Data Analytics at Texas A&M (DATA) Lab

Chasing Fairness under Distribution Shift: a Model Weight Perturbation Approach

Zhimeng Jiang*[1], Xiaotian Han*[1], Hongye Jin[1], Guanchu Wang[2], Rui Chen[3], Na Zou[1], Xia Hu[2]

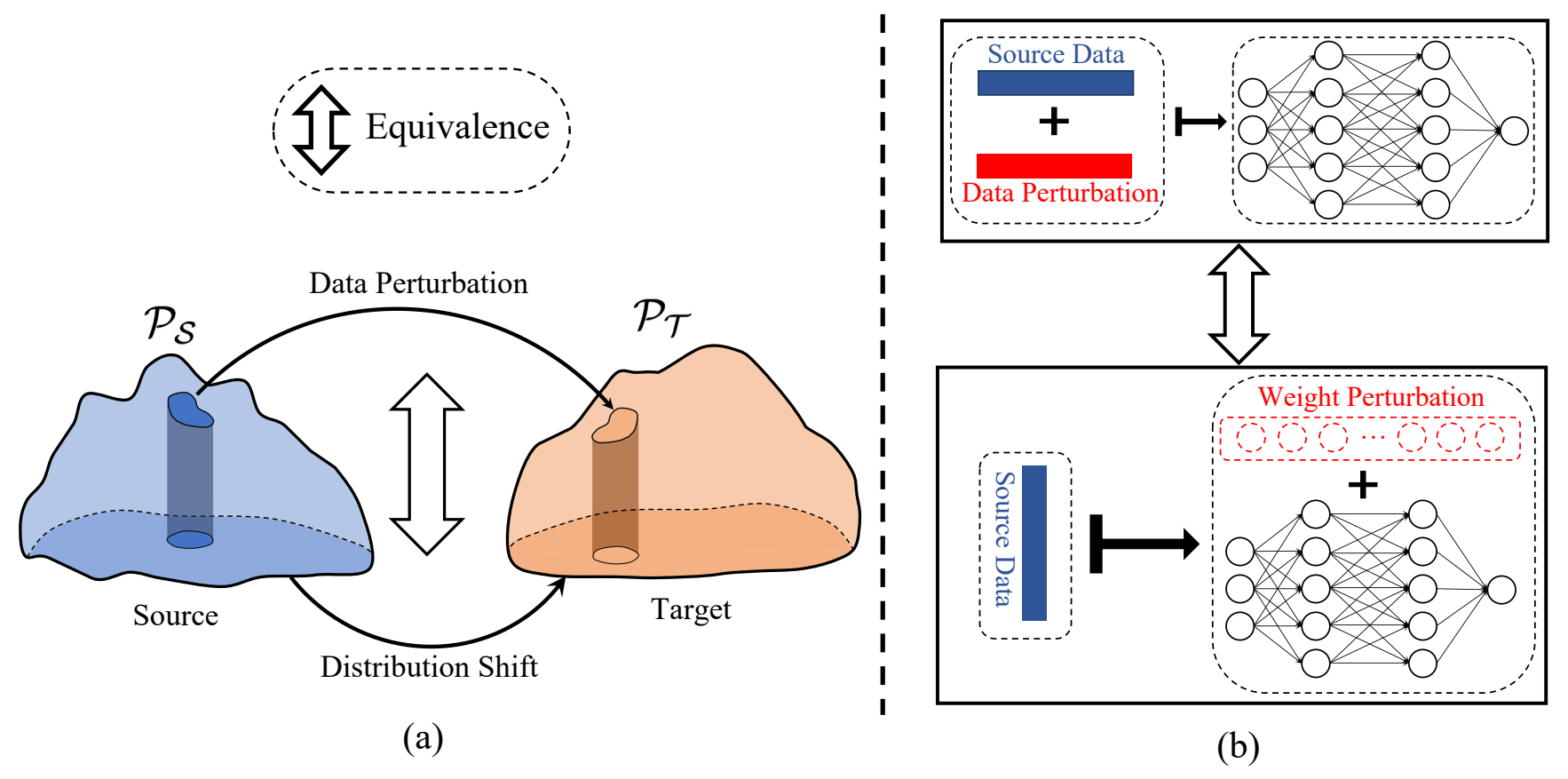
*Equal contribution, [1]Texas A&M University, [2]Rice University, [3]Samsung Electronics America

Research Motivation

- The fairness performance of machine learning model experiences performance degradation under the distribution shifts problem.
- The existing fairness methods mainly focus on algorithmic fairness for in-distribution data.
- **Key question:** Can we design an effective method chasing fairness under distribution shift? If yes, How?

Understanding Distribution Shift

We first theoretically demonstrate the inherent connection between distribution shift, data perturbation, and model weight perturbation. The left part demonstrates distribution shift can be transformed as data perturbation, while the right part shows that data perturbation and model weight perturbation are equivalent.



- **Distribution Shift is Data Perturbation**
  - **Statement:** There exists data perturbation δ so that the training loss of any neural network $f_\theta(\cdot)$ for target distribution equals that for source distribution with data perturbation δ, i.e.,

$$\mathbb{E}_{(X,Y)\sim\mathcal{P}_T}[l(f_\theta(X),Y)] = \mathbb{E}_{\delta_X(X),\delta_Y(Y)}\mathbb{E}_{(X,Y)\sim\mathcal{P}_S}[l(f_\theta(X+\delta_X(X)),Y+\delta_Y(Y))]. \quad (1)$$

  - **Insight:** For the model trained with loss minimization on source data, the deteriorated performance on the target dataset stems from the perturbation of features and labels.
- **Data Perturbation Equals Model Weight Perturbation**
  - **Statement:** For general case, there exists model weight perturbation $\Delta\theta$ so that the training loss on perturbed source dataset is the same with that for model weight perturbation $\Delta\theta$ on source distribution:

$$\mathbb{E}_{\delta_X(X),\delta_Y(Y)}\mathbb{E}_{(X,Y)\sim\mathcal{P}_S}[l(f_\theta(X+\delta_X(X)),Y+\delta_Y(Y))] = \mathbb{E}_{(X,Y)\sim\mathcal{P}_S}[l(f_{\theta+\Delta\theta}(X),Y)]. \quad (2)$$

  - **Insight:** Chasing a robust model over data perturbation can be achieved via model weight perturbation, i.e., finding a "flattened" local minimum in terms of the target objective.

Robust Fairness Regularization

- **Two conditions for robust fairness under distribution shift:** (1) Low demographic parity on source dataset; (2) Low average prediction gap for each demographic group.
- **Proposed Regularization:** The overall objective of RFR is given by $\mathcal{L}_{all} = \mathcal{L}_{CLF} + \lambda \cdot (\mathcal{L}_{DP} + \mathcal{L}_{RFR})$, where

$$\mathcal{L}_{RFR} = \max_{\|\epsilon_0\|_p \le \rho} \mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] + \max_{\|\epsilon_1\|_p \le \rho} \mathbb{E}_{\mathcal{S}_1}[f_{\theta+\epsilon_1}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})] \quad (3)$$

Experiment Results

- **Synthetic Distribution Shift.** RFR outperforms several baselines in terms of fairness and tradeoff performance across different distribution shift intensities.

(α,β)	Methods	Adult			ACS-I			ACS-E		
		Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
(1.0, 2.0)	MLP	82.09±0.05	15.11±0.04	14.33±0.05	77.95±0.52	3.51±0.59	3.77±0.55	80.95±0.10	1.10±0.06	1.43±0.06
	REG	80.60±0.05	3.79±0.06	3.27±0.08	77.77±0.09	2.28±0.32	2.59±0.23	80.44±0.07	0.86±0.09	1.05±0.10
	ADV	78.80±0.68	0.83±0.26	0.79±0.14	75.72±0.63	1.96±0.38	2.00±0.35	79.39±0.15	1.09±0.26	0.95±0.26
	FCR	79.06±0.09	9.98±0.06	9.47±0.07	76.99±0.47	2.94±0.34	2.95±0.28	79.74±0.11	0.97±0.21	1.00±0.22
	RFR	78.84±0.09	**0.44**±0.05	**0.12**±0.06	74.15±0.81	**1.84**±0.27	**1.60**±0.33	80.08±0.08	**0.71**±0.10	**0.06**±0.11
(1.5, 3.0)	MLP	82.05±0.05	15.16±0.09	14.33±0.09	77.85±0.25	3.73±0.53	3.70±0.56	80.42±0.10	1.14±0.07	1.10±0.07
	REG	80.64±0.08	3.74±0.11	3.23±0.10	77.87±0.18	2.25±0.28	2.37±0.27	80.21±0.13	**0.72**±0.04	0.75±0.03
	ADV	78.71±0.41	1.07±0.87	0.87±0.96	75.79±0.68	2.22±0.53	2.44±0.48	79.58±0.13	1.07±0.19	1.26±0.18
	FCR	79.05±0.12	10.01±0.07	9.51±0.06	77.06±0.68	3.39±0.33	3.10±0.36	79.59±0.26	1.17±0.24	1.08±0.23
	RFR	78.91±0.03	**0.46**±0.10	**0.16**±0.09	74.19±0.58	**1.82**±0.29	**2.17**±0.32	80.47±0.03	**0.72**±0.04	**0.71**±0.05
(3.0, 6.0)	MLP	82.07±0.05	15.23±0.14	14.45±0.15	77.89±0.45	3.35±0.36	3.47±0.41	80.30±0.04	1.17±0.04	1.13±0.04
	REG	80.62±0.07	3.72±0.05	3.21±0.04	78.19±0.12	**1.60**±0.48	**1.84**±0.44	80.36±0.09	**0.70**±0.09	0.68±0.11
	ADV	78.97±0.49	**1.28**±0.74	**1.09**±0.50	75.71±0.68	2.28±0.39	2.24±0.41	79.66±0.16	1.34±0.14	1.16±0.13
	FCR	79.03±0.13	10.00±0.05	9.50±0.05	76.71±0.39	2.97±0.34	3.28±0.31	79.89±0.22	1.06±0.14	1.14±0.18
	RFR	80.15±0.07	1.75±0.15	1.30±0.14	74.22±0.56	1.80±0.26	1.89±0.24	80.28±0.12	0.74±0.04	**0.51**±0.04

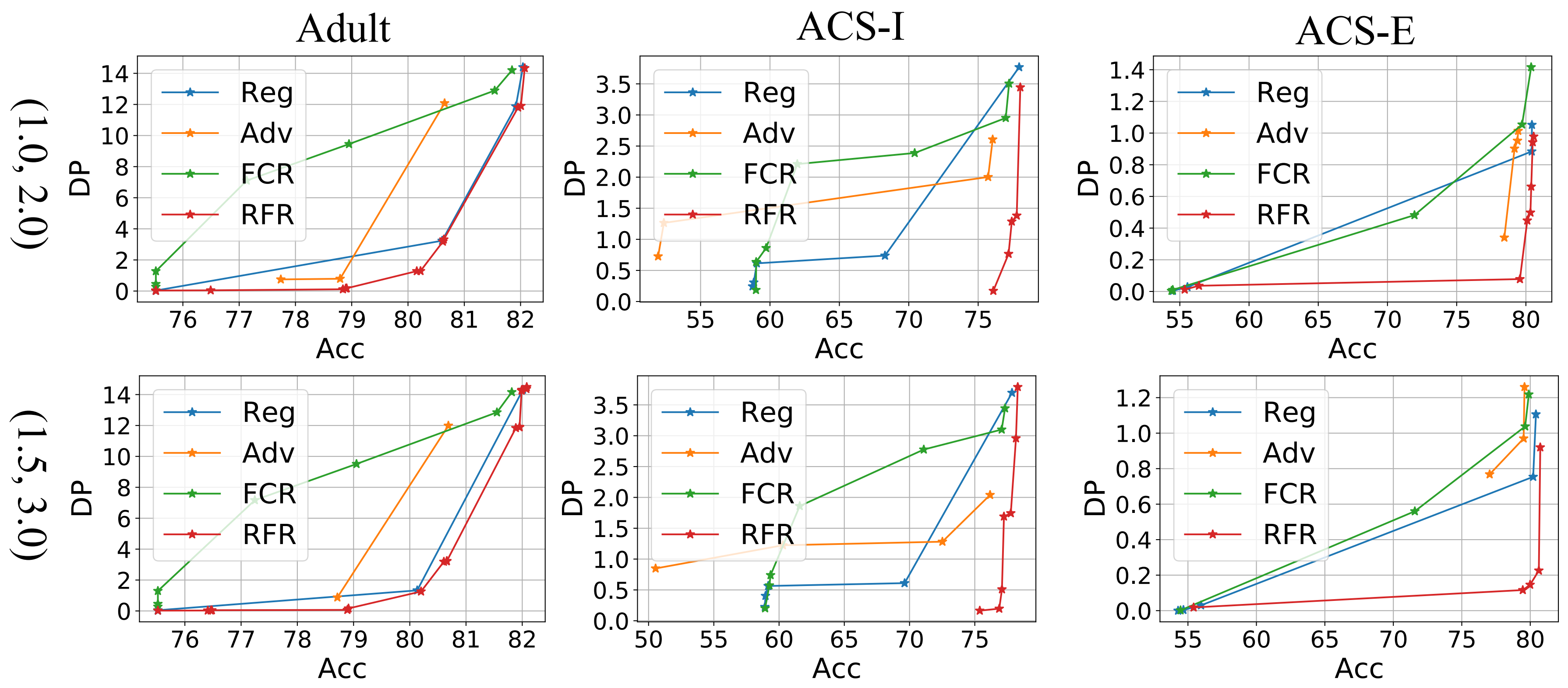
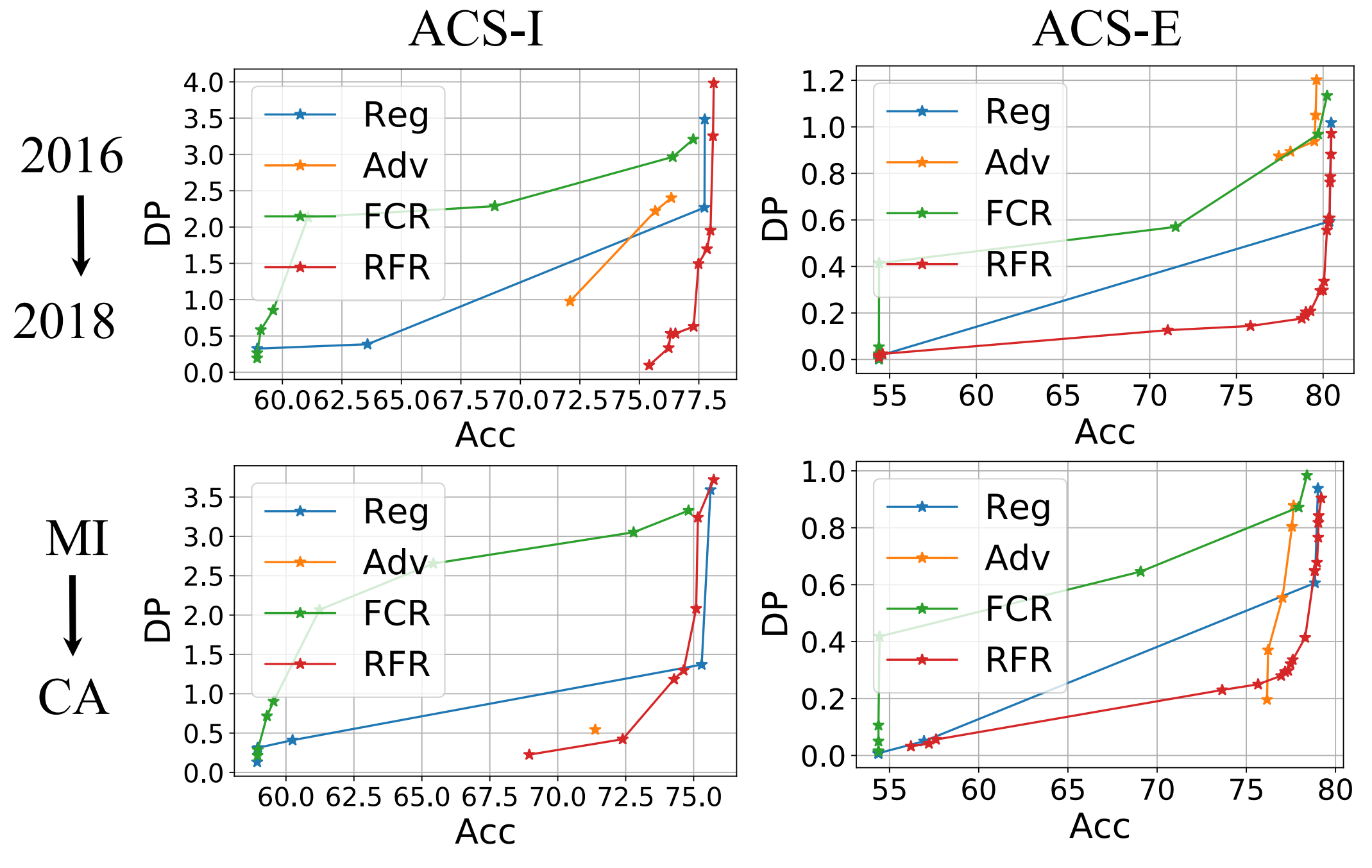


Figure: The fairness (DP) and prediction (Acc) trade-off performance on three datasets with different synthetic distribution shifts. The units for x- and y-axis are percentages (%).

Real-World Distribution Shift RFR outperforms several baselines for spatial and temporal distribution shifts. DP and Acc trade-off performance on three real-world datasets with temporal (Top) and spatial (Bottom) distribution shifts.



Acknowledgement: The authors thank the anonymous reviewers for their helpful comments. The work is in part supported by Samsung Electronic America and NSF grants NSF IIS-2224843.