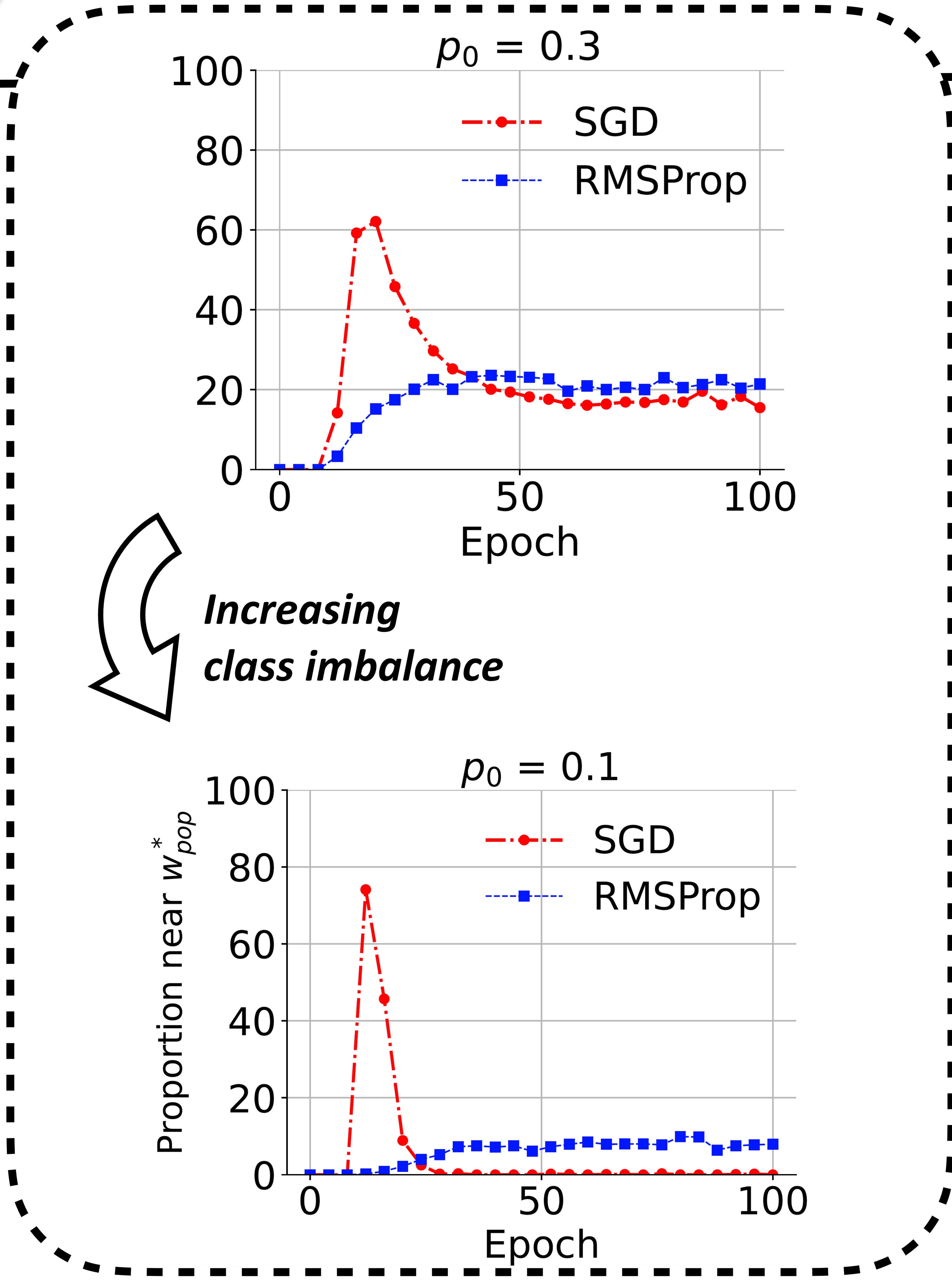
Some Optimizers are More Equal:

Understanding the Role of Optimizers in Group Fairness

Mojtaba Kolahdouzi Hatice Gunes Ali Etemad



Introduction

TL;DR: We show that adaptive optimizers like RMSProp lead to fairer minima more often than SGD, both theoretically and empirically.

Motivation: Deep learning models are widely used in socially impactful domains. Fairness research has mostly focused on external interventions like reweighting to enhance group fairness. Yet, the fairness implications of core training components remain less understood. We ask this question: ***Does the choice of optimizer impact group fairness? And how?***

Contributions: We demonstrate, for the first time, that optimizer choice alone can meaningfully affect group fairness. Through stochastic differential equation (SDE) analysis and new theoretical guarantees, we show that adaptive optimizers, such as RMSProp, are more likely to converge to fairer minima than SGD, particularly under class imbalance. Extensive experiments across datasets, tasks, backbones, and fairness metrics consistently validate these theoretical insights.

SDE Analysis of Tractable Setup

We consider a tractable training scenario with two demographic subgroups, each represented by quadratic loss: $\mathscr{L}_0 = \frac{1}{2}(w-1)^2$ and $\mathscr{L}_1 = \frac{1}{2}(w+1)^2$. , whose optimal solutions lie at $+1$ and -1, respectively. The population objective is their weighted sum, $\mathscr{L}_{pop}$, whose minimizer at $w = 0$ corresponds to the fairest solution. During mini-batch training, samples from subgroups are drawn with probabilities p_0 and p_1, and any imbalance biases the optimization trajectory toward one subgroup's optimum. This controlled setup allows us to analytically compare how fair the optimizers are under class imbalance.

Theorem1. Let $p_0, p_1 \in (0,1)$ with $p_0 + p_1 = 1$ be the subgroup sampling probabilities for the loss functions $\mathscr{L}_0(w)$ and $\mathscr{L}_1(w)$. Suppose we optimize the empirical objective $\mathscr{L}_{emp}(w) = \frac{1}{N}\sum_{r\in\Omega}\mathscr{L}_{q_r}(w)$, where each sample $q_r \in \{0,1\}$ is drawn i.i.d. with probability p_0 for subgroup 0 or p_1 for subgroup 1. Consider mini-batch gradient updates of size 1 using SGD and RMSProp optimization algorithms. Then there exists a constant $\Delta(p_1 p_2, \eta) > 0$ such that, whenever $|p_0 - p_1| > \Delta(p_1 p_2, \eta)$, we have: $p_{rms}(w^*_{pop}) > p_{sgd}(w^*_{pop})$, where $p_{rms}(w^*_{pop})$ and $p_{sgd}(w^*_{pop})$ are the probabilities of RMSProp and SGD converging to fair minima $w^*_{pop} = 0$, respectively.

Group Fairness Analysis

Beyond a tractable setup, we analyze how optimizers affect group fairness in a more general case. We show that adaptive methods, by scaling gradients, naturally shrink update disparities between subgroups, leading to fairer optimization dynamics (***Theorem 2***). Moreover, in a single iteration, their worst-case increase in demographic parity gap is upper-bounded by that of SGD (***Theorem 3***).

Theorem 2. Consider a population that consists of two subgroups with subgroup-specific loss functions $\mathscr{L}_0(w)$ and $\mathscr{L}_1(w)$, sampled with probabilities p_0 and p_1, respectively. Suppose a online training regime, in which each parameter update is computed from a sample drawn from one of the two subgroups. Suppose the gradients $\nabla\mathscr{L}_0(w_k)$and $\nabla\mathscr{L}_1(w_k)$ are well-behaved anisotropic NGOs. Then, the difference in parameter updates between subgroups 0 and 1 under RMSProp has an upper bound given by the corresponding difference under SGD.

Theorem 3. Suppose the same setup as Theorem 2, with subgroup-specific loss functions $\mathscr{L}_0(w)$ and $\mathscr{L}_1(w)$, sampled with probabilities p_0 and p_1, respectively. Then in expectation, the worst-case increase in the demographic parity gap after one iteration of RMSProp has an upper-bound no greater than the corresponding increase under SGD.

Fairness for ViT

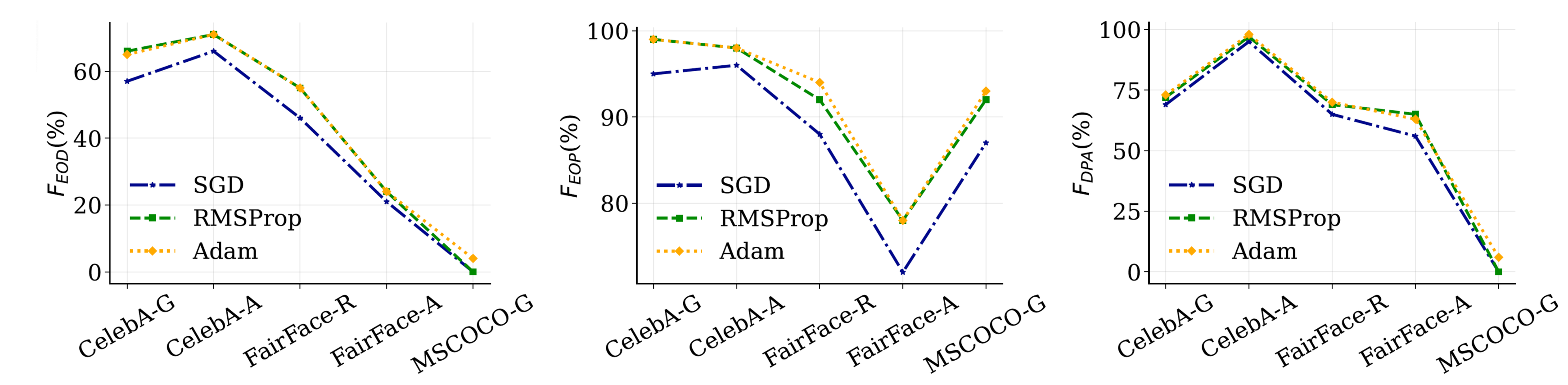


Fig 2. Fairness for ViT across different datasets, attributes (G: gender, A: age, R: race), and metrics.

Optimizers & fairness-enhancing Methods

To test whether our findings extend beyond standard training, we pair each optimizer with an established fairness-enhancing method on tabular benchmarks.

Table 1. Comparison of fairness metrics across optimizers, with and without fairness-enhancing methods.

	Gap in Equal Opportunity			Gap in Equalized Odds			Gap in Demographic Parity		
Dataset	Adam	RMSProp	SGD	Adam	RMSProp	SGD	Adam	RMSProp	SGD
With fairness-enhancing									
ProPublica COMPAS	0.45	0.48	0.71	2.79	2.78	2.90	0.86	0.86	2.60
AdultCensus	3.15	3.16	3.38	2.39	2.41	4.28	7.00	6.80	11.79
Without fairness-enhancing									
ProPublica COMPAS	13.99	13.90	15.19	13.99	13.95	14.98	11.49	11.45	11.80
AdultCensus	21.04	20.91	21.29	20.90	20.91	21.14	12.19	12.23	12.42

Effect Grows with Imbalance

By varying the subgroup imbalance in CelebA, we observe that the fairness gap between RMSProp and SGD widens as the dataset becomes more imbalanced, matching the prediction of our theoretical analysis.

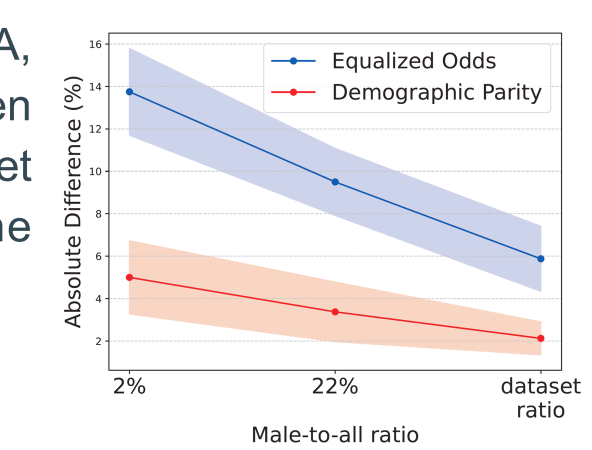


Fig 3. Difference of RMSProp and SGD's fairness in different male-to-all ratios

Fairness Gains are statistically Significant

Repeated runs on CelebA with Wilcoxon tests confirming that the improvements are statistically significant.

Table 2. p-values from the Wilcoxon test comparing SGD vs. RMSProp and SGD vs. Adam optimizers on the CelebA

Metric	Gender		Age	
	SGD-RMSProp	SGD-Adam	SGD-RMSProp	SGD-Adam
F_{EOD}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
F_{EOP}	1×10^{-3}	1×10^{-3}	1×10^{-3}	5×10^{-3}
F_{DPA}	2×10^{-3}	1×10^{-3}	7×10^{-3}	3×10^{-3}

Questions or Interested? Connect with Mojtaba Kolahdouzi
Email: m.kolahdouzi@queensu.ca

SCAN ME