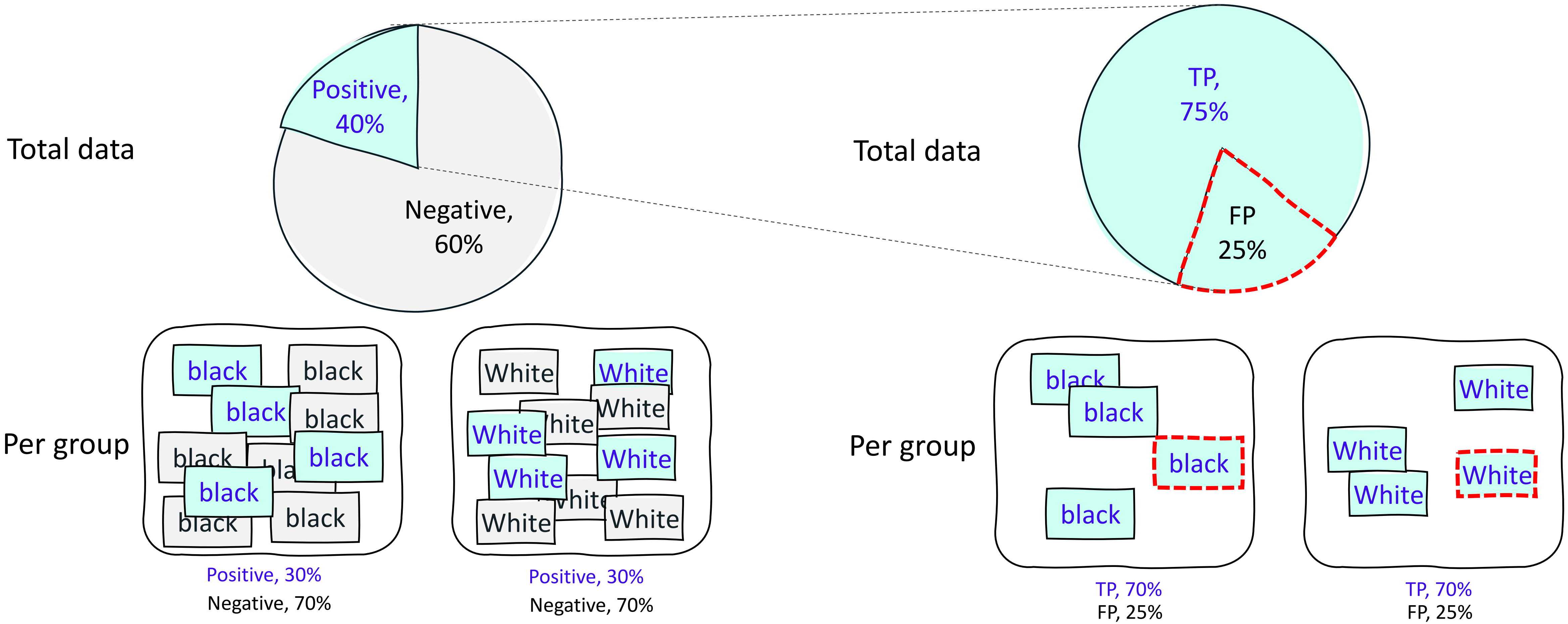
Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness

ICML 2018

Metric of fairness

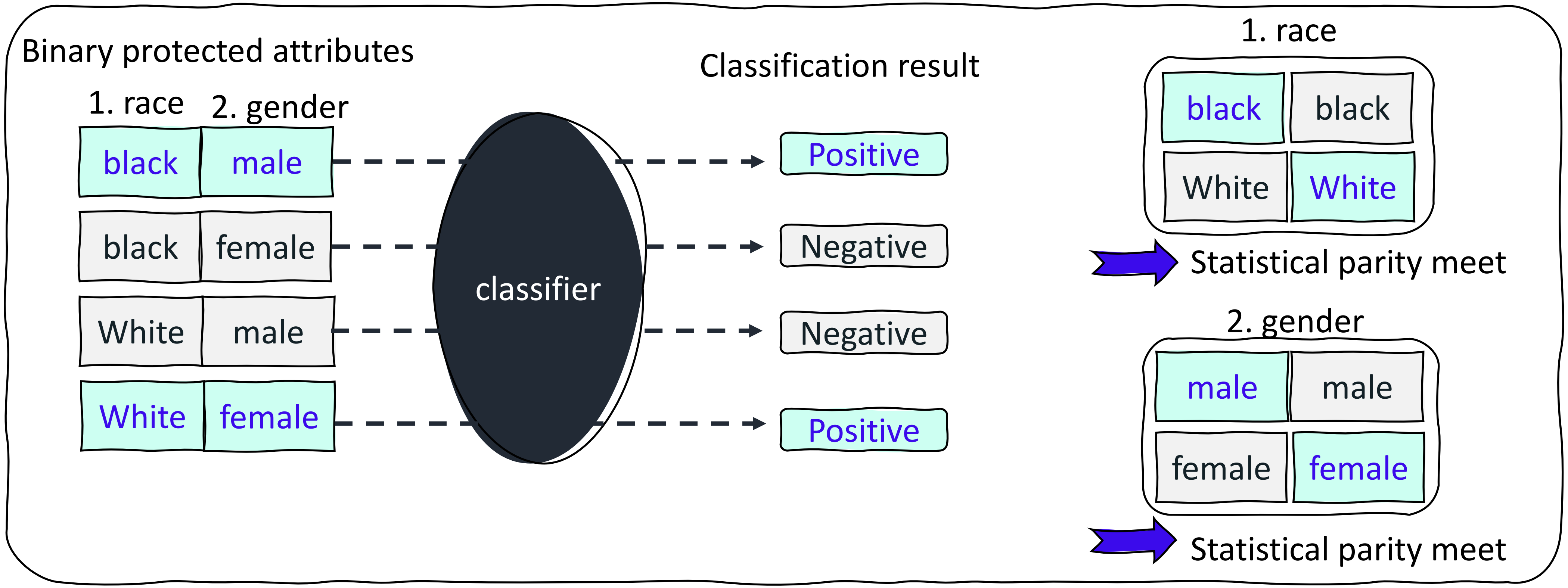
SP, equality of classification rates (statistical parity)

FP, equality of false positive rates (equal opportunity)



Defining of Subgroup

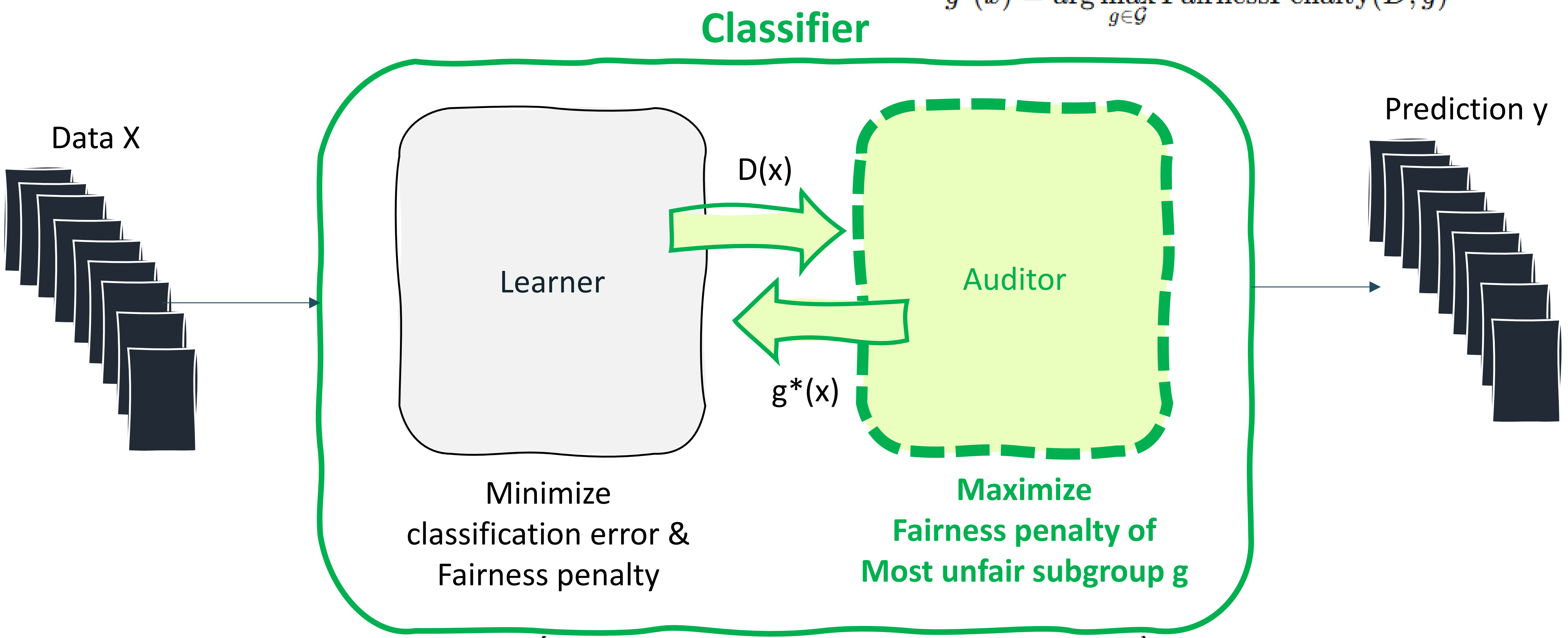
1. group: fairness gerrymandering problem



2. Need of **Sub group** : Fairness Violation evaluation → Auditing (subgroup searching)

(Learner–Auditor) Zero-sum game formulation

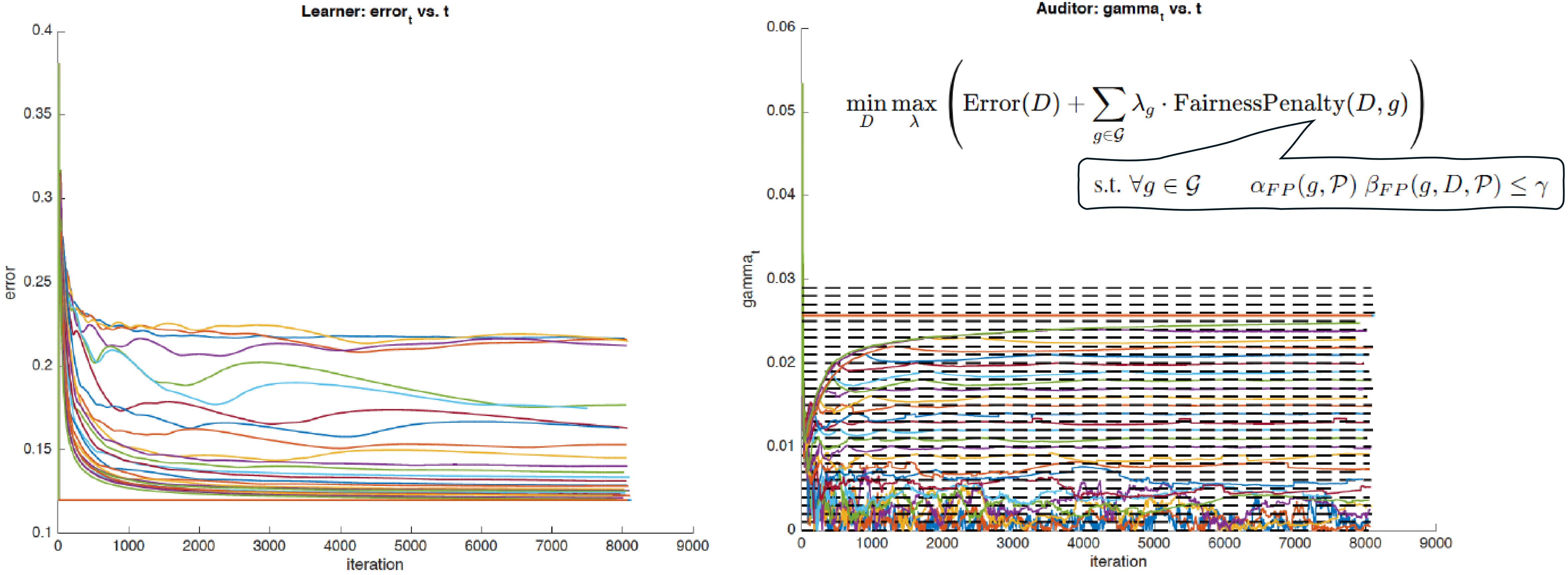
$$g^*(x) = \arg\max_{g \in \mathcal{G}} \text{FairnessPenalty}(D, g)$$



$$\min_{D} \max_{\lambda} \left(\text{Error}(D) + \sum_{g \in \mathcal{G}} \lambda_g \cdot \text{FairnessPenalty}(D, g) \right)$$

Appendix

Experimental result



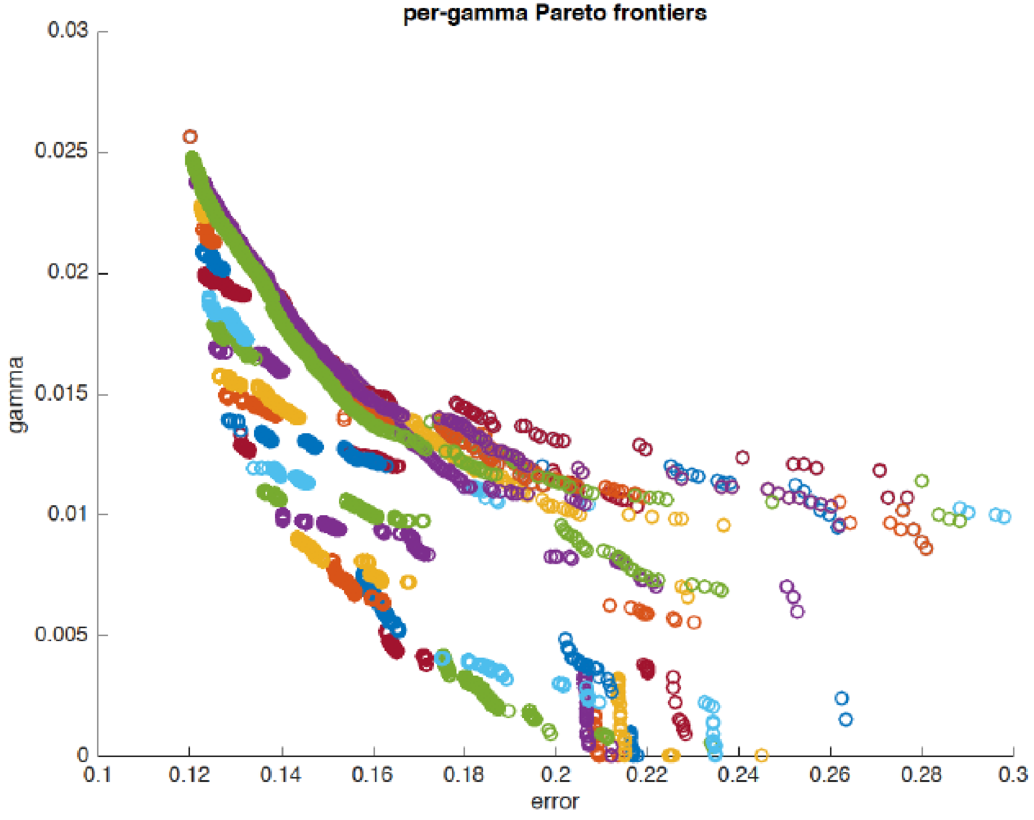
$$\min_{D} \max_{\lambda} \left(\mathrm{Error}(D) + \sum_{g \in \mathcal{G}} \lambda_g \cdot \mathrm{FairnessPenalty}(D, g) \right)$$

$$\text{s.t. } \forall g \in \mathcal{G} \qquad \alpha_{FP}(g, \mathcal{P})\, \beta_{FP}(g, D, \mathcal{P}) \leq \gamma$$

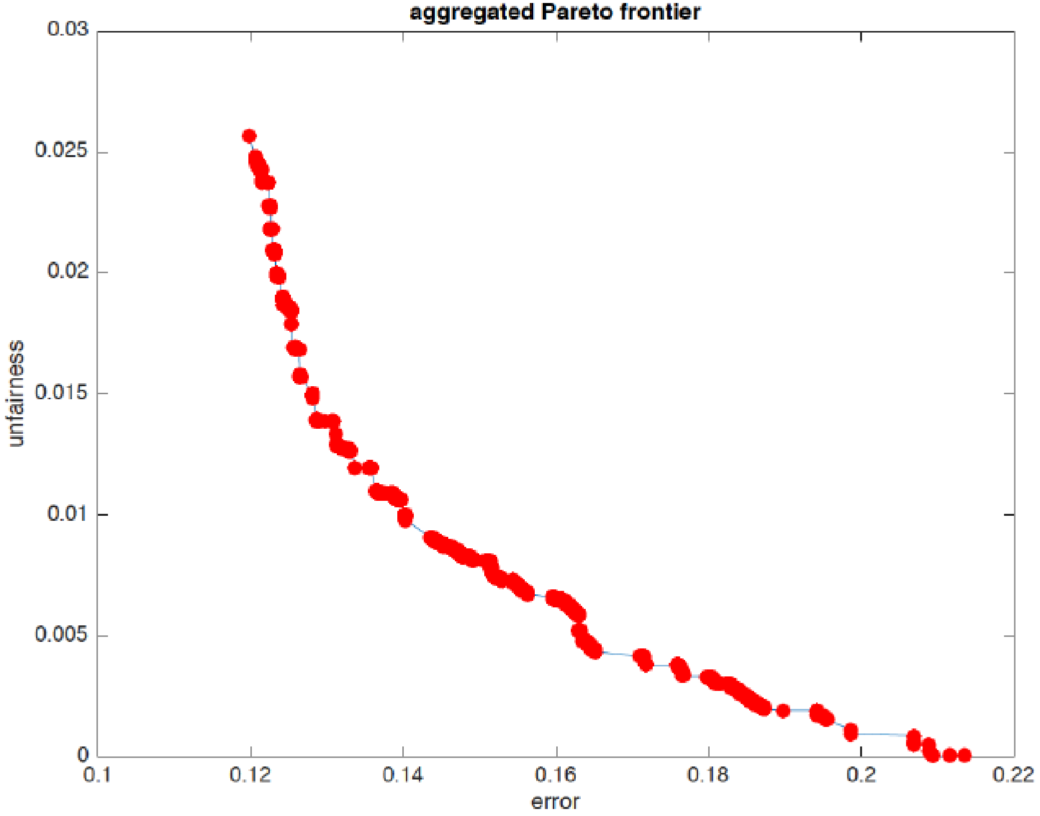
(a) (b)

Figure 1. Evolution of the error and unfairness of Learner's classifier across iterations, for varying choices of r.
(a) Error "t of Learner's model vs iteration t. (b) Unfairness t of subgroup found by Auditor vs. iteration t,
as measured by Definition 2.3. See text for details.

Experimental result



(a)



(b)

Figure 2. (a) Pareto-optimal error-unfairness values, color coded by varying values of the input parameter r.
(b) Aggregate Pareto frontier across all values of r.
Here the r values cover the same range but are sampled more densely to get a smoother frontier. See text for details.

Auditing ~ Weak Agnostic Learning reduction theorem

Proving Computational equivalence

Theorem 3.1. *Fix any distribution $\mathcal{P}$, and any set of group indicators $\mathcal{G}$. Then for any $\gamma, \varepsilon > 0$, the following relationships hold:*

- *If there is a $(\gamma/2, (\gamma/2 - \varepsilon))$ auditing algorithm for $\mathcal{G}$ for all D such that $\mathrm{SP}(D) = 1/2$, then the class $\mathcal{G}$ is $(\gamma, \gamma/2 - \varepsilon)$-weakly agnostically learnable under $\mathcal{P}^D$.*
- *If $\mathcal{G}$ is $(\gamma, \gamma - \varepsilon)$-weakly agnostically learnable under marginal distribution $\mathcal{P}^D$ on $(x, D(X))$ for all D such that $\mathrm{SP}(D) = 1/2$, then there is a $(\gamma, (\gamma - \varepsilon)/2)$ auditing algorithm for $\mathcal{G}$ for SP fairness under $\mathcal{P}$.*

Corollary 3.2. *Fix any distribution $\mathcal{P}$, and any set of group indicators $\mathcal{G}$. The following two relationships hold:*

- *If there is a $(\gamma/2, (\gamma/2 - \varepsilon))$ auditing algorithm for $\mathcal{G}$ for all D with $\mathrm{FP}(D) = 1/2$, then $\mathcal{G}$ is $(\gamma, \gamma/2 - \varepsilon)$-weakly agnostically learnable under $\mathcal{P}^D_{y=0}$.*
- *If $\mathcal{G}$ is $(\gamma, \gamma - \varepsilon)$–weakly agnostically learnable under the conditional distribution $\mathcal{P}^D_{y=0}$ of (X, y) conditioned on the event that $D(X) = 1$ for all D with $\mathrm{FP}(D) = 1/2$, then there is a $(\gamma, (\gamma - \varepsilon)/2)$ auditing algorithm for FP subgroup fairness for $\mathcal{G}$ under distribution $\mathcal{P}$.*

Mathematical theorem

Theorem 3.3. *Under standard complexity-theoretic intractability assumptions, for $\mathcal{G}$ the classes of conjunctions of boolean attributes, linear threshold functions, or bounded-degree polynomial threshold functions, there exist distributions P such that the auditing problem cannot be solved in polynomial time, for either SP or FP fairness.*

Theorem 4.1. *Fix any $\nu, \delta \in (0, 1)$. Then given an input of n data points and accuracy parameters ν, δ and access to oracles $\mathrm{CSC}(\mathcal{H})$ and $\mathrm{CSC}(\mathcal{G})$, there exists an algorithm runs in time $\mathrm{poly}(1/\nu, \log(1/\delta))$, and with probability at least $1 - \delta$, output a randomized classifier $\hat{\mathrm{D}}$ such that $err(\hat{\mathrm{D}}, \mathcal{P}) \leq \mathrm{OPT} + \nu$, and for any $g \in \mathcal{G}$, the fairness constraint violations satisfies*

$$\alpha_{FP}(g, \mathcal{P})\ \beta_{FP}(g, \hat{\mathrm{D}}, \mathcal{P}) \leq \gamma + O(\nu).$$