On the Alignment of Group Fairness with Attribute Privacy

Jan Aalmoes
jan.aalmoes@insa-lyon.fr
Univ Lyon, INSA Lyon, Inria, CITI
Lyon, France

Vasisht Duddu
vasisht.duddu@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Antoine Boutet
antoine.boutet@insa-lyon.fr
Univ Lyon, INSA Lyon, Inria, CITI
Lyon, France

ABSTRACT

Group fairness and privacy are fundamental aspects in designing trustworthy machine learning models. Previous research has highlighted conflicts between group fairness and different privacy notions. We are the first to demonstrate the alignment of group fairness with the specific privacy notion of attribute privacy in a blackbox setting. Attribute privacy, quantified by the resistance to attribute inference attacks (AIAs), requires indistinguishability in the target model’s output predictions. Group fairness guarantees this thereby mitigating AIAs and achieving attribute privacy. To demonstrate this, we first introduce AdaptAIA, an enhancement of existing AIAs, tailored for real-world datasets with class imbalances in sensitive attributes. Through theoretical and extensive empirical analyses, we demonstrate the efficacy of two standard group fairness algorithms (i.e., adversarial debiasing and exponentiated gradient descent) against AdaptAIA. Additionally, since using group fairness results in attribute privacy, it acts as a defense against AIAs, which is currently lacking. Overall, we show that group fairness aligns with attribute privacy at no additional cost other than the already existing trade-off with model utility.

1 INTRODUCTION

Machine learning (ML) has been adopted for several high-stakes decision-making applications, such as criminal justice and healthcare. This raises concerns about the model’s discriminatory behaviour across different demographic subgroups and compromising data privacy. Several regulations require ML practitioners to train models satisfying both privacy and group fairness. To design such models, practitioners have to first understand the interactions between group fairness and privacy.

Prior works have shown that group fairness conflicts with different privacy notions such as differential privacy, membership inference, and distribution inference. It is unclear about how group fairness interacts with privacy of sensitive attributes (hereafter referred to as attribute privacy) which ensures that an adversary cannot exploit distinguishability in target model’s predictions to infer sensitive attributes. In this work, we ask: How does group fairness relate to attribute privacy?

To answer this, we first quantify attribute privacy as the resistance to attribute inference attacks (AIAs) where an adversary infers sensitive attributes (e.g., Race and Sex), from model observables (e.g., predictions, intermediate outputs). These attacks pose a practical threat to data privacy and confidentiality. For example, inferring Race from a model predicting mortgage eligibility, even if not part of the training data, could violate customer privacy and potentially lead to discrimination. We focus on representation-based AIAs in a practical blackbox setting where an adversary exploits distinguishability in model observables for different sensitive attribute values. For instance, Figure 1 depicts distinguishable prediction distributions for two sub-groups.

Figure 1: Privacy is violated if an (potentially sensitive) attribute is inferred from the output predictions of a learning model if these predictions are distinguishable for different attribute values (Appendix B).

Ensuring attribute privacy requires the success of AIAs to be close to random guessing, meaning model observables should be indistinguishable for different sensitive attribute values. Group fairness achieves this by ensuring indistinguishability in predictions. Implementations of such group fairness include adversarial debiasing (AdvDebias) and exponentiated gradient descent (EGD). Hence, we conjecture that group fairness is aligned with attribute privacy and can be used as a defense against AIAs, which is currently lacking.

Contributions. We are the first to show that the objective of group fairness aligns with attribute privacy. In other words, using group fairness results in attribute privacy at no additional cost except for the already existing trade-off with utility. To demonstrate this, we first propose a state-of-the-art AIA, AdaptAIA, designed to infer sensitive attributes while accounting for class imbalance (Section 5). Finally, we demonstrate, theoretically and through extensive empirical evaluation, the effectiveness of EGD and AdvDebias against AdaptAIA (Sections 6 and 7). Our source code will be publicly available upon publication.

2 BACKGROUND

We present a background on ML and notations used in the rest of the paper (Section 2.1), followed by a discussion on AIAs (Section 2.2), and group fairness (Section 2.3).

2.1 Machine Learning Classifiers

Training. ML classifiers are functions \( f^\theta_{tr} \) (omit \( \theta \) for simplicity) parameterized by \( \theta \) that map inputs with corresponding classification labels. \( \theta \) is updated using a training dataset \( (\mathcal{D}_{tr}) \) with the objective to minimize the loss incurred on predicting the classification label for inputs from \( \mathcal{D}_{tr} \). We remove sensitive attributes such...
Table 1: Comparison of prior AIs: attack vector exploited (e.g., \( f_{trg}(X(\omega)) \), \( X(\omega) \), \( Y(\omega) \), distribution over \( S(P_S) \) and confusion matrix \( C(Y, f_{trg} \circ X) \)), whether \( S \) is censored, i.e., included in \( D_{tr} \) and inputs, whether AIs account for class imbalance in \( S \), whether \( Ado \) is active or passive and whether the threat model is blackbox or whitebox.

| Literature | Attack Vector | Is \( S \) censored? | Imbalance in \( S \)? | \( Ado \) | Threat Model |
|------------|---------------|---------------------|---------------------|--------|--------------|
| Fredrikson et al. [21] | \( X, Y, f_{trg} \circ X, P_S, C(Y, f_{trg} \circ X) \) | ✓ | × | Passive | Blackbox |
| Yeom et al. [52] | \( X, Y, f_{trg}, P_S \) | ✓ | × | Passive | Blackbox |
| Mehnaz et al. [38] | \( X, Y, f_{trg}, P_S, C(Y, f_{trg} \circ X) \) | ✓ | × | Passive | Blackbox |
| Jayaraman and Evans [30] | \( X, Y, f_{trg}, P_S, C(Y, f_{trg} \circ X) \) | ✓, ✓ | ✓ | Passive | Whitebox |

Representation-based Attacks

| Song et al. [46] | \( f_{trg} \circ X \) | ✓ | × | Passive | Both |
| Mahajan et al. [11] | \( f_{trg} \circ X \) | ✓ | × | Passive | Blackbox |
| Malekzadeh et al. [37] | \( f_{trg} \circ X \) | × | × | Active | Blackbox |
| Our Work | \( f_{trg} \circ X \) | ✓, ✓ | ✓ | Passive | Blackbox |

as Race or Sex from \( D_{tr} \) to censor them [46]. Consequently, \( f_{trg} \) is trained on non-sensitive attributes. Formally, consider a probability space \( (\Omega, T, \mathcal{P}) \), measurable spaces \( (E, \mathcal{U}) \), \( (\{0, 1\}, \mathcal{P}((0, 1))) \) and \( (\{0, 1\}, \mathcal{B}) \) where \( \mathcal{B} \) is the Borel tribe on \([0, 1]\). We define random variables \( X \) for the input data, \( Y \) for the classification labels and \( S \) for the sensitive attributes:

- \( X : (\Omega, T, P) \rightarrow (E, \mathcal{U}) \),
- \( Y : (\Omega, T, P) \rightarrow (\{0, 1\}, \mathcal{P}((0, 1))) \),
- \( S : (\Omega, T, P) \rightarrow (\{0, 1\}, \mathcal{P}((0, 1))) \).

Then, \( f_{trg} \) is a measurable function \( f_{trg} : (E, \mathcal{U}) \rightarrow (\{0, 1\}, \mathcal{B}) \) which is used to build the statistical approaching \( Y \) by updating the parameters \( \theta \) on \( D_{tr} \). The prediction of \( f_{trg} \) on \( X \) is a random variable: \( \hat{Y}_h = 1_{\tau(1)} \circ \hat{Y}_t \) where \( \hat{Y}_t = f_{trg} \circ X \) and \( \tau \in [0, 1] \). We present a background on probability and measured spaces in Appendix A.

**Inference.** Once training is completed, \( X(\omega) \) is passed to \( f_{trg} \) to obtain a prediction score \( f_{trg}(X(\omega)) \) (aka soft labels). The attributes during inference, are sampled from an unseen test dataset \( D_{te} \) disjoint from \( D_{tr} \) to evaluate how well \( f_{trg} \) generalizes. We refer to \( f_{trg} \)'s final predictions and intermediate outputs as model observables. Sensitive attributes, although available for different data records, play no role in training or inference. They are reserved solely for designing and evaluating attacks.

### 2.2 Attribute Inference Attacks

An attack constitutes a privacy risk if \( Ado \) learns something about \( D_{tr} \) or the inputs which would be impossible to learn without access to \( f_{trg} \). This differentiates between a privacy risk and simple statistical inference [7]. AIs infer the specific attribute for a specific input to ML model given some model observables and background information [11, 21, 37, 38, 46, 52]. \( Ado \) has access to auxiliary data \( D_{aux} \) which is sampled from the same distribution as \( D_{tr} \), a standard assumption across all AIs. Based on \( Ado \)'s knowledge and access, AIs can be categorized into (a) imputation-based and (b) representation-based attacks.

**Imputation-based attacks** assume \( Ado \) has access to non-sensitive attributes and background information (e.g., marginal prior over sensitive attribute and confusion matrix) in addition to model’s predictions. Fredrikson et al. [21], Yeom et al. [52] and Mehnaz et al. [38] assume that \( S \) is part of the input of \( f_{trg} \) and the targeted data point belongs to \( D_{tr} \). Fredrikson et al. [21] and Mehnaz et al. [38] for a targeted data point, compute \( f_{trg} \) for different values of the sensitive attribute to find the most likely one. Yeom et al. [52] predict \( S \) using the output of a membership oracle or assuming it follows some distribution. However, these attacks perform no better than data imputation and does not pose an actual privacy risk [30]. Jayaraman and Evans [30] propose a whitebox AIA which is a privacy risk in the setting where \( Ado \) has limited knowledge. We omit a comparison with this work due to difference in threat model.

**Representation-based attacks** exploit the distinguishability in model observables for different values of sensitive attributes [11, 37, 46]. For instance, the distribution of \( f_{trg} \circ X \) for \( S=\text{males} \) is different from \( S=\text{females} \). Song et al. [46] / Mahajan et al. [11] assume that \( S \) is not in the input. \( Ado \) only observes \( f_{trg} \circ X \). \( Ado \) trains an ML attack model \( f_{att} \) to map the output predictions \( f_{trg}(X(\Omega)) \) to \( S(\Omega) \). Malekzadeh et al. [37] assume that \( Ado \) can actively introduce a “backdoor” and train \( f_{trg} \) to explicitly encode information about \( S \) in \( f_{trg} \circ X \). We omit comparison with Malekzadeh et al. [37] due to difference in threat model and focus on Song et al. [46] / Mahajan et al. [11].

We empirically measure attribute privacy using the resistance to AIs which exploit distinguishability in model predictions for different values of sensitive attributes. Specifically, \( f_{trg} \) satisfies attribute privacy if the success of AIA is random guess.

### 2.3 Group Fairness

Generally, data records in the minority subgroup, identified by some sensitive attribute (e.g., Race or Sex), face unfair prediction behaviour compared to data records in the majority subgroup. For instance in criminal justice, Race plays a non-negligible role in predicting the likelihood of them re-offending [2]. Group fairness algorithms add constraints during training such that different subgroups (i.e., \( S : \Omega \rightarrow \{0, 1\} \)) are treated equally (e.g., AdvDEbias [55] and EGD [1]). \( S \) is either Sex or Race (i.e., \( S(\omega) \in \{0 \) for woman and 1 for man, or 0 for black and 1 for white). There are different definitions of group fairness which have been introduced in prior work.
We discuss two well-established definitions: demographic parity (DemPar) and equalized odds (EqOdds).

**Definition 2.1.** \( \hat{Y}_h \) satisfies DemPar for \( S \) if and only if: \( P(\hat{Y}_h = 0 | S = 0) = P(\hat{Y}_h = 0 | S = 1) \).

DemPar ensures that the number of correct predictions is the same for each subgroup. However, this may result in different false positive (FPR) and true positive rates (TPR) if the true outcome varies with \( S \) [14]. EqOdds is a modification of DemPar to ensure that both TPR and FPR are the same for each subgroup [27].

**Definition 2.2.** \( \hat{Y}_h \), classifier of \( Y \), satisfies EqOdds for \( S \) if and only if: \( P(Y_h = y | S = 0, Y = y) = P(Y_h = y | S = 1, Y = y) \) \( \forall (y, y) \in \{0, 1\}^2 \).

We consider two algorithms: (a) adversarial debiasing (AdvDEbias) [35, 55] and (b) exponentiated gradient descent (EGD) [1]. AdvDEbias achieves fairness by training \( f_{frg} \) to have indistinguishable output predictions in the presence of a discriminator network \( f_{disc} \). \( f_{disc} \) infers \( S \) corresponding to a target data point given \( f_{frg} \circ X \). \( f_{frg} \) is then trained to minimize the success of \( f_{disc} \). AdvDEbias outputs soft labels (i.e., a probability attached to each value of the sensitive attribute). EGD solves an under-constraint optimization problem to find a collection of optimal measurable functions \( \tau_0, \cdots, \tau_{N-1} \) and threshold \( \omega \). One of the measurable functions and generates a randomized classifier: \( \hat{Y}_h = 1_{[\tau_1, \cdots, \tau_N]} \circ \omega \circ X \). EGD can satisfy different fairness constraints (e.g., DemPar or EqOdds). EGD outputs hard labels (i.e., a binary assignment to the sensitive attribute).

### 3 PROBLEM STATEMENT

Understanding the interaction between group fairness and privacy notions is crucial for practitioners to address conflicts before model deployment [13]. This is essential for regulatory compliance, as both fairness and privacy are mandated by regulations [28, 29, 33, 34, 48]. Previous research highlights a conflict between group fairness and differential privacy showing an impossibility to train a differentially private model while ensuring group fairness [8]. Additionally, group fairness increases susceptibility to membership inference attacks [6] and distribution inference attacks [47]. However, the interaction of group fairness with attribute privacy remains unclear.

**Challenge 1:** Our goal is to better understand and evaluate the relation between group fairness and attribute privacy. We conjecture that group fairness aligns with attribute privacy by ensuring indistinguishability in the predictions for different values of \( S(o) \).

**Challenge 2:** We evaluate group fairness algorithms against AIs. If the success of AIs is random guess, it indicates that attribute privacy is satisfied. Appendix B shows that distinguishable output predictions for two sub-groups could be exploited to infer a sensitive attribute even when it is not included in \( f_{frg} \)’s \( D_{frg} \). However, none of the current AIs in literature are effective as they fail to account for real-world datasets with significant class imbalance in sensitive attributes. For instance, the fraction of males and whites in different datasets are 68% and 90% (CENSUS), 81% and 51% (COMPAS), 53% and 36% (MEPS), and 78% and 96% (LFW). We have to design AIs to account for this class imbalance to make them effective.

**Challenge 3:** Group fairness algorithms can output either soft labels (probability scores indicating that an input belongs to different classes) or hard labels (most likely class from soft labels). We have to design AIs for both.

After addressing Challenge 1 and Challenge 2, we can then revisit Challenge 1 to show that group fairness aligns with attribute privacy by mitigating our proposed AIs. We now present our threat model and assumptions about \( \mathcal{A} \)’s knowledge and capabilities.

---

**Figure 2:** \( \mathcal{A} \) wants to infer \( S(o) \) for \( X(o) \) given \( f_{frg}(X(o)) \). \( \mathcal{A} \) uses \( f_{att} \) which is trained on \( D_{aux} \) to map \( S(o) \) from \( f_{frg}(X'(o)) \).

---

**Threat Model:** We assume a blackbox \( \mathcal{A} \) with no knowledge of \( f_{frg} \)’s parameters or architecture. \( \mathcal{A} \) can query \( f_{frg} \) and obtain corresponding predictions. This is the most practical setting typically seen in ML as a service. Additionally, \( \mathcal{A} \) has access to \( D_{aux} \) sampled from the same distribution as \( D_{frg} \) similar to prior works [11, 46, 52]. \( D_{aux} \) is split into two disjoint datasets: \( D_{frg}^{tr} \), used for designing the attack and \( D_{frg}^{te} \) to evaluate the attack. This is a strong assumption, but it was made to favor \( \mathcal{A} \) since, in practice, \( \mathcal{A} \) is likely to have a different distribution. We revisited the case where \( D_{aux} \) is sampled from a different distribution in Appendix F. \( \mathcal{A} \) trains \( f_{att} \) on \( D_{frg}^{tr} \) to infer \( S \) using \( f_{frg} \circ X \). \( f_{att} \) is then evaluated on \( D_{frg}^{te} \) (Figure 2). This attack is applicable in both cases when \( f_{frg} \)’s outputs are hard and soft labels:

**TM1** For hard labels, \( \mathcal{A} \) builds a statistic \( \hat{S} \) to infer \( S: \hat{S} = f_{att} \circ \hat{Y}_h \circ X \).

**TM2** For soft labels, \( \mathcal{A} \) builds a statistic \( \hat{S} \) to infer \( S: \hat{S} = 1[v, 1] \circ f_{att} \circ \hat{Y}_h \circ X \). Here, \( v \in [0, 1] \) is a threshold which can be adapted to improve the attack.

### 4 EXPERIMENTAL SETUP

**Datasets:** We consider four real-world datasets covering different domains: criminal justice (COMPAS), income prediction (CENSUS), healthcare (MEPS), and face recognition (LFW), to illustrate the effectiveness of the proposed AIs. These have been used as benchmarks for privacy [11, 38, 46] and fairness [39].

CENSUS comprises 30,940 data records with 95 attributes about individuals from 1994 US Census data. The attributes include marital status, education, occupation, job hours per week among others. The classification task is to estimate whether an individual makes an income of 50k per annum. For \( f_{frg} \), we use 24,752 data records for \( D_{frg}^{tr} \) and 6,188 data records as \( D_{frg}^{te} \). We use 9,450 data records from \( D_{te} \) as \( D_{aux}^{frg} \) for training \( f_{att} \) and evaluate it on 1,238 unseen data records in \( D_{aux}^{te} \).
COMPAS is used for commercial algorithms by judges and parole officers for estimating the likelihood of a criminal re-offending using seven attributes. The classification task is whether a criminal will re-offend or not, and contains 6,172 criminal defendants in Florida. The dataset contains 7 attributes and we use 4,937 data records for training \( f_{\text{req}} \) and 1,235 data records as testing dataset. We use 988 data records from \( D_{\text{te}} \) as \( D_{\text{aux}}^t \) for training \( f_{\text{att}} \), i.e., \( D_{\text{aux}}^t \) and evaluate it on 247 unseen data records in \( D_{\text{aux}}^e \).

MEPS contains 15,830 records of different patients using medical services by capturing the trips made to clinics and hospitals. The classification task is to predict the utilization of medical resources as 'High' if the sum of the number of office based visits, outpatient visits, ER visits, inpatient nights and home health visits, is greater than ten. We use 12,664 data records for training \( f_{\text{req}} \) and 3,166 \( D_{\text{te}} \) records as testing dataset. We use 2,532 data records from \( D_{\text{te}} \) as \( D_{\text{aux}}^t \) for training \( f_{\text{att}} \) and evaluate it on 634 unseen data records in \( D_{\text{aux}}^e \).

LFW has 8,212 images of people with the classification task to predict whether their age is > 35 years. In all datasets, we consider Race and Sex as binary sensitive attributes to be inferred.

We use 80% of the dataset as \( D_{\text{te}} \) and the remaining 20% for \( D_{\text{te}} \). We use \( D_{\text{te}} \) as \( D_{\text{aux}} \) and ensure that the distribution of S is uniform between them. We use 80% of \( D_{\text{aux}} \) for training \( f_{\text{att}} \) and 20% for evaluation of the attack. The impact of an adversary using auxiliary knowledge (\( D_{\text{aux}} \)) which differs from the target model training data \( D_{\text{te}} \) is evaluated in the Appendix F. We use cross validation where each split is done five times without any overlap. \( f_{\text{req}} \) is trained and evaluated five times and \( f_{\text{att}} \) is trained and validated ten times. We check for statistical significance for the results (i.e., p-value < 0.05).

**Model Architectures:** For all the datasets, we use neural networks with four hidden layers with the following dimensions: [32, 32, 32, 32] and ReLU activation functions.

**Metrics:** We use standard classification accuracy between predicted labels and ground truth labels for evaluating \( f_{\text{req}} \)’s performance. We refer to this as *utility*. We evaluate attack success using balanced accuracy which is the average of the proportion of correct predictions of each class of the sensitive attribute individually: 

\[
\frac{1}{2} \left( P(\hat{S} = S | S = 0) + P(\hat{S} = S | S = 1) \right) \text{.}
\]

This metric accounts for class imbalance and simple accuracy is misleading when the datasets have significant class imbalance. An accuracy of 50% corresponds to random guesses. To evaluate fairness, we use \( \text{DEMPAR} \_\text{Level} \) given by \( |P(\hat{Y} = 0 | S = 0) - P(\hat{Y} = 0 | S = 1)| \). \( \text{DEMPAR} \_\text{Level} \) close to zero indicates that \( f_{\text{req}} \) is fair.

**Baselines:** To evaluate the impact of group fairness on attribute privacy, we compare the attack success of \( \text{ADAPTAIA} \) with and without using group fairness (the former case is referred to as \( \text{Empirical} \) while the latter is referred to as \( \text{Baseline} \)). For classifiers that output hard labels (i.e., using \( \text{ADAPTAIA-H} \)), we also indicate the theoretical bound from Theorem (referred to as \( \text{Theoretical} \)).

5 ADAPTAIA: AN EFFECTIVE AIA

We first address Challenge 2 and Challenge 3 from Section 3. Our goal is to design an AIA which accounts for class imbalance in \( S \), typical of real-world applications [25, 44], and is applicable for soft and hard labels. We present \( \text{ADAPTAIA} \), an AIA with adaptive threshold and its two variants of the attack, \( \text{ADAPTAIA-S} \) for soft labels, and \( \text{ADAPTAIA-H} \) for hard labels. We describe our proposed attacks below.

**ADAPTAIA-S.** Recall from Section 3 that datasets have significant class imbalance in sensitive attributes. We conjecture that this skews \( f_{\text{req}} \)’s predictions, thereby requiring us to adapt \( f_{\text{att}} \) to correctly infer the value of sensitive attributes [4, 25, 44]. Hence, the default threshold of 0.5 over \( f_{\text{att}} \)’s soft labels will not result in accurate estimation of \( S \). \( \text{Adv} \)’s attack success can be improved by using an adaptive threshold [4, 25, 44]. However, none of the prior AIA accounts for this.

Hence, we can improve AIAIs with an adaptive threshold instead of using a default threshold of 0.5 as in prior AIAIs. \( \text{Adv} \) computes the optimized threshold \( v^* \) on \( D_{\text{aux}}^t \) which is later used for the attack on \( D_{\text{aux}}^e \). We compute \( v^* \) to balance TPR and FPR. Ideally, a perfect attack would result in no FPR and only TPR. \( \text{Adv} \)’s goal is thus to approach this optimal value of TPR and FPR. Formally, 

\[
v^* \in [0, 1] \text{ where } v^* = \arg\min_v (1 - \text{TPR}_v)^2 + \text{FPR}_{v}^2 .
\]

![Figure 3: ROC curve: \( v^* \) can lower FPR to infer Race.](image)

For illustration purposes, we plot the ROC curve for inferring Race in Figure 3. We observe that \( v^* \) does not correspond to the default classification threshold of 0.5 used in the literature and results in lower FPR resulting in a more confident attack.

**ADAPTAIA-H.** For hard labels, it is not necessary to train \( f_{\text{att}} \). Instead, we consider a set of functions from \{0, 1\} to \{0, 1\} containing four elements: \( x \mapsto 0, x \mapsto x, x \mapsto 1 - x \), and \( x \mapsto 1 \). We optimize the attack by finding the functions which give the best balanced accuracy on \( D_{\text{aux}}^t \) which is used to evaluate attack success on \( D_{\text{aux}}^e \).

**Evaluating ADAPTAIA-H and ADAPTAIA-S.** For hard labels (TM1), we consider the baseline of training a neural network over hard labels to infer the value of S given X. We then compare this with \( \text{ADAPTAIA-H} \). For soft labels (TM2), we consider the attacks by Song et al. [46]/Mahajan et al. [11] as the prior state-of-the-art baselines with the default \( v = 0.5 \) over \( f_{\text{att}}(f_{\text{req}}(X(\omega))) \). We then compare this with \( \text{ADAPTAIA-S} \).

We present the comparison of the baseline attack with \( \text{ADAPTAIA-H} \) and \( \text{ADAPTAIA-S} \) in Table 2. For all datasets, we see that \( \text{ADAPTAIA-H} \) and \( \text{ADAPTAIA-S} \) are significantly better on average than prior AIAIs. Having successfully addressed Challenge 2 and Challenge 3 by designing \( \text{ADAPTAIA-H} \) and \( \text{ADAPTAIA-S} \), we use the AIAIs to evaluate the alignment with two well-established group fairness algorithms: EGD and AdvDebias.
Table 2: Comparing attack accuracy (average over ten runs) for baseline with \( v = 0.5 \) \cite{11, 46} and both ADAPTIA-H and ADAPTIA-S: an adaptive threshold improves the success of the attack for hard and soft labels.

| Dataset | Baseline (\( v = 0.5 \)) | ADAPTIA-H | ADAPTIA-S |
|---------|--------------------------|-----------|-----------|
| CENSUS | 0.50 ± 0.0 | 0.50 ± 0.0 | 0.56 ± 0.0 |
| COMPAS | 0.62 ± 0.3 | 0.55 ± 0.0 | 0.62 ± 0.0 |
| MEPS | 0.51 ± 0.0 | 0.55 ± 0.0 | 0.53 ± 0.0 |
| LFW | 0.59 ± 0.0 | 0.64 ± 0.1 | 0.61 ± 0.1 |

6 ALIGNMENT OF EGD

We first consider EGD under the threat model TM1 as it outputs hard labels. We evaluate if EGD is aligned with attribute privacy by mitigating ADAPTIA-H. We also quantify the impact on both the utility and group fairness of \( f_{pre} \). As described in Section 2.3, \( f_{pre} \) can be trained with EGD to satisfy either DemPar or EqOdds. Here, we focus on EGD + DemPar but revisit EGD + EqOdds in Appendix E and demonstrate that EqOdds cannot mitigate ADAPTIA-H and by consequence ADAPTIA-S. We present first theoretical results followed by empirical evaluation.

**EGD: Theoretical Guarantees** We theoretically compute the bound on the attack accuracy under EGD + DemPar (proof in Appendix C).

Theorem 6.1. The maximum attack accuracy achievable by ADAPTIA-H is equal to \( \frac{1}{2} (1 + \text{DemPar-Level of } f_{pre}) \).

Hence, we obtain a bound for ADAPTIA-H without any conditions on \( f_{pre} \) or datasets. Additionally, we observe that DemPar-Level=0. Consequently, if \( f_{pre} \) satisfies DemPar then no \( f_{acc} \) will perform better than a random guess. Hence, EGD + DemPar satisfies attribute privacy.

**EGD: Empirical Evaluation** We will now empirically validate the above theoretical guarantee by evaluating \( f_{pre} \) trained with EGD + DemPar against ADAPTIA-H. Due to space limitation, we only present results for CENSUS and MEPS and move the results for other datasets in Appendix C.

To evaluate alignment of EGD + DemPar with attribute privacy, we compare in Figure 11 the result of ADAPTIA-H on a classification without using EGD + DemPar and on using EGD + DemPar. Results show that ADAPTIA-H demonstrates significantly lower effectiveness (approaching random guessing, 50%) when utilizing EGD + DemPar compared to the baseline. Additionally, for ADAPTIA-H, we note that the theoretical bound on attack accuracy matches with the empirical attack accuracy. The theoretical accuracy is equal to the empirical accuracy when the values are > \( \frac{1}{2} \). But DemPar – Level \( \geq 0 \) implies that \( (1 + \text{DemPar} – \text{Level}) \geq \frac{1}{2} \). Hence, we observe that the theoretical accuracy is not equal to the experimental when \( f_{acc} \)'s attack accuracy is random guess (under \( \frac{1}{2} \)). This happens when \( f_{pre} \) nearly follows DemPar where \( \text{Adv's } f_{acc} \) is optimal on \( D_0 \), but worse than random guess for \( D_{1e} \).

![Figure 4: For ADAPTIA-H, we observe that EGD reduces the attack accuracy to random guess (~50%).](image)

**Trade-off with \( f_{pre} \)’s Utility.** Using group fairness improves the attribute privacy, but comes at the cost of \( f_{pre} \)’s utility \cite{1}. To quantify this impact, we compare in Figure 12 the accuracy of \( f_{pre} \) with and without using EGD + DemPar. Results show that the utility is reduced by 15% CENSUS and 5% for MEPS datasets. This trade-off with the utility of \( f_{pre} \) is inherent to the group fairness algorithms \cite{42, 45, 50, 54}.

![Figure 5: Imposing fairness with EGD + DemPar has a significant impact on the accuracy of \( f_{pre} \), matching with the observation from prior work \cite{1}](image)

**Sanity Check for Fairness.** We now quantify group fairness as measured by DemPar-level with and without training \( f_{pre} \) with EGD + DemPar (the lower the better). As seen in Figure 6, we observe that \( f_{pre} \) with EGD + DemPar has significantly lower DemPar-level which is closer to zero as compared to baseline. Hence, EGD + DemPar is effective in achieving group fairness.

**Summary:** Considering our theoretical and empirical evaluation, we conclude that EGD + DemPar induces attribute privacy.
We will now empirically validate our theoretical conjecture both theoretically and empirically.

**AdvDebias: Theoretical Guarantees** We first theoretically show that AdvDebias bounds the balanced attack accuracy to random guess (proof in Appendix D).

**Theorem 7.1.** The following propositions are equivalent:
- \( \hat{Y}_s \) is independent of \( S \); 
- Balanced accuracy of AdaptAIA-S is \( \frac{1}{2} \).

**AdvDebias: Empirical Evaluation** We will now empirically validate the above theoretical guarantee by evaluating \( f_{\text{disc}} \) trained with AdvDebias against AdaptAIA-S. This evaluation of AdvDebias using the soft labels can be converted to hard labels for using AdaptAIA-H. Due to space constraints, this section only reports the results of AdaptAIA-S, leaving the results of AdaptAIA-H in Appendix D. Similarly to Section 6, we only report results for CENSUS and MEPS and move the results for other datasets in Appendix D.

To evaluate the alignment of the fairness constraint imposed by AdvDebias with attribute privacy, we compare in Figure 7 the attack success of AdaptAIA-S on the target model trained with and without group fairness using AdvDebias. Results show that for all datasets AdvDebias reduces the attack accuracy close to random guess (i.e., 50%) regardless of the value of the attack accuracy without group fairness (i.e., Baseline).

**Trade-off with \( f_{\text{reg}} \)'s Utility.** We now quantify the impact on \( f_{\text{reg}} \)'s utility on using AdvDebias (Figure 8). We report the model accuracy on \( D_{te} \) for \( f_{\text{reg}} \) with and without AdvDebias. Results that there is a significant decrease in utility on using AdvDebias across all datasets. However, this trade-off between group fairness and utility is consistent with prior works [55, 57].

**Sanity Check for Fairness.** We now quantify the fairness as measured by DemPar-level with and without training \( f_{\text{reg}} \) with AdvDebias (Figure 9). Results show that \( f_{\text{reg}} \) with AdvDebias has significantly lower DemPar-level which is closer to zero as compared to baseline. Hence, AdvDebias is effective for group fairness.

**Summary:** Similarly to EGD, considering our theoretical and empirical evaluation, we conclude that AdvDebias induces attribute
Figure 9: By reducing DemPar-Level close to 0 indicates that $f_{\text{src}}$ is fair after AdvDebias.

privacy. This alignment allows us to account attribute privacy while only calibrating the two dimensional trade-off between group fairness and utility.

8 RELATED WORK
AIA have been extensively studied in the context of online social media where $Ado$ can leverage user’s behavioural information [5, 22–24, 32, 58]. In ML, AIA are approached as data imputation challenges [21, 38, 52] and by leveraging intermediate model representations [11, 37, 46], which we extensively cover in Section 2.2. Moreover, $Ado$ can execute AIA using membership inference attacks as a subroutine [52]. Here, the success of AIA mirrors that of membership inference attacks, hinging on the degree of $f_{\text{src}}$ overfitting. Nevertheless, Zhao et al.[56] establish this link only holds when membership inference attacks are highly effective. If not, AIA offer minimal advantage, despite the model’s vulnerability to membership inference attacks. Additionally, AIA are effective against model explanations, presenting a compromise between attribute privacy and transparency [12].

Defenses against AIA have limited prior literature. Attriguard [31] safeguards sensitive attributes in social networks by introducing adversarial noise to the output predictions, compelling $f_{\text{att}}$ to make incorrect classifications. For ML, adversarial training has proven effective in diminishing the inference of sensitive attributes within predictions [57]. This technique can additionally disentangle sensitive features from model representations [17, 26, 36, 40, 51]. An alternative avenue, variational autoencoders, was explored for minimizing AIA[53]; however, Song et al.[46] found these defences to be ineffective.

Interactions between Privacy and Fairness have been explored in prior work [3, 6, 15, 20, 43, 49]. Chang et al. [6] demonstrate that applying in-processing fairness algorithms escalates overfitting, heightening vulnerability to MIAs, particularly affecting minority subgroups. Additionally, differential privacy and individual fairness share a common goal, with individual fairness encompassing differential privacy as a broader concept [15]. However, differential privacy and group fairness are at odds when there’s performance discrepancy between minority and majority subgroups [3, 20, 43, 49]. Ferry et al. [18] introduced AIA aimed at predicting sensitive attributes from a fair model. Their approach assumes that the group fairness algorithm for $f_{\text{src}}$ is known to $Ado$, enabling calibration of the attacks. However, in practice, revealing model training specifics for confidentiality is improbable. In distinction from their approach, our attacks function within the context of $Ado$'s blackbox understanding of $f_{\text{src}}$, and the defenses we present effectively counter such attacks.

9 DISCUSSION AND CONCLUSION
This paper shows that there are no conflicts between group fairness and the specific notion of attribute privacy, which is lacking in the literature. Specifically, through an extensive empirical evaluation and theoretical guarantees, we show that group fairness imposed through the use of AdvDebias and EGD satisfying DemPar is aligned with attribute privacy. This alignment means that ensuring group fairness also ensures a protection against the attribute inference attack, which is lacking in the literature. However, ensuring fairness remains at the cost of model utility. To perform our extensive evaluation, we also propose new AIA which outperform prior works.

Output indistinguishability is a general framework which can be applied to satisfy group fairness instead of explicitly optimizing for different fairness metrics [16]. Any algorithm which falls into the category of output indistinguishability with respect to sensitive attributes is likely to satisfy attribute privacy as it will reduce the ability of $Ado$ to perform accurate representation-based AIA in a blackbox setting.

Following the setting of prior AIA and fairness [1, 11, 21, 35, 37, 38, 46, 52, 55], we focus on the case where the sensitive attributes are binary. However, for ADAPTAIA-S, the $f_{\text{att}}$ can be trained to learn to infer non-binary attributes as well. For ADAPTAIA-H, if we have $q$ classes and there are $p$ sensitive attributes, then there are $p^q$ functions to select from. While this is reasonable for small values, efficiently finding functions for large values is left as future work.

REFERENCES
[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In International Conference on Machine Learning, volume 80, pages 60–69, 2018.
[2] Kiana Alikhadeh, Emma Drohina, Diandra Prideau, Brianna Richardson, Duncan Purves, and Juan E. Gilbert. A review of predictive policing from the perspective of fairness. Artif. Intell. Law, 30(1):1–17, mar 2022.
[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In Advances in Neural Information Processing Systems, pages 15479–15488. 2019.
[4] Jason Brownlee. A Gentle Introduction to Threshold-Moving for Imbalanced Classification. MachineLearningMastery.com — machinelearningmastery.com. [Accessed 31-08-2023]
[5] Abdelberi Chaabane, Gergely Acs, and Mohamed Ali Kasfar. You Are What You Like! Information Leakage Through Users’ Interests. In Network and Distributed System Security Symposium, pages 1–14, 2012.
[6] Hongyang Chang and R. Shokri. On the privacy risks of algorithmic fairness. 2021 IEEE European Symposium on Security and Privacy, pages 292–303, 2021.
[7] Graham Cormode. Personal privacy vs population privacy: Learning to attack anonymization. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’11, page 1251–1261. Association for Computing
Machinery, 2011.

[8] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morganstern. On the compatibility of privacy and fairness. In Conference on User Modeling, Adaptation and Personalization, UMAP 19 Adjunct, pages 309–315, 2019.

[9] Emiliano De Cristofaro. An overview of privacy in machine learning. arXiv preprint arXiv:2005.08679, 2020.

[10] Francois Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. Advances in Neural Information Processing Systems, 34, 2021.

[11] Amit Sharma Dreyat Mahajan, Shruti Tope. Does learning stable features provide privacy benefits for machine learning models? In NeurIPS PAML Workshop, 2020.

[12] Vaishali Duddu and Antoine Boulot. Inferring sensitive attributes from model explanations. arXiv preprint arXiv:2208.09967, 2022.

[13] Yu-Feng Li. Class-imbalanced semi-supervised learning with IMBAL-NN. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[14] Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. Attriinfer: Learning adversarially fair and transferable representations. In Proceedings of Machine Learning Research, volume 80, pages 3384–3393, 2018.

[15] Mohammad Malekzadeh, Anastasia Borovoykh, and Deniz Gündüz. Honest-but-curious nets: Sensitive attributes of private inputs can be secretly coded into the classifiers’ outputs. In ACM SIGSAC Conference on Computer and Communications Security, CCS ’21, page 825–844, 2021.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM Comput. Surv., 54(6), jul 2021.

[17] Daniel Moyer, Shuyang Gao, Rob Berkelmans, Greg Ver Steeg, and Aram Galstyan. Invariant representations without adversarial training. In Advances in Neural Information Processing Systems, 2018.

[18] Neha Patki, Roy Wedge, and Kalyan Vezzamachaneni. The synthetic data vault. In IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 399–410, Oct 2016.

[19] Carlos Pinzon, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the incompatibility of accuracy and equal opportunity. Machine Learning, pages 1–30, 05 2023.

[20] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. Nature Machine Intelligence, 1(10) :896–904, 2020.

[21] Congheng Song and Vitaly Shmatikov. Overlearning reveals sensitive attributes. In International Conference on Learning Representations, 2020.

[22] Anshuman Suri, Tifin Lu, Yanjun Chen, and David Evans. Dissecting distribution inference. In 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pages 150–164, IEEE, 2023.

[23] Elham Tabassi, Kevin J. Burns, M. Hadjimichael, Andres Molina-Markham, and Gerome Miklau. Fair decision making using privacy-protected data. In Conference on Fairness, Accountability, and Transparency, FAT ’20, page 189–199, 2020.

[24] Sivaramakrishnan Rajaraman, Prasanth Ganesan, and Sameer K. Antani. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS ONE, 17, 2021.

[25] Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui J. Ferry, U. Aivodji, S. Gambs, M. Huguet, and M. Siala. Exploiting fairness to improve generalization in machine learning. In Proceedings of the 39th International Conference on Artificial Intelligence in JCAI-22, pages 5470–5477. International Joint Conference on Artificial Intelligence Organization, 7 2022. Survey Track.

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. 2011.

[27] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science Conference, ITCS ’12, page 214–226, 2012.

[28] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk explanations. In Proceedings of the 34th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2021, page 1095–1108, New York, NY, USA, 2021. Association for Computing Machinery.

[29] Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. Attriinfer: Differential privacy and fairness in decisions and learning tasks: A survey. In Lud De Raedt, editor, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-21, pages 5470–5477. International Joint Conference on Artificial Intelligence Organization, 7 2022. Survey Track.

[30] David Pujol, Ryan McKeeman, Satya Kuppam, Michael Hay, Adwitm Machanava-jihala, and Gerome Miklau. Fair decision making using privacy-protected data. In Conference on Fairness, Accountability, and Transparency, FAT ’20, page 189–199, 2020.
APPENDIX

We provide more details about the notations (Appendix A), plots for distinguishing in predictions (Section B), proofs and empirical results for EGD (Appendix C) and AdvDEbias (Appendix D), proof for showing EGD + EqODds is not aligned (Appendix E), and finally impact of Adv’s distribution on attack success.

A NOTATIONS

We introduce some notations from probability theory which use in the paper. For a set $A$, the set of subsets of $A$ is given by $\mathcal{P}(A)$. Each element $a \in \mathcal{P}(A)$ is such that $a \subset A$. A tribe $\mathcal{A}$ is a subset of $\mathcal{P}(A)$ that contains: 0, $A$, and is stable by complementary and countable union. Then, we call $(A, \mathcal{A})$ a measure space. A measure $d$ is a function $d: \mathcal{A} \rightarrow [0, +\infty]$ such that $d(\emptyset) = 0$ and $d\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} d(A_i)$ for any $(A_1, A_2, \ldots) \in \mathcal{A}^\mathbb{N}$ with $\forall (i, j), A_i \cap A_j = \emptyset$. We then call $(A, \mathcal{A}, d)$ a measurable space. Any function mapping $A$ to $B$ is called a measurable function if $\forall b \in B, f^{-1}(b) \in \mathcal{A}$ and we note $f: (A, \mathcal{A}) \rightarrow (B, \mathcal{B})$ or $f: (A, \mathcal{A}, d) \rightarrow (B, \mathcal{B})$.

In the special case where $d(A) = 1$ we call $d$ a probability measure. We then call $(A, \mathcal{A}, d)$ a probability space and the measurable functions on this space are random variables.

$S = 1_{[0.5,1]} \circ f_{\text{att}} \circ f_{\text{reg}} \circ X$ translates to $S$ is equal to the composition of four functions: the random variable $X$, $f_{\text{reg}}$, $f_{\text{att}}$ and the indicator function of the set $[0.5,1]$. We summarize all important notations used in our paper in Table 3.

**Table 3:** Summary of notations used in the paper.

| Notation | Meaning |
|----------|---------|
| $\text{Adv}$ | Adversary |
| $f_{\text{reg}}$ | Target model being attacked |
| $f_{\text{reg}} \circ X$ or $f_{\text{reg}}(X(\omega))$ | Predictions |
| $Y_{\text{h}}(\omega)$ | Prediction (Hard label) |
| $Y_{\text{s}}(\omega)$ | Prediction (Soft label) |
| $f_{\text{att}}$ | Adv’s Attack model |
| $f_{\text{disc}}$ | Discriminator network for AdvDEbias |
| $D_{\text{tr}}$ | Data used to train the target model |
| $D_{\text{re}}$ | Data used to test the target model |
| $D_{\text{aux}}$ | Auxiliary data available to the adversary |
| $D'_{\text{aux}}$ | Adv’s dataset to train attack model |
| $D''_{\text{aux}}$ | Adv’s dataset to evaluate attack model |
| $v$ | Threshold on $f_{\text{att}}$’s output |
| $X$ | Random variable of non-sensitive attributes |
| $Y$ | Random variable of sensitive attributes |
| $S$ | Random variable of sensitive attributes |
| $X(\omega)$, $Y(\omega)$, $S(\omega)$ | Specific instances parameterized by $\omega$ |
| $(\Omega, \mathcal{A}, \mathcal{F})$ | Probability space |
| $(E, \mathcal{U})$ | Measurable space |
| $B$ | Borel tribe ($\sigma$-algebra) |

B PREDICTION DISTINGUISHABILITY

To better understand the impact of bias in $D_{\text{tr}}$, on predictions, we plot the conditional densities of $f_{\text{reg}} \circ X$ for Race and Sex (each taking values $S = 0$ and $S = 1$) in Figure 10. We observe that $f_{\text{reg}}(X(\omega))$ are distinguishable for different values of $S(\omega)$, even when it is not included in the input. For CENSUS, the distributions for both Race and Sex corresponding to the minority group (e.g., blacks and females) have a larger peak for lower output probability values. This indicates that the classifier is likely to predict $<50K$ salary for non-white and females compared to members of the majority group. For COMPAS, $f_{\text{reg}}$ is likely to predict males and blacks to re-offend more compared to females or whites. Furthermore, in MEPS the distributions look similar but the use of medical resources is lower for blacks and females. Finally for LFW, males and blacks are likely to be predicted as being $>35$. To account for the skew in predictions and class imbalance, we want to use an adaptive threshold.

Figure 10: Distinguishable output predictions for two subgroups identified by Race and Sex. Adv exploits this to infer $S(\omega)$ for an arbitrary input data record.
C EGD + DEMPAR
We first present the proof of the Theorem 6.1. We then present empirical results to the success of ADAPTAIA showing that \( f_{\text{reg}} \) trained with EGD + DEMPAR is fair compared to the baseline as a sanity check. Finally, we present the impact on \( f_{\text{reg}} \)'s utility on using EGD + DEMPAR.

**Theorem 6.1.** The maximum balanced accuracy achievable by AH is equal to \( \frac{1}{2} (1 + \text{DEMPAR-level of } f_{\text{reg}}) \).

**Proof.** The set \( B \) of function from \( \{0, 1\} \) to \( \{0, 1\} \) contains four elements: \( b_0 = 0, b_1 = id, b_2 = 1 - id \) and \( b_3 = 1 \), where \( \forall x, id(x) = x \). For every \( b \in B \) the balanced AIA accuracy is \( BA(b) = \frac{1}{2} (P(b \circ \hat{Y} = 0 | S = 0) + P(b \circ \hat{Y} = 1 | S = 1)) \). We have \( BA(b_0) = BA(b_3) = \frac{1}{2} \), hence, we can discard those elements when solving the attack optimisation problem. This problem writes \( \max_{b \in B} BA(A(b)) = \max(ba(b_1), BA(b_2)) \). We remark that \( b_1 \circ \hat{Y} = \hat{Y} \) and \( b_2 \circ \hat{Y} = 1 - \hat{Y} \). Hence,

\[
BA(b_1) = \frac{1}{2} (P(\hat{Y} = 0 | S = 0) + P(\hat{Y} = 1 | S = 1))
\]
\[
= \frac{1}{2} (1 + P(\hat{Y} = 1 | S = 1) - P(\hat{Y} = 1 | S = 0))
\]
\[
BA(b_2) = \frac{1}{2} (1 + P(\hat{Y} = 1 | S = 0) - P(\hat{Y} = 1 | S = 1))
\]

Thus,

\[
\max_{b \in B} BA(b) = \frac{1}{2} (1 + \max(P(\hat{Y} = 0 | S = 0) - P(\hat{Y} = 1 | S = 1)))
\]
\[
= \frac{1}{2} (1 + P(\hat{Y} = 1 | S = 0) - P(\hat{Y} = 0 | S = 1))
\]

\[
\frac{1}{2} (1 + |P(\hat{Y} = 1 | S = 1) - P(\hat{Y} = 1 | S = 0)|)
\]

\[\square\]

**Additional Results: ADAPTAIA-H.** We present additional results for COMPAS and LFW evaluating the effectiveness of EGD + DEMPAR against ADAPTAIA-H. Similar to Section 6, we observe that the success of ADAPTAIA-H decreases on using EGD + DEMPAR compared to Baseline (Figure 11). Additionally, the theoretical bound on attack accuracy (Theoretical) matches with the empirical results (Empirical).

**Additional Results: Impact on Utility.** We now present the impact on \( f_{\text{reg}} \)'s utility on using EGD + DEMPAR for COMPAS and LFW (Figure 12). We find that the utility degrades on using EGD + DEMPAR which is in line with findings from prior work [1]. In other words, group fairness comes at the cost of accuracy.

**Additional Results: Fairness.** For COMPAS and LFW, we present the DEMPAR-level with and without training \( f_{\text{reg}} \) with EGD + DEMPAR (Figure 13). We observe that \( f_{\text{reg}} \) with EGD + DEMPAR has significantly lower DEMPAR-level which is closer to zero as compared to Baseline. Hence, EGD + DEMPAR is effective to achieve group fairness.

D ADVDEBIAS
We first present the proof of Theorem 7.1 followed by results showing that models trained with ADVDEBIAS is indeed fair as a sanity check. Finally, we show the impact of using ADVDEBIAS on \( f_{\text{reg}} \)'s utility.

We present the theorem and proof which bounds AIA accuracy to random guess on using ADVDEBIAS. We remark that the definition 2.1 of DEMPAR can be written synthetically as the following property: \( P_{\hat{Y}, S} = P_{\hat{Y}} \otimes P_S \). Where \( P_{\hat{Y}} \otimes P_S \) is the product measure defined as the unique measure on \( \mathcal{P}(Y) \times \mathcal{P}(S) \) such that \( \forall y \in \mathcal{P}(Y) \forall s \in \mathcal{P}(S) \ P_{\hat{Y}} \otimes P_S(y \times s) = P_{\hat{Y}}(y)P_S(s) \). This definition of DEMPAR is equivalent to definition 2.1 for binary labels and sensitive attribute but more general because when \( \hat{Y} \) is not binary as in soft labels, this new definition is well defined. We write formally

**Definition D.1.** \( \hat{Y} \) satisfies extended DEMPAR for \( S \) if and only if: \( P_{\hat{Y}, S} = P_{\hat{Y}} \otimes P_S \).

We remark that we can not derive a quantity similar to DEMPAR-level with this definition but this extended DEMPAR assures indistinguishably of the sensitive attribute when looking at the soft labels. We have the following theorem:

**Theorem 7.1.** The following propositions are equivalent
- \( \hat{Y} \) satisfies extended DEMPAR for \( S \)
- The balanced accuracy of ADAPTAIA-S is \( \frac{1}{2} \)
Figure 13: DemPar-Level for EGD: We observe that DemPar-Level is lower for EGD than the baseline indicating \( f_{\text{reg}} \) is fair after EGD.

Proof.

\[
\forall a \; P(\hat{Y} \in a^{-1}((0))|S = 0) + P(\hat{Y} \in a^{-1}((1))|S = 1) = 1
\]

\[\iff \forall a \; P(\hat{Y} \in a^{-1}((0))|S = 0) = P(\hat{Y} \in a^{-1}((1))|S = 1)\]

\[\iff \forall a \; P(\hat{Y} \in A)|S = 0) = P(\hat{Y} \in A|S = 1)\]

\[\iff P_{Y,S} = P_Y \otimes P_S\]

Adversarial Debiasing aims to imposing \( P_{Y,S} = P_Y \otimes P_S \) hence Adversarial Debiasing mitigates AdaptAA-S.  

Additional Results: AdaptAA-S. We present additional results for COMPAS and LFW in Figure 14 showing that AdvDebias is effective to mitigate AdaptAA-S consistent with the results in Section 7.  

Additional Results: AdaptAA-H. Since, AdvDebias reveals soft labels, we can convert them into hard labels to evaluate against AdaptAA-H. We present eh evaluation for all the datasets in Figure 15. We observe that AdvDebias can successfully lower the attack accuracy of AdaptAA-H. Additionally, we note that the theoretical bound on attack accuracy matches with the empirical attack accuracy.  

Additional Results: Impact on Utility. We now present the impact on \( f_{\text{reg}}'s \) utility on using AdvDebias on remaining datasets: COMPAS and LFW. Similar to results in Section 7, we observe a drop in accuracy in Figure 16 on using AdvDebias which is a known trade-off observed in prior works.  

Additional Results: Fairness. For COMPAS and LFW, we measure DemPar-level with and without AdvDebias (Figure 17). Similar to results in Section 7, we note that there is a decrease in DemPar-level on using AdvDebias indicating that AdvDebias is effective in achieving group fairness.

Figure 14: We observe that AdvDebias reduces the attack accuracy to random guess (~50%).

E EGD + EqOdds

In this section, we theoretically show that EqOdds cannot mitigate AdaptAA-H and by consequence AdaptAA-S.  

Theorem E.1. If \( \hat{Y} \) satisfies EqOdds for \( Y \) and \( S \) then the balanced accuracy of AdaptAA-H is \( \frac{1}{2} \) iff \( Y \) is independent of \( S \) or \( \hat{Y} \) is independent of \( Y \).

Theorem E.1. If \( \hat{Y} \) satisfies EqOdds for \( Y \) and \( S \) then the balanced accuracy of AdaptAA-H is \( \frac{1}{2} \) if and only if \( \hat{Y} \) is independent of \( S \) or \( \hat{Y} \) is independent of \( Y \).

Proof. Let \( f_{\text{att}} \) be the attack model trained for AS: \( \hat{S} = a \circ \hat{Y} \). By the total probability formula

\[
P(\hat{S} = 0|S = 0) = P(\hat{S} = 0|S = 0, Y = 0)P(Y = 0|S = 0)
\]

\[+ P(\hat{S} = 0|S = 0, Y = 1)P(Y = 1|S = 0)\]

and as well

\[
P(\hat{S} = 1|S = 1) = P(\hat{S} = 1|S = 1, Y = 0)P(Y = 0|S = 1)
\]

\[+ P(\hat{S} = 1|S = 1, Y = 1)P(Y = 1|S = 1)\]

Then we substitute those terms in the definition of the balanced accuracy of the target model:

\[
P(\hat{S} = 0|S = 0) + P(\hat{S} = 1|S = 1)
\]

\[= \frac{1}{2} + \frac{1}{2} (P(Y = 0|S = 0) - P(Y = 0|S = 1))
\]

\[
= (P(\hat{Y} \in f_{\text{att}}^{-1}(0)|S = 1, Y = 0) - P(\hat{Y} \in f_{\text{att}}^{-1}(0)|S = 1, Y = 1))
\]

The balanced accuracy is equal to 0.5 if and only if \( P(Y = 0|S = 0) = P(Y = 0|S = 1) \) or \( \forall \text{att} P(Y \in f_{\text{att}}^{-1}(0)|S = 1, Y = 0) = P(\hat{Y} \in f_{\text{att}}^{-1}(0)|S = 1, Y = 1) \). The first term indicates that \( Y \) is independent of \( S \) and the second term indicates that \( S = 1 \) the \( f_{\text{reg}} \) random guess utility. We can do the same computing for \( S = 0 \) and obtain a similar conclusion.
those two conditions are unlikely to happen in the real-world. The condition of $Y$ being independent of $S$ was not observed for our datasets. We evaluate $|P(Y = 0|S = 0) - P(Y = 0|S = 1)|$ where a high value indicates $Y$ and $S$ are dependent. For Race and Sex, we found these values to be 0.05 and 0.27 (COMPAS), 0.20 and 0.13 (CENSUS) and 0.07 and 0.13 (MEPS) respectively. Further, the independence between $Y$ and $S$ means that $f_{\text{freq}}$ has random guess utility. Hence, in practice, EqODos aligns by reducing the risk to AIs but does not perfectly align as seen in DemPar by reducing AIs to random guessing. The choice of fairness metric is important for EGD for perfect alignment.

**Figure 15:** For AdaptAIA-H, we observe that AdvDebias reduces the attack accuracy to random guess (~50%).

**Figure 16:** Utility degradation for AdvDebias: We observe a statistically significant drop in $f_{\text{freq}}$’s accuracy on using AdvDebias.

**Figure 17:** DemPar-Level for AdvDebias: We observe that DemPar-Level is lower for AdvDebias indicating $f_{\text{freq}}$ is fair.

**F IMPACT OF $Ado$’S $D_{\text{aux}}$ DISTRIBUTION**

Recall from Section 4, the distribution of $D_{\text{aux}}$ is the same as $D_{\text{tr}}$. However, in practice, $Ado$ can sample $D_{\text{aux}}$ from a different distribution. To evaluate this, we consider CENSUS but from different US states by using [10]. We evaluate the success of AdaptAIA-H and AdaptAIA-S against EGD, AdvDebias and a baseline without any fairness specific algorithm for different samples of $D_{\text{aux}}$. We trained $f_{\text{freq}}$ on the CENSUS data of Alabama and then we set $D_{\text{aux}}$ to be every other state one after the other. That way we obtain an attack balanced accuracy per state.

To evaluate the shift of distribution from Alabama’ CENSUS to the other state, we use the quality report provided by the Synthetic Data Vault [41]. This score leverages multiple statistical metrics to evaluate how close two distributions are. We then scale this score in $[0, 1]$: a score of 1 indicates no shift while a score of 0 indicates the farther away state from Alabama in terms of distribution.
We present our results in Figure 18. As expected, we observe that EGD and AdvDebias are still successful in mitigating ADAPTAIA-H and ADAPTAIA-S regardless of $Adv$’s $D_{aux}$ distribution.

Figure 18: Impact on attack success of distribution shift between the training set of the target model, $D_{tr}$, and the auxiliary knowledge used by the adversary to conduct the attack, $D_{aux}$ (CENSUS dataset).