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

Vishnu Suresh Lokhande [1] lokhande@cs.wisc.edu
Aditya Kumar Akash [1] aakash@wisc.edu
Sathya N. Ravi [2] sathya@uic.edu
Vikas Singh [1] vsingh@biostat.wisc.edu

[1]University of Wisconsin-Madison
[2]University of Illinois at Chicago

Abstract

Algorithmic decision making based on computer vision and machine learning technologies continue to permeate our lives. But issues related to biases of these models and the extent to which they treat certain segments of the population unfairly, have led to concern in the general public. It is now accepted that because of biases in the datasets we present to the models, a fairness-oblivious training will lead to unfair models. An interesting topic is the study of mechanisms via which the de novo *design or training of the model can be informed by fairness measures. Here, we study mechanisms that impose fairness concurrently while training the model. While existing fairness based approaches in vision have largely relied on training adversarial modules together with the primary classification/regression task, in an effort to remove the influence of the protected attribute or variable, we show how ideas based on well-known optimization concepts can provide a simpler alternative. In our proposed scheme, imposing fairness just requires specifying the protected attribute and utilizing our optimization routine. We provide a detailed technical analysis and present experiments demonstrating that various fairness measures from the literature can be reliably imposed on a number of training tasks in vision in a manner that is interpretable. A project page is available on //GitHub.*

1. Introduction

Fairness and non-discrimination is a core tenet of modern society. Driven by advances in vision and machine learning systems, algorithmic decision making continues to permeate our lives in important ways. Consequently, ensuring that the decisions taken by an algorithm do not exhibit serious biases is no longer a hypothetical topic, rather a key concern that has started informing legislation [22] (e.g., Algorithmic Accountability act). On one extreme, some types of biases can be bothersome – a biometric access system could be more error-prone for faces of persons from certain skin tones [9] or a search for `homemaker` or `programmer` may return gender-stereotyped images [8]. But there are serious ramifications as well – an individual may get pulled aside for an intrusive check while traveling [47] or a model may decide to pass on an individual for a job interview after digesting his/her social media content[13, 24]. Biases in automated systems in estimating recidivism within the criminal judiciary have been reported [35]. There is a growing realization that these problems need to be identified and diagnosed, and then promptly addressed. In the worst case, if no solutions are forthcoming, we must step back and reconsider the trade-off between the benefits versus the harm of deploying such systems, on a case by case basis.

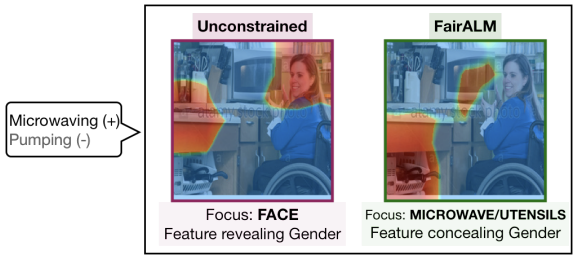


Figure 1: The heat maps of an unconstrained model and a fair model are depicted in this figure. The models are trained to predict the target label *Microwaving* (indicated by a (+)). The fair model attempts to make unbiased predictions with respect to sensitive attribute *gender*. In this example, it is observed that the heat maps of an unconstrained model are concentrated around gender revealing attributes such as the face of person. Alternatively, the heat maps of the fair model are concentrated around non-gender revealing attributes, such as utensils and microwave, which also happen to be more aligned to the target label.

What leads to unfair learning models? One finds that learning methods in general tend to amplify biases that exist in the training dataset [43]. While this creates an incentive for the organization training the model to curate datasets that are "balanced" in some sense, from a practical standpoint, it is often difficult to collect data that is balanced along multiple predictor variables that are "protected", e.g.,

gender, race and age. If a protected feature is correlated with the response variable, a learning model can *cheat* and find representations from other features that are collinear or a good surrogate for the protected variable. A thrust in current research is devoted to devising ways to mitigate such shortcuts. If one does not have access to the underlying algorithm, a recent result [23] shows the feasibility of finding thresholds that can impose certain fairness criteria. Such a threshold search can be post-hoc applied to any learned model. But in various cases, because of the characteristics of the dataset, a fairness-oblivious training will lead to biased models. An interesting topic is the study of mechanisms via which the *de novo* design or training of the model can be informed by fairness measures.

Some general strategies for Fair Learning. Motivated by the foregoing issues, recent work which may broadly fall under the topic of *algorithmic fairness* has suggested several concepts or measures of fairness that can be incorporated within the learning model. While we will discuss the details shortly, these include demographic parity [37], equal odds and equal opportunities [23], and disparate treatment [39]. In general, existing work can be categorized into a few distinct categories. The *first* category of methods attempts to modify the representations of the data to ensure fairness. While different methods approach this question in different ways, the general workflow involves imposing fairness *before* a subsequent use of standard machine learning methods [10, 26]. The *second* group of methods adjusts the decision boundary of an already trained classifier towards making it fair as a *post*-processing step while trying to incur as little deterioration in overall performance as possible [21, 20, 36]. While this procedure is convenient and fast, it is not always guaranteed to lead to a fair model without sacrificing accuracy. Part of the reason is that the search space for a fair solution in the post-hoc tuning is limited. Of course, we may impose fairness during training directly as adopted in the *third* category of papers such as [40, 4], and the approach we take here. Indeed, if we are training the model from scratch and have knowledge of the protected variables, there is little reason not to incorporate this information directly *during* model training. In principle, this strategy provides the maximum control over the model. From the formulation standpoint, it is slightly more involved because it requires satisfying a fairness constraint derived from one or more fairness measure(s) in the literature, while concurrently learning the model parameters. The difficulty varies depending both on the primary task (shallow versus deep model) as well as the specific fairness criteria. For instance, if one were using a deep network for classification, we would need to devise ways to enforce constraints on the *output* of the network, efficiently.

Scope of this paper and contributions. Many studies on fairness in learning and vision are somewhat recent and were partly motivated in response to more than a few controversial reports in the news media. As a result, the literature on mathematically sound and practically sensible fairness measures that can still be incorporated while training a model is still in a nascent stage. In vision, current approaches have largely relied on training adversarial modules in conjunction with the primary classification or regression task, to remove the influence of the protected attribute. In contrast, the **contribution** of our work is to provide a simpler alternative. We show that a number of fairness measures in the literature can be incorporated by viewing them as constraints on the *output* of the learning model. This view allows adapting ideas from constrained optimization, to devise ways in which training can be efficiently performed in a way that at termination, the model parameters correspond to a fair model. For a practitioner, this means that no changes in the architecture or model are needed: imposing fairness only requires specifying the protected attribute, and utilizing our proposed optimization routine.

2. A Primer on Fairness Functions

In this section, we introduce basic notations and briefly review several fairness measures described in the literature.

Basic notations. We denote classifiers using $h : x \mapsto y$ where x and y are random variables that represent the features and labels respectively. A *protected* attribute is a random variable s on the same probability space as x and y – for example, s may be gender, age, or race. Collectively, a training example would be $z := (x, y, s)$. So, our goal is to learn h (predict y given x) while *imposing fairness-type constraints* over s. We will use $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ to denote a set/family of possible classifiers and Δ^N to denote the probability simplex in $\mathbb{R}^N$, i.e., $\Delta := \{q : \sum_{i=1}^N q_i = 1, q_i \geq 0\}$ where q_i is the i-th coordinate of q.

Throughout the paper, we will assume that the distribution of s has finite support. Unless explicitly specified, we will assume that $y \in \{0, 1\}$ in the main paper. For each $h \in \mathcal{H}$, we will use e_h to denote the misclassification rate of h and $e_{\mathcal{H}} \in \mathbb{R}^N$ to be the vector containing all misclassification rates. We will use superscript to denote conditional expectations. That is, if μ_h corresponds to expectation of some function μ (that depends on $h \in \mathcal{H}$), then the conditional expectation/moment of μ_h with respect to s will be denoted by μ_h^s. With a slight abuse of notation, we will use $\mu_h^{s_0}$ to denote the elementary conditional expectation $\mu_h|(s = s_0)$ whenever it is clear from the context. We will use d_h to denote the *difference* between the conditional expectation of the two groups of s, that is, $d_h := \mu_h^{s_0} - \mu_h^{s_1}$. For example, let s be the random variable representing gender, that is, s_0 and s_1 may correspond to male and female. Then, $e_h^{s_i}$ corresponds to the misclassification rate of h on group s_i, and $d_h = e_h^{s_0} - e_h^{s_1}$. Finally, $\mu_h^{s_i,t_j} := \mu_h|(s = s_i, t = t_j)$ denotes the elementary con-

ditional expectation with respect to two random variables s, t.

2.1. Fairness through the lens of Confusion Matrix

Recall that a *fairness* constraint corresponds to a performance requirement of a classifier h on subgroups of features x *induced* by a protected attribute s. For instance, say that h predicts the credit-worthiness y of an individual x. Then, we may require that e_h be "approximately" the same across individuals for different races given by s. Does it follow that functions/metrics that are used to evaluate fairness may be written in terms of the error of a classifier e_h *conditioned* on the protected variable s (or in other words e_h^s)? Indeed, it does turn out to be the case! In fact, many widely used functions in practice can be viewed as imposing constraints on the confusion matrix as our intuition suggests. We will now discuss few common fairness metrics to illustrate this idea.

(a) Demographic Parity (DP) [37]. A classifier h is said to satisfy Demographic Parity (DP) if $h(x)$ is *independent* of the protected attribute s. Equivalently, h satisfies DP if $d_h = 0$ where we set $\mu_h^{s_i} = e_h^{s_i}$ (using notations introduced above). DP can be seen as equating the total false positives and false negatives between the confusion matrices of the two groups. We denote DDP by the difference of the demographic parity between the two groups.

(b) Equality of Opportunity (EO) [23]. A classifier h is said to satisfy EO if $h(x)$ is independent of the protected attribute s for $y \in \{0, 1\}$. Equivalently, h satisfies EO if $d_h^y = 0$ where we set $\mu_h^{s_i} = e_h^{s_i}|(y \in \{0,1\}) =: e_h^{s_i, y_j}$ conditioning on both s and y. Depending on the choice of y in $\mu_h^{s_i}$, we get two different metrics: (i) $y = 0$ corresponds to h with equal *False Positive Rate (FPR)* across s_i [14], whereas (ii) $y = 1$ corresponds to h with equal *False Negative Rate (FNR)* across s_i [14]. Moreover, h satisfies *Equality of Odds* if $d_h^0 + d_h^1 = 0$, i.e., h equalizes both TPR and FPR across s [23]. We denote the difference in EO by DEO.

(c) Predictive Parity (PP) [11]. A classifier h satisfies PP if the likelihood of making a misclassification among the positive predictions of the classifier is independent of the protected variable s. Equivalently, h satisfies PP if $d_h^{\hat{y}} = 0$ where we set $\mu_{h_i}^{s_i} = e_h^{s_i}|(\hat{y} = 1)$. It corresponds to matching the False Discovery Rate between the confusion matrices of the two groups.

3. How to learn fair models?

At a high level, the optimization problem that we seek to solve is written as,

$$\min_{h \in \mathcal{H}} \mathbb{E}_{z:(x,y,s)\sim\mathcal{D}} \mathcal{L}(h; (x, y)) \text{ subject to } h \in \mathcal{F}_{d_h}, \quad (1)$$

where $\mathcal{L}$ denotes the loss function that measures the accuracy of h in predicting y from x, and $\mathcal{F}_{d_h}$ denotes the set of *fair* classifiers. Our approach to solve (1) *provably efficiently* involves two main steps: (i) first, we reformulate problem (1) to compute a posterior distribution q over $\mathcal{H}$; (ii) second, we incorporate fairness as *soft* constraints on the output of q using the augmented Lagrangian of Problem (1). We assume that we have access to sufficient number of samples to approximate $\mathcal{D}$ and solve the empirical version of Problem (1).

3.1. From Fair Classifiers to Fair Posteriors

The starting point of our development is based on the following simple result that follows directly from the definitions of fairness metrics in Section 2:

Observation 1. Fairness metrics such as DP/EO are linear functions of h, whereas PP takes a linear fractional form due to the conditioning on $\hat{y}$, see [11].

Observation 1 immediately implies that $\mathcal{F}_{d_h}$ can be represented using linear (fractional) equations in h. To simplify the discussion, we will focus on the case when $\mathcal{F}_{d_h}$ is given by the DP metric. Hence, we can reformulate (1) as,

$$\min_{q \in \Delta} \quad \sum_i q_i e_{h_i} \text{ s.t. } q_i(\mu_{h_i}^{s_0} - \mu_{h_i}^{s_1}) = 0 \quad \forall i \in [N], \quad (2)$$

where q represents a distribution over $\mathcal{H}$.

3.2. Imposing Fairness via Soft Constraints

In general, there are two ways of treating the N constraints $q_i d_{h_i} = 0$ in Problem (2) viz., (i) as *hard constraints*; or (ii) as *soft constraints*. Algorithms that can handle explicit constraints efficiently require access to an efficient oracle that can minimize a linear or quadratic function over the feasible set in *each* iteration. Consequently, algorithms that incorporate hard constraints come with high per-iteration computational cost since the number of constraints is (at least) linear in N, and is not applicable in large scale settings. Hence, we propose to use algorithms that incorporate fairness as soft constraints. With these two minor modifications, we will now describe our approach to solve problem (2).

4. Fair Posterior from Proximal Dual

Following the reductions approach in [1], we first write the Lagrangian dual problem of DP constrained risk minimization problem (2) using dual variables λ as,

$$\max_{\lambda \in \mathbb{R}^N} \min_{q \in \Delta} L(q, \lambda) := \langle q, e_h \rangle + \lambda \langle q, \mu_h^{s_0} - \mu_h^{s_1} \rangle \quad (3)$$

Interpreting the Lagrangian. Problem 3 can be understood as a game between two players a q-player and a λ-player [16]. We recall an important fact regarding the dual problem (3):

Algorithm 1 FairALM: Linear Classifier

Notations: Dual step size η
  $h_t \in \{h_1, h_2, \ldots, h_N\}$.
Input: Error Vector $e_{\mathcal{H}}$,
  Conditional mean vector $\mu^s_{\mathcal{H}}$
Initializations: $\lambda_0 = 0$
for $t = 0, 1, 2, ..., T$ **do**
  (Primal) $h_t \leftarrow \text{argmin}_i(e_{h_i} + \lambda_t(\mu^{s_0}_{h_i} - \mu^{s_1}_{h_i}))$
  (Dual) $\lambda_{t+1} \leftarrow \lambda_t + \eta(\mu^{s_0}_{h_t} - \mu^{s_1}_{h_t})/t$
end for
Output: h_T

Fact 2. The objective function of the dual problem (3) is *always nonsmooth* with respect to λ because of the inner minimization problem in q.

Technically, there are two main reasons why optimizing nonsmooth functions can be challenging [18]: (i) finding a descent direction in high dimensions N can be challenging; and (ii) subgradient methods can be slow to converge in practice. Due to these difficulties arising from Fact 2, using a first order algorithm such as gradient descent to solve the dual problem in (3) directly can be problematic, and may be suboptimal.

Accelerated optimization using Dual Proximal Functions. To overcome the difficulties due to the nonsmoothness of the dual problem, we propose to *augment* the Lagrangian with a proximal term. Specifically, for some λ_T, the augmented Lagrangian function can be written as,

$$L_T(q, \lambda) = \langle q, e_h \rangle + \lambda \langle q, \mu^{s_0}_h - \mu^{s_1}_h \rangle - \tfrac{1}{2\eta}(\lambda - \lambda_T)^2 \quad (4)$$

Note that, as per our simplified notation, $L_T \equiv L_{\lambda_T}$. The following lemma relates the standard Lagrangian in (3) with its proximal counterpart in (4).

Lemma 3. *At the optimal solution (q^*, λ^*) to L, we have* $\max_\lambda \min_{q \in \Delta} L = \max_\lambda \min_{q \in \Delta} L_{\lambda^*}$.

This is a standard property of proximal objective functions, where λ^* forms a fixed point of $\min_{q \in \Delta} L_{\lambda^*}(q, \lambda^*)$ (section 2.3 of [30]). Intuitively, Lemma 3 states that L and L_T are not at all different for optimization purposes.

Remark 4. While the augmented Lagrangian L_T still may be nonsmooth, the proximal (quadratic) term can be exploited to design *provably* faster optimization algorithms as we will see shortly.

5. Our Algorithm – FairALM

It is common [1, 16, 27] to consider the minimax problem in (4) as a zero sum game between the λ-player and the q-player. The Lagrangian(s) L_T (or L) specify the cost which the q-player pays to the λ-player after the latter makes its choice. An iterative procedure leads to a regret minimizing strategy for the λ-player [34] and a best response strategy for the q-player [1]. While the q-player's move relies on the availability of an efficient *oracle* to solve the minimization problem, $L_T(q, \lambda)$, being a linear program in q makes it less challenging. We describe our algorithm in Alg. 1 and call it *FairALM: Linear Classifier*.

5.1. Convergence Analysis

As the game with respect to λ is a maximization problem, we get a reverse regret bound as shown in the following Lemma. Due to space, proofs appear in the Appendix.

Lemma 5. *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 round $T+1$ to maximize the cumulative reward: $\lambda_{T+1} = \text{argmax}_\lambda \sum_{t=1}^T f_t(\lambda)$. Define $L = \max_t |r_t|$. The following bound on the cumulative reward holds, for any λ*

$$\textstyle\sum_{t=1}^T \left(\lambda r_t - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \leq \sum_{t=1}^T \lambda_t r_t + \frac{\eta}{2} L^2 \mathcal{O}(\log T) \quad (5)$$

The above lemma indicates that the cumulative reward grows in time as $\mathcal{O}(\log T)$. The proximal term in the augmented Lagrangian gives us a *better* bound than an ℓ_2 or an entropic regularizer (which provides a $\sqrt{T}$ bound [34]).

Next, we evaluate the cost function $L_T(q, \lambda)$ after T rounds of the game. We observe that the average play of both the players converges to a saddle point with respect to $L_T(q, \lambda)$. We formalize this in the following theorem,

Theorem 6. *Recall that d_h represents the difference of conditional means. Assume that $||d_h||_\infty \leq L$ and consider T rounds of the game described above. Let the average plays of the q-player be $\bar{q} = \frac{1}{T}\sum_{t=1}^T q_t$ and the λ-player be $\bar{\lambda} = \frac{1}{T}\sum_{t=1}^T \lambda_t$. Then under the following conditions on q, λ and η, we have $L_T(\bar{q}, \bar{\lambda}) \leq L_T(q, \bar{\lambda}) + \nu$ and $L_T(\bar{q}, \bar{\lambda}) \geq L_T(\bar{q}, \lambda) - \nu$*

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

The above theorem indicates that the average play of the q-player and the λ-player reaches a ν-approximate saddle point. Our bounds for $\nu = \frac{1}{T}$ and $\lambda \in \mathbb{R}$ are strictly better than [1].

5.2. Can we train Fair Deep Neural Networks by adapting Alg. 1?

The key difficulty from the analysis standpoint we face in extending these results to the deep networks setting is

Algorithm 2 FairALM: DeepNet Classifier

Notations: Dual step size η, Primal step size τ
Input: Training Set D
Initializations: $\lambda_0 = 0$, w_0
for $t = 0, 1, 2, ..., T$ **do**
  Sample $z \sim D$
  Pick $v_t \in \partial\Big(\hat{e}_{h_w}(z) + (\lambda_t + \eta)\hat{\mu}^{s_0}_{h_w}(z) - (\lambda_t - \eta)\hat{\mu}^{s_1}_{h_w}(z)\Big)$
  (Primal) $w_t \leftarrow w_{t-1} - \tau v_t$
  (Dual) $\lambda_{t+1} \leftarrow \lambda_t + \eta(\hat{\mu}^{s_0}_{h_{w_t}}(z) - \hat{\mu}^{s_1}_{h_{w_t}}(z))$
end for
Output: w_T

that the number of classifiers $|\mathcal{H}|$ may be exponential in number of nodes/layers. This creates a potential problem in computing Step 5 of Algorithm 1 – if viewed mechanistically, is not practical since an epsilon net over the family $\mathcal{H}$ (representable by a neural network) is exponential in size. Interestingly, notice that we often use over-parameterized networks for learning. This is a useful fact here because it means that there exists a solution where $\mathrm{argmin}_i(e_{h_i} + \lambda_t d_{h_i})$ is 0. While iterating through all h_is will be intractable, we may still able to obtain a solution via standard stochastic gradient descent (SGD) procedures [42]. The only unresolved question then is if we can do posterior inference and obtain classifiers that are "fair". It turns out that the above procedure provides us an approximation if we leverage two facts: first, SGD can find the minimum of $L(h, \lambda)$ with respect to h and second, recent results show that SGD, in fact, performs variational inference, implying that the optimization provides an approximate posterior [12]. Having discussed the issue of the exponential sized $|\mathcal{H}|$ – for which we settle for an approximate posterior – we make three additional adjustments to the algorithm to make it suitable for training deep networks. First, the non-differentiable indicator function $\mathbb{1}[\cdot]$ is replaced with a smooth surrogate function (such as a logistic function). Second, as it is hard to evaluate e_h/μ^s_h due to unavailability of the true data distribution, we instead calculate their empirical estimates $z = (x; y; s)$, and denote it by $\hat{e}_h(z)/\hat{\mu}^s_h(z)$. Third, by exchanging the "$\max$" and "$\min$" in (3), we obtain an objective that *upper-bounds* our current objective in (3). This provides us with a closed-form solution to λ thus reducing the minmax objective to a single simpler minimization problem. We present the algorithm for deep neural network training in Alg. 2 and call it *FairALM: DeepNet Classifier.*

6. Experiments

A central theme in our experiments is to assess whether our proposed algorithm, FairALM, can indeed obtain meaningful fairness measure scores *without* compromising the test set performance. We evaluate FairALM on a number of problems where the dataset reflects certain inherent societal/stereotypical biases. Our evaluations are also designed with a few additional goals in mind.

Overview. Our **first** experiment on the CelebA dataset seeks to predict the value of a label for a face image while controlling for certain protected attributes (gender, age). We discuss how prediction of some labels is *unfair* in an unconstrained model and contrast with our FairALM. Next, we focus on the label where predictions are the most unfair and present comparisons against methods available in the literature. For our **second** experiment, we use the ImSitu dataset where images correspond to a situation (activities, verb). Expectedly, some activities such as driving or cooking are more strongly associated with a specific gender. We inspect if an unconstrained model is *unfair* when we ask it to learn to predict two gender correlated activities/verbs. Comparisons with baseline methods will help measure FairALM's strengths/weaknesses. We can use heat map visualizations to qualitatively interpret the value of adding fairness constraints. We threshold the heat-maps to get an understanding of a general behavior of the models. Our **third** experiment addresses an important problem in medical/scientific studies. Small sample sizes necessitate pooling data from multiple sites or scanners [46], but introduce a site or scanner specific nuisance variable which must be controlled for – else a deep (also, shallow) model may cheat and use site specific (rather than disease-specific) artifacts in the images for prediction even when the cohorts are age or gender matched [19]. We study one simple setting here: we use FairALM to mitigate site (hospital) specific differences in predicting "tuberculosis" from X-ray images acquired at two hospitals, Shenzhen and Montgomery (and recently made publicly available [25]).

In all the experiments, we impose Equality of Opportunity (EO) constraint (defined in Section 2.1). We adopt NVP (novel validation procedure) used in [17] to evaluate FairALM. It is a two-step procedure: first, we search for the hyper-parameters that achieve the best accuracy, and then, we report the minimum fairness measure (DEO) for accuracies within 90% of the highest accuracy. This offers some robustness of the reported numbers to hyper-parameter selection. We describe these experiments one by one.

Remark from authors. Certain attributes such as *attractiveness*, obtained via crowd-sourcing, may have socio-cultural ramifications. Similarly, the gender attribute in the dataset is binary (male versus female) which may be insensitive to some readers. We clarify that our goal is to present evidence showing that our algorithm can impose fairness in a sensible way on datasets used in the literature rather than the higher level question of whether

Protected: **GENDER**			Protected: **YOUNG**		
Label	**U**	**F**	Label	**U**	**F**
Attractive	28	3	Attractive	8	1
Bangs	4	2	Heavy Makeup	11	1
High Cheekbones	18	0	High Cheekbones	7	0
Mouth Slightly open	11	3	Male	6	0
Smiling	10	0	Wearing Lipstick	12	4

Table 1: **Identifying Unfair Labels in CelebA dataset.** We report the DEO measure for the Unconstrained model (**U**) and FairALM model (**F**). Using a 3-layers ReLU network, we determine the labels in CelebA dataset that are biased with respect to gender (left) and the attribute young (right). Labels with a precision of at least 70% and a DEO of at least 4% on the unconstrained model are reported here.

our community needs to invest in culturally sensitive datasets with more societally relevant themes.

6.1. CelebA dataset

Data and Setup. CelebA [28] consists of 200K celebrity face images from the internet annotated by a group of paid adult participants [7]. There are up to 40 labels available in the dataset, each of which is binary-valued.

Quantitative results. We begin our analysis by predicting each of the 40 labels with a 3-layer ReLU network. The protected variable, s, are the binary attributes like *Male* and *Young* representing gender and age respectively. We train the SGD algorithm for 5-epochs and select the labels predicted with at least at 70% precision and with a DEO of at least 4% across the protected variables. The biased set of labels thus estimated are shown in Table 1. These labels are consistent with other reported results [32]. It is important to bear in mind that the bias in the labels should not be attributed to its relatedness to a specific protected attributed alone. The cause of bias could also be due to the skew in the label distributions. When training a 3-layer ReLU net with FairALM, the precision of the model remained almost the same ($\pm 5\%$) while the DEO measure reduced significantly as indicated in the Table 1. Next, choosing the most unfair label in Table 1 (i.e., attractive), we train a ResNet18 for a longer duration of about 100 epochs and contrast the

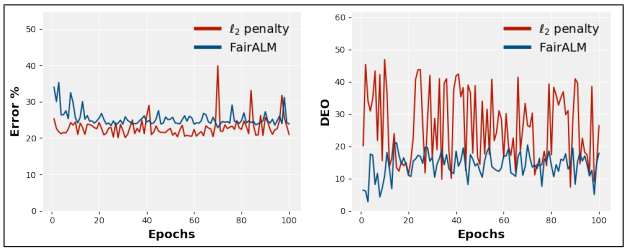


Figure 2: **Comparison to ℓ_2 penalty .** FairALM has a stable training profile in comparison to naive ℓ_2 penalty. The target label is *attractiveness* and protected attribute is *gender*.

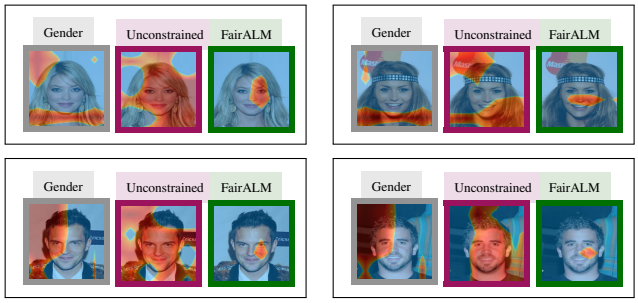


Figure 3: **Interpretable Models for CelebA.** Unconstrained/FairALM predict label *attractiveness* while controlling *gender*. The heatmaps of Unconstrained model overlaps with gender classification task indicating gender leak. FairALM consistently picks non-gender revealing features of the face. Interestingly, these regions are on the left side in accord with psychological studies that the Face's left side is more attractive [6].

performance with a simple ℓ_2-penalty baseline. The training profile is observed to be more stable for FairALM as indicated in Fig. 2. This finding is consistent with the seminal works such as [5, 29] that discuss the ill-conditioned landscape of non-convex penalties. Comparisons to more recent works such as [33, 31] is provided in Table 2. Here, we present a new state-of-the-art result for the DEO measure with the label *attractive* and protected attribute *gender*.

Qualitatively assessing Interpretability. While the DEO measure obtained by FairALM is lower, we can ask an interesting question: when we impose the fairness constraint, precisely which aspects of the image are no longer "legal" for the neural network to utilize? This issue can be approached via visualizing activation maps from models such as CAM [45]. As a representative example, our analysis suggests that in general, an unconstrained model uses the entire face image (including the gender-revealing parts). We find some consistency between the activation maps for *attractiveness* and activation maps of an unconstrained model trained to predict *gender*! In contrast, when we impose the fairness constraint, the corresponding activation maps turn out to be clustered around specific regions of the face which are *not* gender revealing. In particular, a surprising finding was that the left regions in the face were far more prominent which turns out to be consistent with studies in psychology [6].

	Fairness GAN[33]	Quadrianto etal[31]	**FairALM**
ERR	26.6	24.1	24.5
DEO	22.5	12.4	**10.4**
FNR Female	21.2	12.8	**6.6**
FNR Male	43.7	25.2	**17.0**

Table 2: **Quantitative Results on CelebA.** FairALM attains a lower DEO measure and improves the testset errors (ERR). The target label is *attractiveness* and protected attribute is *gender*.

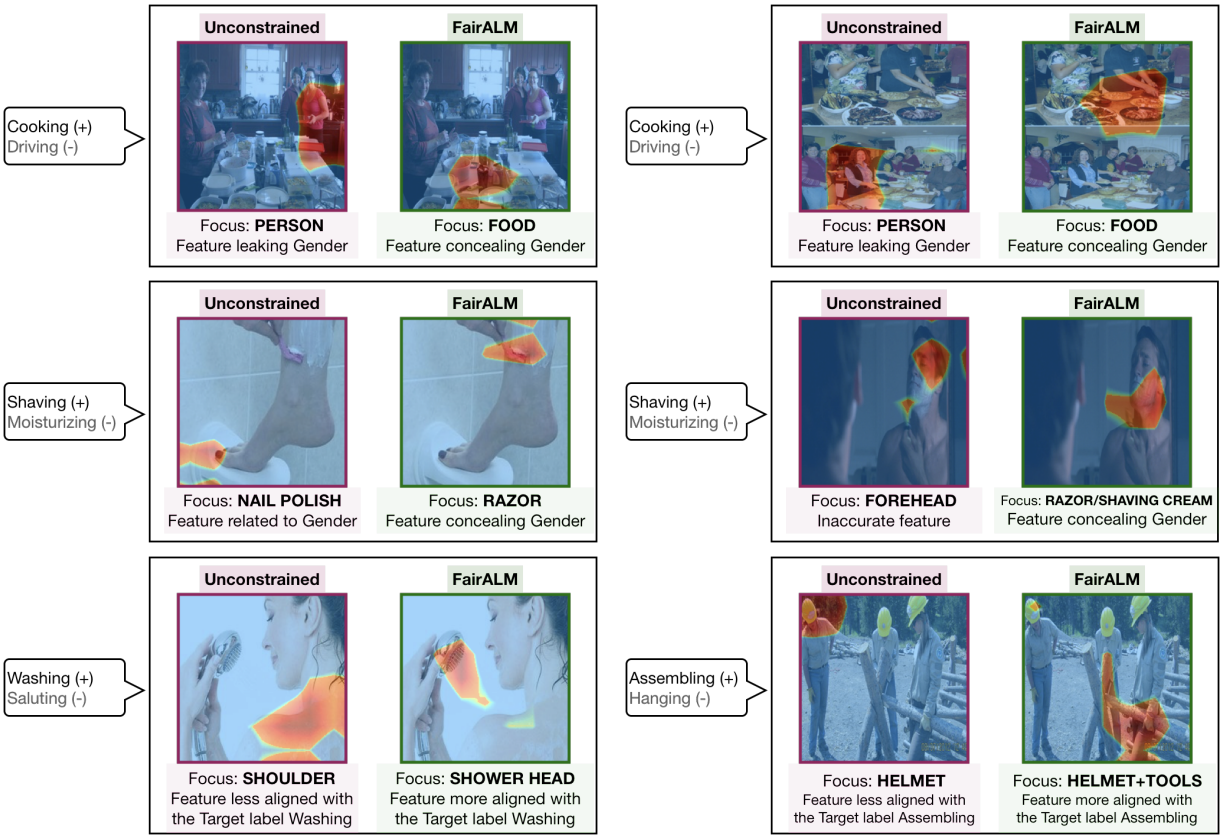


Figure 4: **Interpretability in ImSitu.** The activation maps indicate that FairALM conceals gender revealing attributes in an image. Moreover, the attributes are more aligned with label of interest. The target class predicted is indicated by a $+$. The activation maps in the examples shown in this figure are representative of the general behavior on this dataset. More examples can be found in the Appendix.

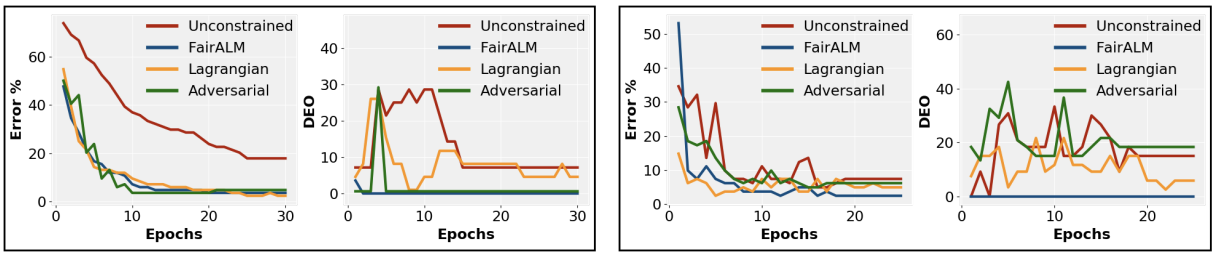
Summary. FairALM minimized the DEO measure without compromising the test error. It has a more stable training profile than an ℓ_2 penalty and is competitive with recent fairness methods in vision. The activation maps in FairALM concentrate on non-gender revealing features of the face when controlled for gender.

6.2. Imsitu Dataset

Data and Setup. ImSitu [38] is a situation recognition dataset consisting of $\sim$ 100K color images taken from the web. The annotations for the image is provided as a summary of the activity in the image and includes a verb describing it, the interacting agents and their roles. The protected variable in this experiment is gender. Our objective is to classify a pair of verbs associated with an image. The pair is chosen such that if one of the verbs is biased towards males then the other would be biased towards females. The authors in [44] report the list of labels in the ImSitu dataset that are gender biased: we choose our verb pairs from this list. In particular, we consider the verbs *Cooking vs Driving*, *Shaving vs Moisturizing*, *Washing vs Saluting* and *Assembling vs Hanging*. We compare our results against multiple baselines such as **(1)** Unconstrained **(2)** ℓ_2*-penalty*, the penalty applied on the DEO measure **(3)** *Re-weighting*, a weighted loss functions where the weights account for the dataset skew **(4)** *Adversarial* [41] **(5)** *Lagrangian* [44] **(6)** *Proxy-Lagrangian* [15]. The supplement includes more details of the baseline methods.

Quantitative results. From Fig. 5, it can be seen that FairALM reaches a zero DEO measure very early in training and attains better test errors than an unconstrained model. Within the family of Lagrangian methods such as [44, 15], FairALM performs better on verb pair 'Shaving vs Moisturizing' in both test error and DEO measure as indicated in Table 3 . While the results on the other verb pairs are comparable, FairALM was observed to be more stable to different hyper-parameter choices. This finding is in accord with recent studies by [2] who prove that proximal function models are robust to step-size selection. Detailed analysis is provided in the supplement. Turning now to an adversarial method such as [44], results in Table 3 show that the DEO measure is not controlled as competently as FairALM. Moreover, complicated training routines and unreliable convergence [3] makes model-training harder.

Interpretable Models. We used CAM [45] to inspect the image regions used by the model for target prediction.



(a) Cooking (+) Driving (-)

(b) Assembling (+) Hanging (-)

Figure 5: **Training Profiles.** FairALM achieves minimum DEO early in training and remains competitive on testset errors. More plots in appendix.

	Cooking(+) Driving(-)		Shaving(+) Moisturize(-)		Washing(+) Saluting(-)		Assembling(+) Hanging(-)	
	ERR	DEO	ERR	DEO	ERR	DEO	ERR	DEO
Unconstrained	17.9	7.1	23.6	4.2	12.8	25.9	7.5	15.0
ℓ_2 Penalty	14.3	14.0	23.6	1.3	10.9	0.0	5.0	21.6
Reweight	11.9	3.5	19.0	5.3	10.9	0.0	4.9	9.0
Adversarial	4.8	0.0	13.5	11.9	14.6	25.9	6.2	18.3
Lagrangian	2.4	3.5	12.4	12.0	3.7	0.0	5.0	5.8
Proxy-lagragn.	2.4	3.5	12.4	12.0	3.7	0.0	14.9	26.0
FairALM	3.6	0.0	20.0	0.0	7.3	0.0	2.5	0.0

Table 3: **Quantitative Results on ImSitu.** Test errors (ERR) and DEO measure are reported in %. The target class that is to be predicted in is indicated by a +. FairALM always achieves a zero DEO while remaining competitive in ERR with the best method for a given verb-pair.

We observe that the unconstrained model ends up picking features from locations that may not be relevant for the task description but merely co-occur with the verbs in this particular dataset (and are gender-biased). Fig. 4 highlights this observation for the selected classification tasks. Overall, we observe that the semantic regions used by the constrained model are more aligned with the action verb present in the image, and this adds to the qualitative advantages of the model trained using FairALM in terms of interpretability.

Limitations. We also note that there are cases where both the unconstrained model and FairALM look at incorrect image regions for prediction, owing to the small dataset sizes. However, the number of such cases are far fewer for FairALM than the unconstrained setup.

Summary. FairALM successfully minimizes the fairness measure while classifying verb/action pairs associated with an image. FairALM uses regions in an image that are more relevant to the target class and less gender revealing.

6.3. Chest X-Ray datasets

Data and Setup. The datasets we examine here are publicly available from the U.S. National Library of Medicine [25]. The images come from two sites/sources. Images for the first site are collected from patients in Montgomery county, USA and includes 138 x-rays. The second set of images includes 662 images collected at a hospital in Shenzhen, China. Our task is to predict pulmonary tuberculosis (TB) from the x-ray images. The images are collected from different x-ray machines with different characteristics, and have site-specific markings or artifacts, see Fig 6. 25% of the samples from the pooked dataset are set aside for testing.

Quantitative Results. We treat the site information, Montgomery or Shenzhen, as a nuisance/protected variable and seek to decorrelate it from the TB labels. We train a ResNet18 network and compare an unconstrained model with FairALM model. Our datasets of choice are small in size, and so deep models easily overfit to site-specific biases present in the training data. Our results corroborate this conjecture, the training accuracies reach 100% very early and the test set accuracies for the unconstrained model has a large variance over multiple experimental runs. Conversely, as depicted in Fig. 7, a FairALM model not only maintains a lower variance in the test set errors and DEO measure but also attains improved performance on these measures. What stands out in this experiment is that the number of

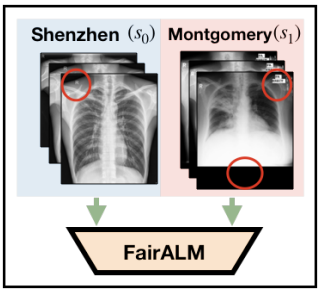


Figure 6: **FairALM for dataset pooling.** Data is pooled from two sites/hospitals, Shenzhen s_0 and Montgomery s_1.

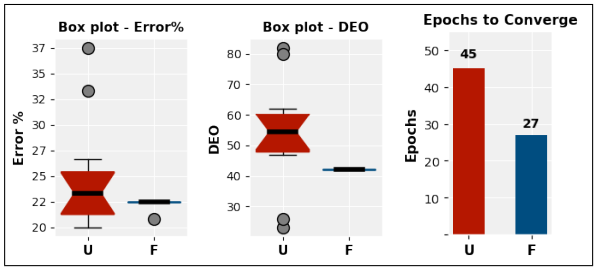


Figure 7: **Better Generalization with FairALM.** We compare Unconstrained mode (**U**) and FairALM (**F**) Box-plots indicate a lower variance in testset error and the DEO measure for FairALM. Moreover, FairALM reaches 20% testset error in fewer epochs.

epochs to reach a certain test set error is lower for FairALM indicating that the model generalizes faster compared to an unconstrained model.

Summary. FairALM is effective at learning from datasets from two different sites/sources, minimizes site-specific biases and accelerates generalization.

7. Conclusion

We introduced FairALM, an augmented Lagrangian framework to impose constraints on fairness measures studied in the literature. On the theoretical side, we provide strictly better bounds – $\mathcal{O}\left(\frac{\log^2 T}{T}\right)$ versus $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$, for reaching a saddle point. On the application side, we provide extensive evidence (qualitative and quantitative) on image datasets commonly used in vision to show the potential benefits of our proposal. Finally, we use FairALM to mitigate site specific differences when performing analysis of pooled medical image datasets. In applying deep learning to scientific/biomedical problems, this is an important issue since sample sizes at individual sites/institutions are often smaller. The overall procedure is simple which we believe will lead to broader adoption and follow-up work on this socially relevant topic.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[2] Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[3] Samuel A Barnett. Convergence problems with generative adversarial networks (gans). *arXiv preprint arXiv:1806.11382*, 2018.

[4] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

[5] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[6] Kelsey Blackburn and James Schirillo. Emotive hemispheric differences measured in real-life portraits using pupil diameter and subjective aesthetic preferences. *Experimental Brain Research*, 219(4):447–455, Jun 2012.

[7] Marc Böhlen, Varun Chandola, and Amol Salunkhe. Server, server in the cloud. who is the fairest in the crowd? *arXiv preprint arXiv:1711.08801*, 2017.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[10] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[11] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328. ACM, 2019.

[12] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*. IEEE, 2018.

[13] Caitlin Chin. Assessing employer intent when ai hiring tools are biased, Dec 2019.

[14] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[15] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. *arXiv preprint arXiv:1804.06500*, 2018.

[16] Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Maya Gupta, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *arXiv preprint arXiv:1809.04198*, 2018.

[17] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[18] John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[19] Alhussein Fawzi and Pascal Frossard. Measuring the effect of nuisance variables on classifiers. pages 137.1–137.12, 01 2016.

[20] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

[21] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[22] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, 38(3):50–57, Oct. 2017.

[23] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[24] Rebecca Heilweil. Artificial intelligence will help determine if you get your next job, Dec 2019.

[25] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest

x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[26] Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6. Citeseer, 2010.

[27] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.

[29] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[30] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[31] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8227–8236, 2019.

[32] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017.

[33] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.

[34] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[35] Berk Ustun and Cynthia Rudin. Learning optimized risk scores from large-scale datasets. *stat*, 1050:1, 2016.

[36] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

[37] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2921–2930, 2017.

[38] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[39] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

[41] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[43] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.

[44] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[45] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

[46] Hao Henry Zhou, Vikas Singh, Sterling C Johnson, Grace Wahba, Alzheimers Disease Neuroimaging Initiative, et al. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proceedings of the National Academy of Sciences*, 115(7):1481–1486, 2018.

[47] Ortrun Zuber-Skerritt and Eva Cendon. Critical reflection on professional development in the social sciences: interview results. *International Journal for Researcher Development*, 5(1):16–32, 2014.

8. Appendix

8.1. Experiments on *FairALM: Linear Classifier*

Data. We consider four standard datasets, `Adult`, `COMPAS`, `German` and `Law Schools` [17, 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 $\$50$K 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 ~ 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 4. 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 [17, 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.* [39]	22.0	5.0	31.0	10.0	38.0	13.0	–	–
Hardt *et al.* [23]	18.0	11.0	29.0	8.0	29.0	11.0	4.5	0.0
Donini *et al.* [17]	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 ± 1	0.7 ± 0.6	34.7 ± 1	0.1 ± 0.1	24.3 ± 2.7	10.8 ± 4.5	4.8 ± 0.1	0.4 ± 0.2

Table 4: **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.

8.2. 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 7. *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,*

$$\sum_{t=1}^{T} \left(\lambda r_t - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \leq \sum_{t=1}^{T} \lambda_t r_t + \eta L^2 \mathcal{O}(\log T) \tag{6}$$

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 [34]. Emulating the proof of Lemma 2.1 in [34], a bound on the negative regret of FTL, for any $\lambda \in \mathbb{R}$, can be derived

due to the concavity of $f_t(\lambda)$,

$$\sum_{t=1}^{T} f_t(\lambda) - \sum_{t=1}^{T} f_t(\lambda_t) \leq \sum_{t=1}^{T} f_t(\lambda_{t+1}) - \sum_{t=1}^{T} f_t(\lambda_t) \tag{7}$$

Our objective, now, is to obtain a bound on RHS of (7). Solving $\text{argmax}_\lambda \sum_{i=1}^{t} f_i(\lambda)$ for λ will show us how λ_t and λ_{t+1} are related,

$$\lambda_{t+1} = \frac{\eta}{t}\sum_{i=1}^{t} r_i + \frac{1}{t}\sum_{i=1}^{t}\lambda_i \quad \implies \lambda_{t+1} - \lambda_t = \frac{\eta}{t} r_t \tag{8}$$

Using (8), 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 $\sum_{i=1}^{T} \frac{1}{i} \leq (\log T + 1)$,

$$\sum_{t=1}^{T} \Big(f_t(\lambda_{t+1}) - f_t(\lambda_t) \Big) \leq \eta L^2 (\log T + 1) \tag{9}$$

Let us denote $\xi_T = \eta L^2(\log T + 1)$, we bound (7) with (9),

Cumulative Reward Bound

$$\forall \lambda \in \mathbb{R} \quad \sum_{t=1}^{T} \Big(\lambda r_t - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \Big) \leq \Big(\sum_{t=1}^{T} \lambda_t r_t \Big) + \xi_T \tag{10}$$

□

Next, using the *Cumulative Reward Bound* (10), 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 8. *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 $L_T(\bar{q}, \lambda) \leq L_T(q, \bar{\lambda}) + \nu$ and $L_T(\bar{q}, \lambda) \geq L_T(\bar{q}, \lambda) - \nu$, 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,

$$L_T(q, \lambda) = \big(\sum_i q_i e_{h_i} \big) + \lambda \big(\sum_i q_i d_{h_i} \big) - \frac{1}{2\eta} \big(\lambda - \lambda_T \big)^2 \tag{11}$$

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

$$\zeta_T(\lambda) = \frac{1}{2\eta} \sum_{t=1}^{T} \Big((\lambda - \lambda_t)^2 - (\lambda - \lambda_T)^2 + (\lambda_t - \lambda_T)^2 \Big) \tag{12}$$

Recollect from (10) 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

$$L_T(\bar{q}, \bar{\lambda}) \leq L_T(q, \bar{\lambda}) + \frac{\zeta_T(\bar{\lambda})}{T} + \frac{\xi_T}{T} \quad \forall q \in \Delta \tag{13}$$

$$\text{Also, } L_T(\bar{q}, \lambda) \leq L_T(q, \bar{\lambda}) + \frac{\zeta_T(\lambda)}{T} + \frac{\xi_T}{T} \quad \forall \lambda \in \mathbb{R}, \forall q \in \Delta \tag{14}$$

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

Average Play Lower Bound

$$L_T(\bar{q}, \bar{\lambda}) \geq L_T(\bar{q}, \lambda) - \frac{\zeta_T(\lambda)}{T} - \frac{\xi_T}{T} \quad \forall \lambda \in \mathbb{R} \tag{15}$$

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*

$$L_T(q, \bar{\lambda}) = \textstyle\sum_i q_i e_{h_i} + \Big(\frac{\sum_t \lambda_t}{T}\Big)\Big(\sum_i q_i d_{h_i}\Big) - \frac{1}{2\eta}\Big(\frac{\sum_t \lambda_t}{T} - \lambda_T\Big)^2 \tag{16}$$

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

$$\geq \frac{1}{T}\sum_t \Big(\textstyle\sum_i q_i e_{h_i} + \lambda_t \sum_i q_i d_{h_i} - \frac{1}{2\eta}(\lambda_t - \lambda_T)^2\Big) \tag{17}$$

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,

$$\geq \frac{1}{T}\sum_t \Big(e_{h_t} + \lambda_t d_{h_t} - \frac{1}{2\eta}(\lambda_t - \lambda_T)^2\Big) \tag{18}$$

Using the *Cumulative Reward Bound* (10),

$$\geq \frac{\sum_t e_{h_t}}{T} + \frac{\lambda \sum_t d_{h_t}}{T} - \frac{1}{T}\sum_t \Big(\frac{(\lambda - \lambda_t)^2}{2\eta} + \frac{(\lambda_t - \lambda_T)^2}{2\eta}\Big) - \frac{\xi_T}{T} \tag{19}$$

Add and subtract $\frac{1}{T}\sum_{t=1}^{T} \frac{1}{2\eta}(\lambda - \lambda_T)^2$, use ζ_T from (12) and regroup the terms,

$$= (\textstyle\sum_i \bar{q}_i e_{h_i}) + (\lambda \sum_i \bar{q}_i d_{h_i}) - \frac{1}{2\eta}(\lambda - \lambda_T)^2 - \frac{\zeta_T(\lambda)}{T} - \frac{\xi_T}{T} \tag{20}$$

$$= L_T(\bar{q}, \lambda) - \frac{\zeta_T(\lambda)}{T} - \frac{\xi_T}{T} \tag{21}$$

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* (10). 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 $|\lambda| \leq B$

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

$$\frac{\zeta_T(\lambda)}{T} + \frac{\xi_T}{T} \leq \frac{B^2}{\eta} + \frac{\eta L^2(\log T + 1)}{T} \tag{22}$$

Minimizing R.H.S in (22) over η gives us a $\nu-$ approximate saddle point,

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

$$L_T(\bar{q}, \bar{\lambda}) \leq L_T(q, \bar{\lambda}) + \nu \quad \text{and} \quad L_T(\bar{q}, \bar{\lambda}) \geq L_T(\bar{q}, \lambda) - \nu \tag{23}$$

$$\text{where } \nu = 2\sqrt{\frac{B^2 L^2(\log T + 1)}{T}} \quad \text{and } \eta = \sqrt{\frac{B^2 T}{L^2(\log T + 1)}} \tag{24}$$

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

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,

$$\frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} = \frac{1}{2\eta}\frac{1}{T}\sum_t \Big(2(\lambda_t - \lambda_T)^2 + 2\eta(\lambda_T - \lambda_t)(\textstyle\sum_i \bar{q}_i d_{h_i})\Big) + \frac{\xi_T}{T} \tag{25}$$

Recollect that $\lambda_{t+1} - \lambda_t = \frac{\eta}{t} d_{h_t}$ (from (8)). 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 (25),

$$\frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} \leq \eta L^2(\log T + 1)^2 + \eta L^2(\log T + 1) + \frac{\eta L^2(\log T + 1)}{T} \tag{26}$$

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

$$\frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} \leq \mathcal{O}(\frac{L^2(\log T + 1)^2}{T}) := \nu \tag{27}$$

Using (27), we prove the convergence of λ in the following way,

$$L_T(\bar{q}, \lambda) \leq L_T(\bar{q}, \lambda_*) \quad \Big(\text{as } \lambda_* \text{ is the maximizer of } L_T(\bar{q}, \lambda)\Big) \tag{28}$$

$$\leq L_T(\bar{q}, \bar{\lambda}) + \frac{\zeta_T(\lambda_*)}{T} + \frac{\xi_T}{T} \quad \Big(\textit{Average Play Lower Bound}\Big) \tag{29}$$

$$\leq L_T(\bar{q}, \bar{\lambda}) + \nu \quad \Big(\text{from (27)}\Big) \tag{30}$$

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

$$L_T(q, \bar{\lambda}) \geq L_T(\bar{q}, \lambda_*) - \frac{\zeta_T(\lambda_*)}{T} - \frac{\xi_T}{T} \quad \Big(\textit{Average Play Upper Bound (21)}\Big) \tag{31}$$

$$\geq L_T(\bar{q}, \lambda_*) - \nu \quad \Big(\text{from (27)}\Big) \tag{32}$$

$$\geq L_T(\bar{q}, \bar{\lambda}) - \nu \quad \Big(\text{as } \lambda_* \text{ is the maximizer of } L_T(\bar{q}, \lambda)\Big) \tag{33}$$

Therefore,

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

$$L_T(\bar{q}, \bar{\lambda}) \leq L_T(q, \lambda) + \nu \quad \text{and} \quad L_T(\bar{q}, \bar{\lambda}) \geq L_T(\bar{q}, \lambda) - \nu \tag{34}$$

$$\text{where } \nu = \mathcal{O}\Big(\frac{L^2(\log T + 1)^2}{T}\Big) \quad \text{and} \quad \eta = \frac{1}{T} \tag{35}$$

□

8.3. 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,

$$\max_{\lambda} \min_{w} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \tag{36}$$

We use DP constraint in (36), 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 (36) obtained by exchanging the "max"/"min" operations.

$$\max_{\lambda} \min_{w} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \tag{37}$$

$$\leq \min_{w} \max_{\lambda} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right) \tag{38}$$

Substituting the closed form solution $\lambda = \lambda_t + \eta(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1})$ in (38),

$$\leq \min_{w} \left(\hat{e}_{h_w} + +\lambda_t(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1}) + \frac{\eta}{2}(\hat{\mu}_{h_w}^{s_0} - \hat{\mu}_{h_w}^{s_1})^2 \right) \tag{39}$$

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 (39) to give us a convenient upper bound,

$$\leq \min_{w} \left(\hat{e}_{h_w} + (\lambda_t + \eta)\hat{\mu}_{h_w}^{s_0} - (\lambda_t - \eta)\hat{\mu}_{h_w}^{s_1} \right) \tag{40}$$

In order to obtain a good minima in (40), it may be essential to run the SGD on (40) a few times: for ImSitu experiments, SGD was run on (40) 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 [5] (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.

8.4. 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 [15] 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: $\eta_i \propto 1/(\#\text{ samples in } s_i)$

Lagrangian [44]

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 [15]

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

8.5. 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 8 and Figure 9 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*.

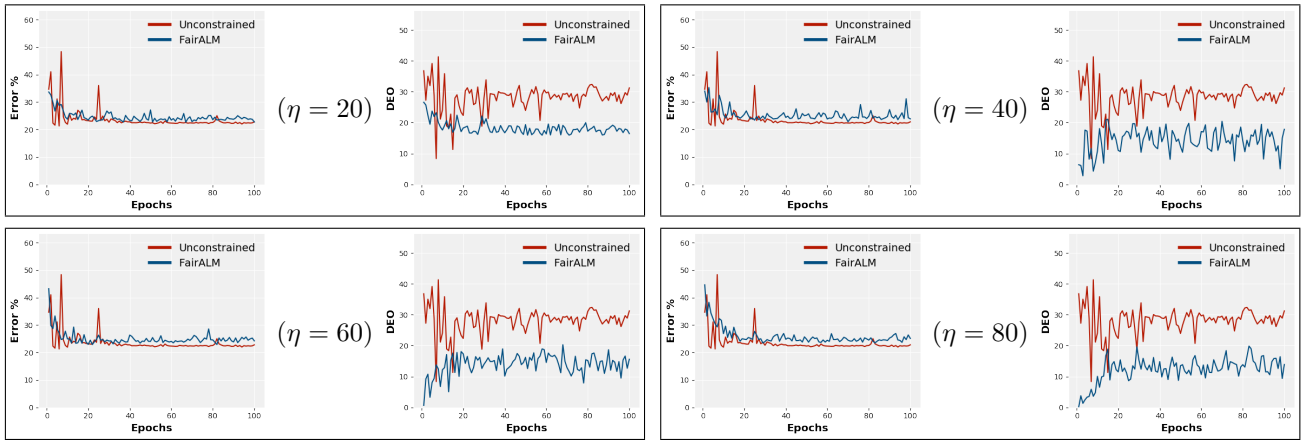


Figure 8: **FairALM Ablation on CelebA.** For a given η, the left image represents the test error and the right image shows the DEO measure. We study the effect of varying the dual step size η on FairALM. We observe that the performance of FairALM is consistent over a wide range of η values.

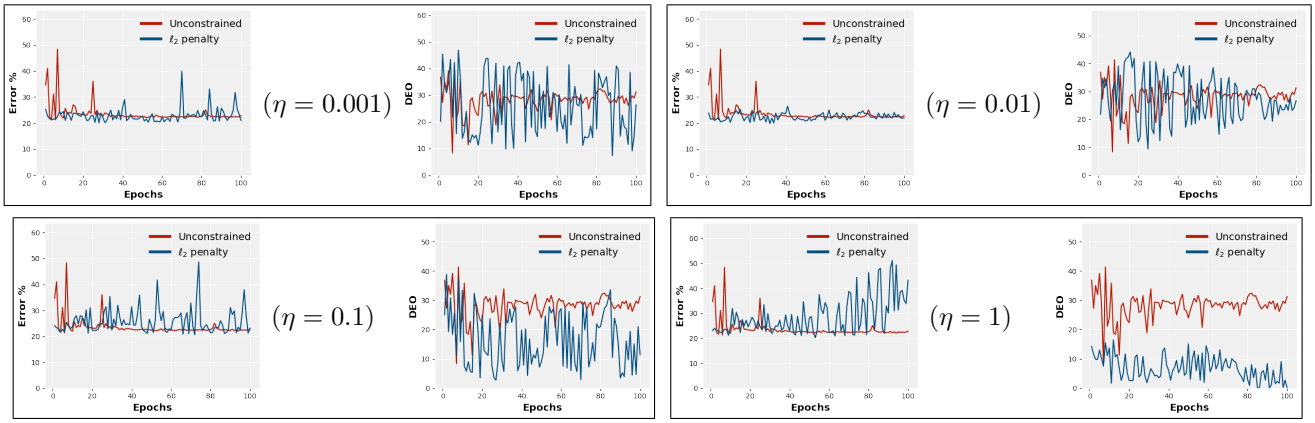


Figure 9: **ℓ_2 Penalty Ablation on CelebA** For each η value, the left image represents the test set errors and the right image shows the fairness measure (DEO). We investigate a popular baseline to impose fairness constraint which is the ℓ_2 penalty. We study the effect of varying the penalty parameter 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 10. 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.

Figure 10: **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.

8.6. 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 $t_j = \textit{cooking}$ 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. 11 below supplement Fig. 5 in the paper.

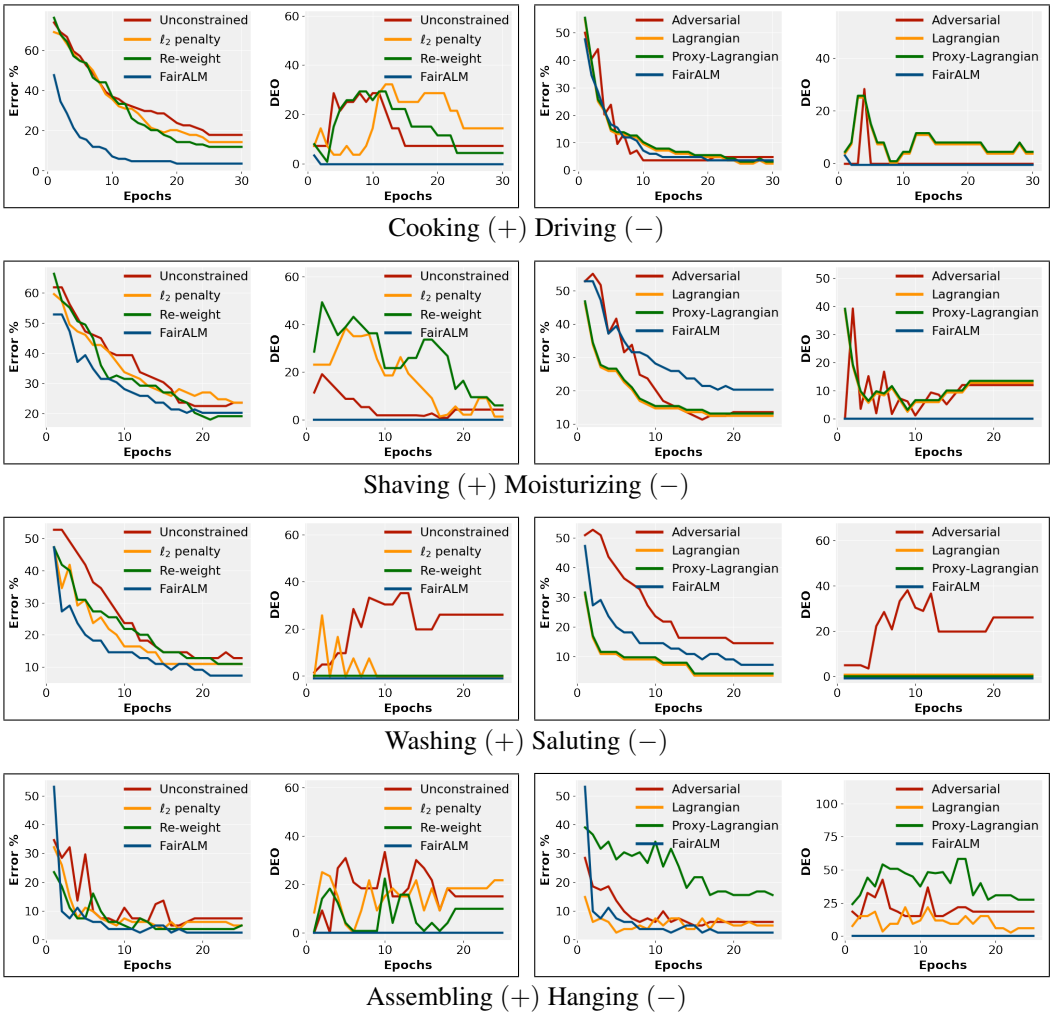


Figure 11: **Supplementary Training Profiles.** FairALM consistently achieves minimum DEO across different verb pair classifications.

Additional qualitative results We show the activation maps in Fig. 12 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 [44]. 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.

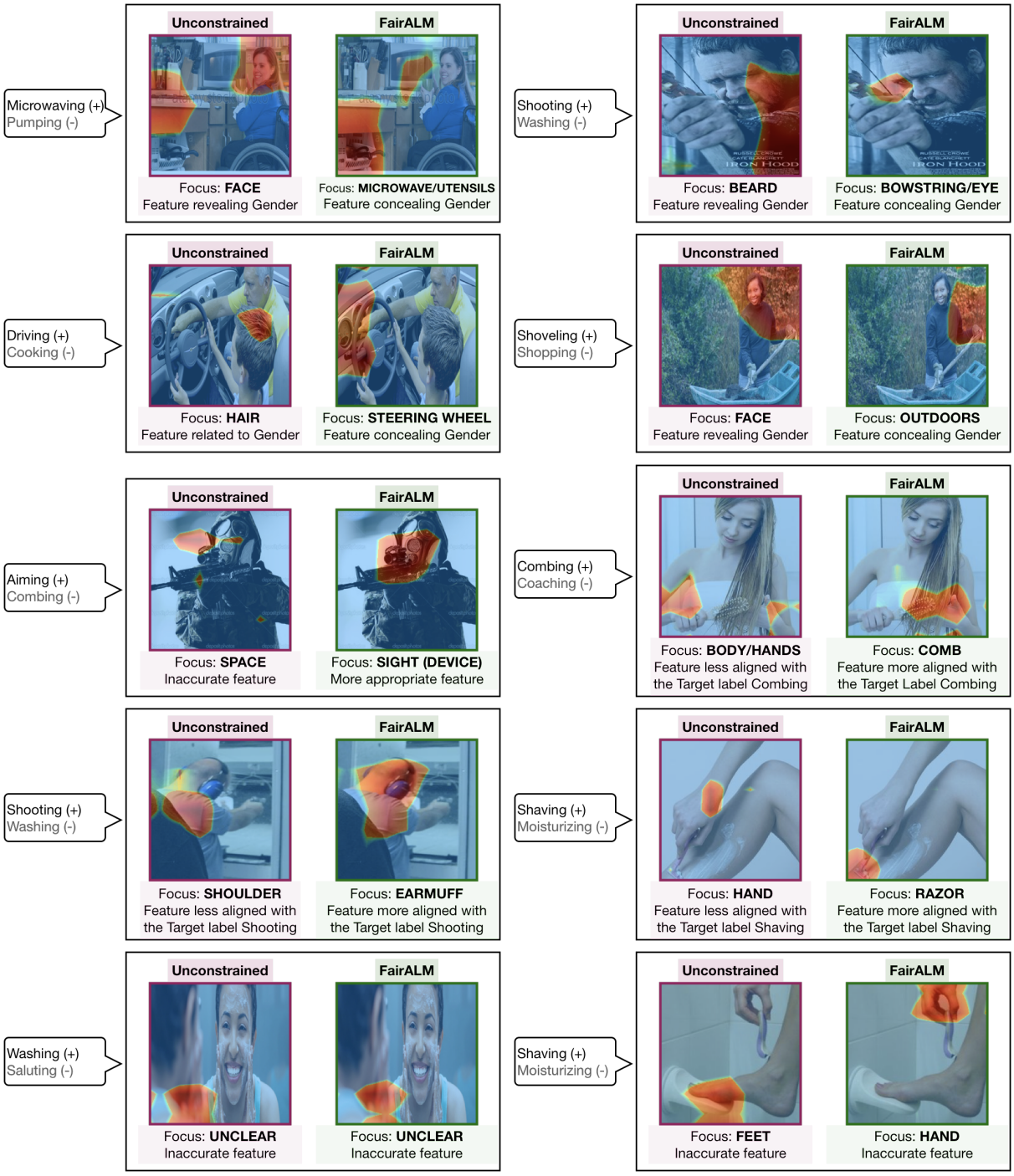


Figure 12: **Additional qualitative Results in ImSitu dataset.** Models predict the target class (+). FairALM consistently avoids gender revealing features and uses features that are more relevant to the target class. Due to the small dataset sizes, a *limitation* of this experiment is shown in the last 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.