FairALM: Augmented Lagrangian Method for Training Fair Models with Little Regret
SUPPLEMENTARY MATERIAL

Vishnu Suresh Lokhande[1], Aditya Kumar Akash[1], Sathya N. Ravi[2], and Vikas Singh[1]

[1] University of Wisconsin-Madison, Madison WI, USA
lokhande@cs.wisc.edu, aakash@wisc.edu, vsingh@biostat.wisc.edu
[2] University of Illinois at Chicago, Chicago IL, USA
sathya@uic.edu

Table of Contents

1 Experiments on *FairALM: Linear Classifier*

Data. We consider four standard datasets, `Adult`, `COMPAS`, `German` and `Law Schools` [4, 1]. The `Adult` dataset is comprised of demographic characteristics where the task is to predict if a person has an income higher (or lower) than \$50K per year. The protected attribute here is gender. In `COMPAS` dataset, the task is to predict the recidivism of individuals based on features such as age, gender, race, prior offenses and charge degree. The protected attribute here is race, specifically, whether the individual is white or black. The `German` dataset classifies people as good or bad credit risks with the person being a foreigner or not as the protected attribute. The features available in this dataset are credit history, saving accounts, bonds, etc. Finally, the `Law Schools` dataset, which comprises of $\sim$ 20K examples, seeks to predict a person's passage of the bar exam. Here, a binary attribute race is considered as the protected attribute.

Setup. We use Alg. 1 in the paper for experiments in this section. Recall from § 3 of the paper that Alg. 1 requires the specification of $\mathcal{H}$. We use the space of logistic regression classifiers as $\mathcal{H}$. At the start of the algorithm we have an empty set of classifiers. In each iteration, we add a newly trained classifier $h \in \mathcal{H}$ to the set of classifiers only if h has a smaller Lagrangian objective value among all the classifiers already in the set.

Quantitative Results. For the `Adult` dataset, FairALM attains a smaller test error and smaller DEO compared to the baselines considered in Table 1. We see big improvements on the DEO measure in `COMPAS` dataset and test error in `German` dataset using FairALM. While the performance of FairALM on `Law Schools` is comparable to other methods, it obtains a better false-positive rate than [1] which is a better metric as this dataset is skewed towards it's target class.

Summary. We train Alg. 1 on standard datasets specified in [4, 1]. We observe that FairALM is competitive with the popular methods in the fairness literature.

	Adult		COMPAS		German		Law Schools	
	ERR	DEO	ERR	DEO	ERR	DEO	ERR	DEO
Zafar *et al.* [7]	22.0	5.0	31.0	10.0	38.0	13.0	–	–
Hardt *et al.* [5]	18.0	11.0	29.0	8.0	29.0	11.0	4.5	0.0
Donini *et al.* [4]	19.0	1.0	27.0	5.0	27.0	5.0	–	–
Agarwal *et al.* [1]	17.0	1.0	31.0	3.0	–	–	4.5	1.0
FairALM	15.8 $\pm$1	0.7 $\pm$0.6	34.7 $\pm$1	0.1 $\pm$0.1	24.3 $\pm$2.7	10.8 $\pm$4.5	4.8 $\pm$0.1	0.4 $\pm$0.2

Table 1. Standard Datasets. We report test error (ERR) and DEO fairness measure in %. FairALM attains minimal DEO measure among the baseline methods while maintaining a similar test error.

2 Proofs for theoretical claims in the paper

Prior to proving the convergence of primal and dual variables of our algorithm with respect to the augmented lagrangian $L_T(q, \lambda)$, we prove a regret bound on the function $f_t(\lambda)$ which is defined in the following lemma. As $f_t(\lambda)$ is a strongly concave function (which we shall see shortly), we obtain a bound on the negative regret.

Lemma 1. *Let r_t denote the reward at each round of the game. The reward function $f_t(\lambda)$ is defined as $f_t(\lambda) = \lambda r_t - \frac{1}{2\eta}(\lambda - \lambda_t)^2$. We choose λ in the round $T+1$ to maximize the cumulative reward, i.e., $\lambda_{T+1} = \operatorname{argmax}_\lambda \sum_{t=1}^{T} f_t(\lambda)$. Define $L = \max_t \mid r_t \mid$. We obtain the following bound on the cumulative reward, for any λ,*

$$\sum_{t=1}^{T}\left(\lambda r_t - \frac{1}{2\eta}(\lambda - \lambda_t)^2\right) \leq \sum_{t=1}^{T} \lambda_t r_t + \eta L^2 \mathcal{O}(\log T) \tag{1}$$

Proof. As we are maximizing the cumulative reward function, in the $(t+1)^{th}$ iteration λ_{t+1} is updated as $\lambda_{t+1} = \operatorname{argmax}_\lambda \sum_{i=1}^{t} f_i(\lambda)$. This learning rule is also called the Follow-The-Leader (FTL) principle which is discussed in Section 2.2 of [6]. Emulating the proof of Lemma 2.1 in [6], a bound on the negative regret of FTL, for any $\lambda \in \mathbb{R}$, can be derived due to the concavity of $f_t(\lambda)$,

$$\sum_{t=1}^{T} f_t(\lambda) - \sum_{t=1}^{T} f_t(\lambda_t) \leq \sum_{t=1}^{T} f_t(\lambda_{t+1}) - \sum_{t=1}^{T} f_t(\lambda_t) \tag{2}$$

Our objective, now, is to obtain a bound on RHS of (2). Solving $\operatorname{argmax}_\lambda \sum_{i=1}^{t} f_i(\lambda)$ for λ will show us how λ_t and λ_{t+1} are related,

$$\lambda_{t+1} = \frac{\eta}{t}\sum_{i=1}^{t} r_i + \frac{1}{t}\sum_{i=1}^{t}\lambda_i \quad \implies \lambda_{t+1} - \lambda_t = \frac{\eta}{t} r_t \tag{3}$$

Using (3), we obtain a bound on $f_t(\lambda_{t+1}) - f_t(\lambda_t)$, we have,

$$f_t(\lambda_{t+1}) - f_t(\lambda_t) \leq \frac{\eta}{t} r_t^2$$

With $L = \max_t |r_t|$ and using the fact that $\sum_{i=1}^{T} \frac{1}{i} \leq (\log T + 1)$,

$$\sum_{t=1}^{T}\Big(f_t(\lambda_{t+1}) - f_t(\lambda_t)\Big) \leq \eta L^2(\log T + 1) \tag{4}$$

Let us denote $\xi_T = \eta L^2(\log T + 1)$, we bound (2) with (4),

Cumulative Reward Bound

$$\forall \lambda \in \mathbb{R} \quad \sum_{t=1}^{T}\left(\lambda r_t - \frac{1}{2\eta}(\lambda - \lambda_t)^2\right) \leq \Big(\sum_{t=1}^{T}\lambda_t r_t\Big) + \xi_T \tag{5}$$

□

Next, using the *Cumulative Reward Bound* (5), we prove the theorem stated in the paper. The theorem gives us the number of iterations required by Alg. 1 (in the paper) to reach a $\nu-$approximate saddle point. Our bounds for $\eta = \frac{1}{T}$ and $\lambda \in \mathbb{R}$ are strictly better than [1]. We re-state the theorem here,

Theorem 1. *Recall that d_h represents the difference of conditional means. Assume that $||d_h||_\infty \leq L$ and consider T rounds of Alg 1 (in the paper). Let $\bar{q} := \frac{1}{T}\sum_{t=1}^{T} q_t$ and $\bar{\lambda} := \frac{1}{T}\sum_{t=1}^{T} \lambda_t$ be the average plays of the q-player and the λ-player respectively. Then, we have $L_T(\bar{q}, \bar{\lambda}) \leq L_T(q, \bar{\lambda}) + \nu$ and $L_T(\bar{q}, \bar{\lambda}) \geq L_T(\bar{q}, \lambda) - \nu$, under the following conditions,*

- *If $\eta = \mathcal{O}(\sqrt{\frac{B^2 T}{L^2(\log T+1)}})$, $\nu = \mathcal{O}(\sqrt{\frac{B^2 L^2(\log T+1)}{T}})$; $\forall |\lambda| \leq B$, $\forall q \in \Delta$*
- *If $\eta = \frac{1}{T}$, $\nu = \mathcal{O}(\frac{L^2(\log T+1)^2}{T})$; $\forall \lambda \in \mathbb{R}$, $\forall q \in \Delta$*

Proof. Recall the definition of $L_T(q, \lambda)$ from the paper,

$$L_T(q, \lambda) = \big(\sum_i q_i e_{h_i}\big) + \lambda\big(\sum_i q_i d_{h_i}\big) - \frac{1}{2\eta}\big(\lambda - \lambda_T\big)^2 \tag{6}$$

For the sake of this proof, let us define ζ_T in the following way,

$$\zeta_T(\lambda) = \frac{1}{2\eta}\sum_{t=1}^{T}\Big((\lambda - \lambda_t)^2 - (\lambda - \lambda_T)^2 + (\lambda_t - \lambda_T)^2\Big) \tag{7}$$

Recollect from (5) that $\xi_T = \eta L^2(\log T + 1)$. We **outline** the proof as follows,

1. First, we compute an upper bound on $L_T(\bar{q}, \bar{\lambda})$,

Average Play Upper Bound

$$L_T(\bar{q}, \bar{\lambda}) \leq L_T(q, \bar{\lambda}) + \frac{\zeta_T(\bar{\lambda})}{T} + \frac{\xi_T}{T} \quad \forall q \in \Delta \tag{8}$$

$$\text{Also, } L_T(\bar{q}, \lambda) \leq L_T(q, \bar{\lambda}) + \frac{\zeta_T(\lambda)}{T} + \frac{\xi_T}{T} \quad \forall \lambda \in \mathbb{R}, \forall q \in \Delta \tag{9}$$

2. Next, we determine an lower bound on $L_T(\bar{q}, \bar{\lambda})$,

Average Play Lower Bound

$$L_T(\bar{q}, \bar{\lambda}) \geq L_T(\bar{q}, \lambda) - \frac{\zeta_T(\lambda)}{T} - \frac{\xi_T}{T} \quad \forall \lambda \in \mathbb{R} \tag{10}$$

3. We bound $\frac{\zeta_T(\lambda)}{T} + \frac{\xi_T}{T}$ for the case $|\lambda| \leq B$ and show that a $\nu-$approximate saddle point is attained.
4. We bound $\frac{\zeta_T(\lambda)}{T} + \frac{\xi_T}{T}$ for the case $\lambda \in \mathbb{R}$ and, again, show that $\nu-$approximate saddle point is attained.

We write the proofs of the above fours parts one-by-one. Steps 1,2 in the above outline are intermediary results used to prove our main results in Steps 3,4. Reader can directly move to Steps 3,4 to see the main proof.

1. Proof for the result on *Average play Upper Bound*

$$L_T(q,\bar{\lambda}) = \textstyle\sum_i q_i e_{h_i} + \Big(\frac{\sum_t \lambda_t}{T}\Big)\Big(\sum_i q_i d_{h_i}\Big) - \frac{1}{2\eta}\Big(\frac{\sum_t \lambda_t}{T} - \lambda_T\Big)^2 \tag{11}$$

Exploiting convexity of $\frac{1}{2\eta}\Big(\frac{\sum_t \lambda_t}{T} - \lambda_T\Big)^2$ via Jenson's Inequality,

$$\geq \frac{1}{T}\sum_t \Big(\textstyle\sum_i q_i e_{h_i} + \lambda_t \sum_i q_i d_{h_i} - \frac{1}{2\eta}(\lambda_t - \lambda_T)^2\Big) \tag{12}$$

As $h_t = \text{argmin}_q L_T(q,\lambda_t)$, we have $L_T(q,\lambda_t) \geq L_T(h_t,\lambda_t)$, hence,

$$\geq \frac{1}{T}\sum_t \Big(e_{h_t} + \lambda_t d_{h_t} - \frac{1}{2\eta}(\lambda_t - \lambda_T)^2\Big) \tag{13}$$

Using the *Cumulative Reward Bound* (5),

$$\geq \frac{\sum_t e_{h_t}}{T} + \frac{\lambda \sum_t d_{h_t}}{T} - \frac{1}{T}\sum_t \Big(\frac{(\lambda-\lambda_t)^2}{2\eta} + \frac{(\lambda_t - \lambda_T)^2}{2\eta}\Big) - \frac{\xi_T}{T} \tag{14}$$

Add and subtract $\frac{1}{T}\sum_{t=1}^T \frac{1}{2\eta}(\lambda - \lambda_T)^2$, use ζ_T from (7) and regroup the terms,

$$= (\textstyle\sum_i \bar{q}_i e_{h_i}) + (\lambda \sum_i \bar{q}_i d_{h_i}) - \frac{1}{2\eta}(\lambda - \lambda_T)^2 - \frac{\zeta_T(\lambda)}{T} - \frac{\xi_T}{T} \tag{15}$$

$$= L_T(\bar{q},\lambda) - \frac{\zeta_T(\lambda)}{T} - \frac{\xi_T}{T} \tag{16}$$

2. Proof for the result on *Average play Lower Bound* Proof is similar to Step 1 so we skip the details. The proof involves finding a lower bound for $L_T(\bar{q},\lambda)$ using the *Cumulative Reward Bound* (5). With simple algebraic manipulations and exploiting the convexity of $L_T(\bar{q},\lambda)$ via the Jenson's inequality, we obtain the bound that we state.

3. Proof for the case $|\boldsymbol{\lambda}| \leq \boldsymbol{B}$

For the case $|\lambda| \leq B$, we have $\zeta_T(\lambda) \leq \frac{B^2 T}{\eta}$, which gives,

$$\frac{\zeta_T(\lambda)}{T} + \frac{\xi_T}{T} \leq \frac{B^2}{\eta} + \frac{\eta L^2(\log T + 1)}{T} \tag{17}$$

Minimizing R.H.S in (17) over η gives us a $\nu-$ approximate saddle point,

$\nu-$ approximate saddle point for $|\lambda| \leq B$

$$L_T(\bar{q},\bar{\lambda}) \leq L_T(q,\bar{\lambda}) + \nu \quad \text{and} \quad L_T(\bar{q},\bar{\lambda}) \geq L_T(\bar{q},\lambda) - \nu \tag{18}$$

$$\text{where } \nu = 2\sqrt{\frac{B^2 L^2(\log T + 1)}{T}} \quad \text{and } \eta = \sqrt{\frac{B^2 T}{L^2(\log T + 1)}} \tag{19}$$

4. Proof for the case $\boldsymbol{\lambda} \in \mathbb{R}$

We begin the proof by bounding $\frac{\zeta_T(\lambda)}{T} + \frac{\xi_T}{T}$. Let $\lambda_* = \text{argmax}_\lambda L_T(\bar{q}, \lambda)$. We have a closed form for λ_* given by $\lambda_* = \lambda_T + \eta \sum_i \bar{q}_i d_{h_i}$. Substituting λ_* in ζ_T gives,

$$\frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} = \frac{1}{2\eta}\frac{1}{T}\sum_t \Big(2(\lambda_t - \lambda_T)^2 + 2\eta(\lambda_T - \lambda_t)(\textstyle\sum_i \bar{q}_i d_{h_i})\Big) + \frac{\xi_T}{T} \tag{20}$$

Recollect that $\lambda_{t+1} - \lambda_t = \frac{\eta}{t} d_{h_t}$ (from (3)). Using telescopic sum on λ_t, we get $(\lambda_T - \lambda_t) \leq \eta L(\log T + 1)$ and $(\lambda_T - \lambda_t)^2 \leq \eta^2 L^2 (\log T + 1)^2$. We substitute these in the previous equation (20),

$$\frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} \leq \eta L^2 (\log T + 1)^2 + \eta L^2 (\log T + 1) + \frac{\eta L^2 (\log T + 1)}{T} \tag{21}$$

Setting $\eta = \frac{1}{T}$, we get

$$\frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} \leq \mathcal{O}\Big(\frac{L^2(\log T + 1)^2}{T}\Big) := \nu \tag{22}$$

Using (22), we prove the convergence of λ in the following way,

$$\begin{aligned}
L_T(\bar{q}, \lambda) &\leq L_T(\bar{q}, \lambda_*) \quad \Big(\text{as } \lambda_* \text{ is the maximizer of } L_T(\bar{q}, \lambda)\Big) && (23)\\
&\leq L_T(\bar{q}, \bar{\lambda}) + \frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} \quad \Big(\textit{Average Play Lower Bound}\Big) && (24)\\
&\leq L_T(\bar{q}, \bar{\lambda}) + \nu \quad \Big(\text{from } (22)\Big) && (25)
\end{aligned}$$

We prove the convergence of q in the following way. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
L_T(q, \bar{\lambda}) &\geq L_T(\bar{q}, \lambda_*) - \frac{\zeta_T(\lambda_*)}{T} - \frac{\xi_T}{T} \quad \Big(\textit{Average Play Upper Bound } (16)\Big) && (26)\\
&\geq L_T(\bar{q}, \lambda_*) - \nu \quad \Big(\text{from } (22)\Big) && (27)\\
&\geq L_T(\bar{q}, \bar{\lambda}) - \nu \quad \Big(\text{as } \lambda_* \text{ is the maximizer of } L_T(\bar{q}, \lambda)\Big) && (28)
\end{aligned}$$

Therefore,

$\nu-$ approximate saddle point for $\lambda \in \mathbb{R}$

$$L_T(\bar{q}, \bar{\lambda}) \leq L_T(q, \lambda) + \nu \quad \text{and} \quad L_T(\bar{q}, \bar{\lambda}) \geq L_T(\bar{q}, \lambda) - \nu \tag{29}$$

$$\text{where } \nu = \mathcal{O}\Big(\frac{L^2(\log T + 1)^2}{T}\Big) \quad \text{and} \quad \eta = \frac{1}{T} \tag{30}$$

□

3 More details on *FairALM: DeepNet Classifier*

Recall that in § 5.2 in the paper, we identified a key difficulty when extending our algorithm to deep networks. The main issue is that the set of classifiers $|\mathcal{H}|$ is not a finite set. We argued that leveraging stochastic gradient descent (SGD) on an over-parameterized network eliminates this issue. When using SGD, few additional modifications of Alg 1 (in the paper) are helpful, such as replacing the non-differentiable indicator function $\mathbb{1}[\cdot]$ with a smooth surrogate function and computing the empirical estimates of the errors and conditional means denoted by $\hat{e}_h(z)/\hat{\mu}_h^s(z)$ respectively. These changes modify our objective to a form that is not a zero-sum game,

$$\max_{\lambda} \min_{w} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \tag{31}$$

We use DP constraint in (31), other fairness metrics discussed in the paper are valid as well. A closed-form solution for λ can be achieved by solving an upper bound to (31) obtained by exchanging the "max"/"min" operations.

$$\max_{\lambda} \min_{w} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \tag{32}$$

$$\leq \min_{w} \max_{\lambda} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \tag{33}$$

Substituting the closed form solution $\lambda = \lambda_t + \eta(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1})$ in (33),

$$\leq \min_{w} \left(\hat{e}_{h_w} + +\lambda_t(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) + \frac{\eta}{2}(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1})^2 \right) \tag{34}$$

Note that the surrogate function defined within $\hat{\mu}_{h_w}^s$ is convex and non-negative, hence, we can exploit Jenson's inequality to eliminate the power 2 in (34) to give us a convenient upper bound,

$$\leq \min_{w} \left(\hat{e}_{h_w} + (\lambda_t + \eta)\hat{\mu}_{h_w}^{s_0} - (\lambda_t - \eta)\hat{\mu}_{h_w}^{s_1} \right) \tag{35}$$

In order to obtain a good minima in (35), it may be essential to run the SGD on (35) a few times: for ImSitu experiments, SGD was run on (35) for 5 times. We also gradually increase the parameter η with time as $\eta_t = \eta_{t-1}(1 + \eta_\beta)$ for a small non-negative value for η_β, e.g., $\eta_\beta \approx 0.01$. This is a common practice in augmented Lagrangian methods, see [2] (page 104). The overall algorithm is available in the paper as Alg. 2. The key primal and dual steps can be seen in the following section.

3.1 Algorithm for baselines

We provide the primal and dual steps used for the baseline algorithms for the ImSitu experiments from the paper. The basic framework for all the baselines remains the same as Alg. 2 in the paper. For Proxy-Lagrangian, only the key ideas in [3] were adopted for implementation.

Unconstrained

PRIMAL: $v_t \in \partial \hat{e}_{h_w}$

DUAL: None

ℓ_2 Penalty

PRIMAL: $v_t \in \partial\Big(\hat{e}_{h_w} + \eta(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w})^2\Big)$

DUAL: None

Parameters: Penalty Parameter η

Reweight

PRIMAL: $v_t \in \partial\Big(\hat{e}_{h_w} + \eta_0\hat{\mu}^{s_0}_{h_w} + \eta_1\hat{\mu}^{s_1}_{h_w}\Big)$

DUAL: None

Parameters: $\eta_i \propto 1/(\#$ samples in $s_i)$

Lagrangian [8]

PRIMAL: $v_t \in \partial\Big(\hat{e}_{h_w} + \lambda_t^{0\setminus 1}(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w} - \epsilon) + \lambda_t^{1\setminus 0}(\hat{\mu}^{s_1}_{h_w} - \hat{\mu}^{s_0}_{h_w} - \epsilon)\Big)$

DUAL: $\lambda^{i\setminus j}_{t+1} \leftarrow \max\big(0, \lambda^{i\setminus j}_t + \eta_{i\setminus j}(\hat{\mu}^{s_i}_{h_w} - \hat{\mu}^{s_j}_{h_w} - \epsilon)\big)$

Parameters: Dual step sizes $\eta_{0\setminus 1}, \eta_{1\setminus 0}$ Tol. $\epsilon \approx 0.05$. $i\setminus j \in \{0\setminus 1, 1\setminus 0\}$

Proxy-Lagrangian [3]

PRIMAL: $v_t \in \partial\Big(\hat{e}_{h_w} + \lambda_t^{0\setminus 1}(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w} - \epsilon) + \lambda_t^{1\setminus 0}(\hat{\mu}^{s_1}_{h_w} - \hat{\mu}^{s_0}_{h_w} - \epsilon)\Big)$

DUAL: $\theta^{i\setminus j}_{t+1} \leftarrow \theta^{i\setminus j}_t + \eta_{i\setminus j}(\hat{\mu}^{s_i}_{h_w} - \hat{\mu}^{s_j}_{h_w} - \epsilon)$

$$\lambda^{i\setminus j}_{t+1} \leftarrow B\frac{\exp\theta^{i\setminus j}_{t+1}}{1 + \exp\theta^{i\setminus j}_{t+1} + \exp\theta^{j\setminus i}_{t+1}}$$

Parameters: Dual step sizes $\eta_{0\setminus 1}/\eta_{1\setminus 0}$. Tol. $\epsilon \approx 0.05$, Hyperparam. B

No surrogates in DUAL for $\hat{\mu}^{s_0}_{h_w}/\hat{\mu}^{s_1}_{h_w}$. $i\setminus j \in \{0\setminus 1, 1\setminus 0\}$

FairALM

PRIMAL: $v_t \in \partial\Big(\hat{e}_{h_w}(z) + (\lambda_t + \eta)\hat{\mu}^{s_0}_{h_w}(z) - (\lambda_t - \eta)\hat{\mu}^{s_1}_{h_w}(z)\Big)$

DUAL: $\lambda_{t+1} \leftarrow \lambda_t + \eta\big(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w}\big)$

Parameters: Dual Step Size η

3.2 Supplementary Results on CelebA

Additional Results. The dual step size η is a key parameter in FairALM training. Analogous to the dual step size η we have the penalty parameter in ℓ_2 penalty training, also denoted by η. It can be seen from Figure 1 and Figure 2 that FairALM is more robust to different choices of η than ℓ_2 penalty. The target class in this section is *attractiveness* and protected attribute is *gender*.

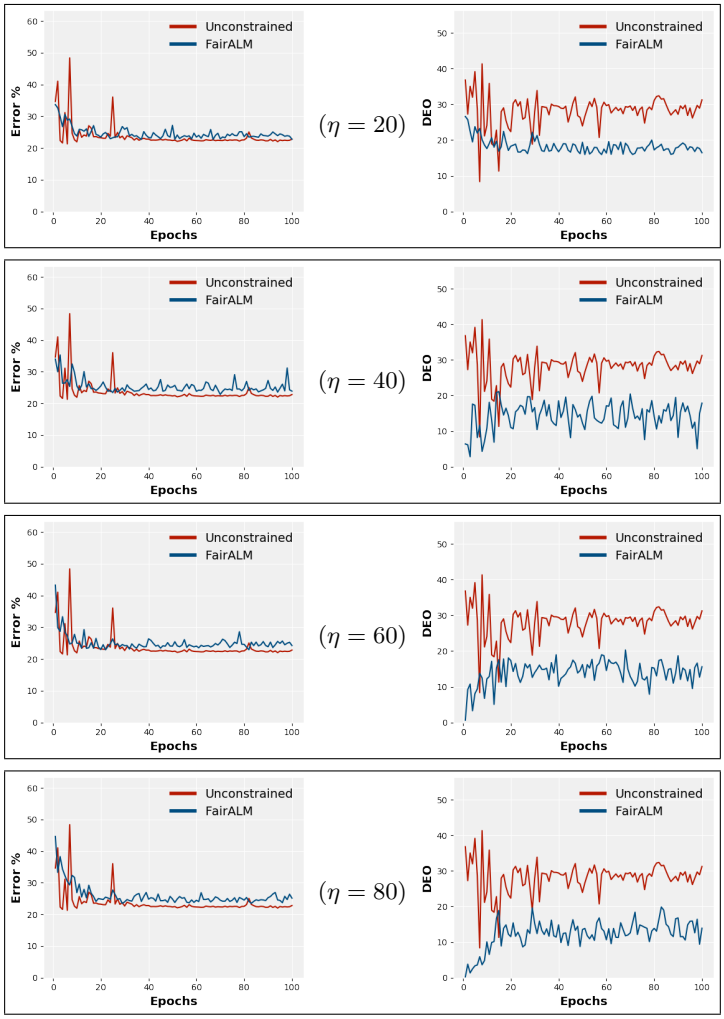


Fig. 1. FairALM Ablation on CelebA. For a given η, the left image represents the test error and the right image shows the DEO measure. We study the effect of varying the dual step size η on FairALM. We observe that the performance of FairALM is consistent over a wide range of η values.

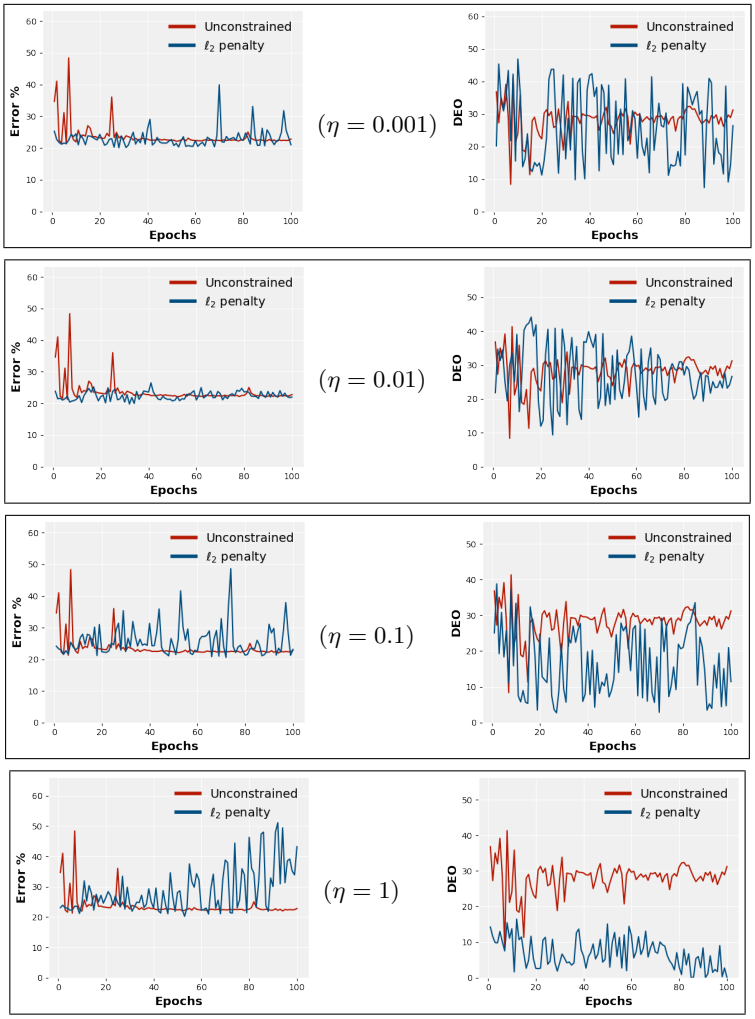


Fig. 2. ℓ_2 Penalty Ablation on CelebA For each η value, the left image represents the test set errors and the right image shows the fairness measure (DEO). We investigate a popular baseline to impose fairness constraint which is the ℓ_2 penalty. We study the effect of varying the penalty parameter η in this figure. We observe that training with ℓ_2 penalty is quite unstable. For $\eta > 1$, the algorithm doesn't converge and raises numerical errors.

More Interpretability Results. We present the activation maps obtained when running the *FairALM* algorithm, unconstrained algorithm and the gender classification task. We show our results in Figure 3. The target class is *attractiveness* and protected attribute is gender. We threshold the maps to show only the most significant colors. The maps from gender classification task look at gender-revealing attributes such as presence of *long-hair*. The unconstrained model looks mostly at the entire image. *FairALM* looks at only a specific region of the face which is not gender revealing.

Fig. 3. Interpretability in CelebA. We find that an unconstrained model picks up a lot of gender revealing attributes however FairALM doesn't. The image labelled `Gender` denotes the map of a gender classification task. We observe overlap between the maps of gender classification task and the unconstrained model. The activation maps are regulated to show colors above a fixed threshold to highlight the most significant regions used by a model.

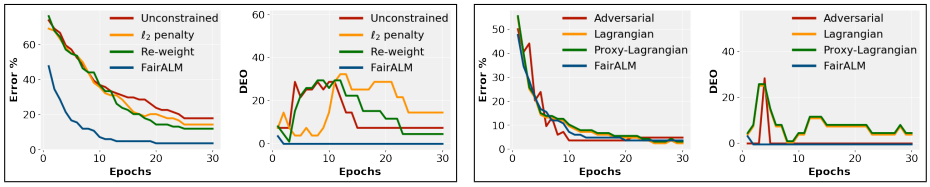
4 Supplementary Results on ImSitu

Detailed Setup. We use the standard ResNet-18 architecture for the base model. We initialize the weights of the conv layers weights from ResNet-18 trained on ImageNet (ILSVRC). We train the model using SGD optimizer and a batch size of 256. For first few epochs (≈ 20) only the linear layer is trained with a learning rate of 0.01/0.005. Thereafter, the entire model is trained end to end with a lower learning rate of 0.001/0.0005 till the accuracy plateaus.

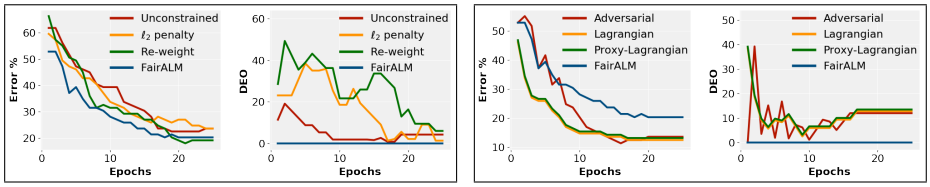
Meaning of Target class (+)**.** Target class (+) is something that a classifier tries to predict from an image. Recall the basic notations § 2 from the paper, $\mu_h^{s_i,t_j} := \mu_h|(s = s_i, t = t_j)$ denotes the elementary conditional expectation of some function μ_h with respect to two random variables s, t. When we say we are imposing DEO for a target class t_j we refer to imposing constraint on the difference in conditional expectation of the two groups of s for the class t_j, that is, $d_h = \mu_h^{s_0,t_j} - \mu_h^{s_1,t_j}$. For example, for *Cooking* (+) vs *Driving* (−) problem when we say *Cooking* (+) is regarded as the target class we mean that $t_j = cooking$ and hence the DEO constraint is of the form $d_h = \mu_h^{s_0,cooking} - \mu_h^{s_1,cooking}$.

Supplementary Training Profiles. We plot the test set errors and the DEO measure during the course of training for the verb pair classifications reported in the paper. We compare against the baselines discussed in Table 1 of the paper. The plots in Fig. 4 below supplement Fig. 5 in the paper.

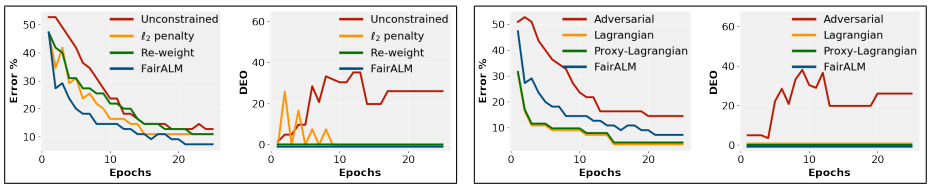
Additional qualitative results We show the activation maps in Fig. 5 to illustrate that the features used by FairALM model are more aligned with the action/verb present in the image and are not gender leaking. The verb pairs have been chosen randomly from the list provided in [8]. In all the cases `Gender` is considered as the protected attribute. The activation maps are regulated to show colors above a fixed threshold in order to highlight the most significant regions used by a model to make a prediction.



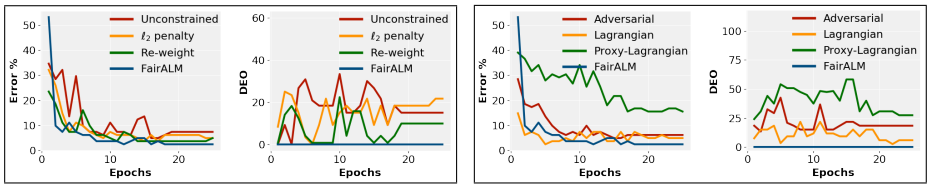
Cooking (+) Driving (−)



Shaving (+) Moisturizing (−)



Washing (+) Saluting (−)



Assembling (+) Hanging (−)

Fig. 4. Supplementary Training Profiles. FairALM consistently achieves minimum DEO across different verb pair classifications.

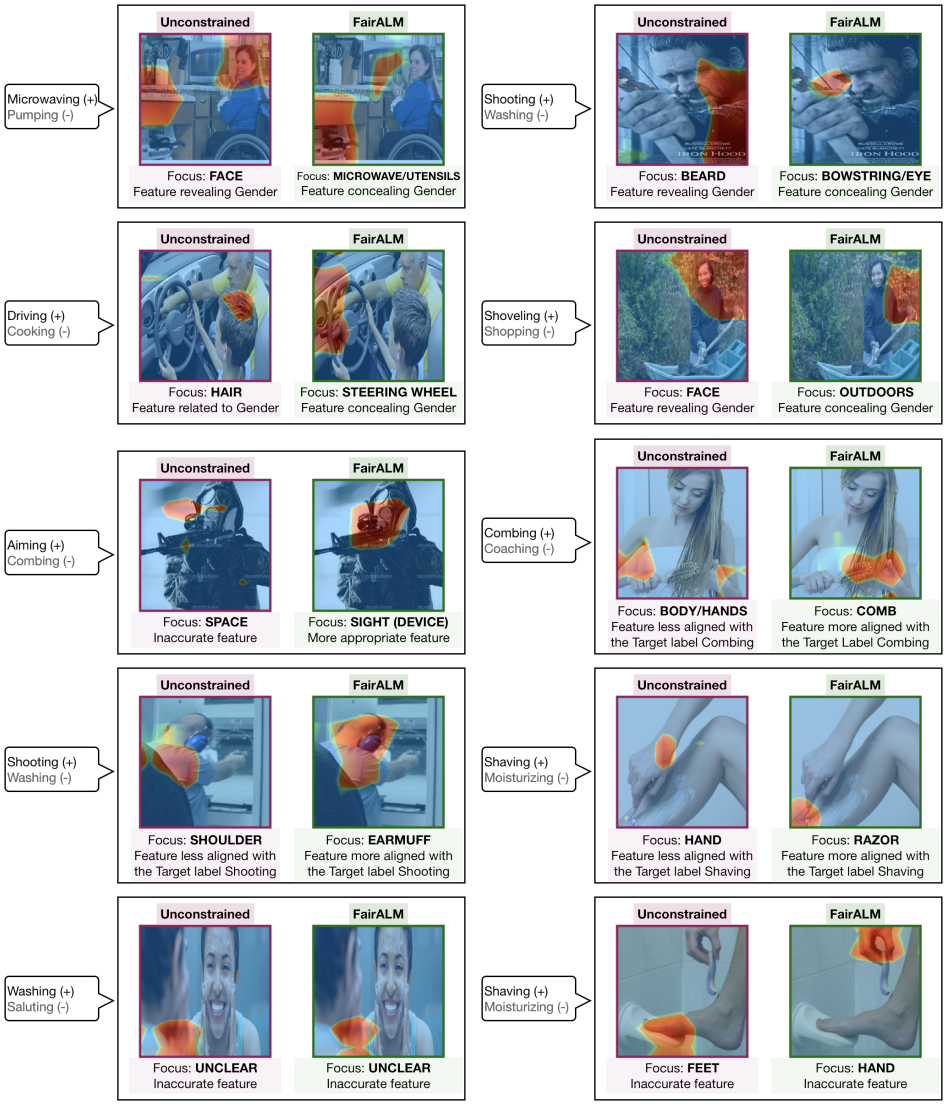


Fig. 5. Additional qualitative Results in ImSitu dataset. Models predict the target class (+). FairALM consistently avoids gender revealing features and uses features that are more relevant to the target class. Due to the small dataset sizes, a *limitation* of this experiment is shown in the last row where both FairALM and Unconstrained model look at incorrect regions. The number of such cases in FairALM is far less than those in the unconstrained model.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. arXiv preprint arXiv:1803.02453 (2018)
2. Bertsekas, D.P.: Constrained optimization and Lagrange multiplier methods. Academic press (2014)
3. Cotter, A., Jiang, H., Sridharan, K.: Two-player games for efficient non-convex constrained optimization. arXiv preprint arXiv:1804.06500 (2018)
4. Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J.S., Pontil, M.: Empirical risk minimization under fairness constraints. In: Advances in Neural Information Processing Systems. pp. 2791–2801 (2018)
5. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems. pp. 3315–3323 (2016)
6. Shalev-Shwartz, S., et al.: Online learning and online convex optimization. Foundations and Trends® in Machine Learning **4**(2), 107–194 (2012)
7. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1171–1180. International World Wide Web Conferences Steering Committee (2017)
8. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457 (2017)