Towards Harmless Rawlsian Fairness Regardless of Demographic Prior

Xuanqian Wang[1] Jing Li[2,3 †] Ivor W. Tsang[2,3,4] Yew-Soon Ong[2,3,4]

[1]School of Computer Science and Engineering, Beihang University, China

[2]Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore

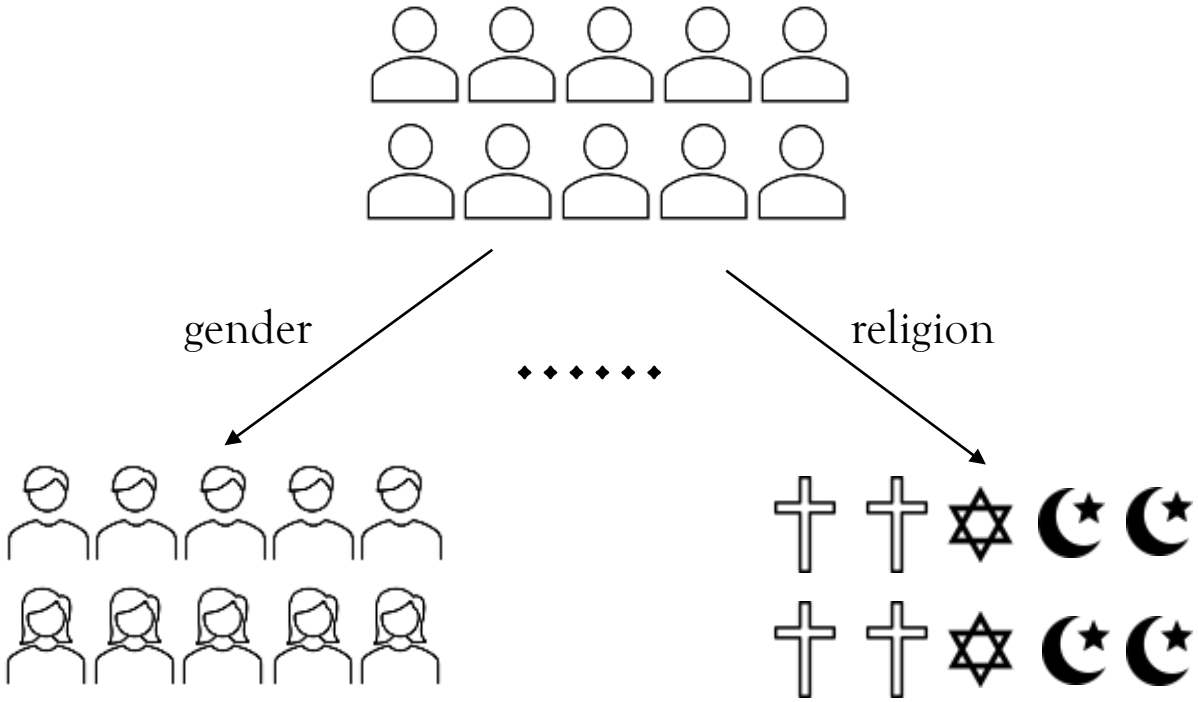
[3]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

[4]College of Computing and Data Science, Nanyang Technological University, Singapore

wwxxqq@buaa.edu.cn {kyle.jingli, ivor.tsang}@gmail.com asysong@ntu.edu.sg

Problem Statement

Rawlsian Fairness without Demographics



- Equal model utility (e.g., accuracy):

  $U(S_1) = U(S_2) = \cdots = U(S_K)$

- Sensitive attributes are private in real cases:

  $S_1, S_2, \ldots, S_K$ *and* K are both unknown

Problem Statement

Given a group ratio bound α, maximizing the worst-off group utility can be approximated:

$$\max_{\theta} \min_{k} U(S_k) \Rightarrow \min_{\theta} \sum_{i}^{N\alpha} \ell(x^{(i)}, \theta)$$

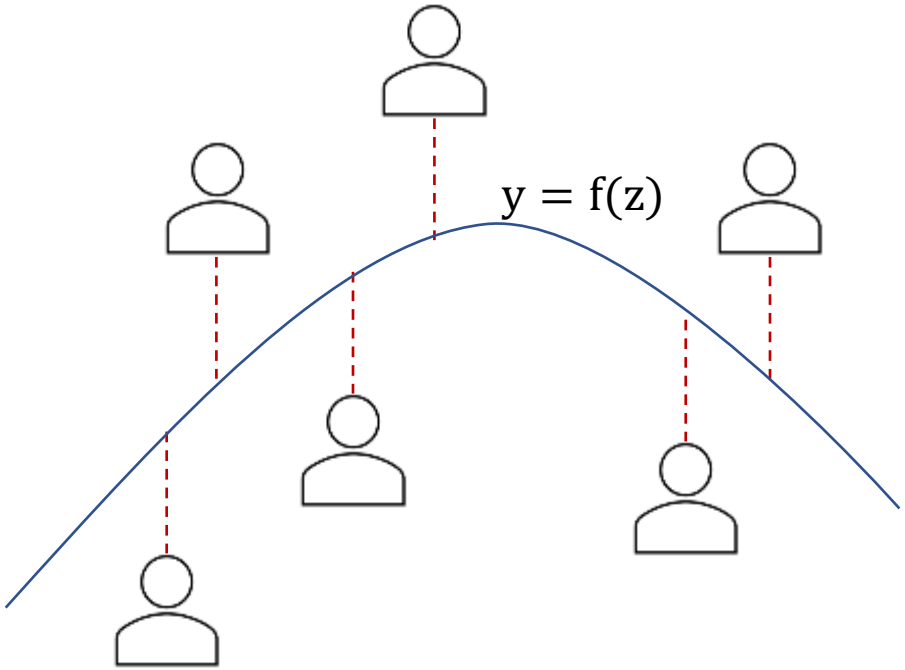
If no demographic prior, we have the same setup with the standard training. We then ask:

To what extent can we improve fairness without hurting the model's overall utility?

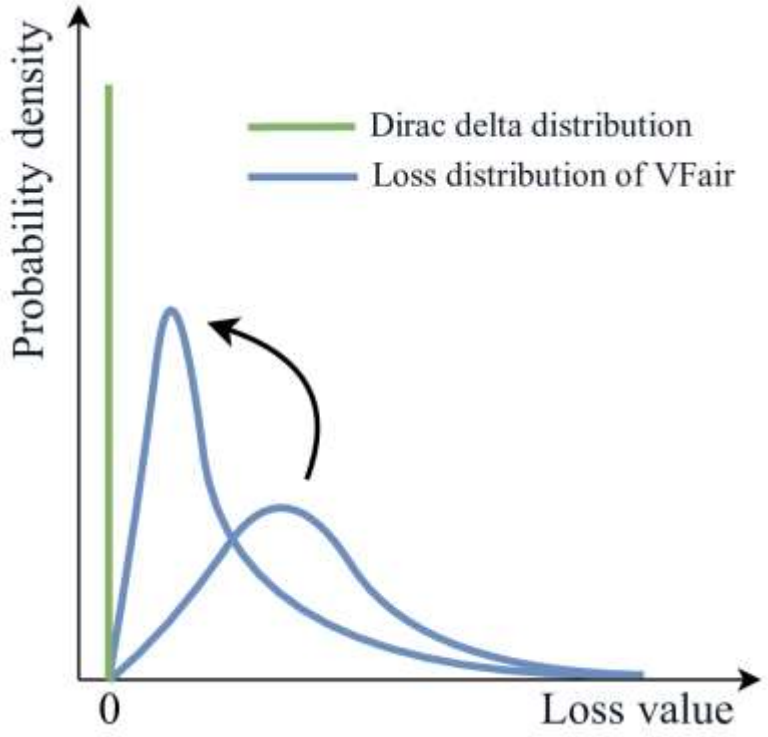
Object Formulation

Proposition 1: *$u \perp s$ holds for any s that splits data into a number of groups, if and only if the loss ℓ is (approximately) independent of the training example z, i.e., $\ell \perp z$.*

Fair regression example



Loss distribution perspective



Object Formulation

Penalty options:

$$\hat{\pi} = \sum_{i=1}^{N-1} |\ell_i - \ell_{i+1}| \qquad \textit{Pairwise}$$

$$\hat{\sigma} = \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^{N} (\ell_i - \hat{\mu})^2} \qquad \underline{\textit{Standard deviation}}$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (\ell_i - \hat{\mu})^2 \qquad \textit{Variance}$$

VFair objective:

$$\min_{\theta} \underbrace{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (\ell(z_i;\theta) - \hat{\mu}(\theta))^2}}_{\hat{\sigma}(\theta)} \quad s.t. \underbrace{\frac{1}{N} \sum_{i=1}^{N} \ell(z_i;\theta)}_{\hat{\mu}(\theta)} \leq \delta$$

- Minimizing both the first and second moment of loss distribution
- δ controls the tolerant harm
- Different motivation from variance reduction [1]

[1] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. Advances in neural information processing systems, 2017.

Optimization

VFair objective:

$$\min_{\theta} \underbrace{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ell(z_i;\theta)-\hat{\mu}(\theta))^2}}_{\hat{\sigma}(\theta)} \quad s.t. \underbrace{\frac{1}{N}\sum_{i=1}^{N}\ell(z_i;\theta)}_{\hat{\mu}(\theta)} \leq \delta$$

Dynamic update scheme:

$$\theta^{t+1} \leftarrow \theta^t - \gamma^t\left(\lambda^t \nabla\hat{\mu}(\theta^t) + \nabla\hat{\sigma}(\theta^t)\right)$$

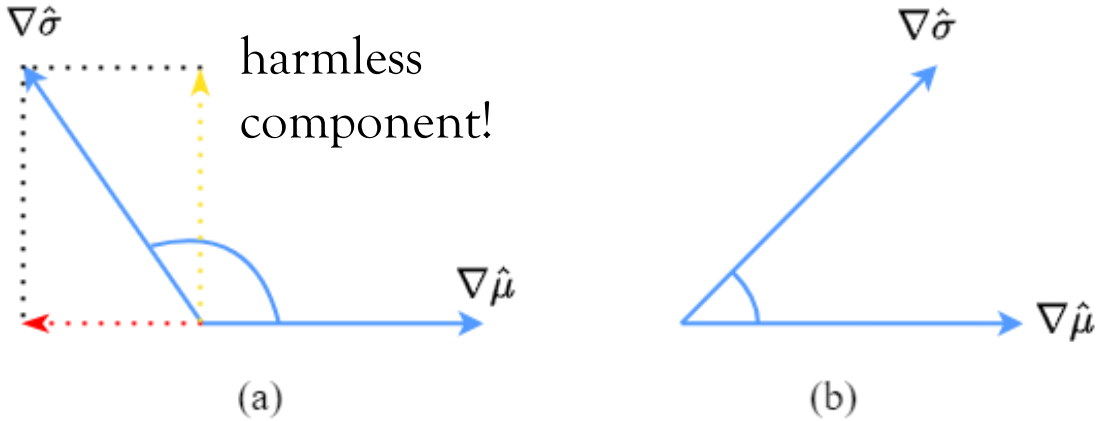
$$\lambda = \max(\lambda_1, \lambda_2) = \max\left(1 - \frac{\nabla\hat{\mu}\cdot\nabla\hat{\sigma}}{\|\nabla\hat{\mu}\|^2}, \frac{\hat{\mu}}{\hat{\sigma}}\right)$$

Gradient view **Loss view**

Optimization

$$\lambda = \max(\lambda_1, \lambda_2) = \max\left(1 - \frac{\nabla\hat{\mu} \cdot \nabla\hat{\sigma}}{\|\nabla\hat{\mu}\|^2}, \frac{\hat{\mu}}{\hat{\sigma}}\right)$$

- **Gradient view**



$$\lambda\nabla\hat{\mu} + \text{Proj}_{\nabla\hat{\mu}}(\nabla\hat{\sigma}) \geq \nabla\hat{\mu} \Rightarrow \lambda \geq 1 - \frac{\nabla\hat{\mu} \cdot \nabla\hat{\sigma}}{\|\nabla\hat{\mu}\|^2} := \lambda_1$$

- **Loss view**

Theorem 1: *The combined gradient derived by the dynamic update scheme can be expressed with an example-reweighting form,*

$$\nabla = \lambda\nabla\hat{\mu} + \nabla\hat{\sigma} = \frac{1}{N}\sum_{i=1}^{N}\underbrace{\left(\lambda + \frac{1}{\hat{\sigma}}(\ell_i - \hat{\mu})\right)}_{w_i(\theta)}\frac{\partial \ell_i}{\partial\theta}$$

$$\forall i \in [N] \;\; \lambda + \frac{1}{\hat{\sigma}}(\ell_i - \hat{\mu}) \geq 0 \Rightarrow$$

$$\lambda \geq \max_{i\in[N]} \frac{\hat{\mu} - \ell_i}{\hat{\sigma}} - \frac{1}{\hat{\sigma}}\left(\hat{\mu} - \min_{i\in[N]} \ell_i\right) \geq \frac{\hat{\mu}}{\hat{\sigma}} := \lambda_2$$

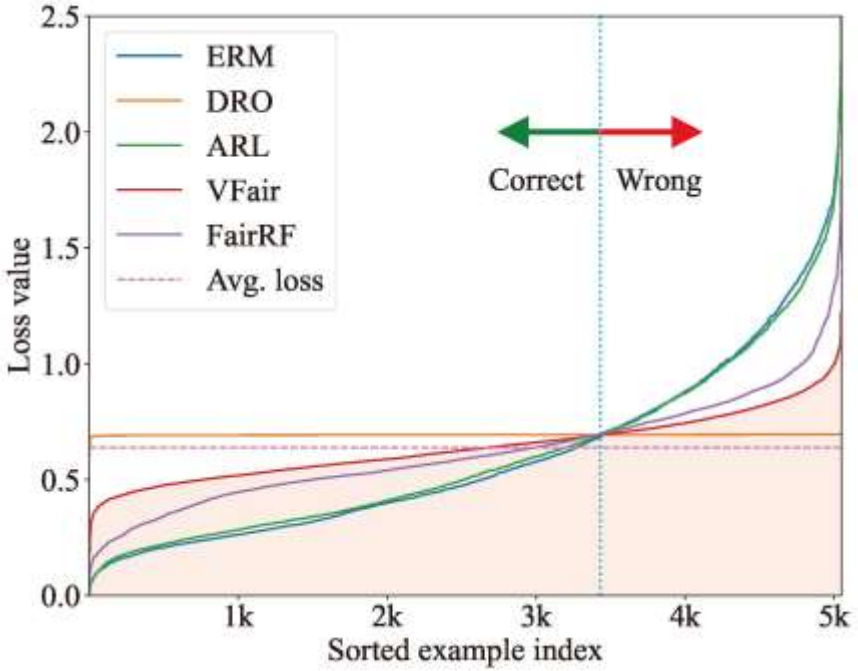
Experiments

		Utility↓	WU↓	MUD↓	TUD↓	VAR↓
Law School	ERM	12.88(0.12)	19.75(0.21)	7.33(0.14)	13.45(0.21)	4.89(0.07)
	DRO	24.85(0.09)	24.98(0.05)	0.14(0.08)	0.23(0.12)	0
	ARL	**12.86**(0.11)	19.72(0.26)	7.33(0.19)	13.54(0.29)	4.86(0.13)
	BPF	18.75(0.50)	43.25(3.44)	26.33(3.20)	47.33(5.16)	3.98(0.29)
	MPFR	13.88	29.39	16.68	32.07	7.15
	FKL	13.1	19.37	6.77	13.42	5.01
	VFair(Ours)	12.95(0.11)	**19.08**(0.22)	**6.63**(0.18)	**12.53**(0.25)	**3.66**(0.12)
	Improved	**-0.07**	**+0.67**	**+0.7**	**+0.92**	**+1.23**
COMPAS	ERM	23.08(0.67)	24.49(0.76)	2.50(1.17)	3.45(1.76)	3.23(0.8)
	DRO	24.97(0.04)	25.05(0.06)	0.12(0.07)	0.17(0.10)	0
	ARL	**22.73**(0.4)	24.26(0.84)	2.92(1.08)	3.78(1.11)	3.19(0.67)
	BPF	50.80(2.18)	63.46(0.99)	22.87(1.69)	37.77(1.84)	11.18(1.11)
	MPFR	36.26	38.36	6.23	9.13	17.33
	FKL	28.56	30.49	3.69	6.47	7.58
	VFair(Ours)	23.15(0.13)	**23.83**(0.21)	**0.93**(0.21)	**1.17**(0.28)	**0.47**(0.07)
	Improved	**-0.07**	**+0.66**	**+1.57**	**+2.28**	**+2.76**
C & C	ERM	41.15(1.25)	109.72(5.60)	106.56(5.67)	337.26(18.15)	87.52(9.43)
	DRO	99.34(3.85)	257.51(40.23)	248.56(48.63)	715.62(189.66)	284.49(72.71)
	ARL	**40.43**(1.14)	109.00(6.06)	106.88(5.70)	331.38(19.63)	83.98(5.55)
	BPF	71.05(1.02)	127.28(4.78)	110.16(10.06)	320.65(26.11)	96.76(8.08)
	MPFR	93.57	296.36	295.59	843.47	375.79
	FKL	83.73	278.3	275.29	794.59	321.42
	VFair(Ours)	41.17(0.64)	**106.40**(2.66)	**104.54**(3.11)	**318.33**(8.96)	**67.44**(3.36)
	Improved	**-0.02**	**+3.32**	**+2.02**	**+18.93**	**+20.08**
AgeDB	ERM	4.25(0.49)	4.32(0.53)	0.15(0.12)	0.15(0.12)	0.57(0.26)
	DRO	17.72(22.59)	17.98(22.65)	0.5(0.37)	0.5(0.37)	5.76(7.67)
	ARL	5.11(1.76)	5.29(1.98)	0.36(0.41)	0.36(0.41)	2.51(3.23)
	VFair(Ours)	**3.57**(0.76)	**3.63**(0.77)	**0.12**(0.09)	**0.12**(0.09)	**0.23**(0.08)
	Improved	**+0.68**	**+0.69**	**+0.03**	**+0.03**	**+0.34**

Harmless fairness can be achieved in regression tasks regardless of any demographic information!

Experiments

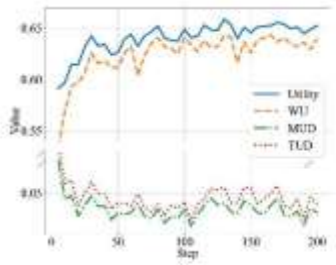
In the context of classification, our VFair achieves limited improvement in terms of fairness metrics.



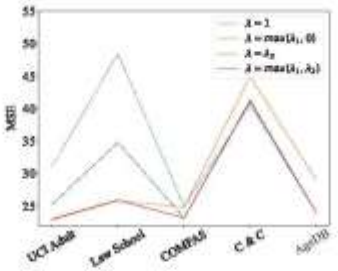
Classification (binary) results are insensitive to the loss values, which is dependent on boundary.

		K=4				K=10				K=20			
		Utility	WU	MUD	TUD	Utility	WU	MUD	TUD	Utility	WU	MUD	TUD
UCI Adult	ERM	2.5	2.5	2.31	2.36	2.5	2.51	2.4	2.46	2.5	2.27	2.44	2.39
	ARL	2.5	2.41	2.48	2.51	2.5	2.35	2.6	2.54	2.5	2.23	2.56	2.61
	VFair	1	1.09	1.21	1.13	1	1.14	1	1	1	1.5	1	1
Law School	ERM	2.7	2.66	2.37	2.36	2.7	2.61	2.52	2.54	2.7	2.52	2.5	2.53
	ARL	2.3	2.34	2.36	2.37	2.3	2.39	2.48	2.46	2.3	2.31	2.5	2.47
	VFair	2.7	1	1.27	1.27	1	1	1	1	1	1.17	1	1
COMPAS	ERM	2.5	2.21	2.56	2.57	2.5	1.89	2.57	2.56	2.5	1.53	2.58	2.62
	ARL	2.5	2.44	2.43	2.42	2.5	1.98	2.43	2.44	2.5	1.56	2.42	2.38
	VFair	1	1.35	1.01	1.01	1	2.13	1	1	1	2.91	1	1

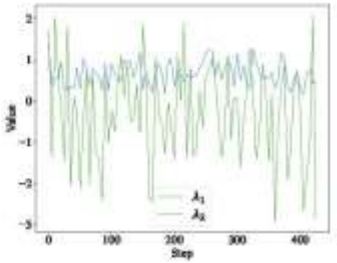
Thank you!
More details to explore...



(a) Training results on COMPAS in the classification task.



(b) MSE on five regression datasets. Lower is better.



(c) The curves of λ_1 and λ_2 during training on C & C.

	UCI Adult				CelebA			
	Utility↑	WU ↑	MUD ↓	TUD ↓	Utility↑	WU ↑	MUD ↓	TUD ↓
ERM	75.02	72.17	6.87	8.88	91.40	70.17	19.39	22.82
DRO	36.27	16.06	23.59	41.17	77.52	74.29	**3.9**	**4.78**
ARL	74.90	71.85	7.32	9.49	91.60	70.39	20.14	24.33
VFair	**75.98**	**72.74**	**5.82**	**7.40**	**91.91**	**75.70**	14.39	18.50

	Objective	Utility ↑	WU↑	MUD↓	TUD↓	VAR↓
UCI Adult	$\hat{\pi}$	82.98	78.35	16.19	21.10	**0**
	$\hat{\sigma}^2$	84.70	80.34	15.72	20.79	7.18
	$\hat{\sigma}$	**84.74**	**80.36**	**15.71**	**20.71**	8.17
Law School	$\hat{\pi}$	84.05	72.96	11.92	22.51	**0.03**
	$\hat{\sigma}^2$	85.33	74.60	11.67	20.91	6.91
	$\hat{\sigma}$	**85.40**	**74.81**	**11.24**	**20.31**	19.35
COMPAS	$\hat{\pi}$	55.78	51.60	8.70	12.24	**0**
	$\hat{\sigma}^2$	63.45	59.14	8.71	11.36	0.04
	$\hat{\sigma}$	**66.80**	**63.86**	**6.25**	**8.47**	1.86
CelebA	$\hat{\pi}$	44.88	20.16	49.16	53.85	**0**
	$\hat{\sigma}^2$	92.45	89.53	3.44	4.63	14.4
	$\hat{\sigma}$	**93.43**	**91.09**	**2.73**	**3.85**	11.7

Algorithm 1 Harmless Rawlsian Fairness without Demographics via VFair.

Input: Training set $\mathcal{D} = \{z_i\}_{i=1}^N$, where $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
Output: Learned model parameterized by $\theta \in \Theta$

Initialize parameters θ
Initialize $\hat{\mu}^0 \leftarrow 0$
for epoch $\leftarrow 1$ **to** N_{epochs} **do**
  for mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    Compute the losses $\{\ell_i\}_{i=1}^b$
    Update $\hat{\mu}^t$ as in Eq. 19
    Update $\hat{\sigma} \leftarrow \sqrt{\frac{1}{b}\sum_{i=1}^b (\ell_i - \hat{\mu}^t)^2}$
    Compute primary gradient $\nabla\hat{\mu}$
    Compute secondary gradient $\nabla\hat{\sigma}$
    Compute λ_1 as in Eq. 6
    Compute λ_2 as in Eq. 8
    Compute dynamic λ^t as in Eq. 9
    Update parameters θ as in Eq. 5
  end for
end for