Sustaining Fairness via Incremental Learning

Somnath Basu Roy Chowdhury, Snigdha Chaturvedi

University of North Carolina at Chapel Hill
{somnath, snigdha}@cs.unc.edu

Abstract

Machine learning systems are often deployed for making critical decisions like credit lending, hiring, etc. While making decisions, such systems can encode the user's demographic information (like gender, age) in their intermediate representations. This can lead to decisions that are biased towards specific demographics. Prior work has focused on debiasing intermediate representations to ensure fair decisions. However, these approaches fail to remain fair with changes in the task or demographic distribution. To ensure fairness in the wild, it is important for a system to adapt to such changes as it accesses new tasks in an incremental fashion. In this work, we propose to address this issue by introducing the problem of learning fair representations in an incremental learning setting. To this end, we present **F**airness-**a**ware **I**ncremental **R**epresentation **L**earning (FaIRL), a representation learning system that can sustain fairness while incrementally learning new tasks. FaIRL is able to achieve fairness and learn new tasks by controlling the rate-distortion function of the learned representations. Our empirical evaluations show that FaIRL is able to make fair decisions while achieving high performance on the target task, outperforming several baselines.

Introduction

An increasing number of organizations are leveraging machine learning solutions for making decisions in critical applications like hiring (Dastin 2018), criminal recidivism (Larson et al. 2016), etc. Machine learning systems can often rely on a user's demographic information, like gender, race, and age (*protected attributes*), encoded in their representations (Elazar and Goldberg 2018) to make decisions, resulting in biased outcomes against certain demographic groups (Mehrabi et al. 2021; Shah, Schwartz, and Hovy 2020). Numerous works try to achieve fairness through unawareness (Apfelbaum et al. 2010) by debiasing model representations from protected attributes (Blodgett, Green, and O'Connor 2016; Elazar and Goldberg 2018; Elazar et al. 2021; Chowdhury and Chaturvedi 2022). However, these techniques are only able to remove in-domain spurious correlations and fail to generalize to new data distributions (Barrett et al. 2019). For example, let us consider a fair resume screening system that was trained only on resumes of software engineering roles. The system may not remain fair while screening for roles like sales or marketing, where the gender demographic distribution may be different. Similarly, a fair system also needs to be robust to shifts in data distribution (e.g. new applicants may report scores on specific tests that didn't appear in the training data) and task changes (e.g. resumes being screened for new roles like social media manager). In such cases, it is not always practical to retrain the system from scratch every time new data comes in because of the resources and environmental impact associated with training modern machine learning systems.

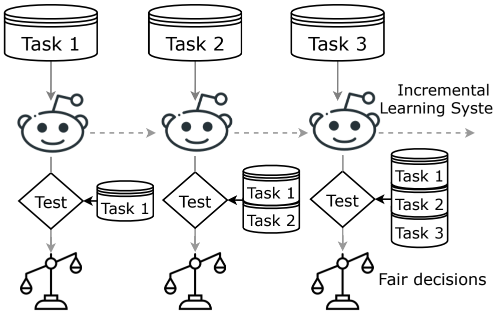

Figure 1: Illustration of a fair representation learning system in an incremental setting. The system is expected to make fair decisions while incrementally learning new tasks.

Previous works focused on improving the robustness of fair learning models by considering shifts in data distribution. These involve learning fair models under covariate shift (Rezaei et al. 2021; Singh et al. 2021) or for streaming data (Zhang et al. 2021; Zhang and Ntoutsi 2019), but these systems do not incrementally learn new tasks. In this work, we introduce the problem of *learning fair representations in an incremental learning setting*. In this setting, data from new tasks, with different underlying demographic distributions, pour in at consecutive training stages and the system has to perform well on all tasks seen so far while making fair decisions (see Figure 1). This setup is quite similar to incremental learning (Rebuffi et al. 2017). However, most works in incremental learning literature focus on target task performance without considering the fairness of their predictions.

To address this problem, we propose a representation learning system – **F**airness-**a**ware **I**ncremental

Representation **L**earning (FaIRL). At its core, FaIRL uses an adversarial debiasing setup for removing demographic information by controlling the number of bits (*rate-distortion*) required to encode the learned representations (Yu et al. 2020; Ma et al. 2007). We leverage this debiasing setup for incremental learning using an exemplar-based approach, by retaining a small set of representative samples from previous tasks, to prevent forgetting. Empirical evaluations show that FaIRL outperforms baseline incremental learning systems in fairness metrics while successfully learning target task information. Our key contributions are:

- We propose FaIRL, a representation learning system that learns fair representations, while incrementally learning new tasks, by controlling their rate-distortion function.
- We show using empirical evaluations that FaIRL outperforms baseline incremental learning systems in making fair decisions while performing well on the target task.
- We also perform extensive analysis experiments to investigate the functioning of FaIRL.

Related Work

In this section, we discuss some of the prior works on fairness in varying setups and incremental learning.

Fair Representation Learning. Zemel et al. (2013) introduced the problem of learning fair representations as an optimization task. Following works (Zhang, Lemoine, and Mitchell 2018; Li, Baldwin, and Cohn 2018; Elazar and Goldberg 2018; Chowdhury et al. 2021) leveraged an adversarial framework (Goodfellow et al. 2014) to achieve fairness, where a discriminator tries to extract demographic information from intermediate representations while performing prediction. Different from these, (Bahng et al. 2020) proposed to learn fair representations, without using protected attribute annotation, by making representations uncorrelated with ones retrieved from a biased classifier. However, these techniques require a target task at hand and are often difficult to train (Elazar and Goldberg 2018). Another line of work introduced by (Bolukbasi et al. 2016), focuses on debiasing representations independent of a target task. These approaches (Ravfogel et al. 2020; Bolukbasi et al. 2016) iteratively identify subspaces that encode protected attribute information, and project vectors onto their corresponding nullspaces. Another line of work (Cheng et al. 2020; Dixon et al. 2018), use counterfactual data augmentation approaches to debias sentence embeddings. Recently, Chowdhury and Chaturvedi (2022) proposed a debiasing framework that makes representations from same protected attribute class uncorrelated by maximizing their rate-distortion function. Despite showcasing promise in a single domain, these frameworks fail to remain fair for out-of-distribution data (Barrett et al. 2019).

Fairness under distribution shift. Several works (Rezaei et al. 2021; Singh et al. 2021) have investigated the robustness of fair classifiers under covariate shift. These works identify conditions where fairness can be sustained given shifts in data and label distribution. Efficacy of fair classifiers has also been studied in online settings (Zhang et al. 2021; Zhang and Ntoutsi 2019), where the data distribution continually evolves depending on the input data stream. However, both lines of work consider a fixed task description at initiation and do not learn new tasks while training.

Incremental Learning. Li and Hoiem (2017) introduced the task of incremental learning and proposed a dynamic architecture leveraging a knowledge distillation loss to prevent catastrophic forgetting (McCloskey and Cohen 1989). Since then, works on incremental learning can be classified into three broad categories: (a) *Regularization-based* approaches (Li and Hoiem 2017; Kirkpatrick et al. 2017; Zenke, Poole, and Ganguli 2017; Castro et al. 2018; Chan et al. 2021) use a penalty measure to ensure model parameters crucial for previous tasks do not change abruptly; (b) *Dynamic architecture-based* approaches (Long et al. 2015; Rusu et al. 2016; Li et al. 2019) introduce new task-specific parameters to prevent interference with parameters from previous tasks. These architectures grow linearly with the number of tasks having a heavy memory footprint; (c) *Exemplar-based* approaches (Rebuffi et al. 2017; Chaudhry et al. 2019a,b; Tong et al. 2022) maintain a small memory of representative samples from previous tasks and replay them to prevent catastrophic forgetting. Our framework FaIRL is similar to that of Tong et al. (2022), as we also control the rate-distortion of learned representations. However, we also consider the fairness of the predictions by ensuring protected information does not get encoded in the representations.

Background

In this section, we discuss the fundamental concepts of rate-distortion theory that form the building blocks of our framework, FaIRL.

Rate Distortion. In information theory (Cover 1999), the compactness of a distribution is measured by their *coding length* – number of bits required to encode it. In lossy data compression, a set of vectors $Z = \{z_1, \ldots, z_n\} \in \mathbb{R}^{n \times d}$, sampled from a distribution $P(Z)$, is encoded using a coding scheme, such that the transmitted vectors $\{\hat{z}_i\}_{i=1}^{n}$ can be recovered up to a distortion ϵ. The minimal number of bits required per vector to encode the sequence Z is defined by the *rate-distortion* function $R(Z, \epsilon)$. The optimal $R(Z, \epsilon)$ for vectors Z sampled from a multivariate Gaussian $\mathcal{N}(0, \Sigma)$ is:

$$R(Z, \epsilon) = \frac{1}{2} \log_2 \det \left(I + \frac{d}{n\epsilon^2} ZZ^T\right) \qquad (1)$$

where n is the number of vectors and d is the dimension of individual vectors. Equation 1 provides a tight bound even in cases where the underlying distribution $P(Z)$ is degenerate (Ma et al. 2007).

In general scenarios, e.g. image representations for multi-label classification, the vector set Z can arise from a mixture of class distributions. In such cases, the overall rate-distortion function can be computed by splitting the vectors into multiple subsets: $Z = Z^1 \cup Z^2 \ldots \cup Z^k$, where Z^j is the subset from the j-th distribution. We can then compute $R(Z^j, \epsilon)$ (Equation 1) for each subset. To facilitate this computation, we leverage a global membership matrix $\Pi = \{\Pi_j\}_{j=1}^{k}$, which is a set of k matrices encoding membership information in each subset. The membership matrix for a subset Z^j is a diagonal matrix defined as: $\Pi_j = \text{diag}(\pi_{1j}, \pi_{2j}, \ldots, \pi_{nj}) \in \mathbb{R}^{n \times n}$, where $\pi_{ij} \in [0, 1]$ is the probability that z_i belongs to Z^j. The matrices satisfy the following constraints: $\sum_j \Pi_j = I_{n\times n}$, $\sum_j \pi_{ij} = 1$, $\Pi_j \succeq 0$. The optimal number of bits to encode Z is given as:

$$R_c(Z, \epsilon|\Pi) = \sum_{j=1}^{k} \frac{\text{tr}(\Pi_j)}{2n} \log_2 \det \left(I + \frac{d}{\text{tr}(\Pi_j)\epsilon^2} Z\Pi_j Z^T\right)$$

The expected number of vectors in a subset Z^j is $\text{tr}(\Pi_j)$ and the corresponding covariance is $\text{cov}(Z_j) = \frac{1}{\text{tr}(\Pi_j)} Z\Pi_j Z^T$. For multi-class data, a vector z_i can only be a member of a single class, we restrict $\pi_{ij} = \{0, 1\}$ and the covariance matrix for j-th subset is $Z^j(Z^j)^T$.

Maximal Coding Rate (MCR^2). Yu et al. (2020) introduced a classification framework by learning discriminative representations using the rate-distortion function. Given n input samples $X = \{x_i\}_{i=1}^n$ belonging to k distinct classes, their representations $Z = \{z_i\}_{i=1}^n$ are obtained using a deep network $f_\theta(x)$. The network parameters (θ) are learned by maximizing a representation-level objective using rate-distortion called maximal coding rate (MCR^2):

$$\max_{\theta} \Delta R(Z, \Pi) = R(Z, \epsilon) - R_c(Z, \epsilon|\Pi) \qquad (2)$$

where Π captures the class label information. To have discriminative representations, same class representations should resemble each other while being different from representations from other classes. This can be achieved by maximizing the overall volume $R(Z, \epsilon)$ and compressing representations within each class by minimizing $R_c(Z, \epsilon|\Pi)$. We provide further details in Appendix B.

Fairness-aware Incremental Representation Learning (FaIRL)

Debiasing Framework

We present a novel adversarial debiasing framework that controls the rate-distortion function of the learned representations. We use rate-distortion in this debiasing framework as it is amenable to incremental learning.

Figure 2 illustrates our proposed adversarial framework. It consists of a feature encoder ϕ and a discriminator D. The feature encoder takes as input a data point x and generates representations $z = \phi(x)$. Its goal is to learn representations that are discriminative for the target attribute $\mathbf{y}$ and not informative about protected attribute $\mathbf{g}$. The discriminator network takes as input the representations produced by feature encoder z and generates $z' = D(z)$. Its goal is to extract protected attribute $\mathbf{g}$ information from z'. The discriminator is trained by maximizing the MCR^2 objective function:

$$\max_{D} \Delta R(Z', \Pi^{\mathbf{g}}) = R(Z', \epsilon) - R_c(Z', \epsilon|\Pi^{\mathbf{g}}) \qquad (3)$$

where $\Pi^{\mathbf{g}}$ is the membership matrix encoding the protected attribute information. The encoder is trained by optimizing the given objective function:

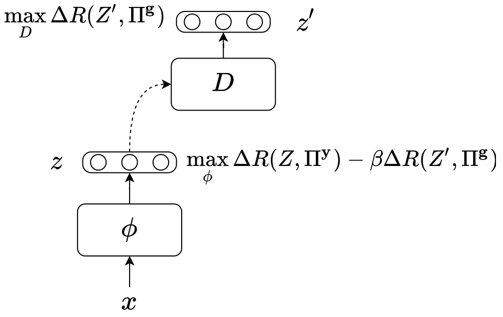
$\max_D \Delta R(Z', \Pi^{\mathbf{g}})$
z'
D
z
$\max_\phi \Delta R(Z, \Pi^{\mathbf{y}}) - \beta\Delta R(Z', \Pi^{\mathbf{g}})$
ϕ
x

Figure 2: Workflow of our debiasing framework. The discriminator tries to extract protected attribute information by optimizing $\Delta R(Z', \Pi^{\mathbf{g}})$. The feature encoder tries to learn discriminative representations for the target task ($\mathbf{y}$) using MCR^2 objective while minimizing the discriminator loss.

$$\max_{\phi} \Delta R(Z, \Pi^{\mathbf{y}}) - \beta\Delta R(Z', \Pi^{\mathbf{g}}) \qquad (4)$$

where $\Pi^{\mathbf{y}}$ is the membership matrix encoding the target attribute information and β is a hyperparameter. Empirically we observed that the proposed debiasing framework is competitive with other debiasing setups in non-incremental learning settings.

FaIRL's debiasing framework leverages the MCR^2 objective (Equation 2) for classification. MCR^2 objective, by itself, is not amenable to incremental learning for reasons discussed below. Yu et al. (2020) showed that using MCR^2 it is possible to learn representations with low-dimensional orthogonal subspaces corresponding to each class. However, naively maximimizing the MCR^2 objective results in representations spanning the complete feature space (an $\mathbb{R}^d$-dimensional feature space can accommodate a maximum of d orthogonal subspaces). This is not ideal for incremental learning as representations from new classes cannot be accommodated in the same feature space. For incremental learning, representations learned at a given training stage should be compact and not span the entire feature space. In FaIRL, we empirically observe that the feature spaces learned at each training stage are compact. This happens because while learning discriminative representations using the MCR^2 objective ($\Delta R(Z, \Pi^{\mathbf{y}})$), the encoder also tries to remove protected information by minimizing $\Delta R(Z', \Pi^{\mathbf{g}})$ (Equation 4). Minimizing $\Delta R(Z', \Pi^{\mathbf{g}})$ makes representations from different protected classes similar, resulting in a compact feature space. The $\Delta R(Z', \Pi^{\mathbf{g}})$ term acts as a natural regularizer to the MCR^2 objective, and prevents the learned representations from expanding in an unconstrained manner, making them suitable for incremental learning. Next, we discuss how we extend this debiasing framework to incremental learning in the following section.

Incremental Learning

For incremental learning, we use an *exemplar-based* approach (Rebuffi et al. 2017; Chaudhry et al. 2019a,b). We store a small set of exemplars from old tasks $\mathcal{X}_{old}$ =

$\{\mathcal{X}^1_{old}, \ldots, \mathcal{X}^m_{old}\}$, where m is the number of target classes ($m = c(t-1) < k$) the system has encountered so far (each training step introduces c target classes, k is the total number of classes). At training stage t, we have a set of new data samples $\mathcal{X}_{new}$ and exemplar set $\mathcal{X}_{old}$ ($\mathcal{X}_{old} = \emptyset$ at $t = 0$). The goal of our system is to learn discriminative representations w.r.t $\mathbf{y}$ for $\mathcal{X}_{new}$ while retaining the old representation subspaces of $\mathcal{X}_{old}$. To ensure fairness, the system also needs to learn representations that are oblivious to the protected attribute $\mathbf{g}$ for both $\mathcal{X}_{new}$ and $\mathcal{X}_{old}$. We will refer to the representations for the old and new data as $Z_{old} = \phi(\mathcal{X}_{old})$ and $Z_{new} = \phi(\mathcal{X}_{new})$ respectively.

Discriminator. In the incremental learning setup, the discriminator tries to extract protected attribute information for $\mathcal{X}_{new}$. This is achieved by maximizing $\Delta R(Z'_{new}, \Pi^{\mathbf{g}}_{new})$, where $Z'_{new} = D(\phi(\mathcal{X}_{new}))$, and $\Pi^{\mathbf{g}}_{new}$ encodes protected attribute $\mathbf{g}$ information for $\mathcal{X}_{new}$.

Feature encoder. The objective of the feature encoder is to learn fair representations that are discriminative for both old and new tasks. To achieve this, the system should have the following properties:

(a) The system should learn representations for $\mathcal{X}_{new}$ that are informative about $\mathbf{y}$. This can be achieved by learning discriminative representations for $\mathcal{X}_{new}$ by *maximizing* the MCR^2 objective: $\Delta R(Z_{new}, \Pi^{\mathbf{y}}_{new})$.

(b) The system should not reveal protected information and learn fair representations for $\mathcal{X}_{new}$. This is achieved by *minimizing* the discriminator loss $\Delta R(Z'_{new}, \Pi^{\mathbf{g}}_{new})$ (Equation 4).

(c) The system should retain knowledge about old tasks encountered in previous training stages. FaIRL maintains an exemplar set $\mathcal{X}_{old}$ and tries to retain the subspace structure learned for these samples. To ensure that encoder ϕ_t at training stage t retains the subspace structure of old representations, we *minimize* the function:

$$\begin{aligned}\Delta R(Z_{old}, \bar{Z}_{old}) &= \sum_{i=1}^{m} \Delta R(Z^i_{old}, \bar{Z}^i_{old}) \\ &= \sum_{i=1}^{m} R(Z^i_{old} \cup \bar{Z}^i_{old}) - \frac{1}{2}\left[R(Z^i_{old}) + R(\bar{Z}^i_{old})\right]\end{aligned} \tag{5}$$

where $\bar{Z}_{old} = \phi_{t-1}(\mathcal{X}_{old})$ are exemplar representations obtained using the encoder at the previous training stage $(t-1)$, and Z^j_{old} are exemplar representations from the j-th target class. $\Delta R(Z^j_{old}, \bar{Z}^j_{old})$ measures the similarity between the representation sets Z^j_{old} and $\bar{Z}^j_{old}$ by computing the difference in the number of bits required to encode them jointly and separately.

(d) The system should learn fair representations for $\mathcal{X}_{old}$. This is achieved by *minimizing* the discriminator loss for the exemplars $\Delta R(Z'_{old}, \Pi^{\mathbf{g}}_{old})$ (Equation 4).

The overall objective function that the encoder optimizes in the incremental learning setup:

$$\begin{aligned}\max_{\phi}\ & \underbrace{\Delta R(Z_{new}, \Pi^{\mathbf{y}}_{new})}_{\text{(a)}} - \beta \underbrace{\Delta R(Z'_{new}, \Pi^{\mathbf{g}}_{new})}_{\text{(b)}} \\ & - \gamma \underbrace{\Delta R(Z_{old}, \bar{Z}_{old})}_{\text{(c)}} - \eta \underbrace{\Delta R(Z'_{old}, \Pi^{\mathbf{g}}_{old})}_{\text{(d)}}\end{aligned} \tag{6}$$

where $Z'_{old} = D(Z_{old})$, $\Pi^{\mathbf{y}}_{new}$ is the membership matrix encoding target class labels for $\mathcal{X}_{new}$, $\Pi^{\mathbf{g}}_{new}$ and $\Pi^{\mathbf{g}}_{old}$ encode protected class labels for $\mathcal{X}_{new}$ and $\mathcal{X}_{old}$ respectively. In the following section, we discuss the selection of representative samples $\mathcal{X}_{old}$ from old classes.

Algorithm 1: Prototype Sampling

```
1: Input: $Z_t = \{\phi(\mathcal{X}^1_t), \ldots, \phi(\mathcal{X}^c_t))\}$ representations of $c$ classes at training stage $t$, reservoir of old samples $\mathcal{X}_{old}$.
2: $\mathcal{X}^t_{old} = \emptyset$        ▷ exemplars for training stage $t$
3: for $i = 1, \ldots, c$ do
4:     $V^i = \texttt{PCA}(Z^i_t)$        ▷ where $Z^i_t = \mathcal{X}^i_t$
5:     $V^i_k = [v_1, \ldots, v_k] \sim$ top-$k(V^i)$        ▷ top-$k$ eigen vectors selected based on singular values
6:     for $j = 1, \ldots, r$ do
7:         $s = v_i^T Z^i_t$        ▷ similarity scores
8:         $\mathcal{X}^i_{old} \sim$ top-$q(\mathcal{X}^i_t)$        ▷ select top $q = \frac{r}{k}$ samples based on similarity scores $s$
9:         $\mathcal{X}^t_{old} = \mathcal{X}^t_{old} \cup \mathcal{X}^i_{old}$
10:    end for
11: end for
12: $\mathcal{X}_{old} = \mathcal{X}_{old} \cup \mathcal{X}^t_{old}$        ▷ add to exemplar set
13: return $\mathcal{X}_{old}$
```

Exemplar Sample Selection

As discussed in previous section, we maintain exemplars $\mathcal{X}_{old} = \{\mathcal{X}^1_{old}, \ldots, \mathcal{X}^m_{old}\}$ belonging to m classes, which is useful for retaining information from previous tasks. For each class, we select r (where $r \ll |\mathcal{X}^i|$) samples $\mathcal{X}^i_{old} \sim \mathcal{X}^i$ by using one of the following sampling techniques:

Random Sampling. We randomly select r samples from each class set $\mathcal{X}^i_{old} \overset{r}{\sim} \mathcal{X}^i$.

Prototype Sampling. We use prototype sampling (Tong et al. 2022) for selecting representative samples for each class. The detailed pseudo-code is presented in Algorithm 1. In this technique, we compute the top k eigenvectors for the set of representations for each class $Z^i_t = \phi(\mathcal{X}^i_t)$ at training stage t. For each eigenvector, we select r/k data samples ($\mathcal{X}^t_{old}$) with the highest similarity scores (line 7). The selected samples are added to $\mathcal{X}_{old}$.

Submodular Optimization. We use submodular optimization (Krause and Golovin 2014) to select representative samples that summarize features of a set. Submodular optimization focuses on set functions which have the diminishing return property. Formally, a submodular function f satifies the property: $f(Z \cup \{s\}) - f(Z) \geq f(Y \cup \{s\}) - f(Y)$, where $Z \subseteq Y \subseteq S$, $s \in S$, and $s \notin Y$.

We construct a submodular function computed using representations Z that capture their diversity. We select r samples that maximizes f. Specifically, we use the facility lo-

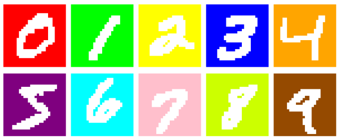

Figure 3: Representative samples from Biased MNIST dataset. We show an example from each class.

cation algorithm (Frieze 1974), which selects r representative samples from a set Z with n elements ($n > r$). For any subset $S \subseteq Z$, the submodular function f is: $f(S) = \sum_{z \in Z} \max_{s \in S} \text{sim}(s, z)$, where $\text{sim}(\cdot, \cdot)$ is the similarity measure between s and z. In our experiments, Z is the set of data representations and we use euclidean distance as our similarity measure $\text{sim}(s, z) = -||s - z||_2^2$.

Evaluation

In this section, we discuss the datasets, experimental setup, and metrics used for evaluating FaIRL. Additional details of our experimental setup can be found in Appendix B. Our implementation of FaIRL is publicly available at `https://github.com/brcsomnath/FaIRL`.

Datasets

We tackle the problem of fairness in an incremental learning setup, where there are no existing benchmarks.[1] We perform evaluations by re-purposing existing datasets.

Biased MNIST. We follow the setup of (Bahng et al. 2020) to generate a synthetic dataset using MNIST (LeCun et al. 1998), by making the background colors highly correlated with the digits. In the training set, the digit category (*target attribute*) is associated with a distinct background color (*protected attribute*) with probability p or a randomly chosen color with probability $1-p$. In the test set, each digit with assigned one of the 10 colors randomly. We evaluate the generalization ability of FaIRL for $p = \{0.8, 0.85, 0.9, 0.95\}$. We simulate incremental learning by providing the system access to 2 classes at each training stage (a total of 5 stages).

Biography classification. We re-purpose the BIOS dataset (De-Arteaga et al. 2019) for incremental learning. BIOS contains biographies of people that are associated with a profession (*target attribute*) and gender label (*protected attribute*). There are 28 different profession categories and 2 gender classes. The demographic distribution can vary vastly depending on the profession (e.g. '*software engineer*' role is skewed towards men while the '*yoga teacher*' role is most associated with females). The detailed demographic distribution is reported in Appendix B.In our setup, the system is presented with samples from 5 classes at each training stage (a total of 6 training stages).

Baselines

We compare FaIRL with the following systems:

- **Incremental Learning systems**. We report the performance of the following incremental systems: (a) LwF (Li and Hoiem 2017) is a dynamic architecture with shared and task specific parameters, with additional parameters being incorporated incrementally for new tasks. LwF uses a knowledge distillation loss along with the current task loss to prevent catastrophic forgetting; (b) Adversarial LwF – we introduced an adversarial head in LwF for fair incremental learning that tries to remove protected attribute information via gradient reversal; (c) iCaRL (Rebuffi et al. 2017) is an exemplar-based approach that uses a knowledge distillation loss to learn representations. iCaRL uses a nearest class mean classifier for performing prediction.
- **Joint learning systems**. We report the performance of the following joint learning systems, where the system has access to the entire dataset in a single training stage: (a) AdS (Chowdhury et al. 2021) is an adversarial debiasing framework that maximizes the entropy of discriminator output. (b) FaRM (Chowdhury and Chaturvedi 2022) is a state-of-the-art system for both constrained and unconstrained debiasing, which performs debiasing by controlling the rate-distortion function of representations; (c) FaIRL (joint). We report the performance of our framework when trained on full data.

Metrics

In this section, we discuss the metrics reported. For each metric, we report the average and the value achieved at the final training stage.

- **Target Accuracy**. We follow (Elazar and Goldberg 2018; Ravfogel et al. 2020; Chowdhury et al. 2021) in evaluating the quality of the learned representations for target task ($\mathbf{y}$) by using a separate probing network.

For Biased MNIST, a fair system would be able to generalize to the test set, therefore target accuracy helps measure the fairness of the system. A *high accuracy* is desired in all settings. For both datasets, we also report group fairness metrics discussed below.

- **Group Fairness Metrics**. We evaluate the fairness of representations using the following metrics. A *low score* on these metrics indicates a fairer system.

(a) **TPR-GAP**. TPR-GAP (De-Arteaga et al. 2019) computes the difference between true positive rates between two protected groups $\text{Gap}_{\mathbf{g},y} = \text{TPR}_{g,y} - \text{TPR}_{\bar{g},y}$, where $g, \bar{g}$ are possible values of the protected attribute. (Romanov et al. 2019) proposed a single fairness score by computing the root mean square of $\text{Gap}_{\mathbf{g},y}$: $\text{Gap}_{\mathbf{g}}^{\text{RMS}} = \sqrt{1/|\mathcal{Y}| \sum_{y \in \mathcal{Y}} (\text{Gap}_{\mathbf{g},y})^2}$, where $\mathcal{Y}$ is the target label set.

(b) **Demographic Parity (DP).** DP measures the difference in target prediction rate w.r.t to protected attribute $\mathbf{g}$. Mathematically, it is expressed as:

$$\text{DP} = \sum_{y \in \mathcal{Y}} |p(\hat{\mathbf{y}} = y | \mathbf{g} = g) - p(\hat{\mathbf{y}} = y | \mathbf{g} = \bar{g})| \quad (7)$$

Zhao and Gordon (2019) illustrated that there is an inherent tradeoff between the utility and fairness in fair representation learning, when $\mathbf{y}$ and $\mathbf{g}$ are correlated. Accordingly, in our experiments, we observe good fairness scores often result in poor target task performance and vice-versa.

[1] Most fairness datasets have target attributes with only 2 classes (along with a binary protected attribute), making them unsuitable for evaluating incremental learning.

Method	$p = 0.8$		$p = 0.85$		$p = 0.9$		$p = 0.95$	
	Last	Avg.	Last	Avg.	Last	Avg.	Last	Avg.
Incremental Systems								
LwF (Li and Hoiem 2017)	10.3	32.4	10.3	31.5	10.6	31.3	10.3	28.6
Adversarial LwF	10.3	32.4	10.3	31.9	10.3	27.1	10.3	25.8
iCaRL (Rebuffi et al. 2017)	62.8	79.2	58.4	72.4	51.1	70.8	47.5	69.9
FaIRL (w/ random)	**81.7**	**90.4**	**77.8**	**88.2**	71.1	83.9	**59.3**	**75.7**
FaIRL (w/ proto.)	80.7	89.8	77.2	87.7	71.0	83.5	57.8	75.3
FaIRL (w/ submod.)	80.5	89.8	77.6	88.0	**72.2**	**84.4**	57.9	73.5
Joint Systems								
FaIRL (joint)	88.08	-	85.64	-	81.94	-	68.85	-
AdS (Chowdhury et al. 2021)	79.98	-	75.39	-	66.46	-	52.49	-
FaRM (Chowdhury and Chaturvedi 2022)	92.44	-	90.54	-	82.55	-	57.09	

Table 1: Evaluation accuracy of incremental and joint learning systems on Biased MNIST dataset. Performance of joint learning systems are reported in gray. FaIRL achieves the best performance among incremental learning baselines (shown in bold). In strongly correlated settings ($p = 0.95$), FaIRL is competitive with joint learning setups.

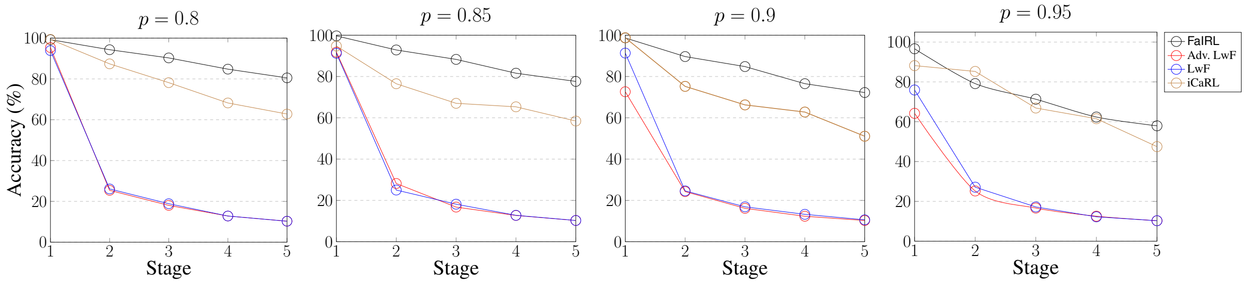

Figure 4: Test accuracy at different training stages of FaIRL and baseline incremental learning systems on Biased MNIST dataset. We observe that FaIRL significantly outperforms baseline approaches in all setups.

Results: Biased MNIST

In Table 1, we report the performance of FaIRL and baseline approaches on Biased MNIST dataset. For this dataset, high target accuracy also implies fair decisions as the training sets are biased. We observe that FaIRL outperforms the incremental learning baselines in all settings (different values of p). FaIRL with prototype and submodular exemplar selection approaches slightly fall behind random sampling. We believe that as the class samples are skewed towards a color, these sampling approaches may have ended up selecting instances based on their color instead of the digit information. It is also interesting to note that FaIRL (joint) is competitive with other state-of-the-art approaches AdS and FaRM, outperforming them when the color and digit information are strongely correlated ($p = 0.95$). In this settings ($p = 0.95$), FaIRL even in the incremental learning setting outperforms joint learning baselines. This shows that FaIRL is able to learn robust representations in challenging scenarios where the bias is highly correlated with the target task. We report the fairness metrics in Appendix D for completeness.

In Figure 4, we report the performance of incremental learning systems at various training stages. We observe that LwF suffers from catastrophic forgetting, achieving near random performance in the final stages. Adversarial LwF achieves a similar performance to LwF. We believe that the adversarial head doesn't provide an added advantage over LwF because it may encounter unseen classes of protected attribute (colors) at later training stages. iCaRL and FaIRL do not suffer from catastrophic forgetting, and in all settings FaIRL consistently outperforms other baselines.

Results: Biography Classification

We present the results of FaIRL on Biography classification in Table 2. We observe that LwF-based systems achieve poor target performance due to catastrophic forgetting. However, as most of their predictions are incorrect these systems end up with good scores on fairness metrics. Adversarial LwF performs slightly better than LwF in terms of target accuracy. iCaRL achieves the best target accuracy but performs the worst on fairness metrics. FaIRL provides a good balance between the two traits – achieving target accuracy close to iCaRL while significantly improving the fairness metrics.

We observe that FaIRL (joint) is competitive with state-of-the-art debiasing frameworks AdS and FaRM. It is interesting to note that incrementally trained FaIRL achieves better DP scores than jointly trained debiasing frameworks. We report the target accuracy and $\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ metric across training stages. In Figure 5(a), FaIRL outperforms most base-

Method	Accuracy (↑)		Fairness			
			DP (↓)		$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ (↓)	
	Last	Avg.	Last	Avg.	Last	Avg.
Incremental						
LwF	17.9	52.1	0.25	0.30	0.05	0.02
Adv. LwF	21.1	54.2	0.31	0.36	0.19	0.05
iCaRL	97.7	99.1	0.45	0.37	0.10	0.05
FaIRL (rand.)	95.1	97.5	0.42	0.35	0.06	0.03
(w/ proto.)	93.9	96.8	0.40	0.34	0.05	0.03
(w/ submod.)	94.4	97.4	0.41	0.35	0.04	0.02
Joint						
FaIRL (joint)	98.5	-	0.43	-	0.06	-
AdS	99.9	-	0.45	-	0.0	-
FaRM	99.9	-	0.42	-	0.0	-

Table 2: Target accuracy and fairness metrics achieved by FaIRL and other baseline approaches on Biographies dataset. FaIRL achieves a good balance between target accuracy and fairness metrics.

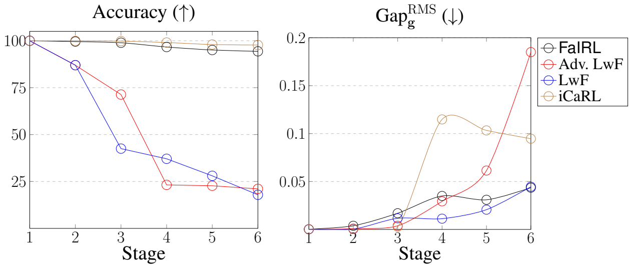

Figure 5: Evolution of target accuracy and $\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ of models at different training stages. We observe that FaIRL achieves a fine balance between accuracy and TPR-GAP.

lines, marginally falling short of iCaRL in the final training stages. However, the $\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ metric (in Figure 5(b)) for FaIRL is much better than iCaRL. LwF-based systems also achieve low scores but this is because of underfitting as evident from their low target accuracies.

Analysis

In this section, we perform several analysis experiments to investigate the functioning of FaIRL.

Task ablations. We vary the number of classes that FaIRL is presented with at a given training stage and report the average accuracy and fairness scores on Biographies dataset. In Table 3, we observe a significant drop in target performance when the number of classes (in a training stage) are reduced accompanied by an improvement in DP, reflecting the tradeoff between fairness and utility as noted by (Zhao and Gordon 2019). The complete results for all sampling strategies are reported in Appendix C.

Visualization. We visualize the UMAP (McInnes et al. 2018) feature projections before and after the debiasing process in Biographies dataset. The feature vectors are color-coded according to the protected attribute (gender). In Figure 6, we observe that before debiasing (left) it is easier to distinguish features, and after debiasing (right) features from both gender encompass similar subspaces.

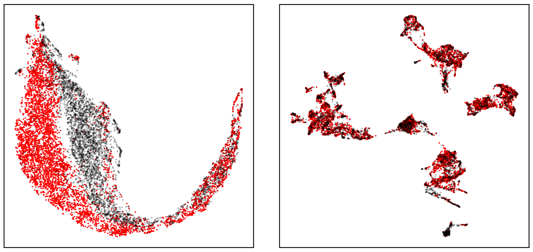

Figure 6: UMAP projections of representations from FaIRL before and after all the training stages.

# classes	Acc. (↑)	DP (↓)	$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ (↓)
2	89.47	0.26	0.046
5	97.49	0.35	0.032
10	98.57	0.39	0.031

Table 3: Performance of FaIRL with varying number of classes per training stage. We observe an improved performance when the class count per training stage is increased.

We report additional analysis experiments to investigate the memory usage, sample efficiency, robustness, and effect of exemplar size on FaIRL's performance in Appendix C.

Conclusion

In this work, we tackle the problem of learning fair representations in an incremental learning setting. To achieve this, we proposed **F**airness-**a**ware **I**ncremental **R**epresentation **L**earning (FaIRL), a representation learning system that can make fair decisions while learning new tasks by controlling the rate-distortion function of representations. Empirical evaluations show that FaIRL is able to make fair decisions outperforming prior baselines, even in scenarios where the target and protected attributes are strongly correlated. Through extensive analysis, we observe that the debiasing framework at the core of FaIRL is able to keep the feature compact, which helps FaIRL to learn new tasks in an incremental fashion. Our framework, FaIRL can make fair decisions with incremental access to unseen tasks. Such systems will be crucial for achieving fairness in the wild, as learning systems are increasingly being deployed to critical applications. Future work can focus on developing incrementally trained fair decision-making systems with minimal reliance on protected attribute annotations.

Acknowledgements

This research project was supported in part by Amazon Research Awards.

References

Apfelbaum, E. P.; Pauker, K.; Sommers, S. R.; and Ambady, N. 2010. In blind pursuit of racial equality? *Psychological science*, 21(11): 1587–1592.

Bahng, H.; Chun, S.; Yun, S.; Choo, J.; and Oh, S. J. 2020. Learning De-biased Representations with Biased Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 528–539. PMLR.

Barrett, M.; Kementchedjhieva, Y.; Elazar, Y.; Elliott, D.; and Søgaard, A. 2019. Adversarial Removal of Demographic Attributes Revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6330–6335. Hong Kong, China: Association for Computational Linguistics.

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1119–1130. Austin, Texas: Association for Computational Linguistics.

Bolukbasi, T.; Chang, K.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 4349–4357.

Castro, F. M.; Marín-Jiménez, M. J.; Guil, N.; Schmid, C.; and Alahari, K. 2018. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, 233–248.

Chan, K. H. R.; Yu, Y.; You, C.; Qi, H.; Wright, J.; and Ma, Y. 2021. ReduNet: A white-box deep network from the principle of maximizing rate reduction. *ArXiv preprint*, abs/2105.10446.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2019a. Efficient Lifelong Learning with A-GEM. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H.; and Ranzato, M. 2019b. On tiny episodic memories in continual learning. *ArXiv preprint*, abs/1902.10486.

Cheng, P.; Hao, W.; Yuan, S.; Si, S.; and Carin, L. 2020. FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders. In *International Conference on Learning Representations*.

Chowdhury, S. B. R.; and Chaturvedi, S. 2022. Learning Fair Representations via Rate-Distortion Maximization. *Transactions of the Association for Computational Linguistics*, 10: 1159–1174.

Chowdhury, S. B. R.; Ghosh, S.; Li, Y.; Oliva, J.; Srivastava, S.; and Chaturvedi, S. 2021. Adversarial Scrubbing of Demographic Information for Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 550–562. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, 296–299. Auerbach Publications.

De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Elazar, Y.; and Goldberg, Y. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 11–21. Brussels, Belgium: Association for Computational Linguistics.

Elazar, Y.; Ravfogel, S.; Jacovi, A.; and Goldberg, Y. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9: 160–175.

Frieze, A. M. 1974. A cost function property for plant location problems. *Mathematical Programming*, 7(1): 245–248.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *ArXiv preprint*, abs/1406.2661.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Krause, A.; and Golovin, D. 2014. Submodular function maximization. *Tractability*, 3: 71–104.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, 9(1): 3–3.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to Grow: A Continual Structure Learning Framework for Overcoming Catastrophic Forgetting. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 3925–3934. PMLR.

Li, Y.; Baldwin, T.; and Cohn, T. 2018. Towards Robust and Privacy-preserving Text Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 25–30. Melbourne, Australia: Association for Computational Linguistics.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning Transferable Features with Deep Adaptation Networks. In Bach, F. R.; and Blei, D. M., eds., *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, 97–105. JMLR.org.

Ma, Y.; Derksen, H.; Hong, W.; and Wright, J. 2007. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9): 1546–1562.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; and Goldberg, Y. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7237–7256. Online: Association for Computational Linguistics.

Rebuffi, S.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental Classifier and Representation Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 5533–5542. IEEE Computer Society.

Rezaei, A.; Liu, A.; Memarrast, O.; and Ziebart, B. D. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9419–9427.

Romanov, A.; De-Arteaga, M.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; Rumshisky, A.; and Kalai, A. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4187–4195. Minneapolis, Minnesota: Association for Computational Linguistics.

Rusu, A. A.; Rabinowitz, N. C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; and Hadsell, R. 2016. Progressive neural networks. *ArXiv preprint*, abs/1606.04671.

Shah, D. S.; Schwartz, H. A.; and Hovy, D. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. Online: Association for Computational Linguistics.

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 3–13.

Tong, S.; Dai, X.; Wu, Z.; Li, M.; Yi, B.; and Ma, Y. 2022. Incremental Learning of Structured Memory via Closed-Loop Transcription. *ArXiv preprint*, abs/2202.05411.

Yu, Y.; Chan, K. H. R.; You, C.; Song, C.; and Ma, Y. 2020. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *ArXiv preprint*, abs/2006.08558.

Zemel, R. S.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, 325–333. JMLR.org.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual Learning Through Synaptic Intelligence. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 3987–3995. PMLR.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, W.; Bifet, A.; Zhang, X.; Weiss, J. C.; and Nejdl, W. 2021. Farf: A fair and adaptive random forests classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 245–256. Springer.

Zhang, W.; and Ntoutsi, E. 2019. FAHT: An Adaptive Fairness-aware Decision Tree Classifier. In *IJCAI*.

Zhao, H.; and Gordon, G. J. 2019. Inherent Tradeoffs in Learning Fair Representations. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 15649–15659.