Human-AI Cooperation for Fairness Elicitation

Cyrus Cousins, Chang Zeng

What is the Justifiable Fairness Concept?

What's the best way to allocate the funds?

Options\Groups			
New Campus Bike Path	8	3	10
Residential Hall Renovation	4	8	8
More Solar Panels	5	10	5

Utilitarian

Nash

Egalitarian

Power Mean (Generalized Mean)

Justifiable Fairness Concepts can be represented by Power Mean (p) !

Groups (G)			
Utility ($\boldsymbol{s}$)	8	3	10
Weights ($\mathbf{w}$) ($\|\mathbf{w}\|_1 = 1$)	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

Power Mean Welfare (Unweighted) (P = 1)	Power Mean Welfare (Weighted) (P = 1)
$\mathrm{M}_1\left(\langle 8,3,10\rangle; \langle \frac{1}{3},\frac{1}{3},\frac{1}{3}\rangle\right) = \sqrt[1]{\frac{1}{3}(8^1+3^1+10^1)} = 7$	$\mathrm{M}_1\left(\langle 8,3,10\rangle; \langle \frac{1}{5},\frac{3}{5},\frac{2}{5}\rangle\right) = \sqrt[1]{\frac{1}{5}8^1+\frac{3}{5}3^1+\frac{1}{5}10^1)} = 5.4$

$$\mathrm{M}_p(\boldsymbol{s};\boldsymbol{w}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{s}_i^{\,p}}$$

Special Cases:

- $\mathrm{M}_0(\boldsymbol{s};\boldsymbol{w}) = \prod_{i=1}^{g} \boldsymbol{s}_i^{\boldsymbol{w}_i}$
- $\mathrm{M}_\infty(\boldsymbol{s};\boldsymbol{w}) = \max_{1<i<g} \boldsymbol{s}_i$
- $\mathrm{M}_{-\infty}(\boldsymbol{s};\boldsymbol{w}) = \min_{1<i<g} \boldsymbol{s}_i$

Power Mean - Fairness Concept

Options\Groups			
New Campus Bike Path	8	3	10
Residential Hall Renovation	4	8	8
More Solar Panels	5	10	5

For the simplicity, assume uniform weights for groups.

Options\Type	Utilitarian (P = 1)	Nash (P = 0)	Egalitarian (P = $-\infty$)
New Campus Bike Path	$\boldsymbol{\sqrt[1]{\frac{1}{3}(8^1+3^1+10^1)} = 7}$	$(8 * 3 * 10)^{\frac{1}{3}} \approx 6.2145$	$\min(8, 3, 10) = 3$
Residential Hall Renovation	$\sqrt[1]{\frac{1}{3}(4^1+8^1+8^1)} \approx 6.7$	$\boldsymbol{(4 * 8 * 8)^{\frac{1}{3}} \approx 6.3496}$	$\min(4, 8, 8) = 4$
More Solar Panels	$\sqrt[1]{\frac{1}{3}(5^1+10^1+5^1)} \approx 6.7$	$(5 * 10 * 5)^{\frac{1}{3}} \approx 6.2996$	$\boldsymbol{\min(5, 10, 5) = 5}$

Why Power Mean

Previous work: An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

Distance Between Power Mean Fairness Concepts

What does it even mean to measure distance between fairness concepts?

Intuitive Solution:
Difference between welfare given same sentiment value and probability measure!

$$|\mathrm{M}_{p_{\uparrow}}(\boldsymbol{s};\boldsymbol{w}) - \mathrm{M}_{p_{\downarrow}}(\boldsymbol{s};\boldsymbol{w})|$$

Distance Between Power Mean Fairness Concepts

$$|\mathrm{M}_1(\boldsymbol{s};\boldsymbol{w}) - \mathrm{M}_{-\infty}(\boldsymbol{s};\boldsymbol{w})|$$

Options\Groups			
New Campus Bike Path	8	3	10
More Solar Panels	5	10	5

For the simplicity, assume uniform weights for groups.

New Campus Bike Path	More Solar Panels
$\left\|\mathrm{M}_1\left(\langle 8,3,10\rangle;\langle\frac{1}{3},\frac{1}{3},\frac{1}{3}\rangle\right) - \mathrm{M}_{-\infty}\left(\langle 8,3,10\rangle;\langle\frac{1}{3},\frac{1}{3},\frac{1}{3}\rangle\right)\right\|$ $= \sqrt[1]{\frac{1}{3}(8^1+3^1+10^1)} - \min(8,3,10) = \mathbf{4}$	$\left\|\mathrm{M}_1\left(\langle 5,10,5\rangle;\langle\frac{1}{3},\frac{1}{3},\frac{1}{3}\rangle\right) - \mathrm{M}_{-\infty}\left(\langle 5,10,5\rangle;\langle\frac{1}{3},\frac{1}{3},\frac{1}{3}\rangle\right)\right\|$ $= \sqrt[1]{\frac{1}{3}(5^1+10^1+5^1)} - \min(5,10,5) \approx \mathbf{1.7}$

Distance Between Power Mean Fairness Concepts

$$\Delta(p_{\uparrow}, p_{\downarrow}; \boldsymbol{w}) \doteq \sup_{\boldsymbol{s} \in [0,1]^g} \left| \mathrm{M}_{p_{\uparrow}}(s; \boldsymbol{w}) - \mathrm{M}_{p_{\downarrow}}(s; \boldsymbol{w}) \right|$$

Properties:

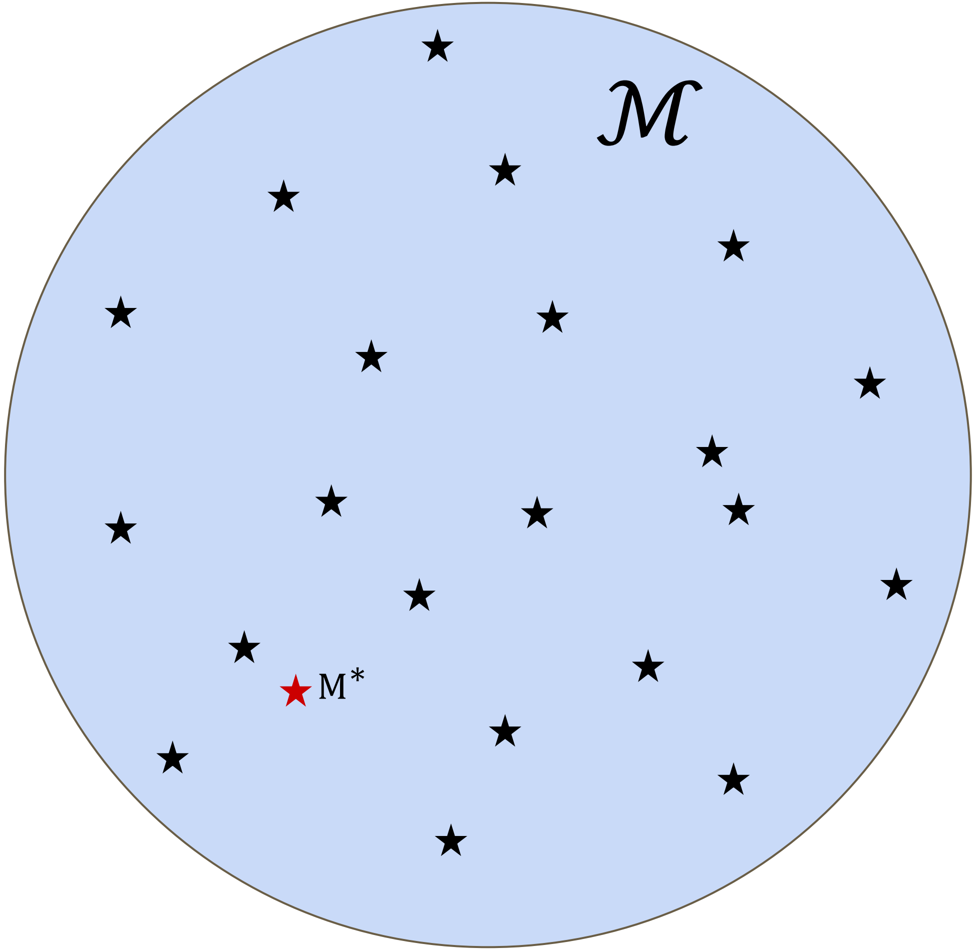
- Triangle Inequality
- Symmetric
- Positive-Definiteness
  - F(x, y) = 0 iff x = y
  - F(x, y) $\geq$ 0

Problem Setup

$\mathcal{M}$: Justifiable Fairness Concept Set

M^* : Human Cardinal Fairness Concept

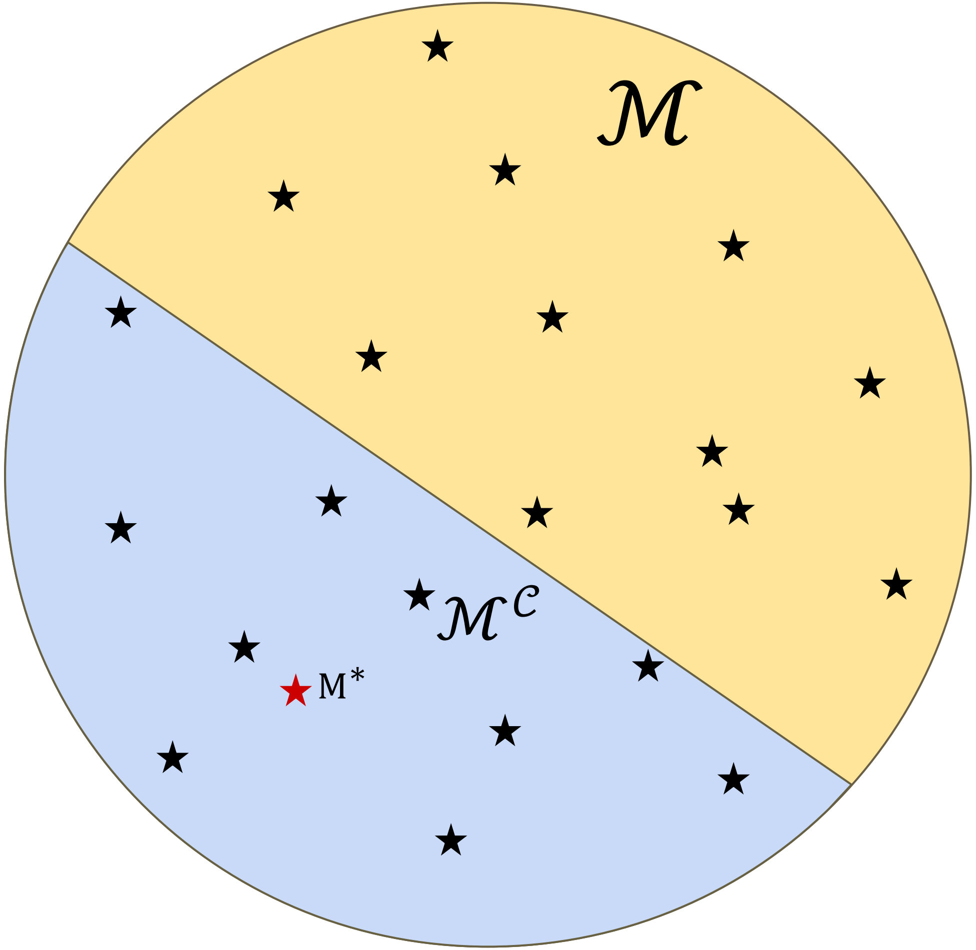
Query: $\mathrm{M}^*(\boldsymbol{s}; \boldsymbol{w}) > \mathrm{M}^*(\boldsymbol{s}'; \boldsymbol{w}')$?



Problem Setup

$\mathcal{M}$: Justifiable Fairness Concept Set
$\mathcal{M}^{\mathcal{C}}$: Concordant Fairness Concept Set
M^* : Human Cardinal Fairness Concept

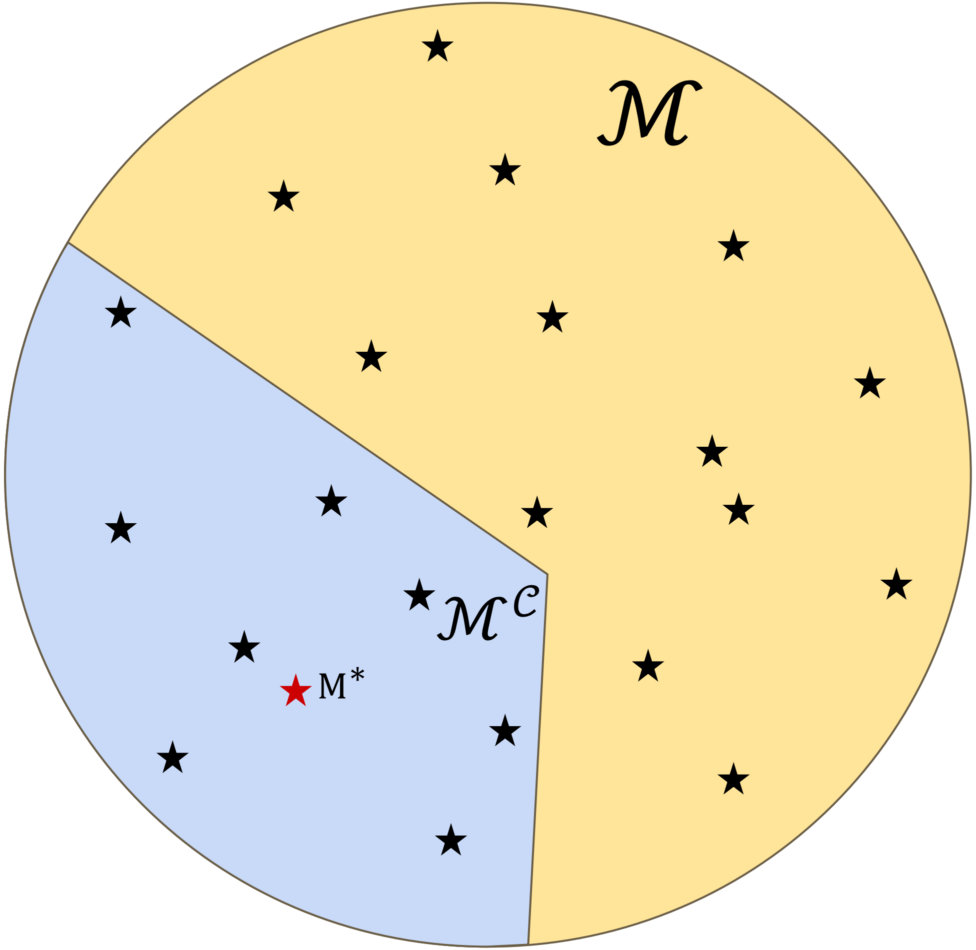
Query: $\mathrm{M}^*(\boldsymbol{s}; \boldsymbol{w}) > \mathrm{M}^*(\boldsymbol{s}'; \boldsymbol{w}')$?



Problem Setup

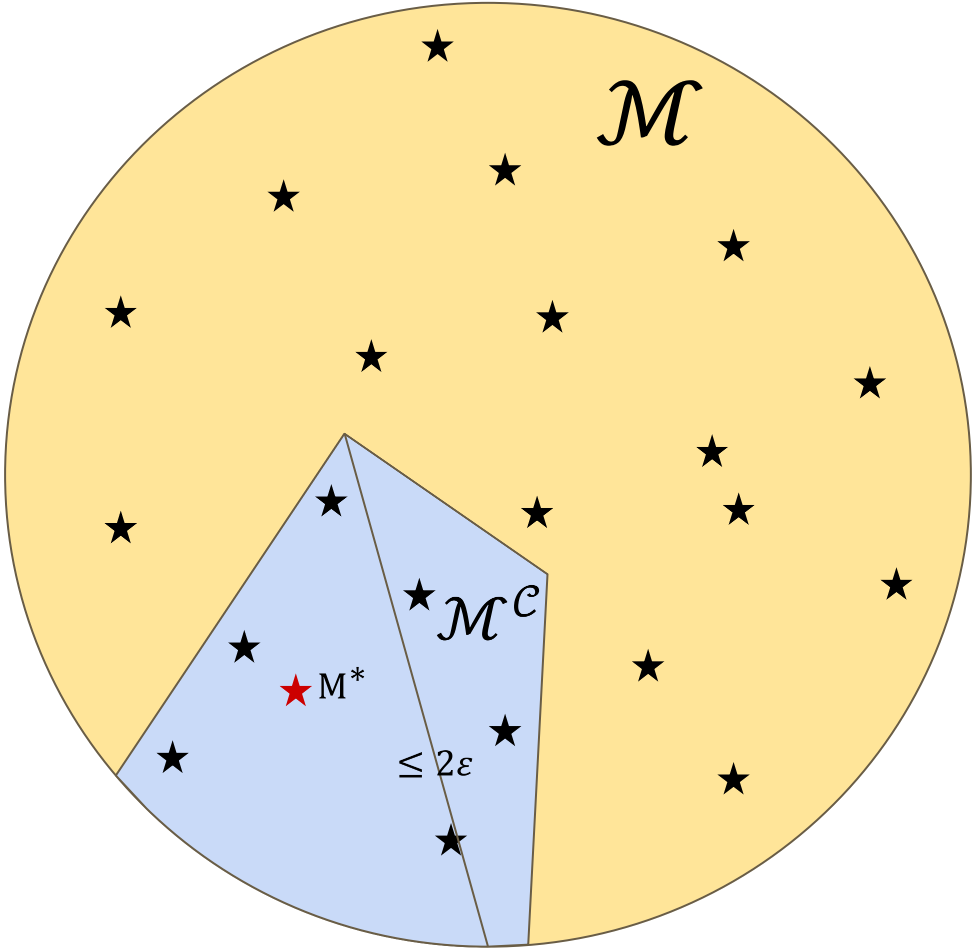
$\mathcal{M}$: Justifiable Fairness Concept Set
$\mathcal{M}^{\mathcal{C}}$: Concordant Fairness Concept Set
M^* : Human Cardinal Fairness Concept

Query: $\mathrm{M}^*(\boldsymbol{s};\boldsymbol{w}) > \mathrm{M}^*(\boldsymbol{s}';\boldsymbol{w}')$?



Problem Setup

$\mathcal{M}$: Justifiable Fairness Concept Set
$\mathcal{M}^{\mathcal{C}}$: Concordant Fairness Concept Set
M^* : Human Cardinal Fairness Concept
ε : Error Tolerance



Recap & Our Contribution

We have introduced:

- Power Mean Fairness Concept p $\left(\mathrm{M}_p(\boldsymbol{s};\boldsymbol{w}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{s}_i^p}\right)$
- Distance Metric on Power Mean Fairness Concept. $(\Delta(p_\uparrow, p_\downarrow; w))$

Our Contribution for this work are:

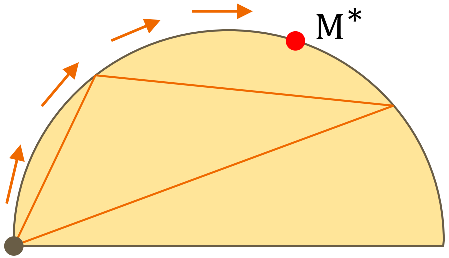
- Upper bound on the distance between Power Mean Fairness Concept. $(\Delta_\uparrow(p_\uparrow, p_\downarrow; w))$
- Search Algorithms on the Justifiable Fairness Concepts set.

Intuition of Upper Bounds

$$\Delta(p_\uparrow, p_\downarrow; \mathrm{w}) \doteq \sup_{s\in[0,1]^g} |\mathrm{M}_{p_\uparrow}(s;w) - \mathrm{M}_{p_\downarrow}(s;w)|$$

$\left(\mathrm{M}_p(\boldsymbol{s};\boldsymbol{w}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{s}_i^{\,p}}\right)$

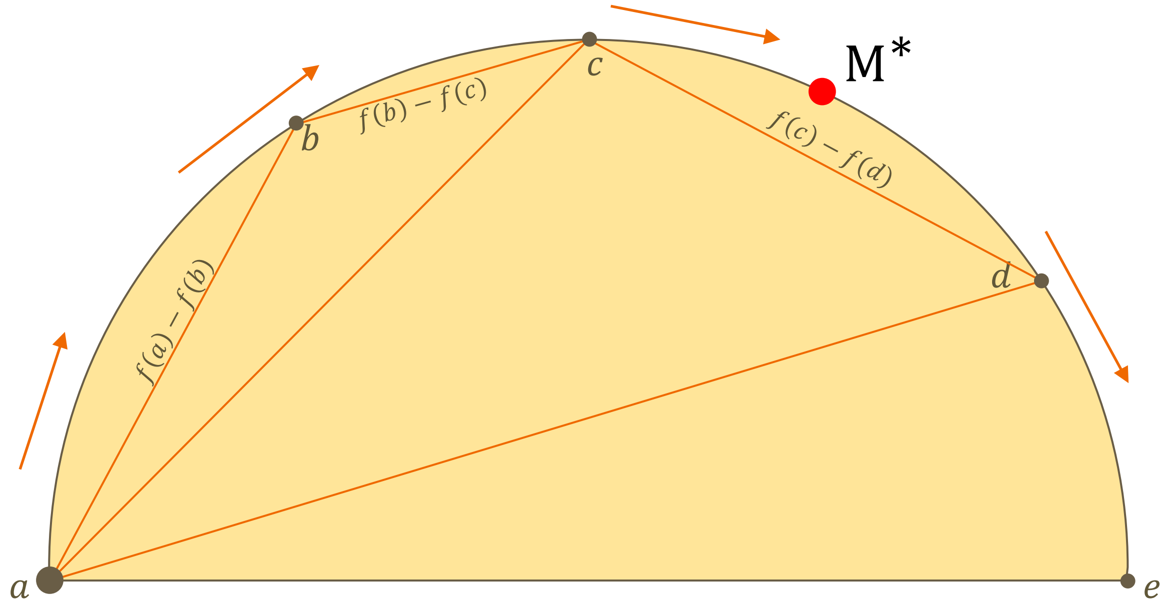
How to compute supremum?



$\Delta(p_\uparrow, p_\downarrow; w)$ obeys Triangle Inequality

(Unable to upper bound time/query complexity)

Optimal Continuous Anti-Triangular Symmetric Bound



$$\sup|f(a) - f(b)| + \sup|f(b) - f(c)| \geq \sup|f(a) - f(c)|$$

$$\lim_{h \to 0} \sup_{s \in [0,1]^g} \left| \mathrm{M}_{p_\downarrow + h}(\boldsymbol{s}; \boldsymbol{w}) - \mathrm{M}_{p_\downarrow}(\boldsymbol{s}; \boldsymbol{w}) \right| + \sup_{s \in [0,1]^g} \left| \mathrm{M}_{p_\downarrow + 2h}(\boldsymbol{s}; \boldsymbol{w}) - \mathrm{M}_{p_\downarrow + h}(\boldsymbol{s}; \boldsymbol{w}) \right| + \cdots$$

Optimal Continuous Anti-Triangular Symmetric Bound

$$\lim_{h\to 0} \sup_{s\in[0,1]^g} \left|\mathrm{M}_{p_\downarrow+h}(\boldsymbol{s};\boldsymbol{w}) - \mathrm{M}_{p_\downarrow}(\boldsymbol{s};\boldsymbol{w})\right| + \sup_{s\in[0,1]^g} \left|\mathrm{M}_{p_\downarrow+2h}(\boldsymbol{s};\boldsymbol{w}) - \mathrm{M}_{p_\downarrow+h}(\boldsymbol{s};\boldsymbol{w})\right| + \cdots$$

$$\lim_{h\to 0} \sum_{i=1}^{\frac{p_\uparrow - p_\downarrow}{h}} \frac{\sup_{s\in[0,1]^g} \left|\mathrm{M}_{p_\downarrow+ih}(\boldsymbol{s};\boldsymbol{w}) - \mathrm{M}_{p_\downarrow+(i-1)h}(\boldsymbol{s};\boldsymbol{w})\right|}{h} * h$$

$$\left|\int_{p_\downarrow}^{p_\uparrow} \sup_{s\in[0,1]^g} \frac{\mathrm{d}}{\mathrm{d}p}[\mathrm{M}_p(\boldsymbol{s};\boldsymbol{w})]\mathrm{d}p\right|$$

Optimal Continuous Anti-Triangular Symmetric Bound

$$\Delta(p_\uparrow, p_\downarrow; \boldsymbol{w}) \doteq \sup_{s \in [0,1]^g} \left| \int_{p_\downarrow}^{p_\uparrow} \frac{\mathrm{d}}{\mathrm{d}p} [\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})] \mathrm{d}p \right|$$

$$\Delta_\uparrow^*(p_\uparrow, p_\downarrow; \boldsymbol{w}) \doteq \left| \int_{p_\downarrow}^{p_\uparrow} \sup_{s \in [0,1]^g} \frac{\mathrm{d}}{\mathrm{d}p} [\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})] \mathrm{d}p \right|$$

Properties:

- Additive (Most Important !!)
- Symmetric
- Positive-Definiteness
  - F(x, y) = 0 iff x = y
  - F(x, y) ≥ 0

Upper Bounds on Power Mean

Upper bounds on Power Mean Fairness Concepts:

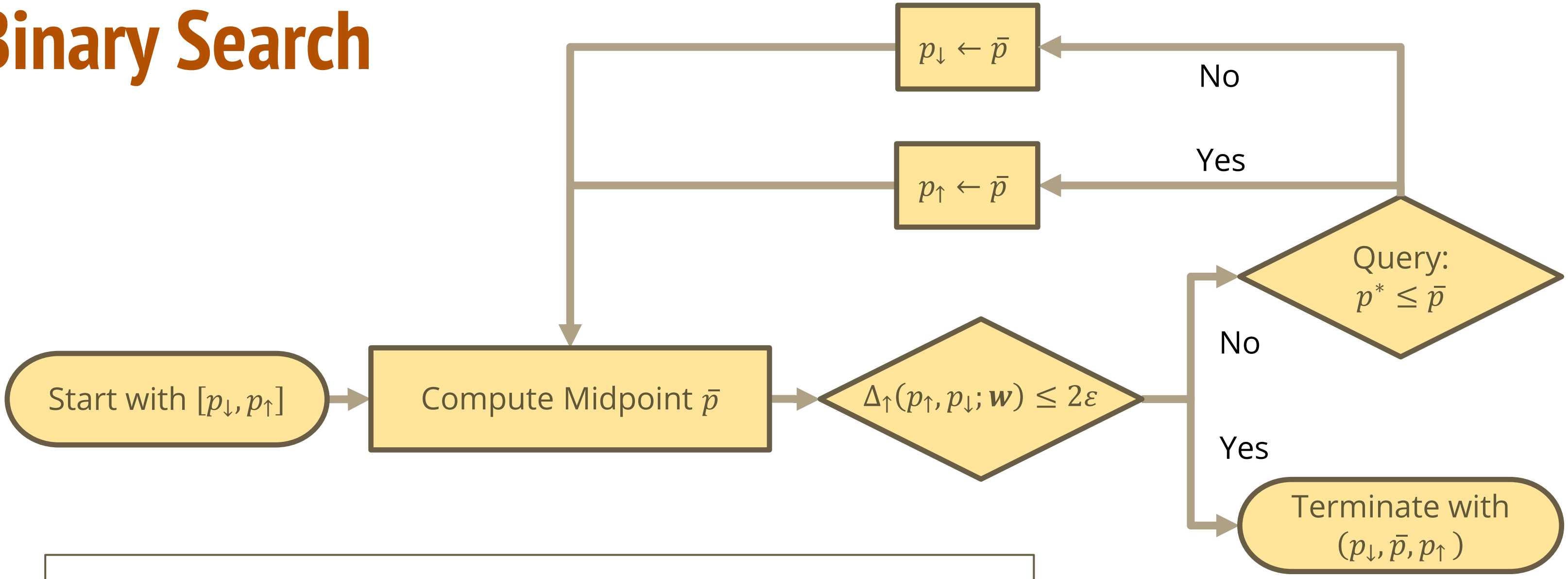
- $\Delta_{\uparrow}(p_{\uparrow}, p_{\downarrow}; \boldsymbol{w}) \doteq \frac{1}{e} \ln \frac{p_{\uparrow}}{p_{\downarrow}}$, for any $p_{\uparrow}, p_{\downarrow} > 0$ (log ratio)
- $\Delta_{\uparrow}(p_{\uparrow}, p_{\downarrow}; \boldsymbol{w}) \doteq \left(\frac{p_{\uparrow} - p_{\downarrow}}{p_{\uparrow} p_{\downarrow}}\right) \ln \frac{1}{\boldsymbol{w}_{\min}}$, for any $p_{\uparrow} p_{\downarrow} > 0$ (harmonic difference)

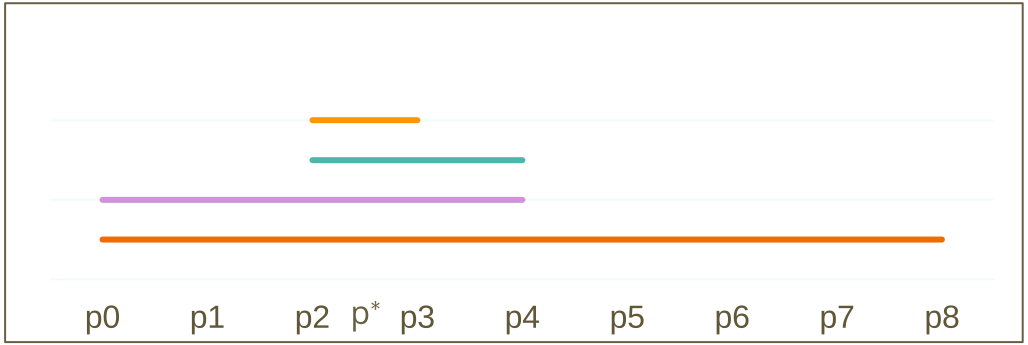
$(\boldsymbol{w}_{\min} = \min_{1 \le i \le g} \boldsymbol{w}_i)$

Extreme case ($p = \pm\infty$):

- $\Delta_{\uparrow}(\infty, p_{\downarrow}; \boldsymbol{w}) = \frac{1}{p_{\downarrow}} \ln \left(\frac{1}{\boldsymbol{w}_{\min}}\right)$
- $\Delta_{\uparrow}(p_{\uparrow}, -\infty; \boldsymbol{w}) = -\frac{1}{p_{\uparrow}} \ln \left(\frac{1}{\boldsymbol{w}_{\min}}\right)$

Binary Search



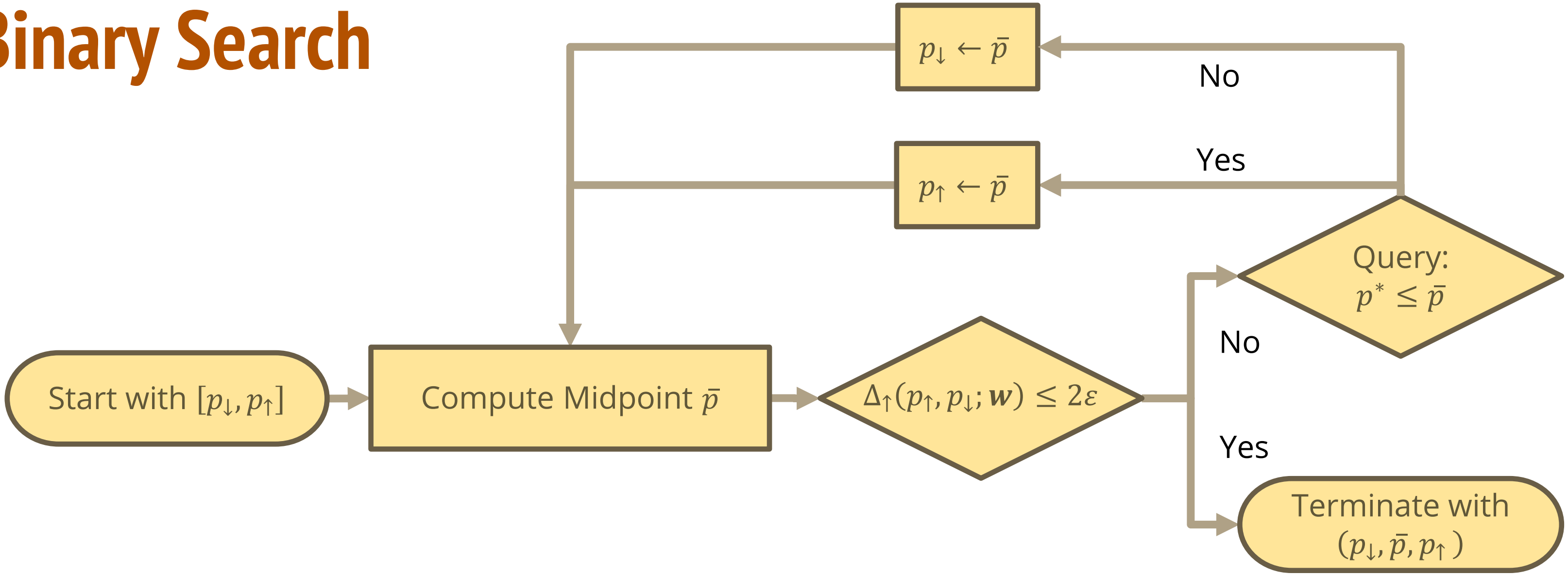


Query Complexity ($\mathrm{N_E}$):

$$\log \frac{\Delta(p_\uparrow, p_\downarrow; \boldsymbol{w})}{2\varepsilon} \le \mathrm{NE} \le \log \frac{\Delta_\uparrow(p_\uparrow, p_\downarrow; \boldsymbol{w})}{2\varepsilon}$$

Suppose $p^* \in [p_2, p_3]$ and $\Delta_\uparrow(p_i, p_{i+1}; \boldsymbol{w}) \le 2\varepsilon$

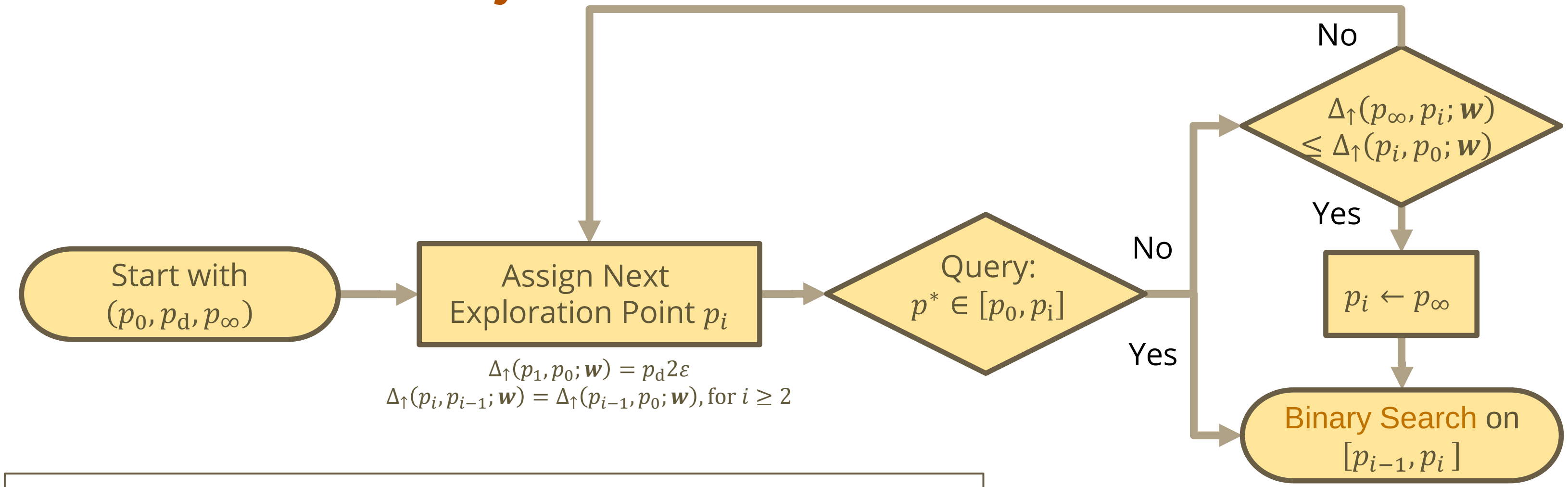
Binary Search

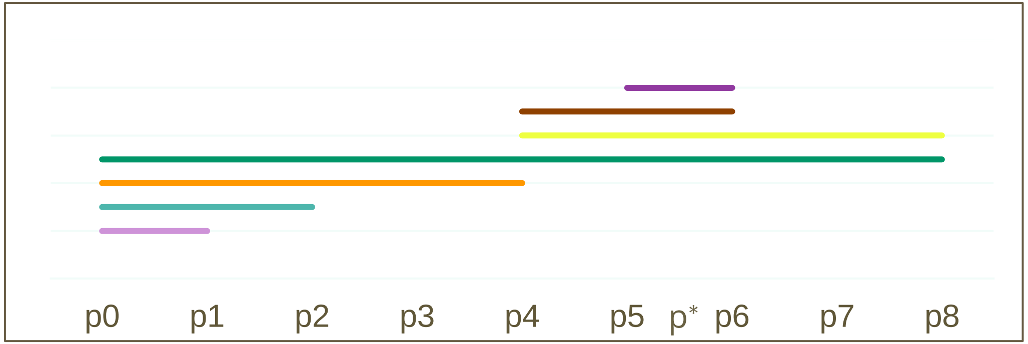


Require Starting Interval $[p_{\downarrow}, p_{\uparrow}]$

Power Mean Fairness Concept $p \in [-\infty, \infty]$

Unbounded Binary Search



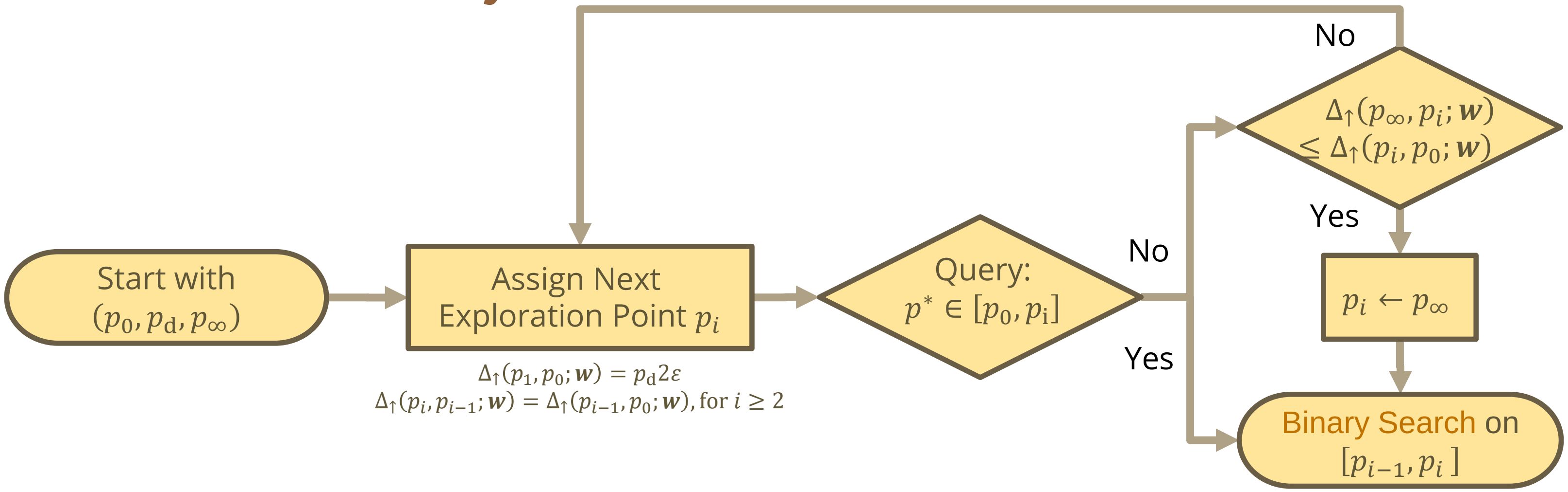


Query Complexity ($\mathrm{N_E}$):

$$2\log\frac{\Delta(p_0, p^*; \boldsymbol{w})}{2\varepsilon} \leq \mathrm{NE} \leq 2\log\frac{\Delta_\uparrow(p_0, p^*; \boldsymbol{w})}{2\varepsilon}$$

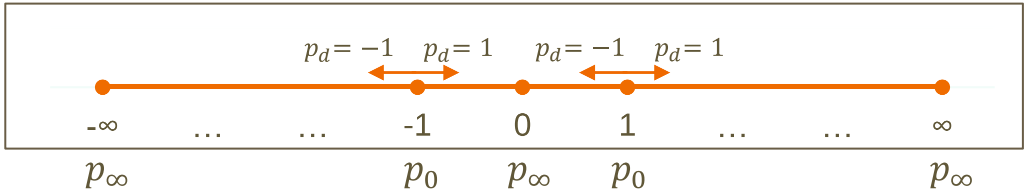
Suppose $p^* \in [p_5, p_6]$ and $\Delta_\uparrow(p_i, p_{i+1}; \boldsymbol{w}) \leq 2\varepsilon$

Unbounded Binary Search



Power Mean Fairness Concept $p \in [-\infty, \infty]$

Also Require Starting points $p_0 \in \pm 1, p_d \in \pm 1, p_\infty \in \{0, \pm\infty\}$



Conclusion

We have presented:

- Upper bound on the distance between Power Mean Fairness Concept.
  - $\Delta_\uparrow^*(p_\uparrow, p_\downarrow; \boldsymbol{w}) \doteq \left| \int_{p_\downarrow}^{p_\uparrow} \sup_{s \in [0,1]^g} \frac{\mathrm{d}}{\mathrm{d}p} [\mathrm{M}p(\boldsymbol{s}; \boldsymbol{w})] \mathrm{d}p \right|$ (Optimal Continuous Anti-Triangular Symmetric Bound)
  - $\Delta_\uparrow(p_\uparrow, p_\downarrow; \boldsymbol{w}) \doteq \frac{1}{e} \ln \frac{p_\uparrow}{p_\downarrow}$, for any $p_\uparrow, p_\downarrow > 0$ (log ratio)
  - $\Delta_\uparrow(p_\uparrow, p_\downarrow; \boldsymbol{w}) \doteq \left(\frac{p_\uparrow - p_\downarrow}{p_\uparrow p_\downarrow}\right) \ln \frac{1}{\boldsymbol{w}_{\min}}$, for any $p_\uparrow p_\downarrow > 0$ (harmonic difference) $(\boldsymbol{w}_{\min} = \min_{1 \le i \le g} \boldsymbol{w}_i)$
- Search Algorithms on the Justifiable Fairness Concepts set.
  - Binary Search $\left(\log \frac{\Delta(p_\uparrow, p_\downarrow; \boldsymbol{w})}{2\varepsilon} \le \mathrm{NE} \le \log \frac{\Delta_\uparrow(p_\uparrow, p_\downarrow; \boldsymbol{w})}{2\varepsilon}\right)$
  - Unbounded Binary Search $\left(2 \log \frac{\Delta(p_\uparrow, p_\downarrow; \boldsymbol{w})}{2\varepsilon} \le \mathrm{NE} \le 2 \log \frac{\Delta_\uparrow(p_\uparrow, p_\downarrow; \boldsymbol{w})}{2\varepsilon}\right)$

Thank You

Constraint on Sentiment Value

$$\Delta(p_\uparrow, p_\downarrow; \boldsymbol{w}) \doteq \sup_{\boldsymbol{s}\in[0,1]^g} |\mathrm{M}_{p_\uparrow}(\boldsymbol{s}; \boldsymbol{w}) - \mathrm{M}_{p_\downarrow}(\boldsymbol{s}; \boldsymbol{w})|$$

$$\Delta'(p_\uparrow, p_\downarrow; \boldsymbol{w}) \doteq \sup_{\boldsymbol{s}\in[0,\alpha]^g} |\mathrm{M}_{p_\uparrow}(\boldsymbol{s}; \boldsymbol{w}) - \mathrm{M}_{p_\downarrow}(\boldsymbol{s}; \boldsymbol{w})|$$

$$\mathrm{M}_p(\alpha\boldsymbol{s}; \boldsymbol{w}) = \alpha\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}) \rightarrow \Delta(p_\uparrow, p_\downarrow; \boldsymbol{w}) = \alpha\Delta'(p_\uparrow, p_\downarrow; \boldsymbol{w})$$

Composite $\Delta_{\uparrow}$ Function

Harmonic difference, $\left(\frac{p_{\uparrow}-p_{\downarrow}}{p_{\uparrow}p_{\downarrow}}\right)\ln\frac{1}{\boldsymbol{w}_{\min}}$, depends on weights

What happen if $\boldsymbol{w}_{\min}$ too small? Super loose bound!!

Compositing log ratio and harmonic difference bound together:

($\boldsymbol{w}_{\min} = \min_{1\leq i\leq g} \boldsymbol{w}_i$, $\tilde{p} = e\ln\frac{1}{\boldsymbol{w}_{\min}}$)

$$\Delta_{\uparrow}(p_{\uparrow}, p_{\downarrow}; \boldsymbol{w}) \doteq \begin{cases} \left(\frac{p_{\uparrow}-p_{\downarrow}}{p_{\uparrow}p_{\downarrow}}\right)\ln\frac{1}{\boldsymbol{w}_{\min}} & \tilde{p} \leq p_{\downarrow} \\ \frac{1}{e}\ln\frac{p_{\uparrow}}{p_{\downarrow}} & \tilde{p} \geq p_{\uparrow} \\ \left(\frac{p_{\uparrow}-\tilde{p}}{p_{\uparrow}p_{\downarrow}}\right)\ln\frac{1}{\boldsymbol{w}_{\min}} + \frac{1}{e}\ln\frac{\tilde{p}}{p_{\downarrow}} & \tilde{p} \in [p_{\downarrow}, p_{\uparrow}] \end{cases}$$