}_p(\boldsymbol{\ell})$

Properties of Welfare and Malfare



1: Monotonicity
2: Symmetry
3: Continuity
4: IUA

5: ICS

6: × Linearity

7: Unit-Scale

Mono. in G. Mean
$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_f)(\boldsymbol{\ell})$

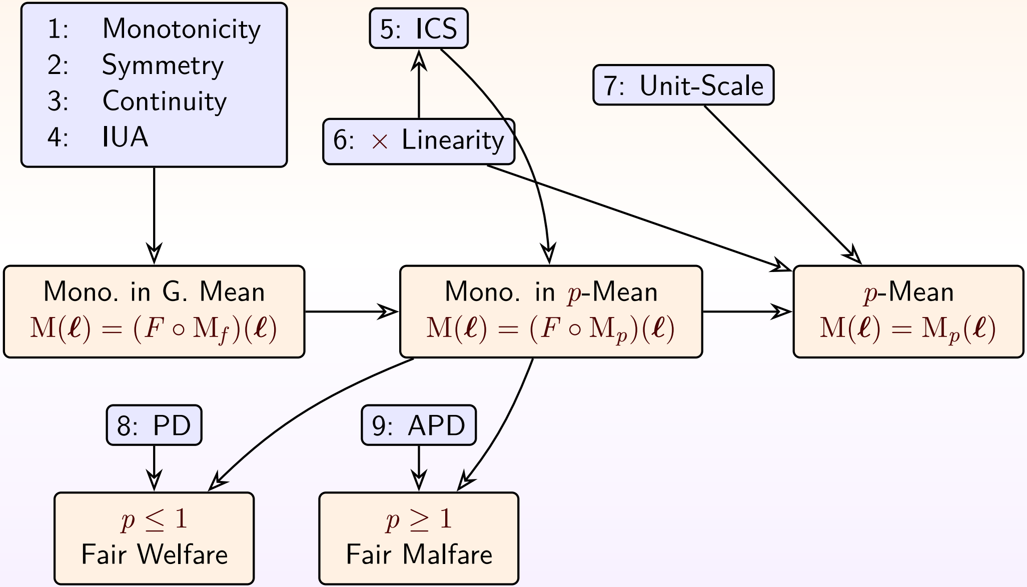
Mono. in p-Mean
$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_p)(\boldsymbol{\ell})$

p-Mean
$\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}_p(\boldsymbol{\ell})$

8: PD

$p \leq 1$
Fair Welfare

Properties of Welfare and Malfare



1: Monotonicity
2: Symmetry
3: Continuity
4: IUA

5: ICS

6: × Linearity

7: Unit-Scale

Mono. in G. Mean
$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_f)(\boldsymbol{\ell})$

Mono. in p-Mean
$\mathrm{M}(\boldsymbol{\ell}) = (F \circ \mathrm{M}_p)(\boldsymbol{\ell})$

p-Mean
$\mathrm{M}(\boldsymbol{\ell}) = \mathrm{M}_p(\boldsymbol{\ell})$

8: PD

9: APD

$p \leq 1$
Fair Welfare

$p \geq 1$
Fair Malfare

Second Canto: Estimation, Inference, and Fair Machine Learning

The Statistics of Fair Machine Learning as Malfare Minimization

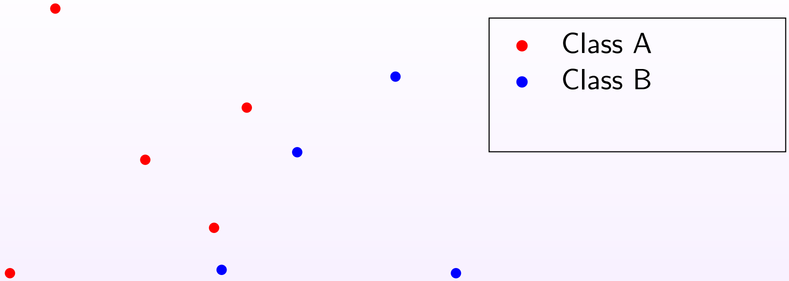
Learning Linear Classifiers

1. *How much* training data do we need?
   - Concentration inequalities
   - Vapnik-Chervonenkis dimension
   - Rademacher averages

Learning Linear Classifiers

❶ *How much* training data do we need?

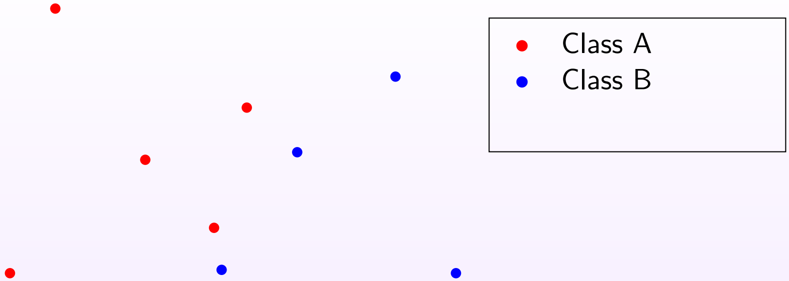
- Concentration inequalities
- Vapnik-Chervonenkis dimension
- Rademacher averages



Learning Linear Classifiers

1. *How much* training data do we need?
   - Concentration inequalities
   - Vapnik-Chervonenkis dimension
   - Rademacher averages
2. How can we *learn* from the data?
   - *Empirical Risk Minimization*: select $h \in \mathcal{H}$ to optimize

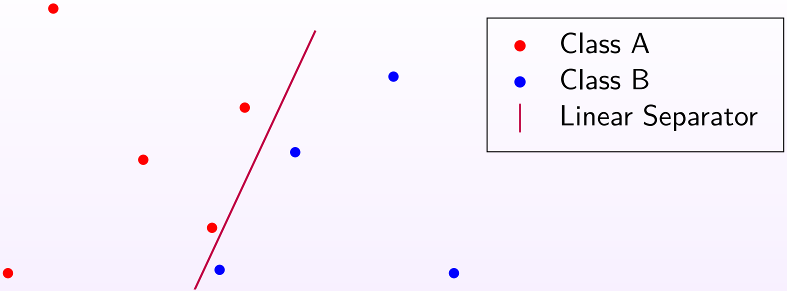
$$\hat{\mathrm{R}}(h) \doteq \frac{1}{m} \sum_{j=1}^{m} \ell(h(\boldsymbol{x}_j), \boldsymbol{y}_j)$$



Learning Linear Classifiers

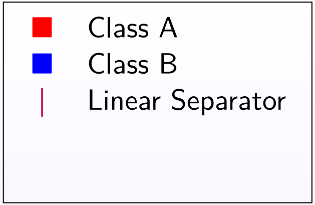
1. *How much* training data do we need?
   - Concentration inequalities
   - Vapnik-Chervonenkis dimension
   - Rademacher averages
2. How can we *learn* from the data?
   - *Empirical Risk Minimization*: select $h \in \mathcal{H}$ to optimize

$$\hat{\mathrm{R}}(h) \doteq \frac{1}{m}\sum_{j=1}^{m} \ell(h(\boldsymbol{x}_j), \boldsymbol{y}_j)$$



Learning Fair Linear Classifiers

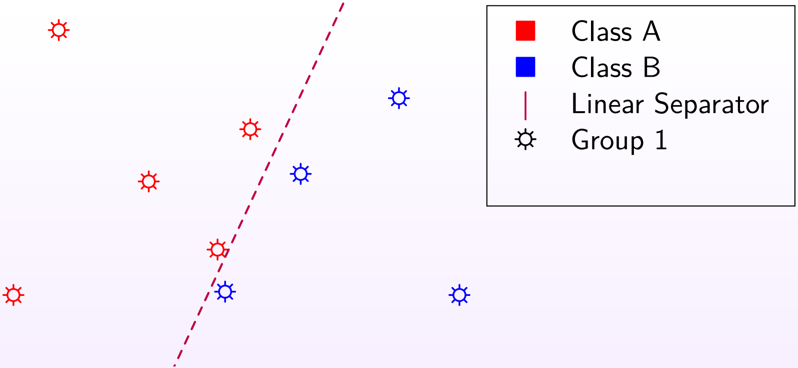
- What changes with *multiple groups* $(\boldsymbol{x}_{1:g,1:m}, \boldsymbol{y}_{1:g,1:m})$?



Learning Fair Linear Classifiers

- What changes with *multiple groups* $(\boldsymbol{x}_{1:g,1:m}, \boldsymbol{y}_{1:g,1:m})$?
- We can handle each group individually:

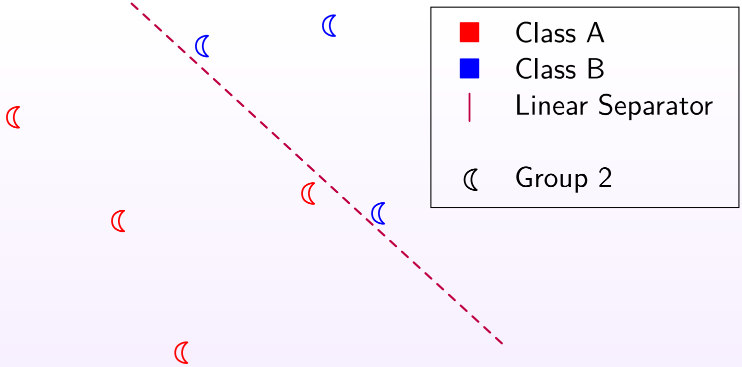
$$\hat{\mathrm{R}}(h;\boldsymbol{x}_i,\boldsymbol{y}_i) \doteq \frac{1}{m}\sum_{j=1}^{m}\ell\big(h(\boldsymbol{x}_{i,j}),\boldsymbol{y}_{i,j}\big); \quad \forall i : \; \hat{h}_i \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \hat{\mathrm{R}}(h;\boldsymbol{x}_i,\boldsymbol{y}_i)$$



Learning Fair Linear Classifiers

- What changes with *multiple groups* $(\boldsymbol{x}_{1:g,1:m}, \boldsymbol{y}_{1:g,1:m})$?
- We can handle each group individually:

$$\hat{\mathrm{R}}(h; \boldsymbol{x}_i, \boldsymbol{y}_i) \doteq \frac{1}{m}\sum_{j=1}^{m} \ell\big(h(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}\big); \quad \forall i: \;\; \hat{h}_i \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\mathrm{R}}(h; \boldsymbol{x}_i, \boldsymbol{y}_i)$$

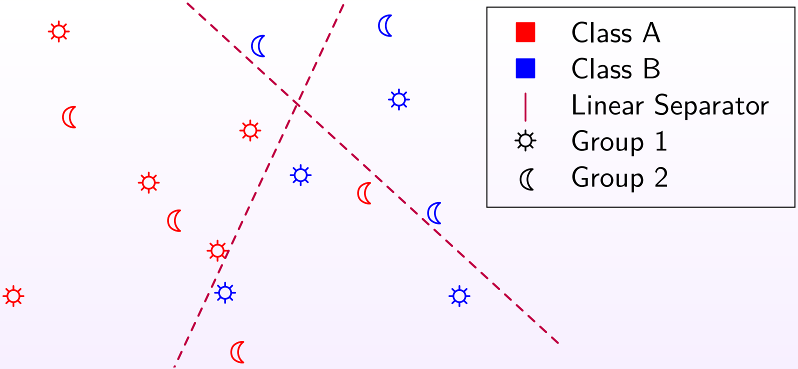


Learning Fair Linear Classifiers

- What changes with *multiple groups* $(\boldsymbol{x}_{1:g,1:m}, \boldsymbol{y}_{1:g,1:m})$?
- We can handle each group individually:

$$\hat{\mathrm{R}}(h; \boldsymbol{x}_i, \boldsymbol{y}_i) \doteq \frac{1}{m}\sum_{j=1}^{m} \ell\big(h(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}\big); \quad \forall i : \; \hat{h}_i \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\mathrm{R}}(h; \boldsymbol{x}_i, \boldsymbol{y}_i)$$

- What is the best classifier *overall?*

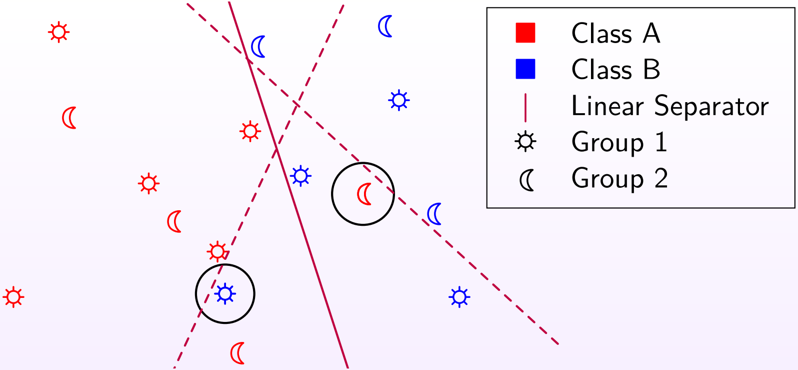


Learning Fair Linear Classifiers

- What changes with *multiple groups* $(\boldsymbol{x}_{1:g,1:m}, \boldsymbol{y}_{1:g,1:m})$?
- We can handle each group individually:

$$\hat{\mathrm{R}}(h; \boldsymbol{x}_i, \boldsymbol{y}_i) \doteq \frac{1}{m}\sum_{j=1}^{m} \ell\big(h(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}\big); \quad \forall i : \ \hat{h}_i \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \hat{\mathrm{R}}(h; \boldsymbol{x}_i, \boldsymbol{y}_i)$$

- What is the best classifier *overall?*
  - *Empirical malfare minimization* $\hat{h} \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{M}\Big(\hat{\mathrm{R}}(h; \boldsymbol{x}_1, \boldsymbol{y}_1), \hat{\mathrm{R}}(h; \boldsymbol{x}_2, \boldsymbol{y}_2)\Big)$



Statistical Estimation

- Suppose *sample mean* $\hat{\boldsymbol{\ell}}_i \doteq \frac{1}{m}\sum_{j=1}^{m} \ell(\boldsymbol{x}_{i,j})$, *true mean* $\boldsymbol{\ell}_i \doteq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i}[\ell(x)]$

Statistical Estimation

- Suppose *sample mean* $\hat{\boldsymbol{\ell}}_i \doteq \frac{1}{m}\sum_{j=1}^{m} \ell(\boldsymbol{x}_{i,j})$, *true mean* $\boldsymbol{\ell}_i \doteq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i}[\ell(x)]$
- By *continuity* and the *law of large numbers*:

$$\lim_{m\to\infty} \mathrm{M}(\hat{\boldsymbol{\ell}}) = \mathrm{M}(\boldsymbol{\ell})$$

Statistical Estimation

- Suppose *sample mean* $\hat{\boldsymbol{\ell}}_i \doteq \frac{1}{m}\sum_{j=1}^{m} \ell(\boldsymbol{x}_{i,j})$, *true mean* $\boldsymbol{\ell}_i \doteq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i}[\ell(x)]$
- By *continuity* and the *law of large numbers*:

$$\lim_{m \to \infty} \mathrm{M}(\hat{\boldsymbol{\ell}}) = \mathrm{M}(\boldsymbol{\ell})$$

- For finite sample size m
  - $\mathbb{E}[\mathrm{M}(\hat{\boldsymbol{\ell}})] \neq \mathrm{M}(\boldsymbol{\ell})$
  - $\mathrm{M}(\hat{\boldsymbol{\ell}})$ is a *biased estimator* of $\mathrm{M}(\boldsymbol{\ell})$!

Statistical Estimation

- Suppose *sample mean* $\hat{\boldsymbol{\ell}}_i \doteq \frac{1}{m}\sum_{j=1}^{m} \ell(\boldsymbol{x}_{i,j})$, *true mean* $\boldsymbol{\ell}_i \doteq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i}[\ell(x)]$
- By *continuity* and the *law of large numbers*:

$$\lim_{m \to \infty} \mathrm{M}(\hat{\boldsymbol{\ell}}) = \mathrm{M}(\boldsymbol{\ell})$$

- For finite sample size m
  - $\mathbb{E}[\mathrm{M}(\hat{\boldsymbol{\ell}})] \neq \mathrm{M}(\boldsymbol{\ell})$
  - $\mathrm{M}(\hat{\boldsymbol{\ell}})$ is a *biased estimator* of $\mathrm{M}(\boldsymbol{\ell})$!

Theorem (A Hoeffding-Type Malfare-Estimation Bound)

Suppose fair malfare $\mathrm{M}_p(\cdot)$ ($p \geq 1$), g groups, and loss range r. Then with probability at least $1 - \delta$

$$\left|\mathrm{M}(\boldsymbol{\ell}) - \mathrm{M}(\hat{\boldsymbol{\ell}})\right| \leq r\sqrt{\frac{\ln \frac{2g}{\delta}}{2m}}$$

Statistical Estimation (contd.)

Theorem (A Bernstein-Type Malfare-Estimation Bound)

Suppose fair malfare $\mathrm{M}_p(\cdot)$ ($p \geq 1$), g groups, loss range r, and maximum variance $v_{\max}$. Then with probability at least $1-\delta$ over sampling, we have

$$\left|\mathrm{M}(\boldsymbol{\ell}) - \mathrm{M}(\hat{\boldsymbol{\ell}})\right| \leq \underbrace{\frac{r \ln \frac{2g}{\delta}}{3m}}_{\textsc{Scale Term}} + \underbrace{\sqrt{\frac{v_{\max} \ln \frac{2g}{\delta}}{2m}}}_{\textsc{Variance Term}}$$

Statistical Estimation (contd.)

Theorem (A Bernstein-Type Malfare-Estimation Bound)

Suppose fair malfare $\mathrm{M}_p(\cdot)$ *(*$p \geq 1$*),* g *groups, loss range* r*, and maximum variance* $v_{\max}$*. Then with probability at least* $1 - \delta$ *over sampling, we have*

$$\left|\mathrm{M}(\boldsymbol{\ell}) - \mathrm{M}(\hat{\boldsymbol{\ell}})\right| \leq \underbrace{\frac{r \ln \frac{2g}{\delta}}{3m}}_{\textsc{Scale Term}} + \underbrace{\sqrt{\frac{v_{\max} \ln \frac{2g}{\delta}}{2m}}}_{\textsc{Variance Term}}$$

- Can show similar bounds for *any concentration inequality*
- *Uniform bounds* for a learnable family $\mathcal{H}$ with Rademacher averages

$$\sup_{h \in \mathcal{H}} \left|\mathrm{M}(\boldsymbol{\ell}(h)) - \mathrm{M}(\hat{\boldsymbol{\ell}}(h))\right| \leq \max_{i \in 1,\dots,g} 2\mathfrak{R}_m(\mathcal{F}, \mathcal{D}_i) + \varepsilon \in \boldsymbol{\Theta}\left(\frac{r \ln \frac{g}{\delta}}{m} + \sqrt{\frac{v_{\max} \ln \frac{g}{\delta}}{m}}\right)$$

Statistical Estimation (contd.)

Theorem (A Bernstein-Type Malfare-Estimation Bound)

Suppose fair malfare $\mathrm{M}_p(\cdot)$ $(p \geq 1)$, g groups, loss range r, and maximum variance $v_{\max}$. Then with probability at least $1-\delta$ over sampling, we have

$$\left|\mathrm{M}(\boldsymbol{\ell}) - \mathrm{M}(\hat{\boldsymbol{\ell}})\right| \leq \underbrace{\frac{r \ln \frac{2g}{\delta}}{3m}}_{\textsc{Scale Term}} + \underbrace{\sqrt{\frac{v_{\max} \ln \frac{2g}{\delta}}{2m}}}_{\textsc{Variance Term}}$$

- Can show similar bounds for *any concentration inequality*
- *Uniform bounds* for a learnable family $\mathcal{H}$ with Rademacher averages

$$\sup_{h \in \mathcal{H}} \left|\mathrm{M}(\boldsymbol{\ell}(h)) - \mathrm{M}(\hat{\boldsymbol{\ell}}(h))\right| \leq \max_{i \in 1,\ldots,g} 2\mathfrak{R}_m(\mathcal{F}, \mathcal{D}_i) + \varepsilon \in \boldsymbol{\Theta}\left(\frac{r \ln \frac{g}{\delta}}{m} + \sqrt{\frac{v_{\max} \ln \frac{g}{\delta}}{m}}\right)$$

- Control *overfitting* in *machine learning*
  - Finite $\mathcal{H}$, bounded Lipschitz families
  - Bounded linear regression, finite-dimensional linear classifiers,

Statistical Estimation (contd.)

Theorem (A Bernstein-Type Malfare-Estimation Bound)

Suppose fair malfare $\mathrm{M}_p(\cdot)$ ($p \geq 1$), g groups, loss range r, and maximum variance $v_{\max}$. Then with probability at least $1-\delta$ over sampling, we have

$$\left|\mathrm{M}(\boldsymbol{\ell}) - \mathrm{M}(\hat{\boldsymbol{\ell}})\right| \leq \underbrace{\frac{r \ln \frac{2g}{\delta}}{3m}}_{\text{Scale Term}} + \underbrace{\sqrt{\frac{v_{\max} \ln \frac{2g}{\delta}}{2m}}}_{\text{Variance Term}}$$

- Can show similar bounds for *any concentration inequality*
- *Uniform bounds* for a learnable family $\mathcal{H}$ with Rademacher averages

$$\sup_{h \in \mathcal{H}} \left|\mathrm{M}(\boldsymbol{\ell}(h)) - \mathrm{M}(\hat{\boldsymbol{\ell}}(h))\right| \leq \max_{i \in 1,\ldots,g} 2\mathfrak{R}_m(\mathcal{F}, \mathcal{D}_i) + \varepsilon \in \boldsymbol{\Theta}\left(\frac{r \ln \frac{g}{\delta}}{m} + \sqrt{\frac{v_{\max} \ln \frac{g}{\delta}}{m}}\right)$$

  - Control *overfitting* in *machine learning*
    - Finite $\mathcal{H}$, bounded Lipschitz families
    - Bounded linear regression, finite-dimensional linear classifiers, Generalized linear models, support vector machines, multiple kernel learning, bounded depth decision trees, rank-constrained matrix factorization, neural networks,

Statistical Estimation (contd.)

Theorem (A Bernstein-Type Malfare-Estimation Bound)

Suppose fair malfare $\mathrm{M}_p(\cdot)$ *(*$p \geq 1$*),* g *groups, loss range* r*, and maximum variance* $v_{\max}$*. Then with probability at least* $1 - \delta$ *over sampling, we have*

$$\left| \mathrm{M}(\boldsymbol{\ell}) - \mathrm{M}(\hat{\boldsymbol{\ell}}) \right| \leq \underbrace{\frac{r \ln \frac{2g}{\delta}}{3m}}_{\textsc{Scale Term}} + \underbrace{\sqrt{\frac{v_{\max} \ln \frac{2g}{\delta}}{2m}}}_{\textsc{Variance Term}}$$

- Can show similar bounds for *any concentration inequality*
- *Uniform bounds* for a learnable family $\mathcal{H}$ with Rademacher averages

$$\sup_{h \in \mathcal{H}} \left| \mathrm{M}(\boldsymbol{\ell}(h)) - \mathrm{M}(\hat{\boldsymbol{\ell}}(h)) \right| \leq \max_{i \in 1,\dots,g} 2\mathfrak{R}_m(\mathcal{F}, \mathcal{D}_i) + \varepsilon \in \boldsymbol{\Theta}\left(\frac{r \ln \frac{g}{\delta}}{m} + \sqrt{\frac{v_{\max} \ln \frac{g}{\delta}}{m}} \right)$$

- Control *overfitting* in *machine learning*
  - Finite $\mathcal{H}$, bounded Lipschitz families
  - Bounded linear regression, finite-dimensional linear classifiers, Generalized linear models, support vector machines, multiple kernel learning, bounded depth decision trees, rank-constrained matrix factorization, neural networks, (constrained) Boolean formulae, boosting methods, convex ensemble methods, learning distance metrics, . . .

Third Canto: Fair Probably Approximately Correct Learning

A generic theory of fair statistical and computational learning

Classical Statistical Learning Theory

- Consider *linear classification*: $\mathcal{H}_d \doteq \left\{ \vec{x} \mapsto \operatorname{sgn}(\vec{w} \cdot \vec{x}) \,\middle|\, \vec{w} \in \mathbb{R}^d \right\}$
  - Optimize *risk* $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(y, h(x))\right]$, for 0-1 loss $\ell(y, \hat{y}) = 1 - \mathbb{1}_y(\hat{y})$

Classical Statistical Learning Theory

- Consider *linear classification*: $\mathcal{H}_d \doteq \left\{ \vec{x} \mapsto \mathrm{sgn}(\vec{w} \cdot \vec{x}) \,\middle|\, \vec{w} \in \mathbb{R}^d \right\}$
  - Optimize *risk* $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(y, h(x))\right]$, for 0-1 loss $\ell(y, \hat{y}) = 1 - \mathbb{1}_y(\hat{y})$
- Can this class be *efficiently learned*?

What does that even mean?

Classical Statistical Learning Theory

- Consider *linear classification*: $\mathcal{H}_d \doteq \left\{ \vec{x} \mapsto \mathrm{sgn}(\vec{w} \cdot \vec{x}) \,\middle|\, \vec{w} \in \mathbb{R}^d \right\}$
  - Optimize *risk* $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(y, h(x))\right]$, for 0-1 loss $\ell(y, \hat{y}) = 1 - \mathbb{1}_y(\hat{y})$
- Can this class be *efficiently learned*? What does that even mean?

Definition (PAC Learning)

Suppose

❶ Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$ ❷ Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$

$\mathcal{H}$ is *PAC-learnable* w.r.t. ℓ iff $\exists$ algorithm A s.t. $\forall$

❶ distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$

❷ additive errors $\varepsilon > 0$

❸ failure probabilities $\delta \in (0, 1)$

A can identify a hypothesis $\hat{h} \in \mathcal{H}$ s.t.

❶ A has $\mathrm{m}(\varepsilon, \delta) < \infty$ sample complexity

❷ with probability at least $1 - \delta$, $\hat{h}$ obeys

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(y, \hat{h}(x))\right] \leq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(y, h^*(x))\right] + \varepsilon$$

Classical Statistical Learning Theory

- Consider *linear classification*: $\mathcal{H}_d \doteq \left\{ \vec{x} \mapsto \operatorname{sgn}(\vec{w} \cdot \vec{x}) \,\middle|\, \vec{w} \in \mathbb{R}^d \right\}$
  - Optimize *risk* $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(y, h(x))\right]$, for 0-1 loss $\ell(y, \hat{y}) = 1 - \mathbb{1}_y(\hat{y})$
- Can this class be *efficiently learned*? What does that even mean?

Definition (PAC Learning)

Suppose

❶ Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$ ❷ Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$

$\mathcal{H}$ is *PAC-learnable* w.r.t. ℓ iff $\exists$ algorithm A s.t. $\forall$

❶ distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ ❸ failure probabilities $\delta \in (0, 1)$

❷ additive errors $\varepsilon > 0$

A can identify a hypothesis $\hat{h} \in \mathcal{H}$ s.t.

❶ A has $\mathrm{m}(\varepsilon, \delta) < \infty$ sample complexity

❷ with probability at least $1 - \delta$, $\hat{h}$ obeys

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(y, \hat{h}(x))\right] \leq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(y, h^*(x))\right] + \varepsilon$$

- May also consider *efficient PAC-learnable*: require *poly-time* A

Fairness and Statistical Learning Theory

Definition (Fair-PAC Learning)

Suppose

1. Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$
2. Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$

$\mathcal{H}$ is *fair PAC-learnable* w.r.t. ℓ iff $\exists$ algorithm A s.t. $\forall$

1. distributions $\mathcal{D}_{1:g}$ over $(\mathcal{X} \times \mathcal{Y})^g$
2. fair malfare functions $\mathcal{M}(\cdot)$
3. additive errors $\varepsilon > 0$
4. failure probabilities $\delta \in (0, 1)$

A can identify a hypothesis $\hat{h} \in \mathcal{H}$ s.t.

1. A has $\mathrm{m}(\varepsilon, \delta, g)$ sample complexity
2. with probability at least $1 - \delta$, $\hat{h}$ obeys

$$\mathcal{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1} \left[\ell(y, \hat{h}(x))\right], \dots \right) \leq \mathop{\mathrm{argmin}}_{h^* \in \mathcal{H}} \mathcal{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1} \left[\ell(y, h^*(x))\right], \dots \right) + \varepsilon$$

Fairness and Statistical Learning Theory

Definition (Fair-PAC Learning)

Suppose

❶ Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$

❷ Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$

$\mathcal{H}$ is *fair PAC-learnable* w.r.t. ℓ iff $\exists$ algorithm A s.t. $\forall$

❶ distributions $\mathcal{D}_{1:g}$ over $(\mathcal{X} \times \mathcal{Y})^g$

❷ fair malfare functions $\mathrm{M}(\cdot)$

❸ additive errors $\varepsilon > 0$

❹ failure probabilities $\delta \in (0, 1)$

A can identify a hypothesis $\hat{h} \in \mathcal{H}$ s.t.

❶ A has $\mathrm{m}(\varepsilon, \delta, g)$ sample complexity

❷ with probability at least $1 - \delta$, $\hat{h}$ obeys

$$\mathrm{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1} \left[\ell(y, \hat{h}(x))\right], \dots \right) \leq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathrm{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1} \left[\ell(y, h^*(x))\right], \dots \right) + \varepsilon$$

- Do we capture a valuable, generic notion of fair learning?
  - Axiomatic social planning problem motivation

Fairness and Statistical Learning Theory

Definition (Fair-PAC Learning)

Suppose

1. Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$
2. Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$

$\mathcal{H}$ is *fair PAC-learnable* w.r.t. ℓ iff $\exists$ algorithm A s.t. $\forall$

1. distributions $\mathcal{D}_{1:g}$ over $(\mathcal{X} \times \mathcal{Y})^g$
2. fair malfare functions $\mathcal{M}(\cdot)$
3. additive errors $\varepsilon > 0$
4. failure probabilities $\delta \in (0, 1)$

A can identify a hypothesis $\hat{h} \in \mathcal{H}$ s.t.

1. A has $\mathrm{m}(\varepsilon, \delta, g)$ sample complexity
2. with probability at least $1 - \delta$, $\hat{h}$ obeys

$$\mathcal{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}\left[\ell(y, \hat{h}(x))\right], \dots\right) \leq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathcal{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}\left[\ell(y, h^*(x))\right], \dots\right) + \varepsilon$$

- Do we capture a valuable, generic notion of fair learning?
  - Axiomatic social planning problem motivation
- Do practical fair PAC-learning algorithms exist?

Fairness and Statistical Learning Theory

Definition (Fair-PAC Learning)

Suppose

1. Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$
2. Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$

$\mathcal{H}$ is *fair PAC-learnable* w.r.t. ℓ iff $\exists$ algorithm A s.t. $\forall$

1. distributions $\mathcal{D}_{1:g}$ over $(\mathcal{X} \times \mathcal{Y})^g$
2. fair malfare functions $\mathcal{M}(\cdot)$
3. additive errors $\varepsilon > 0$
4. failure probabilities $\delta \in (0, 1)$

A can identify a hypothesis $\hat{h} \in \mathcal{H}$ s.t.

1. A has $\mathrm{m}(\varepsilon, \delta, g)$ sample complexity
2. with probability at least $1 - \delta$, $\hat{h}$ obeys

$$\mathcal{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1} \left[\ell(y, \hat{h}(x))\right], \dots \right) \leq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathcal{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1} \left[\ell(y, h^*(x))\right], \dots \right) + \varepsilon$$

- Do we capture a valuable, generic notion of fair learning?
  - Axiomatic social planning problem motivation
- Do practical fair PAC-learning algorithms exist?
- Can we theoretically relate PAC and fair-PAC learning?

Fairness and Statistical Learning Theory

Definition (Fair-PAC Learning)

Suppose

1. Hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$
2. Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$

$\mathcal{H}$ is *fair PAC-learnable* w.r.t. ℓ iff $\exists$ algorithm A s.t. $\forall$

1. distributions $\mathcal{D}_{1:g}$ over $(\mathcal{X} \times \mathcal{Y})^g$
2. fair malfare functions $\mathrm{M}(\cdot)$
3. additive errors $\varepsilon > 0$
4. failure probabilities $\delta \in (0, 1)$

A can identify a hypothesis $\hat{h} \in \mathcal{H}$ s.t.

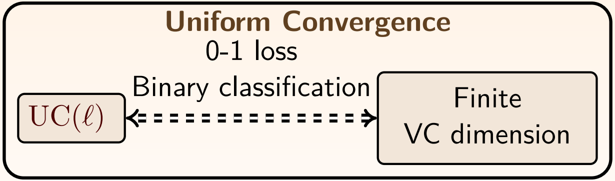
1. A has $\mathrm{m}(\varepsilon, \delta, g)$ sample complexity
2. with probability at least $1 - \delta$, outputs

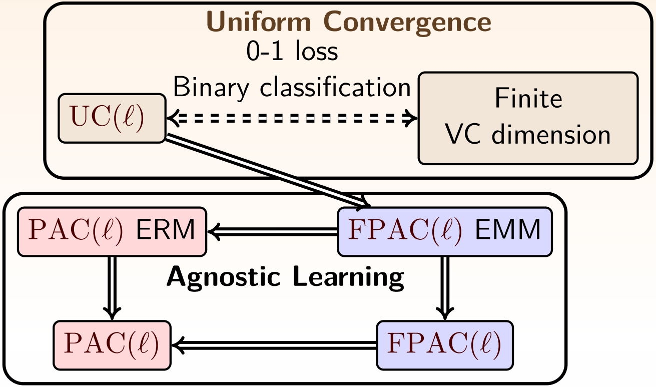
$$\mathrm{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}[\ell(y, \hat{h}(x))], \dots\right) \leq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathrm{M}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_1}[\ell(y, h^*(x))], \dots\right) + \varepsilon$$

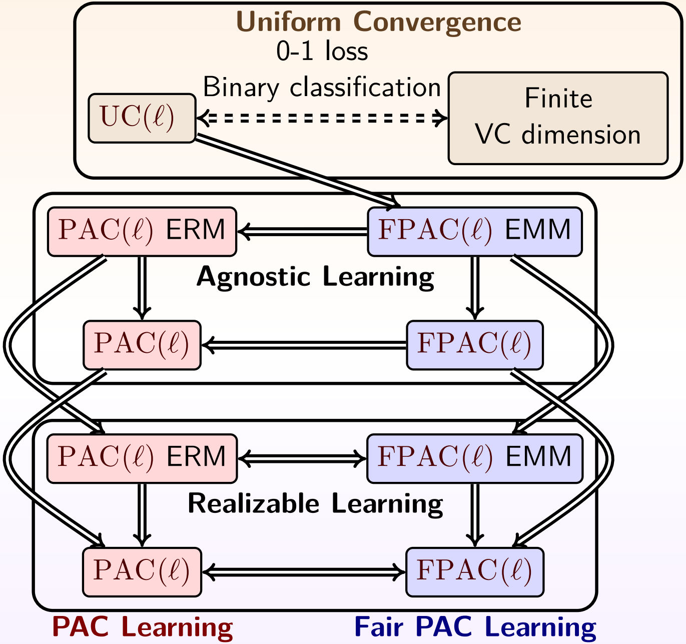
Yes.

Why else would I ask?

- Do we capture a valuable, generic notion of fair learning?
  - Axiomatic social planning problem motivation
- Do practical fair PAC-learning algorithms exist?
- Can we theoretically relate PAC and fair-PAC learning?


Uniform Convergence
0-1 loss
Binary classification
$\mathrm{UC}(\ell)$
Finite
VC dimension


Uniform Convergence
0-1 loss
Binary classification
UC(ℓ)
Finite
VC dimension
PAC(ℓ) ERM
FPAC(ℓ) EMM
Agnostic Learning
PAC(ℓ)
FPAC(ℓ)



Uniform Convergence

0-1 loss

Binary classification

$\mathrm{UC}(\ell)$

Finite VC dimension

$\mathrm{PAC}(\ell)$ ERM

$\mathrm{FPAC}(\ell)$ EMM

Agnostic Learning

$\mathrm{PAC}(\ell)$

$\mathrm{FPAC}(\ell)$

$\mathrm{PAC}(\ell)$ ERM

$\mathrm{FPAC}(\ell)$ EMM

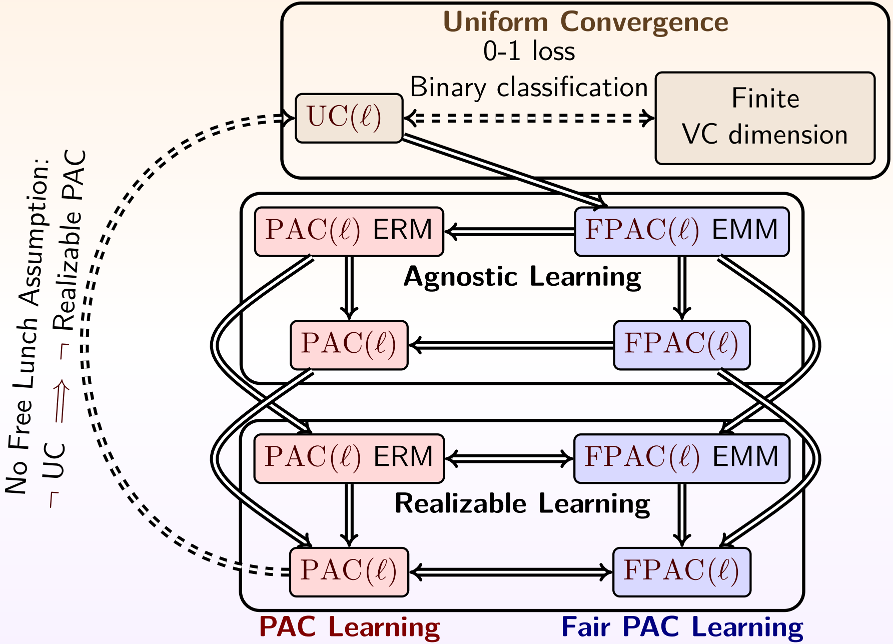
Realizable Learning

$\mathrm{PAC}(\ell)$

$\mathrm{FPAC}(\ell)$

PAC Learning

Fair PAC Learning



No Free Lunch Assumption:

$\neg$ UC $\implies$ $\neg$ Realizable PAC

Uniform Convergence

0-1 loss

Binary classification

$\mathrm{UC}(\ell)$

Finite VC dimension

$\mathrm{PAC}(\ell)$ ERM

$\mathrm{FPAC}(\ell)$ EMM

Agnostic Learning

$\mathrm{PAC}(\ell)$

$\mathrm{FPAC}(\ell)$

$\mathrm{PAC}(\ell)$ ERM

$\mathrm{FPAC}(\ell)$ EMM

Realizable Learning

$\mathrm{PAC}(\ell)$

$\mathrm{FPAC}(\ell)$

PAC Learning

Fair PAC Learning

Computational Learning Theory

- Can we construct *computationally-efficient* FPAC learners from PAC learners?
    - *Efficient* means $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, g)$ sample complexity
    - Realizable case: reduction preserves polynomial-time complexity
    - Agnostic case: Cyrus has no answer (yet)

Computational Learning Theory

- Can we construct *computationally-efficient* FPAC learners from PAC learners?
  - *Efficient* means $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, g)$ sample complexity
  - Realizable case: reduction preserves polynomial-time complexity
  - Agnostic case: Cyrus has no answer (yet)
- Are conditions for efficient PAC sufficient for efficient FPAC?
  - This area looks promising!

Computational Learning Theory

- Can we construct *computationally-efficient* FPAC learners from PAC learners?
  - *Efficient* means $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, g)$ sample complexity
  - Realizable case: reduction preserves polynomial-time complexity
  - Agnostic case: Cyrus has no answer (yet)
- Are conditions for efficient PAC sufficient for efficient FPAC?
  - This area looks promising!
  - Efficiently Coverable Classes:
    - If we can *efficiently approximately enumerate* $\mathcal{H}$
    - And our loss-function is well-behaved
    - Then we can PAC or FPAC-learn in $\mathcal{H}$
    - Think “all separating hyperplanes of bounded dimension”

Computational Learning Theory

- Can we construct *computationally-efficient* FPAC learners from PAC learners?
  - *Efficient* means $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, g)$ sample complexity
  - Realizable case: reduction preserves polynomial-time complexity
  - Agnostic case: Cyrus has no answer (yet)
- Are conditions for efficient PAC sufficient for efficient FPAC?
  - This area looks promising!
  - Efficiently Coverable Classes:
    - If we can *efficiently approximately enumerate* $\mathcal{H}$
    - And our loss-function is well-behaved
    - Then we can PAC or FPAC-learn in $\mathcal{H}$
    - Think “all separating hyperplanes of bounded dimension”
  - Convex optimization:
    - Suppose *bounded* parameter space Θ
    - Assume $\ell \circ h_\theta$ is *convex* + *Lipschitz continuous* in θ
    - Then ε-empirical risk minimization requires *polynomial time*
    - Same for empirical malfare minimization (this work)

Convex Optimization

Strategy: Assume class $\ell \circ \mathcal{H}$ is:

1. Uniformly Convergent
   - Vapnik-Chervonenkis theory: *Uniform* bounds over distribution $\mathcal{D}$
   - *Estimation error*: $\epsilon(m, \delta)$ s.t.

$$\mathbb{P}\left(\sup_{f \in \ell \circ \mathcal{H}} \left| \mathbb{E}[f] - \widehat{\mathbb{E}}[f] \right| \geq \epsilon(m, \delta) \right) \leq \delta$$

   - *Sample complexity*

$$\mathrm{m}(\varepsilon, \delta) \doteq \operatorname{argmin} \left\{ m : \epsilon(m, \delta) \leq \varepsilon \right\} \in \mathrm{Poly}\left(\tfrac{1}{\varepsilon}, \tfrac{1}{\delta}\right)$$

2. Bounded
   - Bounded parameter space $\Theta \in \mathbb{R}^d$
3. Lipschitz Continuous
   - λ_ℓ-Lipschitz loss ℓ, $\lambda_{\mathcal{H}}$-Lipschitz $\mathcal{H}$
4. Convex
   - $\ell(\circ h(x; \theta), y)$ is convex in θ over Θ

The Algorithm

1. Draw $\mathrm{m}(\frac{\varepsilon}{3}, \frac{\delta}{g})$ samples (per group)
2. Define *empirical malfare* objective

$$f(\theta) \doteq \mathrm{M}_p\left(i \mapsto \hat{\mathrm{R}}(h(\cdot; \theta); \ell, \boldsymbol{x}_i, \boldsymbol{y}_i)\right)$$

3. Iterations: $n \doteq \left(\frac{3\,\mathrm{diam}(\Theta)\lambda_\ell\lambda_{\mathcal{H}}}{\varepsilon}\right)^2$
4. Learning rate $\alpha \doteq \frac{\mathrm{diam}(\Theta)}{\lambda_\ell\lambda_{\mathcal{H}}\sqrt{n}} \approx \frac{\varepsilon}{3\lambda_\ell^2\lambda_{\mathcal{H}}^2}$
5. Shor's *projected subgradient* algorithm

$$\hat{\theta} \leftarrow \mathrm{PSG}(f, \Theta, n, \alpha)$$

6. Return $h(\cdot; \hat{\theta})$

W.h.p., *estimation* + *optimization* error don't exceed ε

Polynomial time + sample complexity

Convex Optimization

Strategy: Assume class $\ell \circ \mathcal{H}$ is:

1. Uniformly Convergent
   - Vapnik-Chervonenkis theory: *Uniform* bounds over distribution $\mathcal{D}$
   - *Estimation error*: $\epsilon(m, \delta)$ s.t.

$$\mathbb{P}\left(\sup_{f \in \ell \circ \mathcal{H}} \left| \mathbb{E}[f] - \widehat{\mathbb{E}}[f] \right| \geq \epsilon(m, \delta) \right) <$$

   - *Sample complexity*

$$\mathrm{m}(\varepsilon, \delta) \doteq \operatorname{argmin} \left\{ m : \epsilon(m, \delta) \leq \varepsilon \right\} \in \operatorname{Poly}\left(\tfrac{1}{\varepsilon}, \tfrac{1}{\delta} \right)$$

2. Bounded
   - Bounded parameter space $\Theta \in \mathbb{R}^d$
3. Lipschitz Continuous
   - λ_ℓ-Lipschitz loss ℓ, $\lambda_{\mathcal{H}}$-Lipschitz $\mathcal{H}$
4. Convex
   - $\ell(\circ h(x; \theta), y)$ is convex in θ over Θ

The Algorithm

1. Draw $\mathrm{m}(\frac{\varepsilon}{3}, \frac{\delta}{g})$ samples (per group)

efin *pirical malfare* objective

$\hat{\mathrm{R}}(h(\cdot; \theta); \ell, \boldsymbol{x}_i, \boldsymbol{y}_i))$

$\left(\frac{3\, \mathrm{diam}(\Theta) \lambda_\ell \lambda_{\mathcal{H}}}{\varepsilon} \right)^2$

$\frac{\mathrm{iam}(\Theta)}{\lambda_\ell \lambda_{\mathcal{H}} \sqrt{n}} \approx \frac{\varepsilon}{3 \lambda_\ell^2 \lambda_{\mathcal{H}}^2}$

subgradient algorithm

$\mathrm{G}(f, \Theta, n, \alpha)$

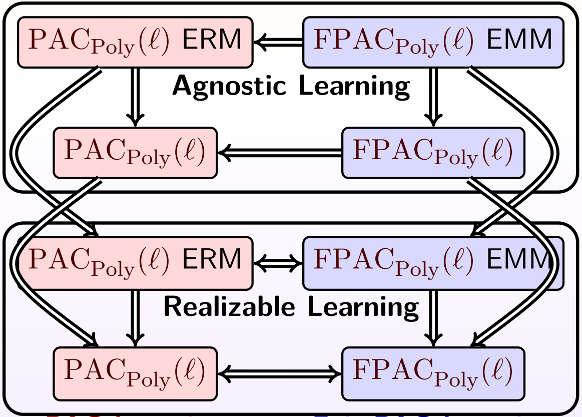
eturn $; \hat{\theta})$

Read the Paper Today!

Now 50% Off!

W.h.p., *estimation* + *optimization* error don't exceed ε

Polynomial time + sample complexity



$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$ ERM

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$ EMM

Agnostic Learning

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$ ERM

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$ EMM

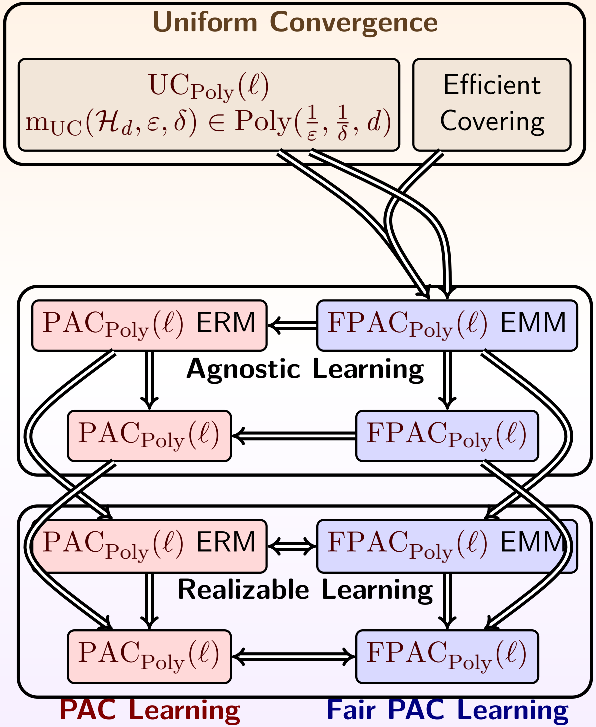
Realizable Learning

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$

PAC Learning

Fair PAC Learning



Uniform Convergence

$\mathrm{UC}_{\mathrm{Poly}}(\ell)$

$\mathrm{m}_{\mathrm{UC}}(\mathcal{H}_d, \varepsilon, \delta) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$

Efficient Covering

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$ ERM

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$ EMM

Agnostic Learning

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$ ERM

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$ EMM

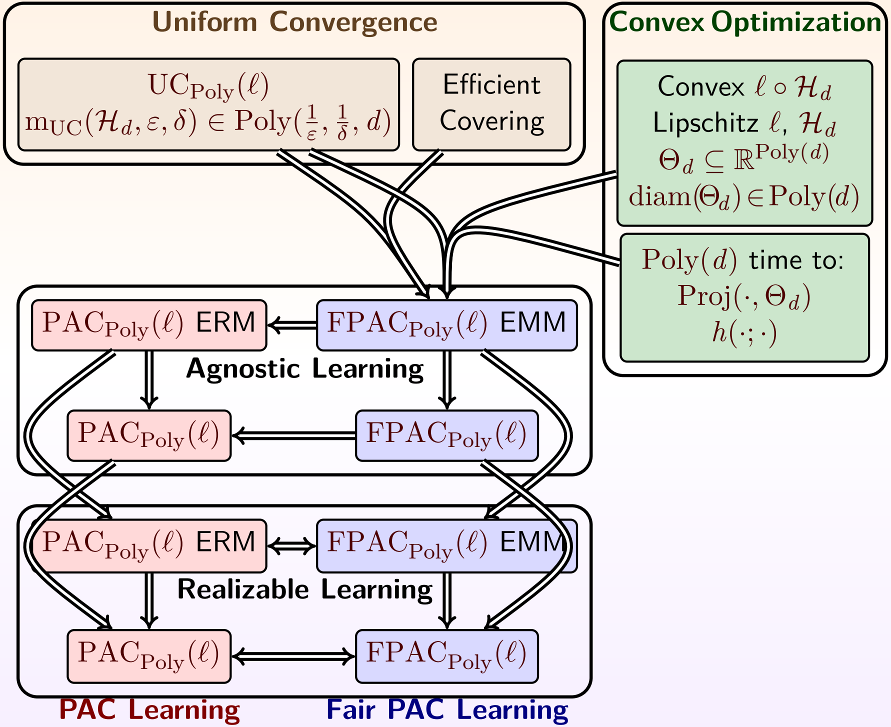
Realizable Learning

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$

PAC Learning

Fair PAC Learning



Uniform Convergence

$\mathrm{UC}_{\mathrm{Poly}}(\ell)$

$\mathrm{m}_{\mathrm{UC}}(\mathcal{H}_d, \varepsilon, \delta) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$

Efficient Covering

Convex Optimization

Convex $\ell \circ \mathcal{H}_d$

Lipschitz ℓ, $\mathcal{H}_d$

$\Theta_d \subseteq \mathbb{R}^{\mathrm{Poly}(d)}$

$\mathrm{diam}(\Theta_d) \in \mathrm{Poly}(d)$

$\mathrm{Poly}(d)$ time to:

$\mathrm{Proj}(\cdot, \Theta_d)$

$h(\cdot; \cdot)$

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$ ERM

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$ EMM

Agnostic Learning

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$ ERM

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$ EMM

Realizable Learning

$\mathrm{PAC}_{\mathrm{Poly}}(\ell)$

$\mathrm{FPAC}_{\mathrm{Poly}}(\ell)$

PAC Learning

Fair PAC Learning

Recap: Characterizing Fair PAC-Learnability

- Classical method: measure population sentiment with welfare
- This work: welfare and malfare on equal axiomatic footing
  - *Malfare minimization* is *fair extension* of *risk minimization*

Recap: Characterizing Fair PAC-Learnability

- Classical method: measure population sentiment with welfare
- This work: welfare and malfare on equal axiomatic footing
  - *Malfare minimization* is *fair extension* of *risk minimization*
- Under some conditions, $\mathrm{PAC} = \mathrm{FPAC}$ (statistical equivalence)
  - $\mathrm{FPAC} \implies \mathrm{PAC}$ (as a special case)
  - Constructive $\mathrm{PAC} \implies \mathrm{FPAC}$ reduction in realizable case
  - General case is non-constructive, assumes no-free-lunch argument

Recap: Characterizing Fair PAC-Learnability

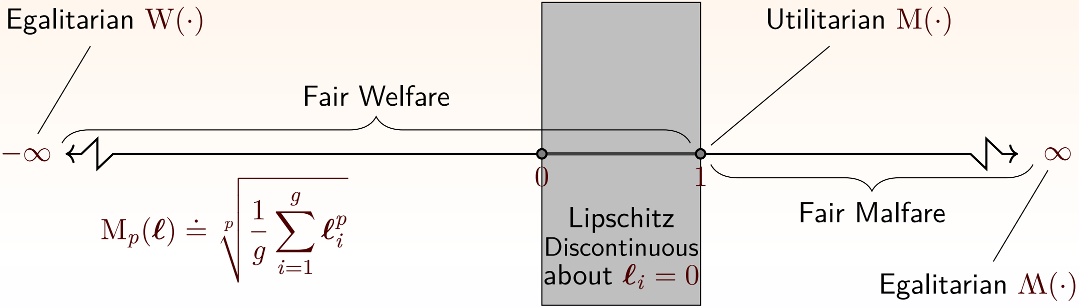
- Classical method: measure population sentiment with welfare
- This work: welfare and malfare on equal axiomatic footing
  - *Malfare minimization* is *fair extension* of *risk minimization*
- Under some conditions, $\mathrm{PAC} = \mathrm{FPAC}$ (statistical equivalence)
  - $\mathrm{FPAC} \implies \mathrm{PAC}$ (as a special case)
  - Constructive $\mathrm{PAC} \implies \mathrm{FPAC}$ reduction in realizable case
  - General case is non-constructive, assumes no-free-lunch argument
- Open research question: does efficient PAC $\implies$ efficient FPAC?
  - Constructive reduction in realizable case
  - Efficient cover enumerability sufficient for both
  - Standard convex optimization assumptions sufficient for both

Recap: Characterizing Fair PAC-Learnability

- Classical method: measure population sentiment with welfare
- This work: welfare and malfare on equal axiomatic footing
    - *Malfare minimization* is *fair extension* of *risk minimization*
- Under some conditions, $\mathrm{PAC} = \mathrm{FPAC}$ (statistical equivalence)
    - $\mathrm{FPAC} \implies \mathrm{PAC}$ (as a special case)
    - Constructive $\mathrm{PAC} \implies \mathrm{FPAC}$ reduction in realizable case
    - General case is non-constructive, assumes no-free-lunch argument
- Open research question: does efficient PAC $\implies$ efficient FPAC?
    - Constructive reduction in realizable case
    - Efficient cover enumerability sufficient for both
    - Standard convex optimization assumptions sufficient for both

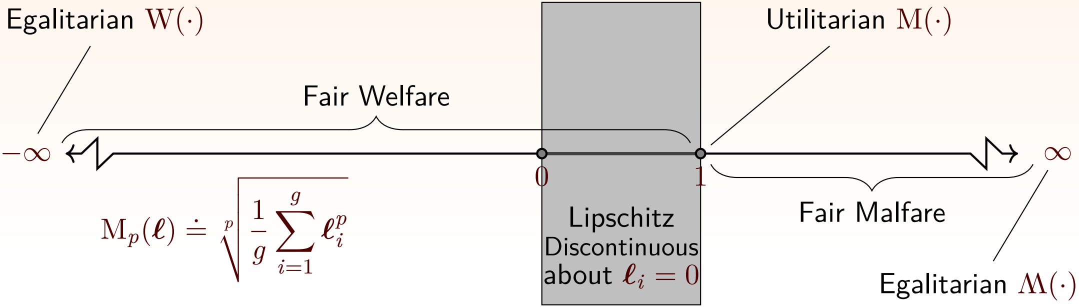
Conjecture: No, $\exists$ PAC-learnable class, where FPAC-learning is NP-hard (and $\mathrm{P} \neq \mathrm{NP}$)

Recap: Malfare, Welfare, and FPAC Learning


Egalitarian $\mathrm{W}(\cdot)$
Fair Welfare
Utilitarian $\mathrm{M}(\cdot)$
$-\infty$
∞
0
1
Lipschitz Discontinuous about $\boldsymbol{\ell}_i = 0$
Fair Malfare
Egalitarian $\mathcal{M}(\cdot)$

$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\ell}_i^p}$$

Recap: Malfare, Welfare, and FPAC Learning

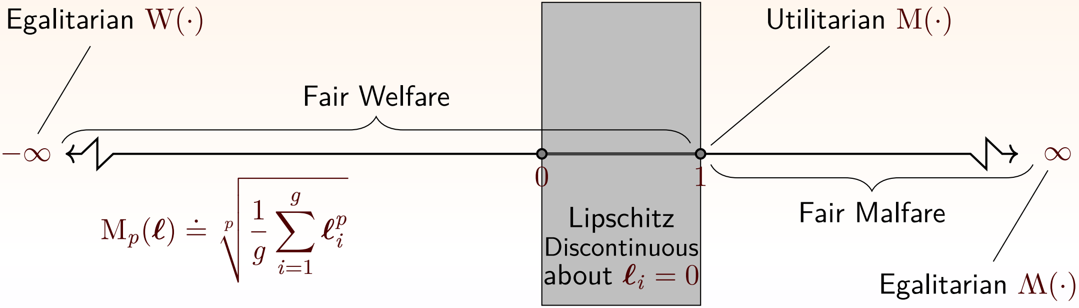

Egalitarian $\mathrm{W}(\cdot)$
Utilitarian $\mathrm{M}(\cdot)$
Fair Welfare
$-\infty$
0
1
∞
Lipschitz Discontinuous about $\boldsymbol{\ell}_i = 0$
Fair Malfare
Egalitarian $\mathcal{M}(\cdot)$

$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\ell}_i^p}$$

Why use malfare instead of welfare?

1. "Most" machine learning tasks more naturally cast as *loss minimization*
   - Exceptions: reward, profit, accuracy maximization

Recap: Malfare, Welfare, and FPAC Learning

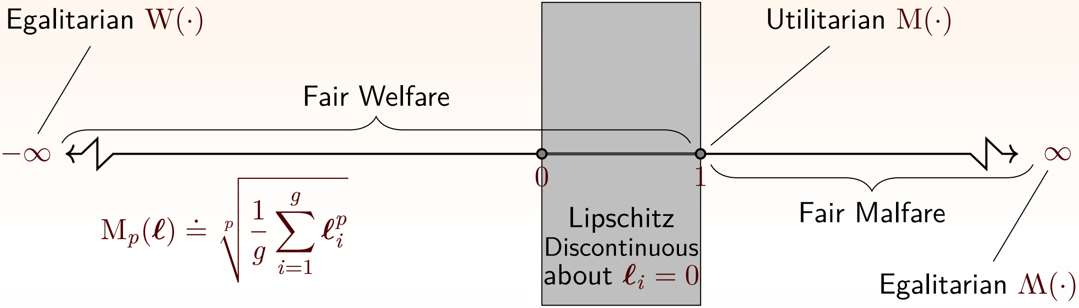

Egalitarian $\mathrm{W}(\cdot)$
Utilitarian $\mathrm{M}(\cdot)$
Fair Welfare
$-\infty$
∞
0
1
Fair Malfare
Lipschitz Discontinuous about $\boldsymbol{\ell}_i = 0$
Egalitarian $\mathcal{M}(\cdot)$

$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\ell}_i^p}$$

Why use malfare instead of welfare?

1. “Most” machine learning tasks more naturally cast as *loss minimization*
   - Exceptions: reward, profit, accuracy maximization
2. Fair PAC-Learning *with welfare targets* is tricky
   - Inherent statistical instability for $p \in [0, 1)$
   - Require additional assumptions, or restricted capabilities

Recap: Malfare, Welfare, and FPAC Learning



Egalitarian $\mathrm{W}(\cdot)$

Utilitarian $\mathrm{M}(\cdot)$

Fair Welfare

$-\infty$ 0 1 ∞

Fair Malfare

Egalitarian $\mathcal{M}(\cdot)$

Lipschitz Discontinuous about $\boldsymbol{\ell}_i = 0$

$$\mathrm{M}_p(\boldsymbol{\ell}) \doteq \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \boldsymbol{\ell}_i^p}$$

Why use malfare instead of welfare?

1. "Most" machine learning tasks more naturally cast as *loss minimization*
   - Exceptions: reward, profit, accuracy maximization
2. Fair PAC-Learning *with welfare targets* is tricky
   - Inherent statistical instability for $p \in [0, 1)$
   - Require additional assumptions, or restricted capabilities
3. **Why not?**