Valentin L. Buchner, Philip O.O. Schutte, Yassin Ben Allal, and Hamed Ahadi

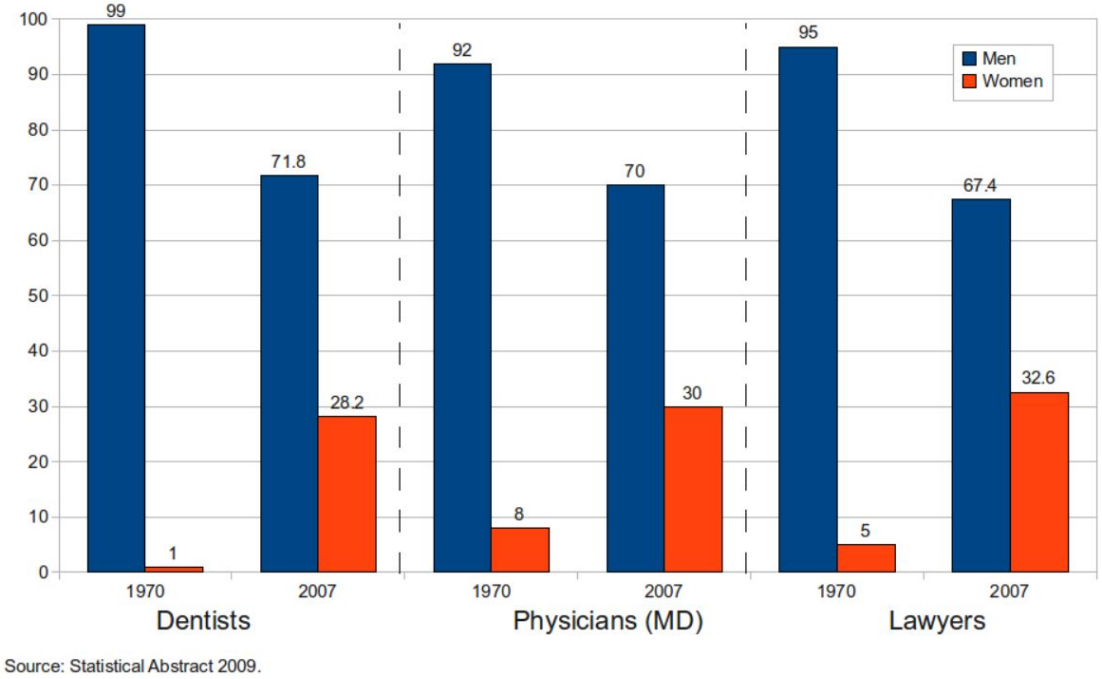
[RE] Fairness Guarantees under Demographic Shift

NeurIPS 2023, Journal Track - ReScience C Vol. 9, Issue 2

UNIVERSITY OF AMSTERDAM

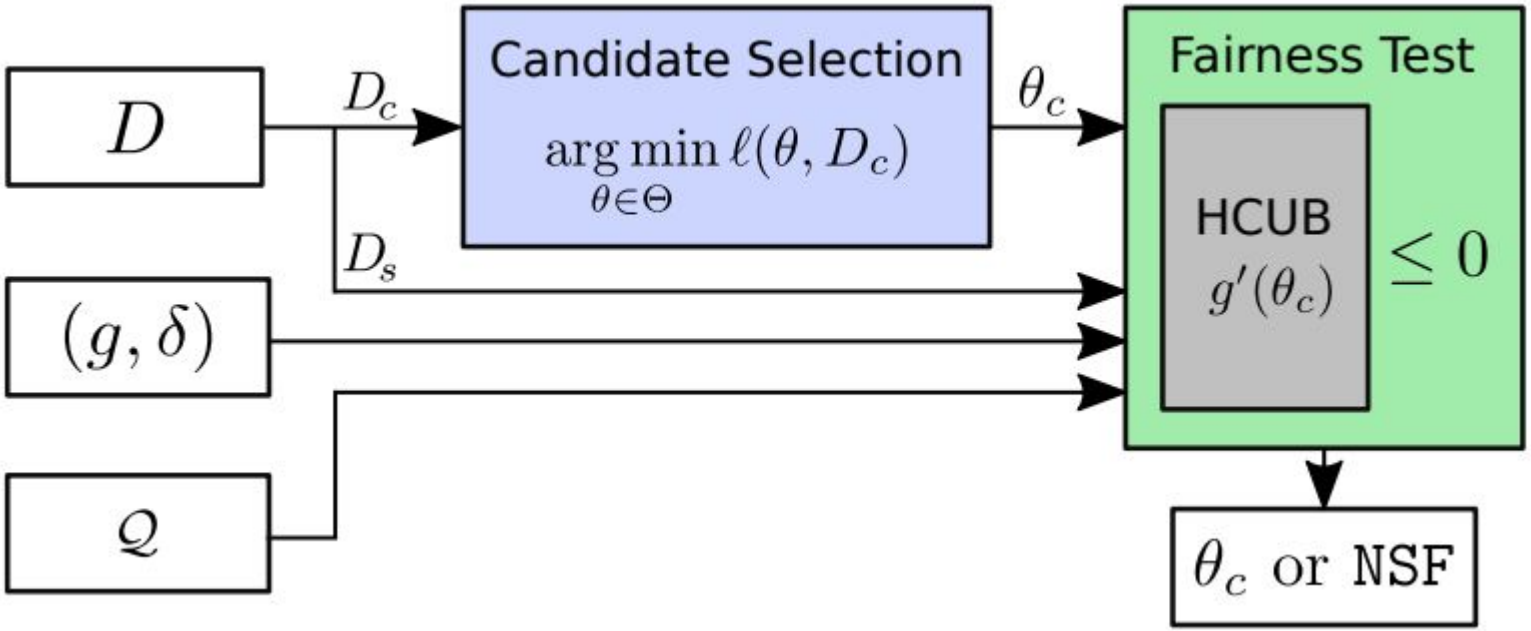
Context

- ML models are frequently biased
- Some models can guarantee to meet given fairness constraints
- However, these guarantees do not hold if demographics shift
- Shifty can guarantee fairness for
  - Known demographic shift
  - Unknown demographic shift

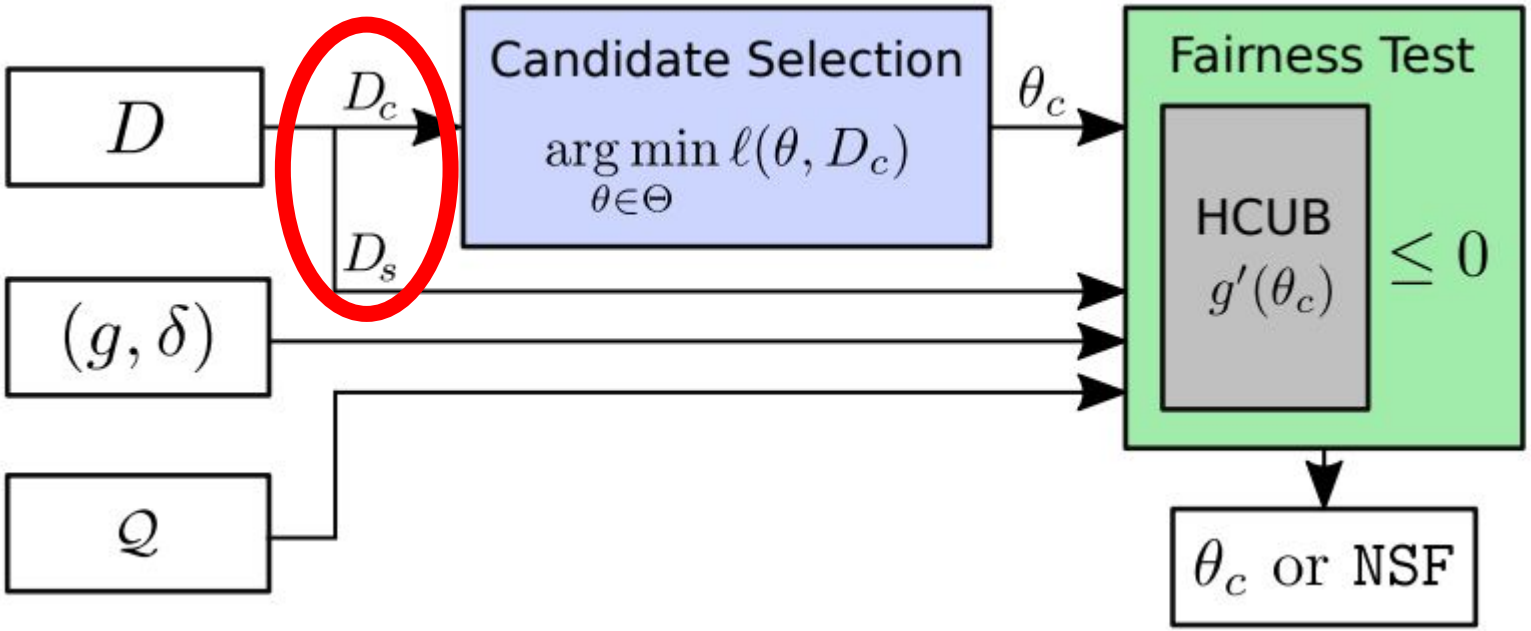


An example of demographic shift over time

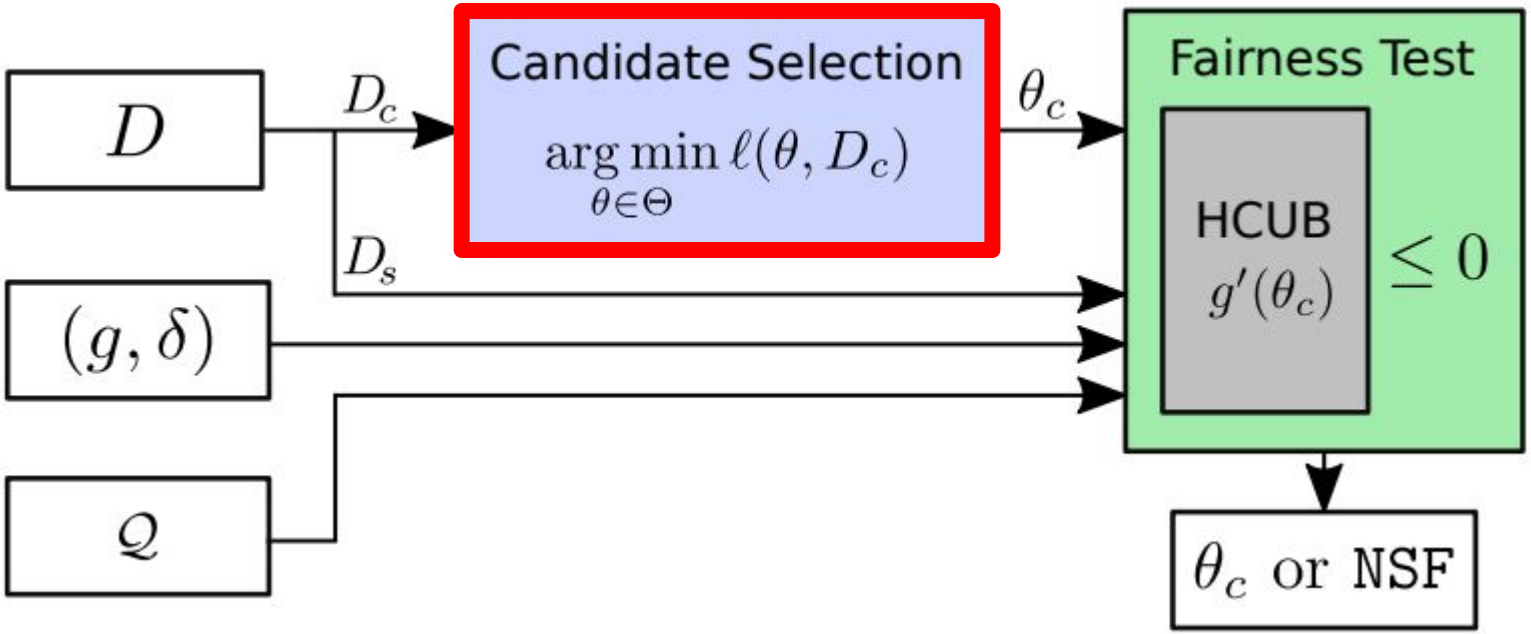
Shifty


D
D_c
Candidate Selection
$\arg\min_{\theta \in \Theta} \ell(\theta, D_c)$
θ_c
Fairness Test
HCUB
$g'(\theta_c)$
≤ 0
D_s
(g, δ)
$\mathcal{Q}$
θ_c or NSF

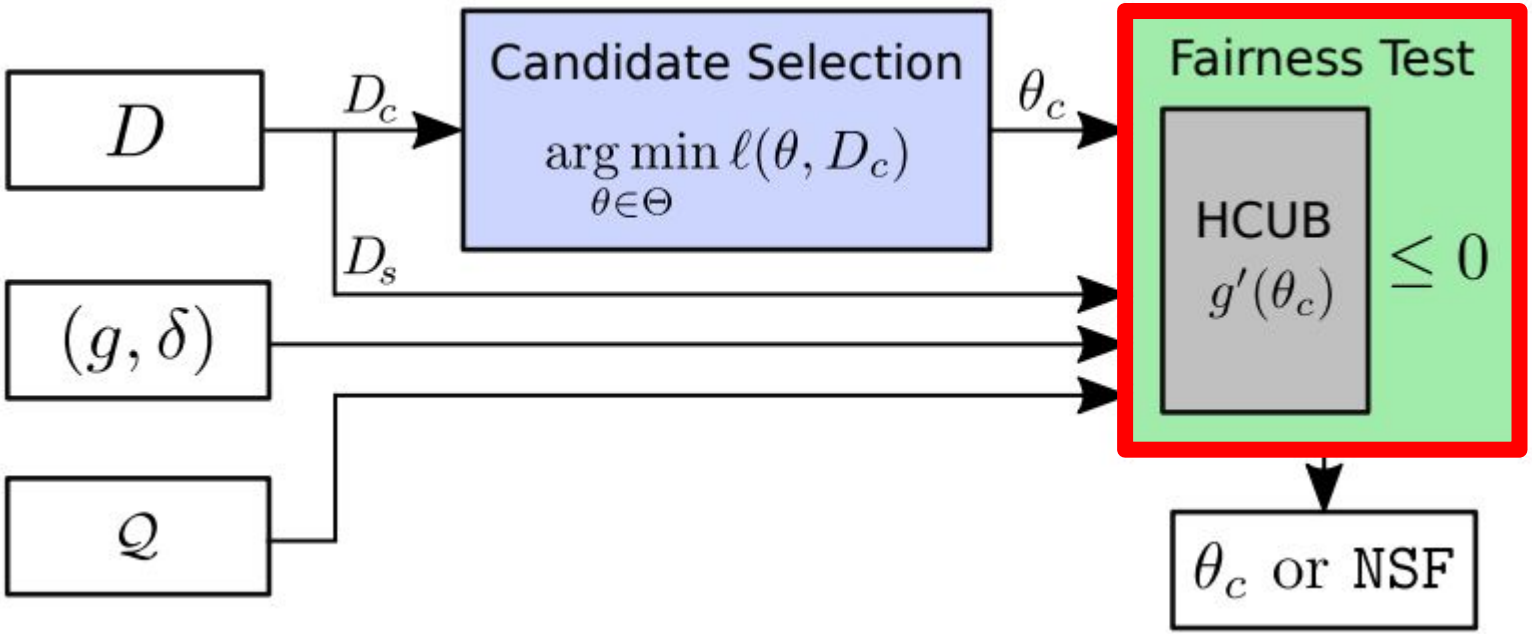
Shifty


D
D_c
D_s
Candidate Selection
$\arg\min_{\theta \in \Theta} \ell(\theta, D_c)$
θ_c
Fairness Test
HCUB
$g'(\theta_c)$
≤ 0
(g, δ)
$\mathcal{Q}$
θ_c or NSF

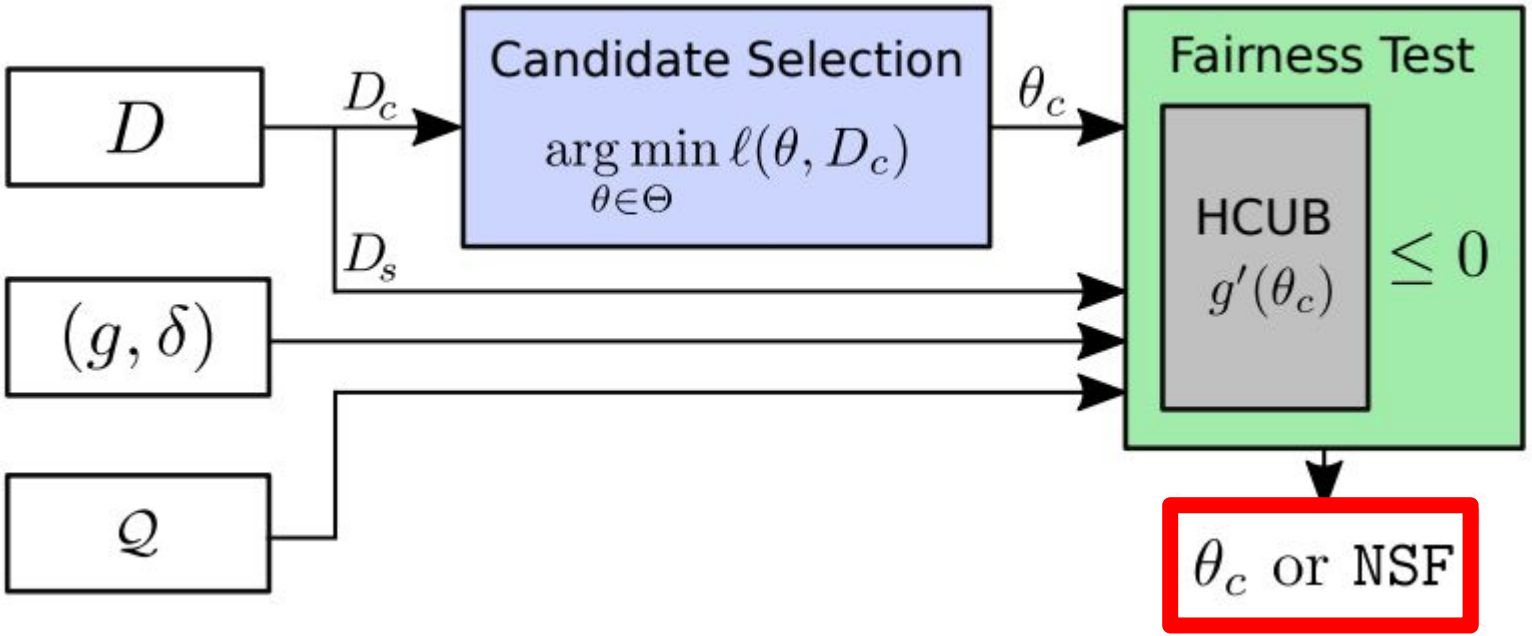
Shifty


D
D_c
Candidate Selection
$\arg\min_{\theta \in \Theta} \ell(\theta, D_c)$
θ_c
Fairness Test
HCUB
$g'(\theta_c)$
≤ 0
D_s
(g, δ)
$\mathcal{Q}$
θ_c or NSF

Shifty


D
D_c
Candidate Selection
$\arg\min_{\theta \in \Theta} \ell(\theta, D_c)$
θ_c
Fairness Test
HCUB
$g'(\theta_c)$
≤ 0
D_s
(g, δ)
$\mathcal{Q}$
θ_c or NSF

Shifty


D
D_c
Candidate Selection
$\arg\min_{\theta \in \Theta} \ell(\theta, D_c)$
θ_c
Fairness Test
HCUB
$g'(\theta_c)$
≤ 0
D_s
(g, δ)
$\mathcal{Q}$
θ_c or NSF

Methodology

- Experiments reproduced with slight adjustments to code provided on GitHub
- Used same datasets (Brazil, Adult) and fairness definitions as original paper
- 80 hours runtime to run the experiments for the original paper

Results for reproducing the paper's claims

- Original paper only includes figures → difficult to compare exact results
- We contribute by providing the raw results

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ*Acc*
FairConst	n/a	0.782	1.000	-0.004	n/a	0.802	1.000	-0.009
RFLearn	n/a	**0.787**	1.000	0.000	n/a	0.823	1.000	0.005
Fairlearn	n/a	0.781	1.000	-0.001	n/a	**0.842**	1.000	0.007
Quasi-SC	0.520	0.762	0.417	0.111	0.600	0.767	0.500	0.139
Shifty	0.720	0.750[1]	**0.000**	0.074	0.400	0.750[2]	**0.000**	0.167
SC	0.680	0.759	**0.000**	0.105	0.500	0.781	**0.000**	0.140

[1] significantly worse than best model, $p < 0.001$, $t = 12.987$, $df = 30$
[2] significantly worse than best model, $p < 0.001$, $t = 32.561$, $df = 24$

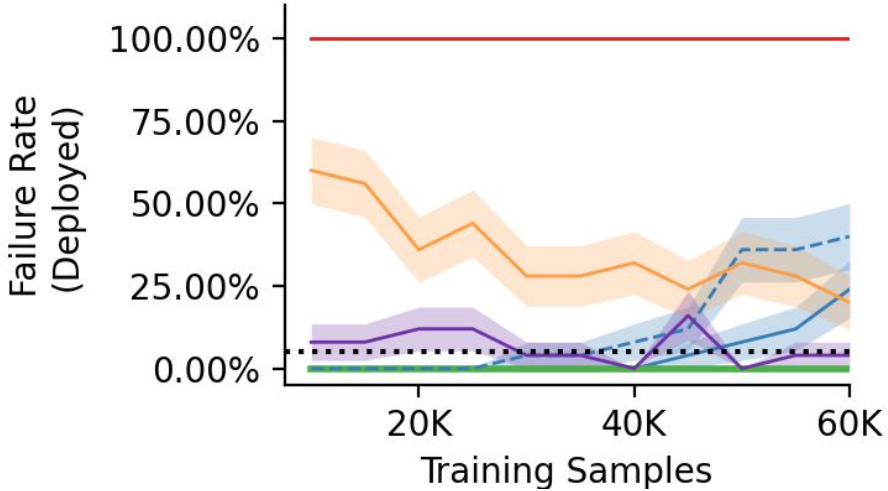
Scope of reproducibility

Main claims of authors:

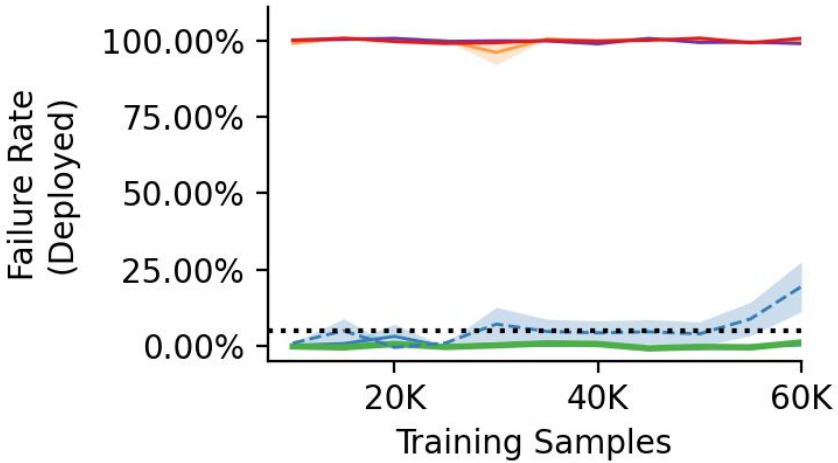
1. *“Shifty provides high-confidence guarantees under demographic shifts”*
2. *“Given sufficient data, no loss in accuracy compared to other models”*
3. *“Returns NSF if fairness constraints not met or too little data”*
4. *“Shifty is model agnostic ”*

Claim 1: “Shifty provides high-confidence guarantees under demographic shifts”

For known bounds:



Brazil dataset

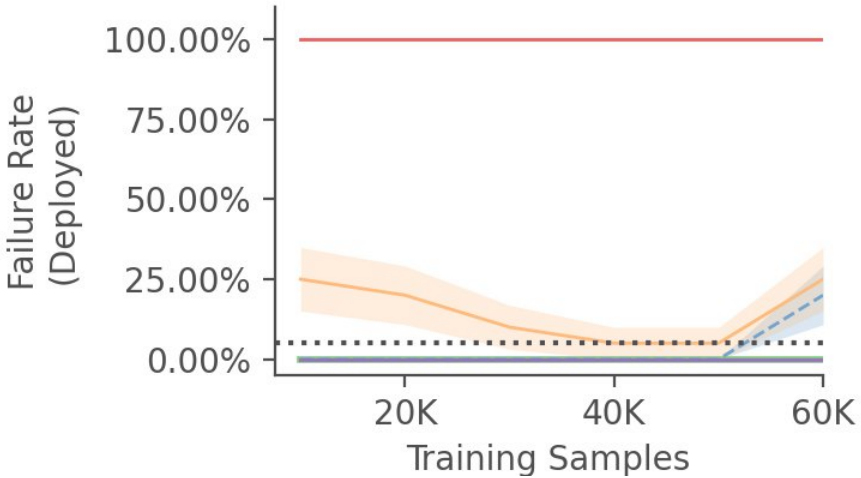


Adult dataset

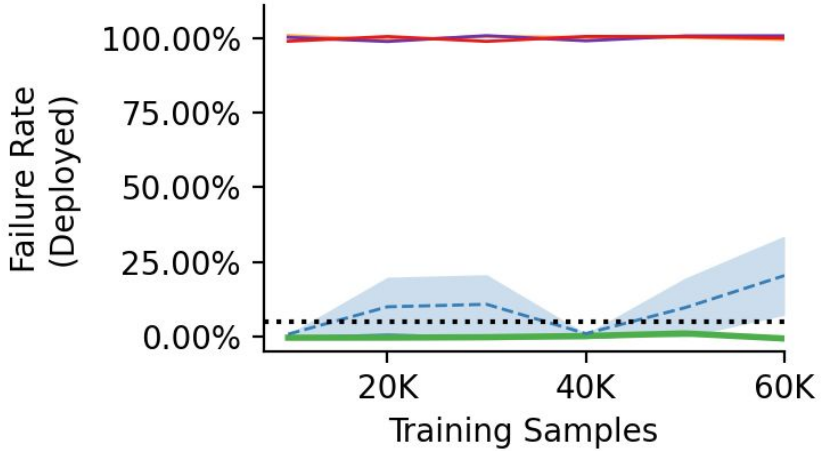
Shifty RFLearn Seldonian Quasi-Seldonian Fairness Constraints Fairlearn

Claim 1: “Shifty provides high-confidence guarantees under demographic shifts”

For unknown bounds:



Brazil dataset

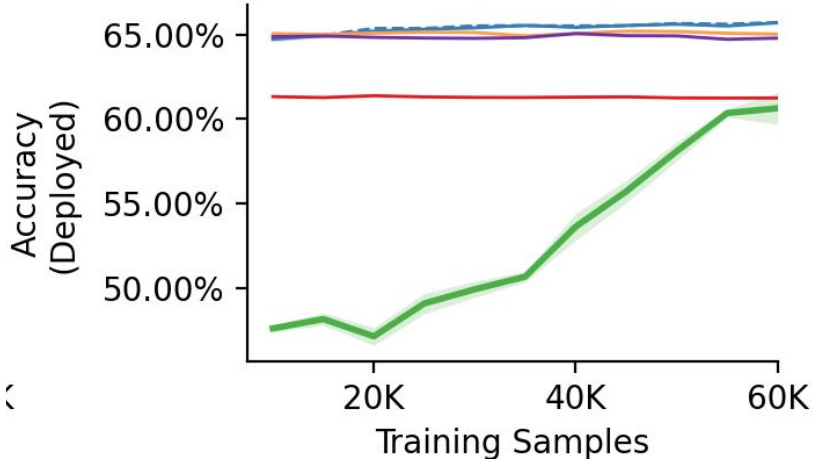


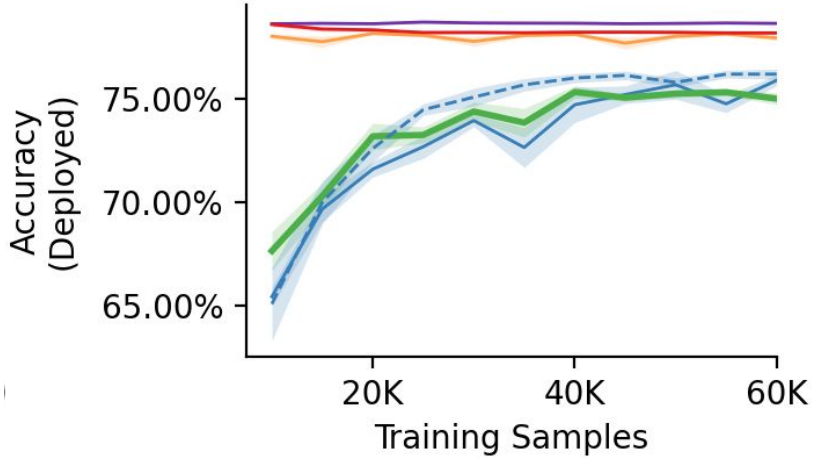
Adult dataset

Shifty RFLearn Seldonian Quasi-Seldonian Fairness Constraints Fairlearn

Claim 2: “Given sufficient data, Shifty shows no loss in accuracy compared to other models”

For known bounds:





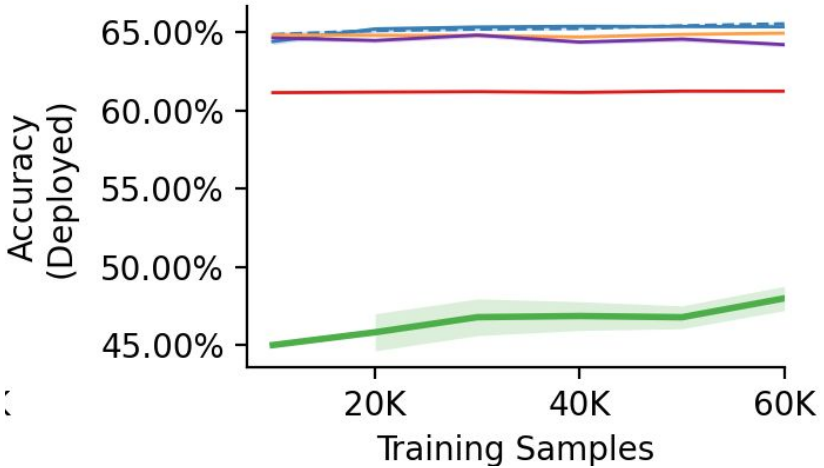
Brazil dataset

Adult dataset

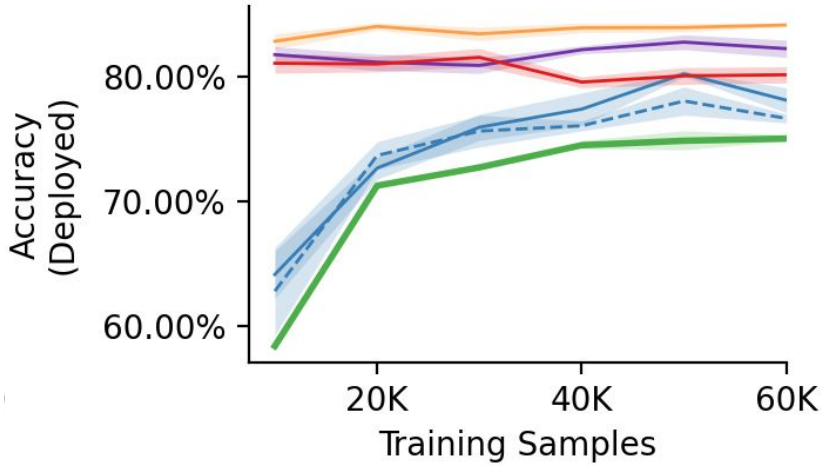
Shifty RFLearn Seldonian Quasi-Seldonian Fairness Constraints Fairlearn

Claim 2: “Given sufficient data, Shifty shows no loss in accuracy compared to other models”

For unknown bounds:



Brazil dataset

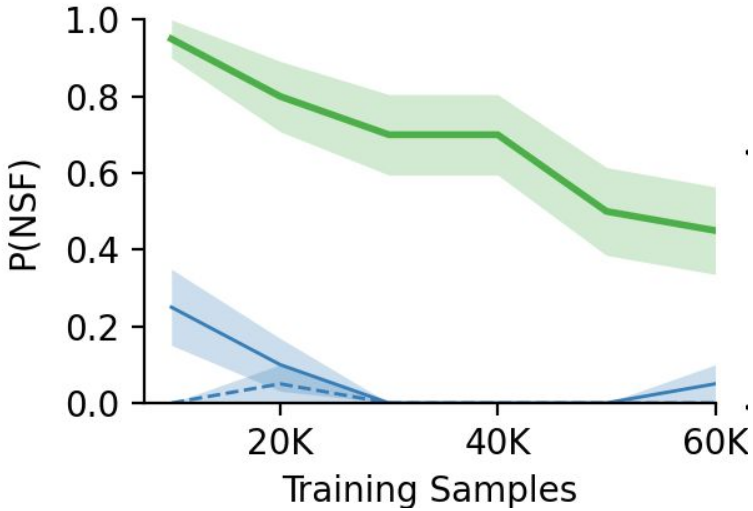


Adult dataset

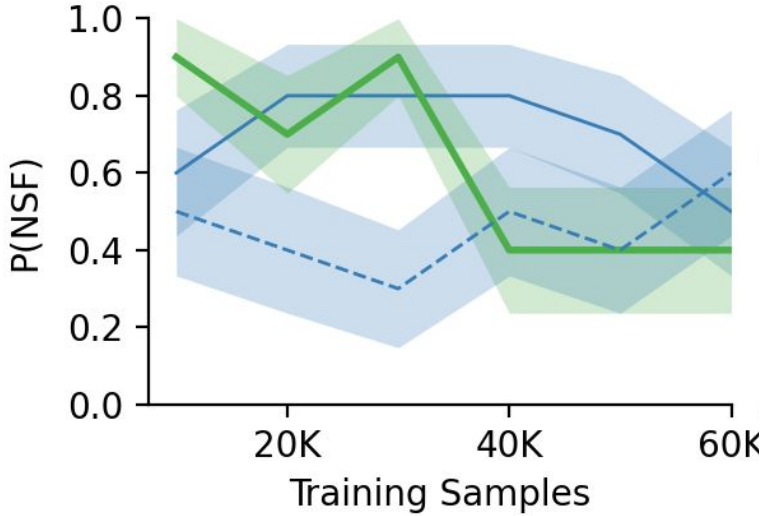
Shifty RFLearn Seldonian Quasi-Seldonian Fairness Constraints Fairlearn

Claim 3: “Shifty returns NSF if fairness constraints not met”

For unknown bounds:



Brazil dataset



Adult dataset

Shifty RFLearn Seldonian Quasi-Seldonian Fairness Constraints Fairlearn

Claim 4: “Shifty is model-agnostic”

- In theory true. In practice… not per se
- Original implementation only applies CMA-ES
  - This avoids backpropagation
  - However, this also requires more time, computational resources, and larger datasets

Space for methodological improvements

Using different classification and optimization methods for Shifty and baselines

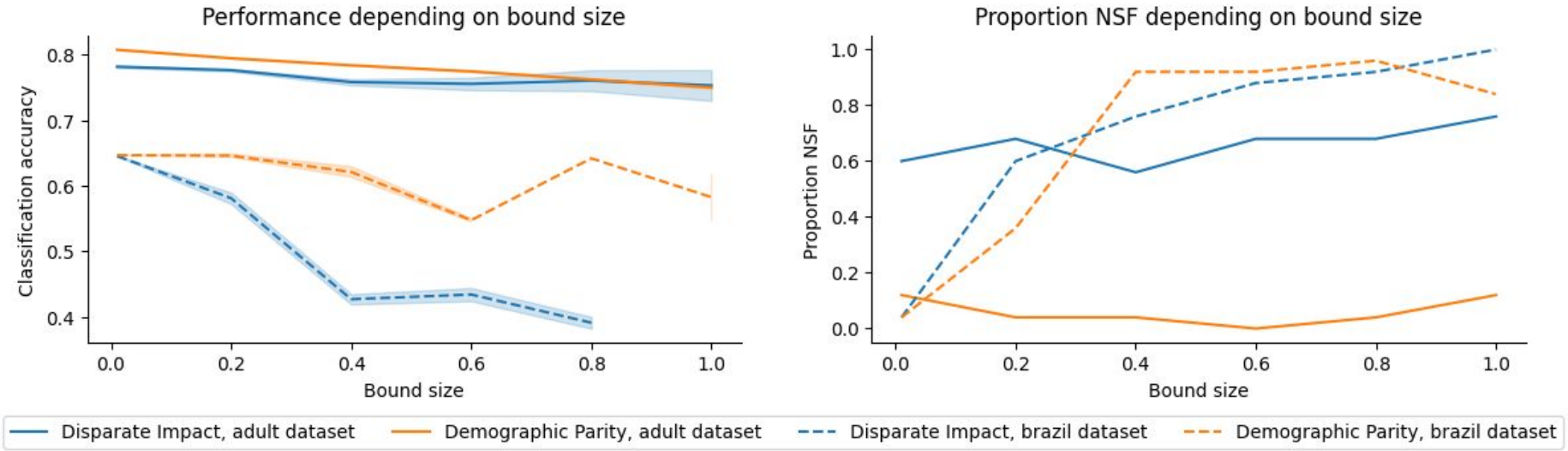
makes comparison difficult

	Classification	Activation	Optimization
Seldonian	1 linear layer, no bias	sign function	SLSQP + CMA-ES
FairConst	1 linear layer, no bias	sign function	SLSQP
Fairlearn	linear SVC[1]	n/a	expgrad[2]
RFLearn	1 linear layer with bias	softmax function	SGD[3]
Shifty	1 linear layer, no bias	sign function	SLSQP + CMA-ES

[1] Support Vector Classifier
[2] exponentiated gradient reduction
[3] Stochastic Gradient Descent

Results beyond the original paper

How does Shifty's performance change with the possible bounds of the demographic shift?



Conclusion

- Results can be approximately reproduced
- Validity of certain claims can be questioned
    - Difficult to compare Shifty and baselines due to differences in classification and optimization methods