Minimax Pareto Fairness:
A Multi-Objective Perspective

Natalia Martinez, Martin Bertran, Guillermo Sapiro
Department of Electrical and Computer Engineering
Duke University

Outline

Minimax Pareto Fairness (MMPF)

- Motivation
- General overview
- Problem formulation
- Pareto solutions
- Optimization
- Experiments
- Conclusions and future work

Motivation

- Machine Learning models may be discriminatory
  [Barocas et al 2016, Buolamwini et al 2018]
- Many fairness notions based on parity
  [Feldman et al 2015, Hardt et al 2016, Zafar et al 2017]
- Perfect Fairness and optimality may not be possible
  [Kaplow et al 1999, Chen et al 2018]
- Less work done on scenarios were optimality is desired
  [Ustun et al 2019]

Motivation

- Machine Learning models may be discriminatory
  [Barocas et al 2016, Buolamwini et al 2018]
- Many fairness notions based on parity
  [Feldman et al 2015, Hardt et al 2016, Zafar et al 2017]
- Perfect Fairness and optimality may not be possible
  [Kaplow et al 1999, Chen et al 2018]
- Less work done on scenarios were optimality is desired
  [Ustun et al 2019]

Our Focus

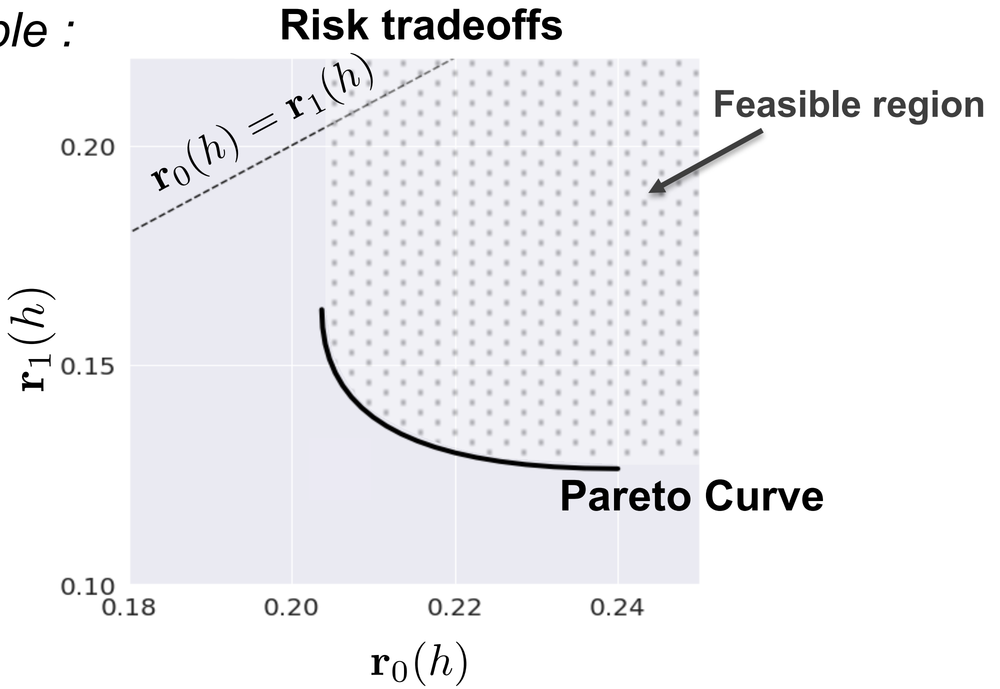
- Characterizing the optimal solutions (Pareto front)
- Fairest model without unnecessary harm (preserve optimality)

General Overview

Minimax Pareto Fairness (MMPF)

- Fairest model without unnecessary harm (preserve optimality)
- Fairness as a multi-objective optimization problem (MOOP)

2 Population example :

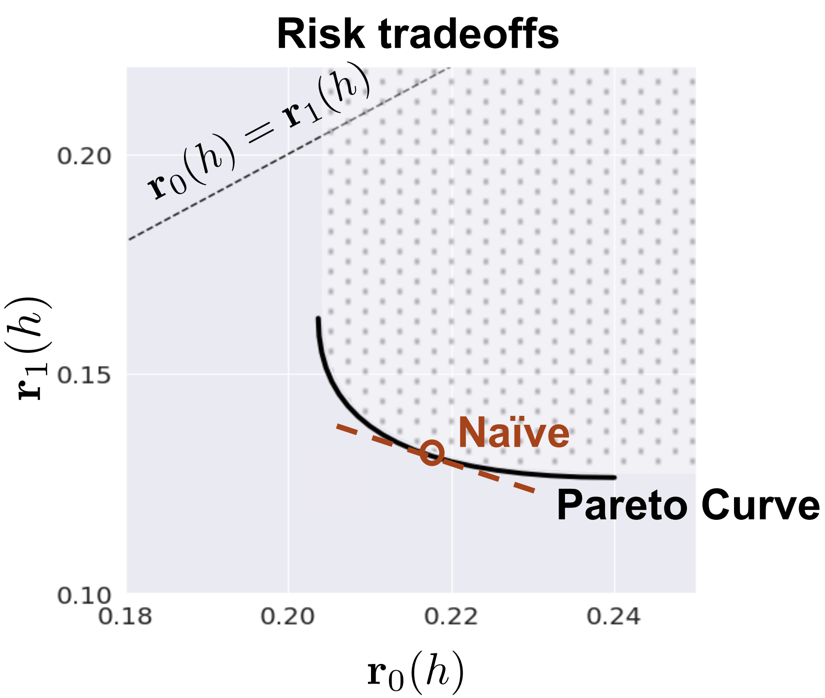


Population risk: $r_a(h) = E_{X,Y|A=a}[\ell(Y, h(X))]$

General Overview

Minimax Pareto Fairness (MMPF)

- Fairest model without unnecessary harm (preserve optimality)
- Fairness as a multi-objective optimization problem (MOOP)

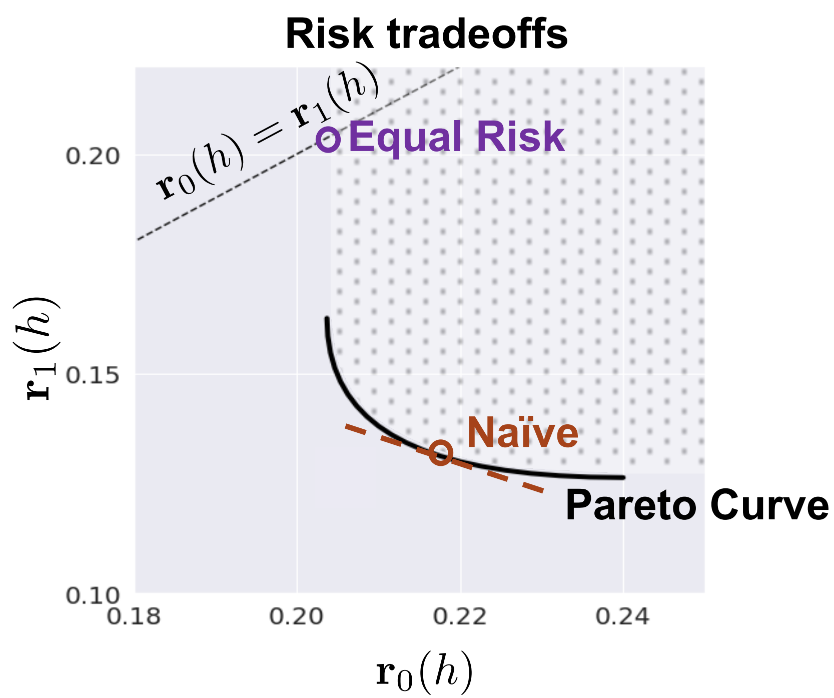


Population risk: $r_a(h) = E_{X,Y|A=a}[\ell(Y, h(X))]$

General Overview

Minimax Pareto Fairness (MMPF)

- Fairest model without unnecessary harm (preserve optimality)
- Fairness as a multi-objective optimization problem (MOOP)

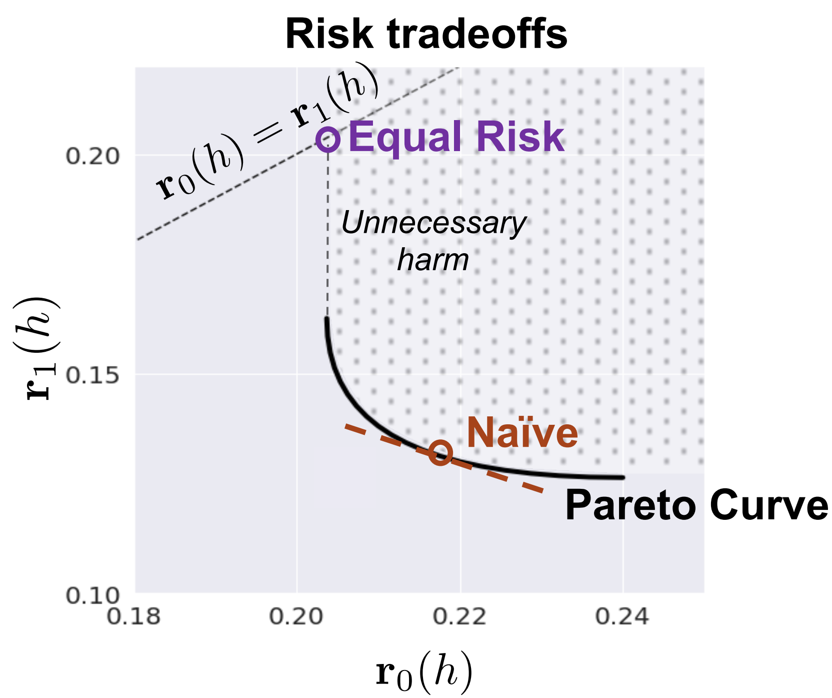


Population risk: $r_a(h) = E_{X,Y|A=a}[\ell(Y, h(X))]$

General Overview

Minimax Pareto Fairness (MMPF)

- Fairest model without unnecessary harm (preserve optimality)
- Fairness as a multi-objective optimization problem (MOOP)

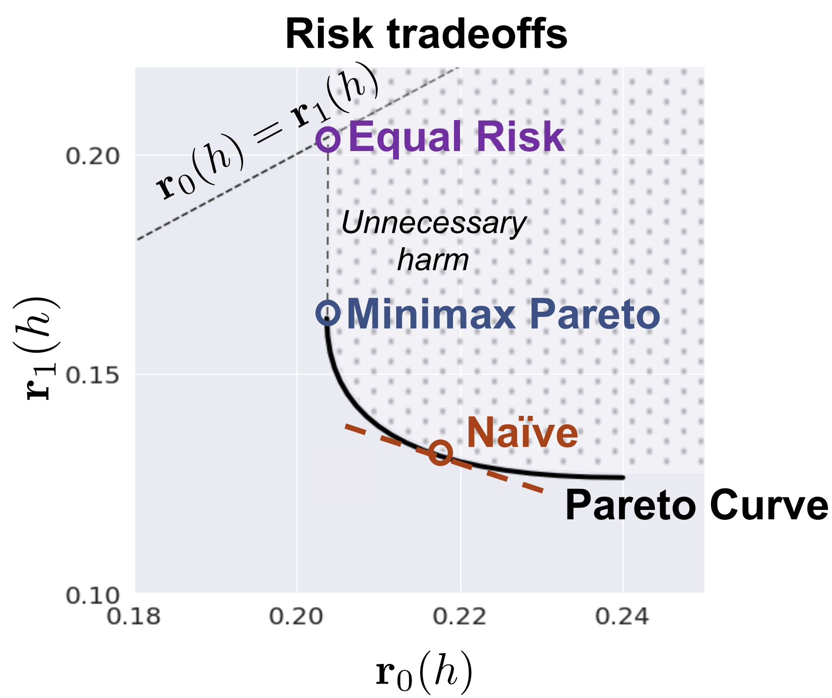


Population risk: $r_a(h) = E_{X,Y|A=a}[\ell(Y, h(X))]$

General Overview

Minimax Pareto Fairness (MMPF)

- Fairest model without unnecessary harm (preserve optimality)
- Fairness as a multi-objective optimization problem (MOOP)



Population risk: $r_a(h) = E_{X,Y|A=a}[\ell(Y, h(X))]$

General Overview

Minimax Pareto Fairness (MMPF)

- Fairest model without unnecessary harm (preserve optimality)
- Fairness as a multi-objective optimization problem (MOOP)

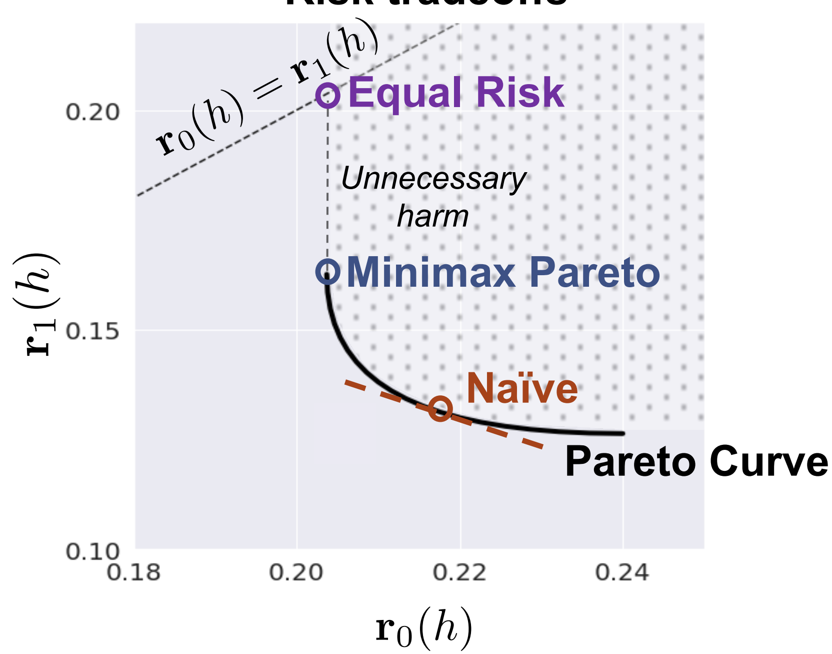
MOOP: $\min_{h\in\mathcal{H}}(r_1(h), ..., r_{|\mathcal{A}|}(h))$

- **MMPF Objective**

$$h^* \in \arg\min_{h\in\mathcal{P}_{\mathcal{A},\mathcal{H}}} ||\mathbf{r}(h)||_\infty$$

$$\mathbf{r}^* = \mathbf{r}(h^*)$$

Risk tradeoffs



Population risk: $r_a(h) = E_{X,Y|A=a}[\ell(Y, h(X))]$

Outline

Minimax Pareto Fairness (MMPF)

- Motivation
- General overview
- Problem formulation
- Pareto solutions
- Optimization
- Experiments
- Conclusions and future work

MMPF: Problem Formulation

Learning Setting

- (X, Y, A) Input, target, and population variables
- $\mathcal{H} = \{h : \mathcal{X} \rightarrow [0,1]^{|\mathcal{Y}|}\}$ Hypothesis class (e.g., DNN Classifier)
- $\ell : [0,1]^{|\mathcal{Y}|} \times [0,1]^{|\mathcal{Y}|} \rightarrow R^+$ Loss function
- $r_a(h) = E_{X,Y|A=a}[\ell(Y, h(X))]$ Population risk

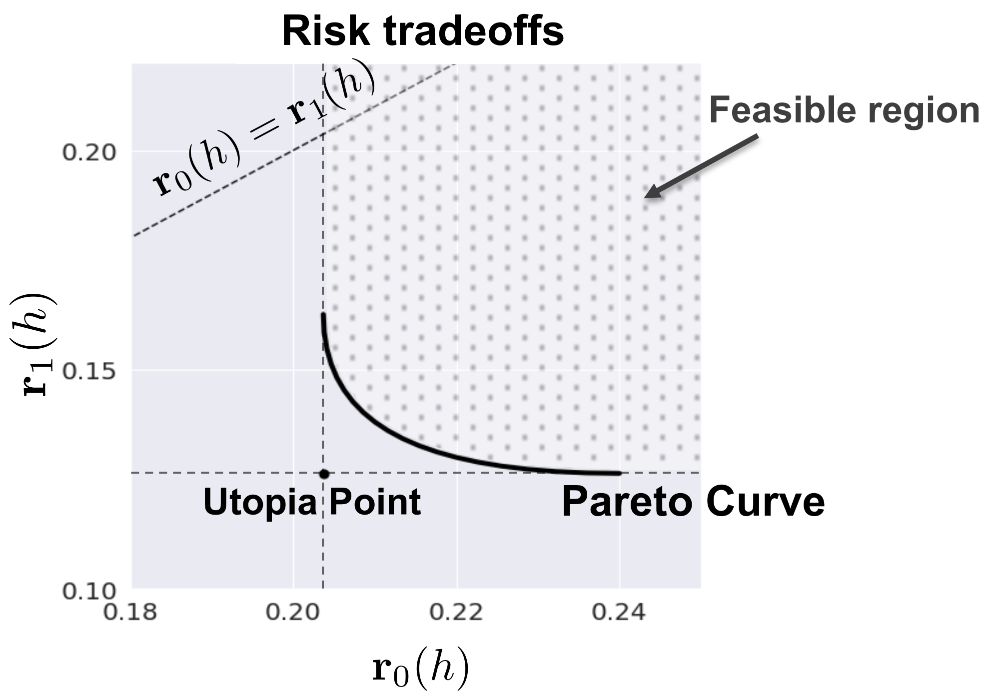
Multi-Objective Optimization Problem

$$\min_{h \in \mathcal{H}} (r_1(h), ..., r_{|\mathcal{A}|}(h))$$

MMPF: Problem Formulation

Optimal Tradeoffs

- Pareto hypotheses $\mathcal{P}_{\mathcal{A},\mathcal{H}} = \{h \in \mathcal{H} : \nexists h' \in \mathcal{H} | \mathbf{r}(h') \prec \mathbf{r}(h)\}$
- Pareto risks $\mathcal{P}^{\mathbf{r}}_{\mathcal{A},\mathcal{H}} = \{\mathbf{r} \in R^{+|A|} : \exists h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}, \mathbf{r} = \mathbf{r}(h)\}$

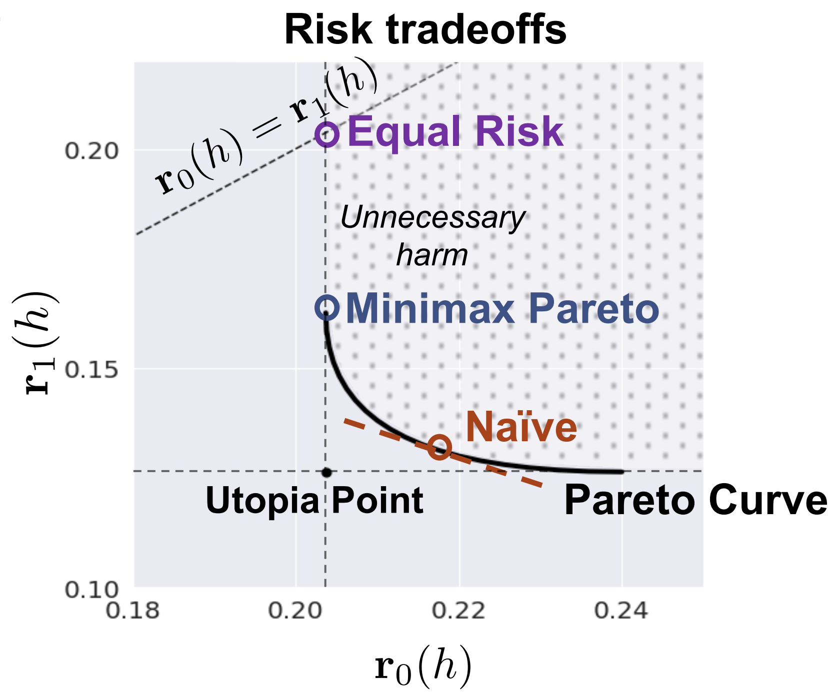


MMPF: Problem Formulation

Minimax Pareto Fair model (MMPF)

$$h^* \in \arg\min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} ||\mathbf{r}(h)||_\infty$$

$$\mathbf{r}^* = \mathbf{r}(h^*)$$

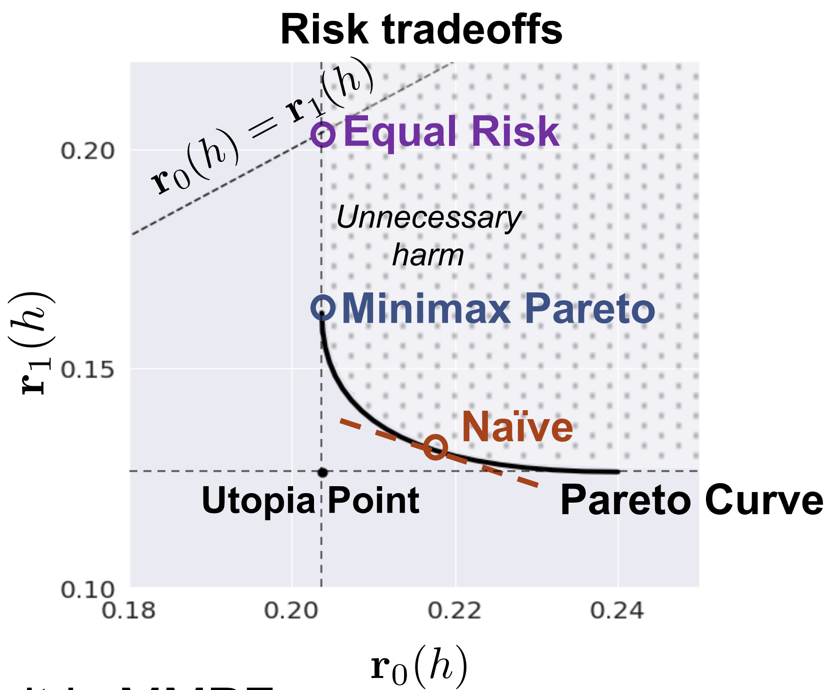


MMPF: Problem Formulation

Minimax Pareto Fair model (MMPF)

$$h^* \in \arg\min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} ||\mathbf{r}(h)||_\infty$$

$$\mathbf{r}^* = \mathbf{r}(h^*)$$



- *Lemma 3.1.* If equal risk is Pareto then it is MMPF.
- *Lemma 3.2.* Best equal risk is obtained by adding noise to MMPF.

MMPF: Pareto Solutions

Analysis of Pareto solutions

Theorem 4.1.

- $\mathcal{H}$ Convex hypothesis class
- $\{\mathbf{r}_a(h)\}_{a\in\mathcal{A}}$ Convex risk functions

→

Convex $\mathcal{P}^{\mathbf{r}}_{\mathcal{A},\mathcal{H}}$ fully characterized by

$$\mathbf{r}(h^{\mu}) : \; h^{\mu} = \underset{h\,\in\,\mathcal{H}}{argmin} \sum_{a=1}^{|\mathcal{A}|} \mu_a \mathbf{r}_a(h)$$

$$||\mu||_1^1 = 1\,,\; \mu_a > 0$$

MMPF: Pareto Solutions

Analysis of Pareto solutions

Theorem 4.1.

- $\mathcal{H}$ Convex hypothesis class
- $\{\mathbf{r}_a(h)\}_{a\in\mathcal{A}}$ Convex risk functions

→

Convex $\mathcal{P}^{\mathbf{r}}_{\mathcal{A},\mathcal{H}}$ fully characterized by

$$\mathbf{r}(h^\mu):\ h^\mu = \underset{h\in\mathcal{H}}{argmin}\sum_{a=1}^{|\mathcal{A}|}\mu_a\mathbf{r}_a(h)$$

$$||\mu||_1^1 = 1\,,\ \mu_a > 0$$

Classification with Cross Entropy *(similar for Brier Score)*:

$Y \in \mathcal{Y}, |\mathcal{Y}| < \infty$

$$h^\mu(x) = \frac{\sum_{a=1}^{|\mathcal{A}|}\mu_a p(x|a)p(y|x,a)}{\sum_{a=1}^{|\mathcal{A}|}\mu_a p(x|a)}$$

$$r_a^{CE}(\mu) = H(Y|X,a) + E_{X|a}\Big[D_{KL}\Big(p(y|X,a)\Big|\Big|h^\mu(X)\Big)\Big]$$

$r^{CE} = E_{X,Y}[-\langle\delta^Y, \ln(h(X))\rangle]$

$p(y|X,a) = \{p(Y=y|X,A=a)\}_{y\in\mathcal{Y}}$

MMPF: Pareto Solutions

Analysis of Pareto solutions

Theorem 4.1.

- $\mathcal{H}$ Convex hypothesis class
- $\{\mathbf{r}_a(h)\}_{a\in\mathcal{A}}$ Convex risk functions

→ Convex $\mathcal{P}^{\mathbf{r}}_{\mathcal{A},\mathcal{H}}$ fully characterized by

$$\mathbf{r}(h^\mu) : \; h^\mu = \underset{h\in\mathcal{H}}{argmin} \sum_{a=1}^{|\mathcal{A}|} \mu_a \mathbf{r}_a(h)$$

$$||\mu||_1^1 = 1\,,\; \mu_a > 0$$

Classification with Cross Entropy *(similar for Brier Score)*:

$Y \in \mathcal{Y}, |\mathcal{Y}| < \infty$

$$h^\mu(x) = \frac{\sum_{a=1}^{|\mathcal{A}|} \mu_a p(x|a) p(y|x,a)}{\sum_{a=1}^{|\mathcal{A}|} \mu_a p(x|a)}$$

$$r_a^{CE}(\mu) = H(Y|X,a) + E_{X|a}\Big[D_{KL}\Big(p(y|X,a)\Big|\Big|h^\mu(X)\Big)\Big]$$

Lemma 4.3. No tradeoffs exist if $Y \perp A|X$ or $H(A|X) \to 0$

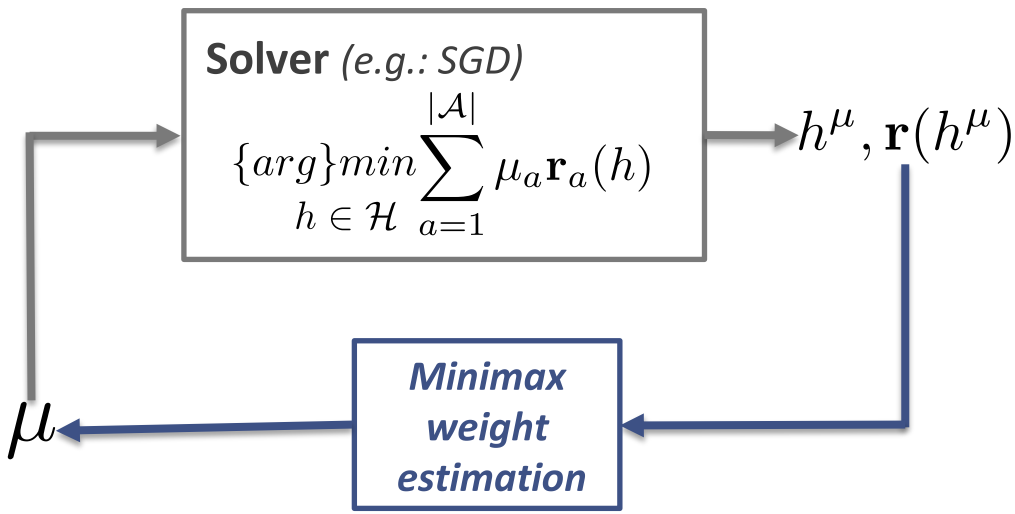
$r^{CE} = E_{X,Y}[-\langle \delta^Y, \ln(h(X))\rangle]$

$p(y|X,a) = \{p(Y=y|X,A=a)\}_{y\in\mathcal{Y}}$

MMPF: Optimization

Objective

- Find minimax weight $\mu^* : h^{\mu^*} \in \underset{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}}{argmin} ||\mathbf{r}(h)||_\infty$

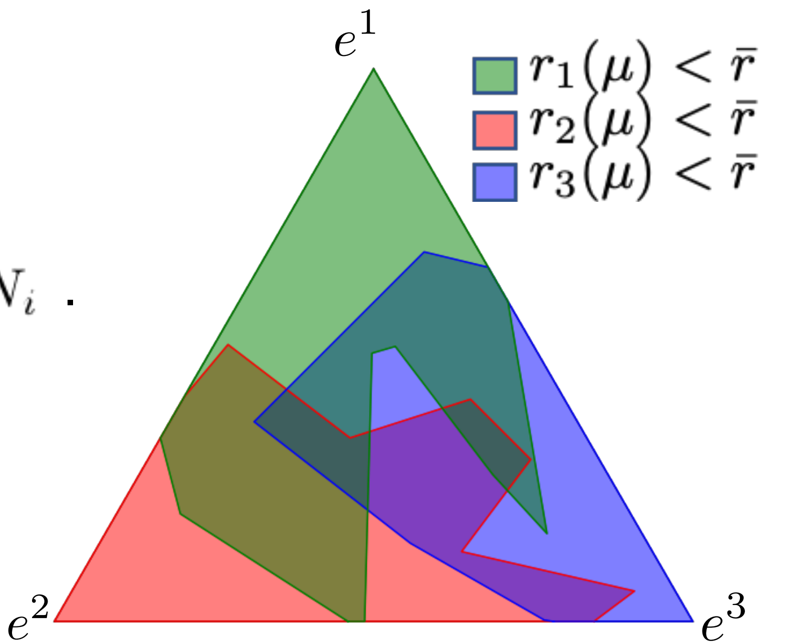


MMPF: Optimization

Minimax weight estimation: APStar

Theorem 5.1.

- Weight loss landscape $N_i = \{\mu : r_i(\mu) < \bar{r}\}, i \in \mathcal{A}$ is star-shaped.
- Minimax weight $\boldsymbol{\mu}^* \in \bigcap_{i \in \mathcal{A}} N_i$.
- $\forall \mathcal{I} \subseteq \mathcal{A}, \boldsymbol{\mu} : \mu_{\mathcal{A} \setminus \mathcal{I}} = 0 \to \boldsymbol{\mu} \in \bigcup_{i \in \mathcal{I}} N_i$.



2D simplex

$\bar{r} > \min_{\mu \in \Delta^{|\mathcal{A}|-1}} ||r(\mu)||_\infty$

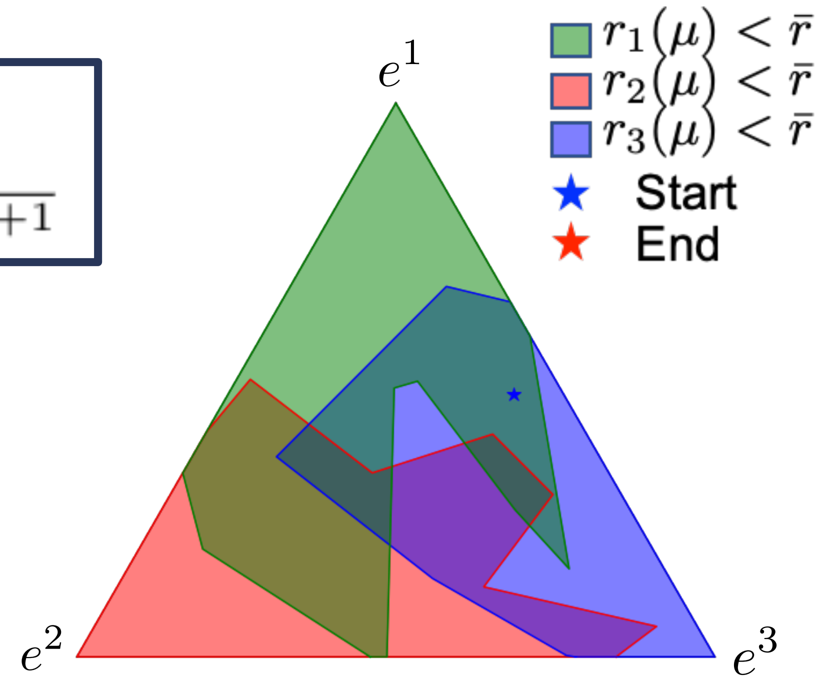
MMPF: Optimization

Minimax weight estimation: APStar

- We propose the following weight update

Current minimax

$$\mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|}$$
$$\boldsymbol{\mu} \leftarrow \left(\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu\right)\frac{K}{(K-1)\alpha+1}$$



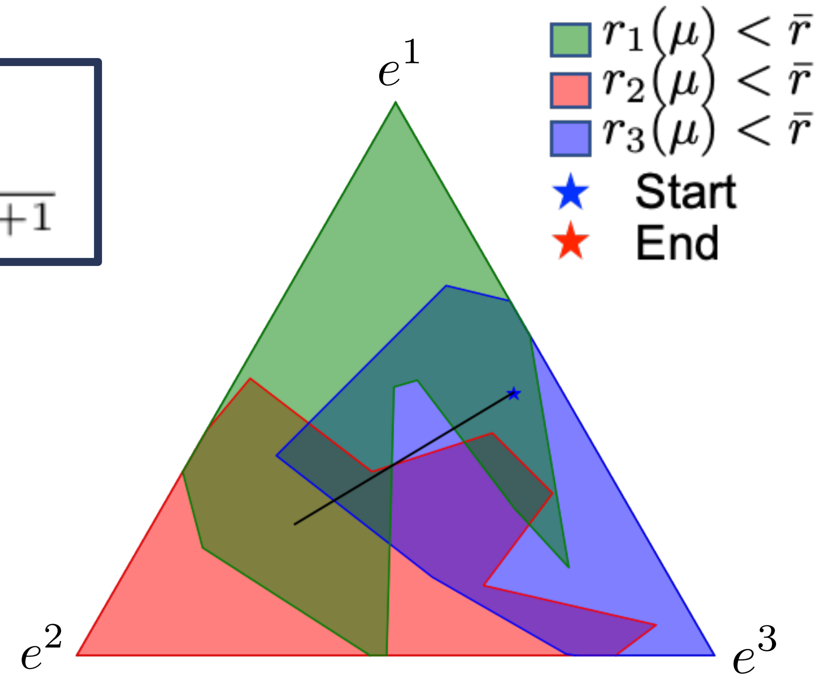
MMPF: Optimization

Minimax weight estimation: APStar

- We propose the following weight update

Current minimax

$$\mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|}$$
$$\boldsymbol{\mu} \leftarrow \left(\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu\right)\frac{K}{(K-1)\alpha+1}$$



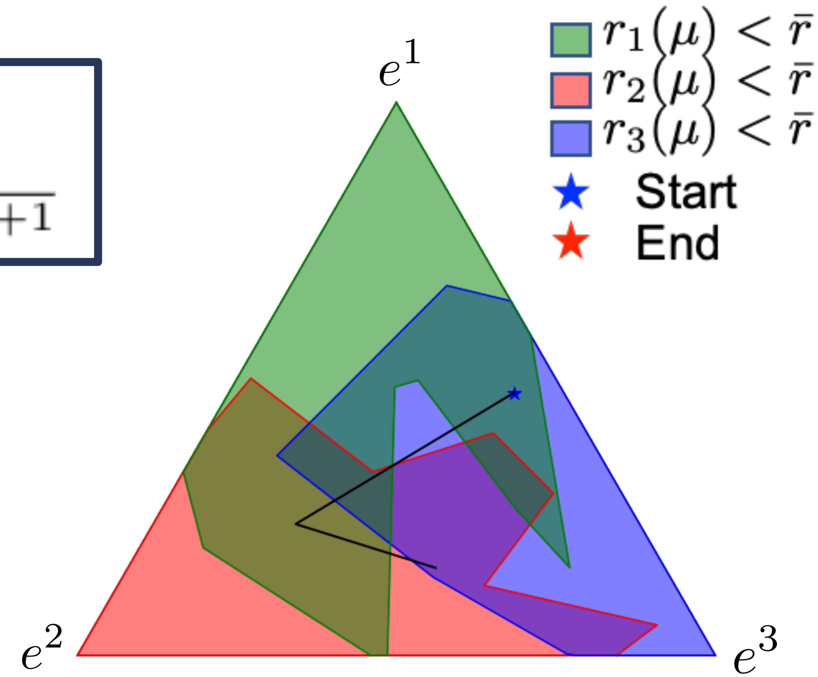
MMPF: Optimization

Minimax weight estimation: APStar

- We propose the following weight update

Current minimax

$$\mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|}$$
$$\boldsymbol{\mu} \leftarrow \left(\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu\right)\frac{K}{(K-1)\alpha+1}$$



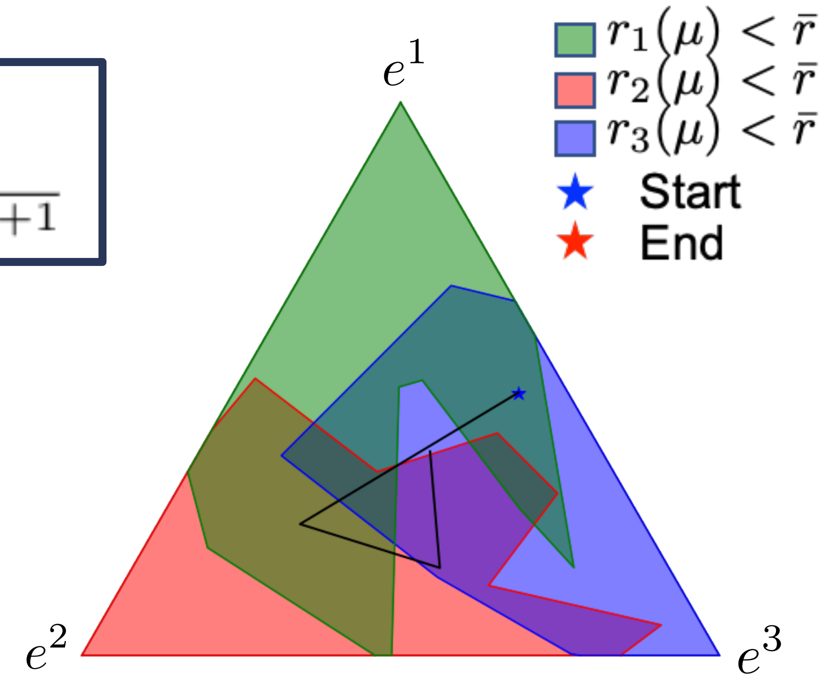
MMPF: Optimization

Minimax weight estimation: APStar

- We propose the following weight update

Current minimax

$$\mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|}$$
$$\boldsymbol{\mu} \leftarrow \left(\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu\right)\frac{K}{(K-1)\alpha+1}$$



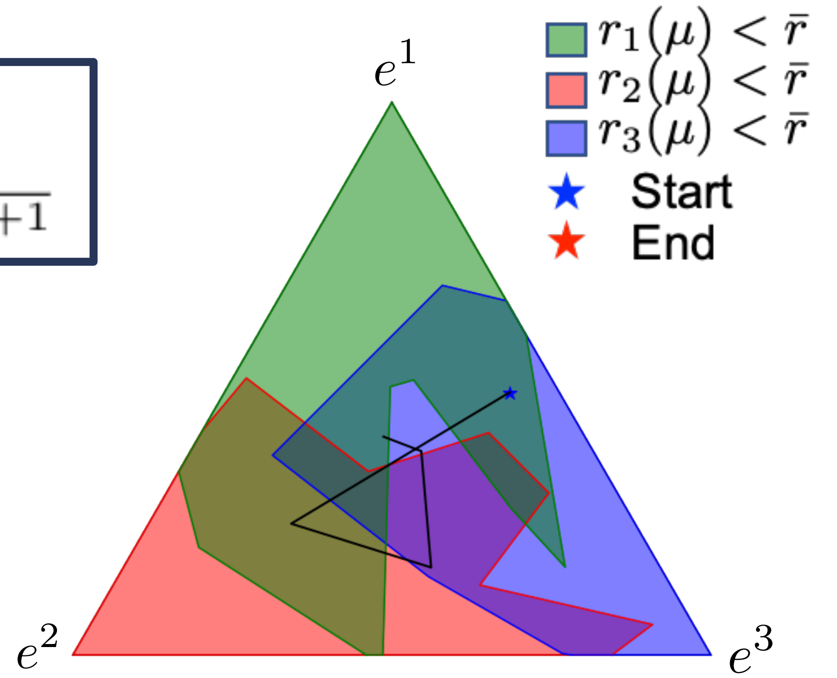
MMPF: Optimization

Minimax weight estimation: APStar

- We propose the following weight update

Current minimax

$$\mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|}$$
$$\boldsymbol{\mu} \leftarrow \left(\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu\right)\frac{K}{(K-1)\alpha+1}$$



MMPF: Optimization

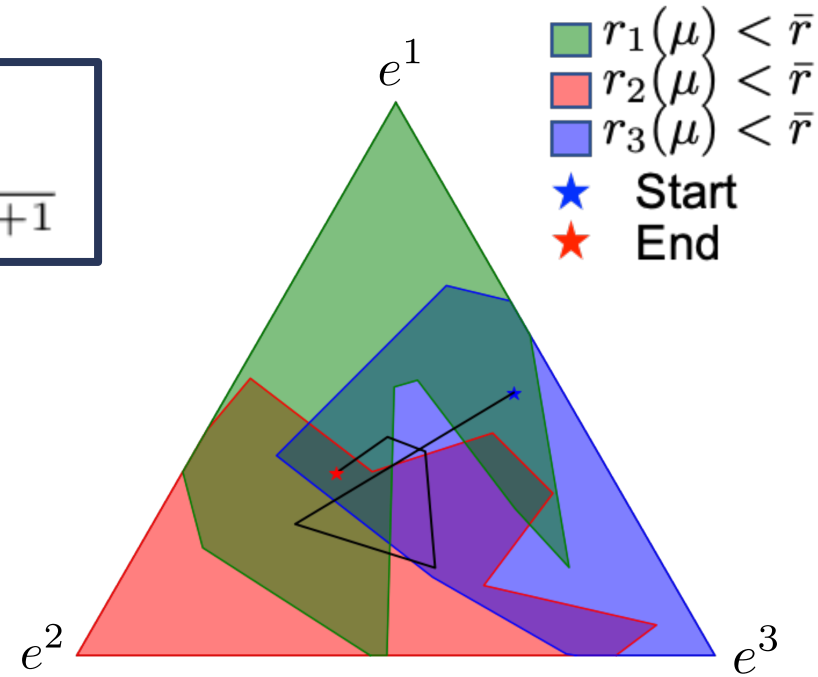
Minimax weight estimation: APStar

- We propose the following weight update

Current minimax

$$\mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|}$$

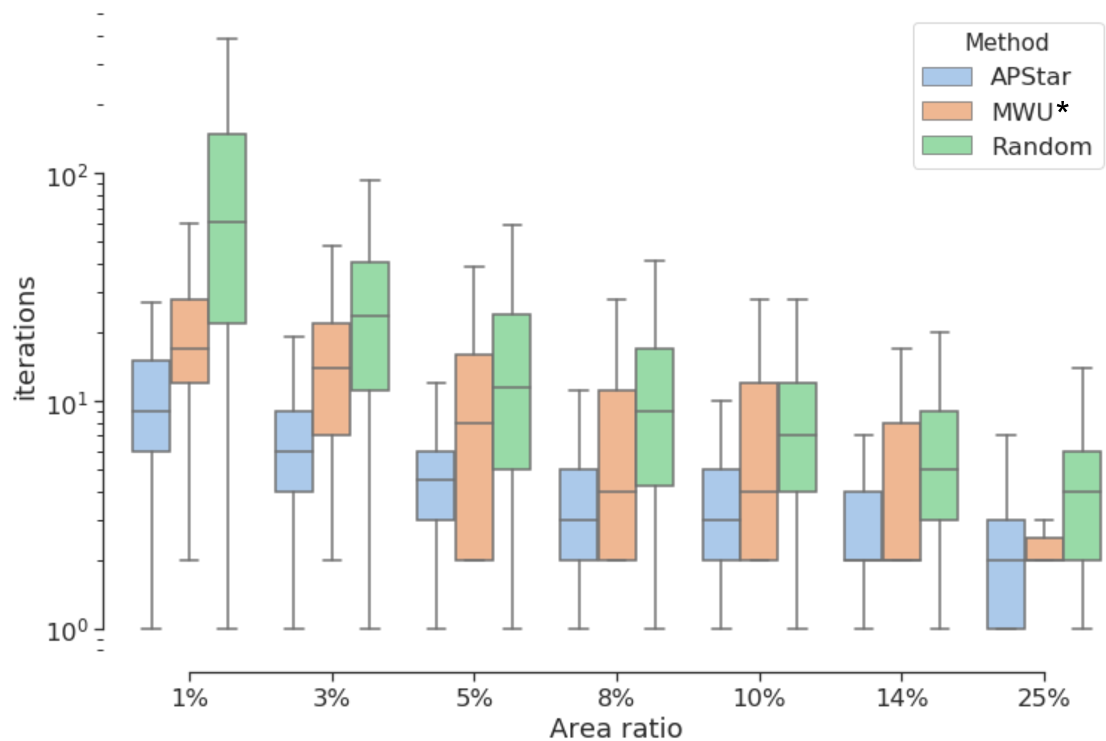
$$\boldsymbol{\mu} \leftarrow \left(\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu\right)\frac{K}{(K-1)\alpha+1}$$

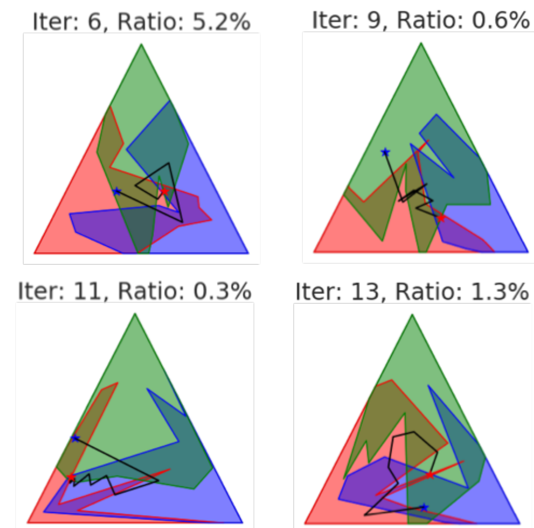


MMPF: Experiments and Results

Synthetic data

- Performance evaluation on sampled star-sets





* Multiplicative Weights Updates by Chen et al 2017: $\mu^{t+1} = e^{\gamma \mathbf{r}(\mu^t)} \mu^t$

MMPF: Experiments and Results

Predicting Mortality in ICU (MIMIC III) from Medical Notes

- 8 Sensitive Groups

Accuracy Comparison

	Sample mean	Group mean	Worst group	Disparity
Naive	89.5 ±0.2	61.9 ±1.7	19.0 ±2.0	80.5 ±1.3
Balanced	79.4±0.6	77.5±1.4	66.8±2.2	22.6±2.3
Zafar	86.2±0.3	65.8±1.8	32.0±2.4	62.9±3.6
Feldman	88.6±2.4	64.4±2.9	28.7±2.4	72.1±5.5
Kamishima	89.3±0.2	63.6±2.0	25.1±5.1	76.4±5.2
MMPF	76.2±0.2	78.3±1.5	**72.6±1.7**	**17.1±3.5**
Balanced+H	75.6±1.1	71.7±1.6	65.6±2.8	19.1±1.8
Zafar+H	62.8±1.6	58.3±2.1	51.5±2.8	17.8±3.1
MMPF+H	72.4±1.1	72.3±1.5	**72.0±3.7**	**11.4±3.5**

- *H = Hardt et al 2016*
- *Comparisons using Friedler et al 2019 benchmark*

MMPF: Experiments and Results

Skin Lesion Classification (HAM10000)

- 7 Classes.
- Imbalanced classification problem $(Y = A)$.

Brier Score Comparison

	Sample mean	Group mean	Worst group	Disparity
Naive	$.31 \pm .01$	$.69 \pm .4$	$1.38 \pm .05$	$1.29 \pm .05$
Balanced	$.41 \pm .02$	$.42 \pm .03$	$0.64 \pm .05$	$0.45 \pm .07$
MMPF P	$.49 \pm .02$	$.46 \pm .05$	**0.56 ± 0.4**	**$0.23 \pm .07$**

Accuracy Comparison

	Sample mean	Group mean	Worst group	Disparity
Naive	78.5 ± 0.6	50.8 ± 2.1	2.6 ± 3.9	91.1 ± 4.4
Balanced	70.1 ± 2.4	70.1 ± 2.5	52.6 ± 5.9	32.5 ± 5.1
MMPF P	64.7 ± 1.4	66.7 ± 3.9	**56.8 ± 3.5**	**19.8 ± 7.3**

MMPF: Conclusion and Future Work

Conclusions

- Recover an efficient model that reduces worst-case group risks.
- We characterized Pareto solutions for DNN models and CE, BS risks.
- APStar improves minimax group risk, with no test-time access to group membership.

Future work

- Convergence proof for APStar algorithm.
- Automatically identify high-risk sub-populations.
- Use the Pareto front to inform fairness policies.

Thanks!

Code: https://github.com/natalialmg/MMPF