Mitigating Information Leakage in Image Representations: A Maximum Entropy Approach

Proteek Chandan Roy and Vishnu Naresh Boddeti
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{royprote, vishnu}@msu.edu

Abstract

Image recognition systems have demonstrated tremendous progress over the past few decades thanks, in part, to our ability of learning compact and robust representations of images. As we witness the wide spread adoption of these systems, it is imperative to consider the problem of unintended leakage of information from an image representation, which might compromise the privacy of the data owner. This paper investigates the problem of learning an image representation that minimizes such leakage of user information. We formulate the problem as an adversarial non-zero sum game of finding a good embedding function with two competing goals: to retain as much task dependent discriminative image information as possible, while simultaneously minimizing the amount of information, as measured by entropy, about other sensitive attributes of the user. We analyze the stability and convergence dynamics of the proposed formulation using tools from non-linear systems theory and compare to that of the corresponding adversarial zero-sum game formulation that optimizes likelihood as a measure of information content. Numerical experiments on UCI, Extended Yale B, CIFAR-10 and CIFAR-100 datasets indicate that our proposed approach is able to learn image representations that exhibit high task performance while mitigating leakage of predefined sensitive information.

1. Introduction

Current day machine learning algorithms based on deep neural networks have demonstrated impressive progress across multiple domains such as image classification, speech recognition etc. By stacking together multiple layers of linear and non-linear operations deep neural networks have been able to learn and identify complex patterns in data. As a by-product of these capabilities, deep neural networks have also become powerful enough to inadvertently identify sensitive information or features of data even in the absence of any additional side information. For example, consider a scenario where a user enrolls their facial image in a face recognition system for the purpose of access control. During enrollment, a feature vector is extracted from the image and stored in a database. Apart from the identity of the user, this feature vector potentially contains information that is sensitive to the user, such as the age, information that the user may never have expressly consented to provide. More generally, learned data representations could leak auxiliary information that the participants may never have intended to release. Information obtained in this manner can be used to compromise the privacy of the user or to be biased and unfair to the user. Therefore, it is imperative to develop representation learning algorithms that can *intentionally* and *permanently* obscure sensitive information while retaining task dependent information. Addressing this problem is the central aim of this paper.

A few recent attempts have been made to study related problems, such as learning censored [3], fair [14], or invariant [18] representations of data. The central idea of these approaches, collectively referred to as *Adversarial Representation Learning* (ARL), is to learn a representation of data in an adversarial setting. These approaches couple together (i) an adversarial network that seeks to classify and extract sensitive information from a given representation, and (ii) an embedding network that is tasked with extracting a compact representation of data while preventing the adversarial network from succeeding at leaking sensitive information. To achieve their respective goals, the adversary is optimized to maximize the likelihood of the sensitive information, while the encoder is optimized to minimize the same likelihood i.e., adversary's likelihood of the sensitive information, thereby leading to a zero-sum game. We will henceforth refer to this formulation as *Maximum Likelihood Adversarial Representation Learning* (ML-ARL).

The zero-sum game formulation of optimizing the likelihood, however, is practically sub-optimal from the perspective of preventing information leakage. As an illustration consider a problem where the sensitive attribute has three categories. Let there be two instances where the adversary's probability distribution of the sensitive label is (0.33, 0.17, 0.5) and (0.33., 0.33., 0.33.) and let the correct label be class 1 for both of them. In each of these cases the likelihood of the discriminator is the same i.e., $\log 0.33$ but the former instance is more informative than the latter. Moreover, the potential of this formulation to prevent information leakage is predicated upon: (i) the existence of an equilibrium, and (ii) the ability of practical optimization procedures to converge to such an equilibrium. As we will show, in practice, the conditions necessary for convergence may not be satisfied. Therefore, when the optimization does not reach the equilibrium, a probability distribution with the minimum likelihood is the distribution that is most certain with the potential to leak the most amount of information. In contrast, the second instance is a uniform distribution over the sensitive labels and provides no information to the adversary. This solution corresponds to the maximum entropy distribution over the sensitive labels.

Contributions: Building on the observations above, we propose a framework, dubbed *Maximum Entropy Adversarial Representation Learning* (MaxEnt-ARL), which optimizes an image representation with two major objectives, (i) maximally retain information pertinent to a given target attribute, and (ii) minimize information leakage about a given sensitive attribute. We pose the learning problem in an adversarial setting as a non-zero sum three player game between an encoder, a predictor and a discriminator (proxy adversary) where the encoder tries to maximize the entropy of the discriminator on the sensitive attribute and maximizes the likelihood of the predictor on the target attribute.

We analyze the equilibrium and convergence properties of the ML-ARL as well as the proposed MaxEnt-ARL formulation using tools from non-linear systems theory. We compare and evaluate the numerical performance of ML-ARL and MaxEnt-ARL for fair classification tasks on the UCI dataset, illumination invariant classification on the Extended Yale B dataset and two fabricated tasks on the CIFAR-10 and CIFAR-100 datasets. On a majority of these tasks MaxEnt-ARL outperforms all other baselines.

2. Related Work

Adversarial Representation Learning: In the context of image classification, adversarial learning has been utilized to learn representations that are invariant across domains [4, 5, 17], thereby enabling us to train classifiers on a source domain and utilize on a target domain.

The entire body of work devoted to learning fair and unbiased representations of data share many similarities to the adversarial representation learning problem. Early work on this topic did not involve an explicit adversary but shared the goal of learning representations with competing objectives. The concept of learning fair representations was first introduced by Zemel et al [19], where the goal was to learn a representation of data by "fair clustering" while maintaining the discriminative features of the prediction task. Building upon this work many approaches have been proposed to learn an unbiased representation of data while retaining its effectiveness for a prediction task. To remove influence of "nuisance variables" Louizos et al [14] proposed variational fair autoencoder (VFAE), a joint optimization framework for learning an invariant representation and a prediction task. In order to improve fairness in the representation, they regularized the marginal distribution $p(z|s)$ through Maximum Mean Discrepancy (MMD).

More recent approaches [3, 20, 1, 18] have used explicit adversarial networks to measure information content of sensitive attributes. These problems are set up as a minimax game between the encoder and the adversary. The encoder is setup to achieve fairness by maximizing the loss of the adversary i.e. minimizing negative log-likelihood of sensitive variables as measured by the adversary. Among these approaches, our proposed MaxEnt-ARL formulation is most directly related to the Adversarial Invariant Feature Learning introduced by Xie et al. [18].

Optimization Theory for Adversarial Learning: The formulation of adversarial representation learning poses unique challenges from an optimization perspective. The parameters of the models in ARL are typically optimized through stochastic gradient descent, either jointly [3, 15] or alternatively [4]. The former is, however, more commonly used in practice and is a generalization of gradient descent. While the convergence properties of gradient descent and its variants are well understood, there is relatively little work on the convergence and stability of simultaneous gradient descent in adversarial minimax problems. Recently, Mescheder *et al.* [15] and Nagarajan *et al.* [16] both leveraged tools from non-linear systems theory [9] to analyze the convergence properties of simultaneous gradient descent in the context of GANs. They show that without the introduction of additional regularization terms to the objective of the zero-sum game, simultaneous gradient descent does not converge. Our convergence analysis of ML-ARL and MaxEnt-ARL also leverages the same non-linear systems theory tools and show the conditions under which they converge.

3. Adversarial Representation Learning

The Adversarial Representation Learning setup involves observational input $\boldsymbol{x}$, a target attribute with n classes $\mathcal{Y} = \{y_1, \dots, y_n\}$ and a sensitive attribute with m classes $\mathcal{S} = \{s_1, \dots, s_m\}$. In this paper, we restrict ourselves to attributes over a discrete space with multiple labels. Our goal is to learn an embedding function that maps $\boldsymbol{x}$ to $\boldsymbol{z}$

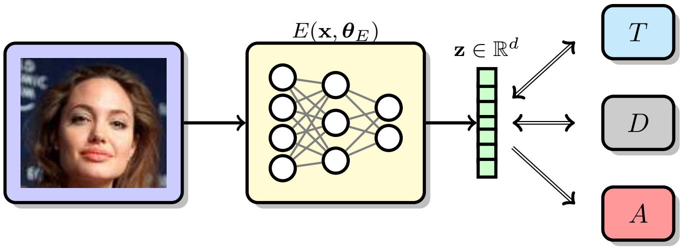

$E(\mathbf{x}, \boldsymbol{\theta}_E)$
$\mathbf{z} \in \mathbb{R}^d$
T
D
A

Figure 1: **Adversarial Representation Learning:** We consider the problem of learning an embedding function $E(\cdot, \boldsymbol{\theta}_E)$ that maps a high-dimensional image to a low-dimensional representation $\mathbf{z} \in \mathbb{R}^d$ while satisfying two competing goals: retain as much image information necessary to accurately predict a target attribute $\mathcal{Y} = \{y_1, \dots, y_n\}$ while simultaneously minimizing information leakage about a sensitive attribute $\mathcal{S} = \{s_1, \dots, s_m\}$ by an unknown adversary $A(\cdot, \boldsymbol{\theta}_A)$. The learning problem is formulated as a minimax game between $\{E(\cdot, \boldsymbol{\theta}_E), T(\cdot, \boldsymbol{\theta}_T)\}$ and a proxy adversary $D(\cdot, \boldsymbol{\theta}_D)$.

from which we can predict a target attribute $\mathcal{Y}$, while also minimizing information leakage about a known sensitive attribute $\mathcal{S}$ i.e. class labels of attribute $\mathcal{S}$.

3.1. Problem Setting

The Adversarial Representation Learning problem is formulated as a game among three players, *encoder* E, a *target predictor* T, and a *discriminator* D that serves as a proxy for an unknown *adversary* A. After E is learned and fixed, we train and evaluate an *adversary* A with the aim of leaking information of the sensitive attribute that we sought to protect. Since the adversary A is unknown to encoder at training, the encoder E is trained against the discriminator D, which thereby acts as a proxy for the unknown A. An illustration of this setting is shown in Fig. 1. The *encoder* is modeled as a deterministic function, $\boldsymbol{z} = E(\boldsymbol{x}; \boldsymbol{\theta}_E)$, the *target predictor* models the conditional distribution $p(t|\boldsymbol{x})$ via $q_T(t|\boldsymbol{z}; \boldsymbol{\theta}_T)$ and the *discriminator* models the conditional distribution $p(s|\boldsymbol{x})$ via $q_D(s|\boldsymbol{z}; \boldsymbol{\theta}_D)$, where $p(t|\boldsymbol{x})$ and $p(s|\boldsymbol{x})$ are the ground truth labels for a given target and sensitive labels t and s, respectively.

3.2. Background

In existing formulations of ARL, the goal of the encoder is to maximize the likelihood of the target attribute, as measured by the *target predictor* T, while minimizing the likelihood of the sensitive attribute, as measured by the *discriminator* D. This problem (henceforth referred to as ML-ARL) was formally defined by Xie *et al.* [18] as a three player zero-sum minimax game:

$$\min_{\boldsymbol{\theta}_E, \boldsymbol{\theta}_T} \max_{\boldsymbol{\theta}_D} J_1(\boldsymbol{\theta}_E, \boldsymbol{\theta}_T) - \alpha J_2(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) \tag{1}$$

where α is a parameter that allows us to trade-off between the two competing objectives for the encoder and,

$$\begin{aligned} J_1(\boldsymbol{\theta}_E, \boldsymbol{\theta}_T) &= KL\left(p\left(t|\boldsymbol{x}\right) \| q_T\left(t|E(\boldsymbol{x}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_T\right)\right) \\ J_2(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) &= KL\left(p\left(s|\boldsymbol{x}\right) \| q_D\left(s|E(\boldsymbol{x}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_D\right)\right) \end{aligned}$$

where the $KL(\cdot\|\cdot)$ terms reduce to the log-likelihood if the label distributions are ideal categorical distributions.

3.3. Maximum Entropy Adversarial Representation Learning

In the MaxEnt-ARL formulation the goal of the encoder is to maximize the likelihood of the target attribute, as measured by the *target predictor*, while maximizing the uncertainty in the sensitive attribute, as measured by the entropy of the *discriminator's* prediction. Formally, we define the MaxEnt-ARL optimization problem as a three player non-zero sum game:

$$\begin{aligned} &\min_{\boldsymbol{\theta}_D} V_1(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) \\ &\min_{\boldsymbol{\theta}_E, \boldsymbol{\theta}_T} V_2(\boldsymbol{\theta}_E, \boldsymbol{\theta}_T) + \alpha V_3(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) \end{aligned} \tag{2}$$

where α allows us to trade-off between the two competing objectives for the encoder and,

$$\begin{aligned} V_1(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) &= KL\left(p\left(s|\boldsymbol{x}\right) \| q_D\left(s|E(\boldsymbol{x}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_D\right)\right) \\ V_2(\boldsymbol{\theta}_E, \boldsymbol{\theta}_T) &= KL\left(p\left(t|\boldsymbol{x}\right) \| q_T\left(t|E(\boldsymbol{x}; \boldsymbol{\theta}_E); \boldsymbol{\theta}_T\right)\right) \\ V_3(\boldsymbol{\theta}_E, \boldsymbol{\theta}_D) &= KL\left(q_D\left(s|E(\boldsymbol{x}; \boldsymbol{\theta}_E)\,; \boldsymbol{\theta}_D\right) \| U\right) \end{aligned}$$

where U is the uniform distribution. The crucial difference between the MaxEnt-ARL formulation and the ML-ARL formulation is the fact that while the encoder and the discriminator have competing objectives, in ML-ARL they directly compete against each other on the same metric (likelihood of sensitive attribute), while in MaxEnt-ARL they are optimizing competing metrics that are related but not the exact same metric.

Optimizing the embedding function to maximize the entropy of the discriminator instead of minimizing its likelihood has one crucial practical advantage. Entropy maximization inherently does not need class labels for training. This is advantageous in settings where it is either, (i) Undesirable for the embedding function to have access to the sensitive label, potentially for privacy reasons., or (ii) Sensitive labels for the data points are unknown. For instance consider, a semi-supervised scenario where only the desired label is known while the sensitive label is unknown. The embedding function can learn from such data by obtaining gradients from the entropy of the discriminator.

4. Theoretical Analysis

In this section we analyze the properties of the MaxEnt-ARL formulation and compare it to the ML-ARL formulation, both in terms of equilibrium as well as convergence dynamics under simultaneous gradient descent.

4.1. Equilibrium

Theorem 1. Given a fixed encoder E, the optimal discriminator is $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*) = p(s|E(\boldsymbol{x};\boldsymbol{\theta}_E))$ and the optimal predictor is $q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*) = p(t|E(\boldsymbol{x};\boldsymbol{\theta}_E))$.

Proof. The proof uses the fact that, given a fixed encoder E, the objective is convex w.r.t. each distribution. Thus we can obtain the stationary point for $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D)$ and $q_T(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T)$ as a function of $p(s|E(\boldsymbol{x};\boldsymbol{\theta}_E))$ and $p(t|E(\boldsymbol{x};\boldsymbol{\theta}_E))$, respectively. The detailed proof is included in the supplementary material. □

Therefore, both the optimal distributions $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)$ and $q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)$ are functions of the encoder parameters $\boldsymbol{\theta}_E$. The objective for optimizing the encoder now reduces to:

$$\begin{aligned}&\min_{\boldsymbol{\theta}_E}\ \mathbb{E}_{\boldsymbol{x},t}\left[-\log q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)\right] + \log m\\ &+\alpha\mathbb{E}_{\boldsymbol{x}}\left[\sum_{i=1}^{m} q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)\log q_D(s_i|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)\right]\end{aligned}$$

where the first term is minimizing the uncertainty (negative log-likelihood) of the true target attribute label and the second term is maximizing unpredictability (as measured by entropy) across all the classes in the discriminator distribution, thereby, preventing leakage of any information about the sensitive attribute label. In contrast the corresponding objective of the ML-ARL problem is [18],

$$\begin{aligned}\min_{\boldsymbol{\theta}_E}\ &\mathbb{E}_{\boldsymbol{x},t}\left[-\log q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)\right]\\ &+\alpha\mathbb{E}_{\boldsymbol{x},s}\left[\log q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)\right]\end{aligned}$$

where the first term is minimizing the uncertainty (negative log-likelihood) of the true target attribute label, while the second term is maximizing uncertainty (log-likelihood) of only the true sensitive attribute label. However, by doing so, the encoder inadvertently becomes more certain about the other labels, and can still be informative to an adversary.

Equilibrium when $s \perp\!\!\!\perp t$**:** When the target and sensitive attributes are independent with respect to each other (e.g., age and gender), the two terms in the encoder optimization can both reach their optima simultaneously. Furthermore, the problem reduces to a non-zero sum two player game between the *encoder* and the *discriminator* in the MaxEnt-ARL case and to a zero-sum two player game between the same players in the case of ML-ARL.

Corollary 1.1. When $s \perp\!\!\!\perp t$, let the optimum discriminator and predictor for an encoder E be $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)$ and $q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)$ respectively. The optimal encoder $E(\cdot)$ in the MaxEnt-ARL formulation induces a uniform distribution in the discriminator $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E^*);\boldsymbol{\theta}_D^*)$ over the classes of the sensitive attribute.

Proof. The proof uses the fact that, given a fixed optimal discriminator D, $q_T(t|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_T^*)$ is independent of $q_D(s|E(\boldsymbol{x};\boldsymbol{\theta}_E);\boldsymbol{\theta}_D^*)$ when $s \perp\!\!\!\perp t$. The detailed proof is included in the supplementary material. □

Equilibrium when $s \not\perp\!\!\!\perp t$**:** When the target and sensitive attributes are related to each other (e.g., beard and gender), the two terms in the encoder optimization cannot reach their optima simultaneously. In both the formulations, ML-ARL and MaxEnt-ARL, the relative optimality of the two objectives depends on the trade-off factor α.

4.2. Convergence Dynamics

We analyze the standard algorithm (simultaneous stochastic gradient descent) for finding the equilibrium solution of such adversarial games. That is, we take simultaneous gradient steps in $\boldsymbol{\theta}_E$, $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_T$, which can be expressed as differential equations of the form:

$$\begin{aligned}\dot{\boldsymbol{\theta}}_D &= f_D(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}_D} V_1(\boldsymbol{\theta}_E,\boldsymbol{\theta}_D)\\ \dot{\boldsymbol{\theta}}_T &= f_T(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}_T} V_2(\boldsymbol{\theta}_E,\boldsymbol{\theta}_T)\\ \dot{\boldsymbol{\theta}}_E &= f_E(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}_E} V_2(\boldsymbol{\theta}_E,\boldsymbol{\theta}_T) + \alpha V_3(\boldsymbol{\theta}_E,\boldsymbol{\theta}_T)\end{aligned} \tag{3}$$

where the gradients $f(\boldsymbol{\theta}) = (f_D(\boldsymbol{\theta}), f_T(\boldsymbol{\theta}), f_E(\boldsymbol{\theta}))$ define a *vector field* over $\boldsymbol{\theta} = (\boldsymbol{\theta}_D, \boldsymbol{\theta}_T, \boldsymbol{\theta}_E)$.

The qualitative behavior of the aforementioned non-linear system near any equilibrium point can be determined via *linearization* with respect to that point [9]. Restricting our attention to a sufficiently small neighborhood of the equilibrium point, the non-linear state equations in (3) can be approximated by a linear state equation:

$$\dot{\boldsymbol{\theta}} = \boldsymbol{J}\boldsymbol{\theta} \tag{4}$$

where, $\boldsymbol{J} = \left.\begin{bmatrix}\frac{\partial f_D(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_D} & \frac{\partial f_D(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_T} & \frac{\partial f_D(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_E}\\ \frac{\partial f_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_D} & \frac{\partial f_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_T} & \frac{\partial f_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_E}\\ \frac{\partial f_E(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_D} & \frac{\partial f_E(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_T} & \frac{\partial f_E(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_E}\end{bmatrix}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is the Jacobian of the vector field evaluated at the chosen equilibrium point $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_D^*, \boldsymbol{\theta}_T^*, \boldsymbol{\theta}_E^*)$. For small neighborhoods

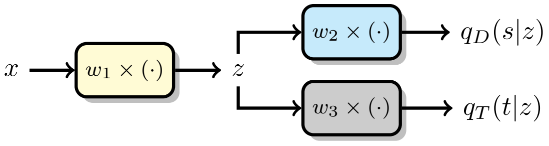


Figure 2: Three Player Game: Linear Example

around an equilibrium, the trajectories of the non-linear system in (3) is expected to be "close" to the trajectories of the linear approximate system in (4).

Theorem 2 (Linearization)**.** Let $\boldsymbol{x} = \mathbf{0}$ be an equilibrium point for the non-linear system, $\dot{\boldsymbol{x}} = f(\boldsymbol{x})$, where $f : \mathcal{D} \rightarrow \mathbb{R}^n$ is continuously differentiable and $\mathcal{D}$ is a neighborhood of the origin. Let, $\boldsymbol{J} = \frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\mathbf{0}}$. Then,

- The origin is asymptotically stable if $\text{Re}(\lambda_i) < 0$ for all eigenvalues of $\boldsymbol{J}$.
- The origin is unstable if $\text{Re}(\lambda_i) \geq 0$ for one or more of the eigenvalues of $\boldsymbol{J}$.

Proof. See Theorem 4.7 of [9]. □

5. Numerical Experiments

In this section we will evaluate the efficacy of the proposed *Maximum Entropy Adversarial Representation Learning* model and compare it with other *Adversarial Representation Learning* baselines.

5.1. Three Player Game: Linear Case

As an illustrative example we analyze the convergence of both ML-ARL and MaxEnt-ARL under the same setting. The encoder, discriminator and predictor are linear models with multiplicative weights w_1, w_2 and w_3, respectively. We limit our model to this three variable setting for ease of analysis and visualization. Both predictor and the discriminator are optimizing cross-entropy loss on binary $\{0, 1\}$ labels. To observe the game between the three players we provide same data sample $x = 1$ yet with different target and sensitive labels i.e., 4 samples with $\{00, 01, 10, 11\}$ for target and sensitive labels. Loss is calculated as the average over all samples and corresponding vector field values are also computed. The stationary point of this game, for both ML-ARL and MaxEnt-ARL, is at $(w_1 = 0, w_2 = 0, w_3 = 0)$ and the gradient of the loss functions are zero at this point. We consider a small ($30 \times 30 \times 30$ grid) neighborhood around the stationary point in the range $[-0.01, 0.01]$ for weights w_1, w_2, w_3 and visualize trajectories by following the vector field of the game.

Figure 3 shows streamline plots of the vector field around $(0, 0, 0)$ for a point starting at the green location. In the ML-ARL case, we observe that when the predictor is fixed at $w_3 = 0$, the trajectory for the encoder and the discriminator does not converge and rotates around the stationary point. In contrast, for the MaxEnt-ARL method converges to the stationary point. When $w_1 = 0$, the streamlines for both ML-ARL and MaxEnt-ARL converge to $(0, 0)$. For an alternate formulation, where the discriminator is of the form $D = z^2 + b_2$, we found convergent behavior for both ML-ARL and MaxEnt-ARL.

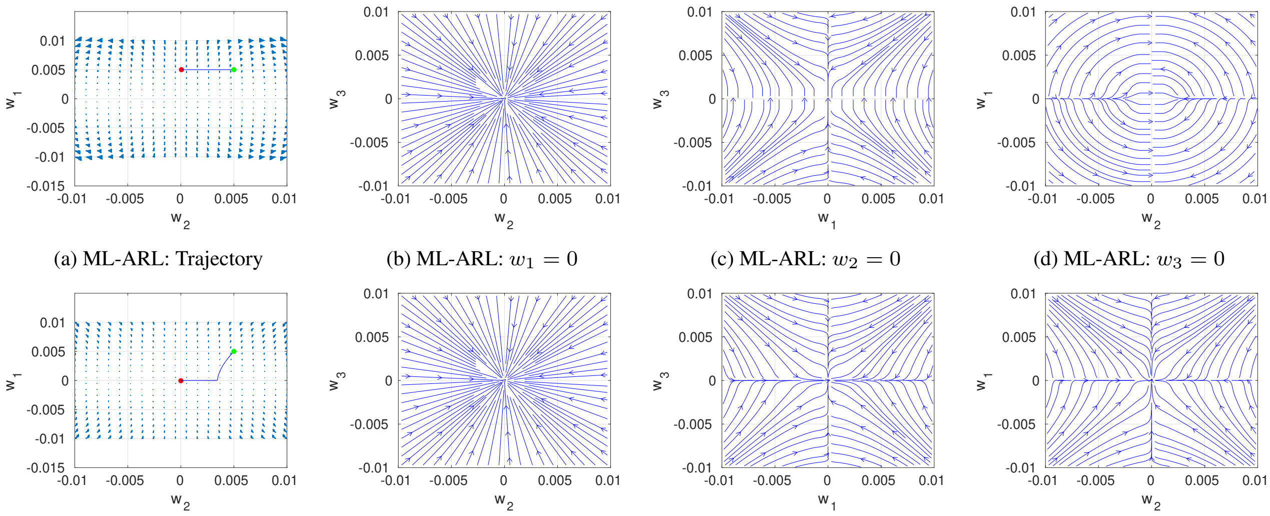
5.2. Mixture of Gaussians

In this experiment we seek to visualize and compare the representation learned by MaxEnt-ARL and ML-ARL. We consider a mixture of 4 Gaussians with means μ at $((1, 1), (2, 1.5), (1.5, 2.5), (2.5, 3))$ and variance $\sigma = 0.3$ in each case. Our model is a neural network with 2 hidden layer with 2 neuron in each layer. Each data sample has two attributes, color and shape. We setup the ARL problem with shape as the target attribute and color as the sensitive attribute. The encoder is a neural network with one hidden layer, mapping the 2-D shape into another 2-D embedding, and both the predictor and discriminator are logistic regression classifiers. The trade-off parameter is set to $\alpha = 0.1$ and the parameters are learned using the Adam optimizer with learning rate of 10^{-4}. After learning the embedding function, we freeze its parameters and learn a logistic classifier as the adversary. The test accuracy of the adversary is 63% for MaxEnt-ARL and 70% for ML-ARL. Therefore, by optimizing the entropy instead of the likelihood MaxEnt-ARL is able to leak less information about the sensitive label compared to ML-ARL. Figure 4 shows the data and the learned embeddings.

5.3. Fair Classification

We consider the setting of fair classification on two datasets from the UCI ML-repository [2], (a) The German credit dataset with 20 attributes for 1000 instances with target label being classifying bank account holders with good or bad credit and gender being the sensitive attribute, (b) The Adult income dataset has 45,222 instances with 14 attributes. The target is a binary label of annual income more or less than $\$50,000$, while gender is the sensitive attribute. For both ML-ARL and MaxEnt-ARL, the encoder is a NN with one hidden layer, discriminator is a NN with 2 hidden layers, and target predictor is linear logistic regression. Following ML-ARL [18] we choose 64 units in each hidden layer. We compare both ARL formulations with state-of-the-art baselines LFR (Learning Fair Representations [19]), VAE (Variational Auto-encoder [11]) and VFAE (Variational Fair Auto-encoder [14]). For MaxEnt-ARL, after learning the embedding, we again learn an adversary to extract the sensitive attribute.

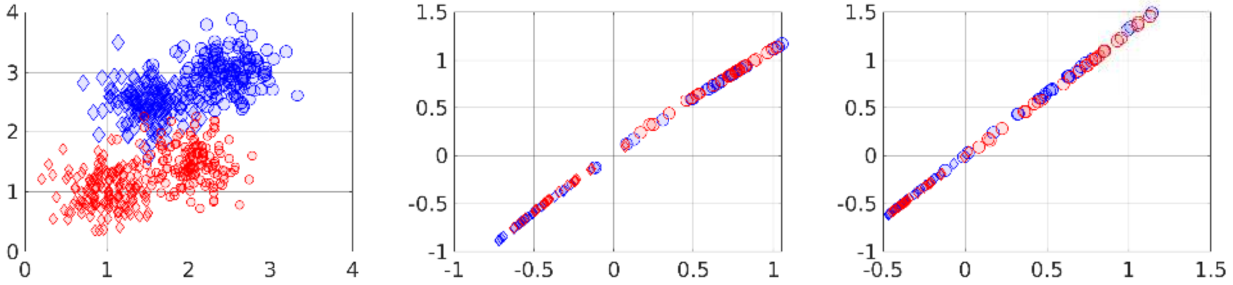
Figure 5 show the results for the German and Adult datasets, for both the target and sensitive attributes. For Ger-



(a) ML-ARL: Trajectory (b) ML-ARL: $w_1 = 0$ (c) ML-ARL: $w_2 = 0$ (d) ML-ARL: $w_3 = 0$

(e) MaxEnt-ARL: Trajectory (f) MaxEnt-ARL: $w_1 = 0$ (g) MaxEnt-ARL: $w_2 = 0$ (h) MaxEnt-ARL: $w_3 = 0$

Figure 3: Streamline plots for linear three-player game near stationary point (0,0,0). Trajectories start at the green point and converge to the red point by following the vector field. (a) and (e) shows the top-view of the 3-D trajectories. When $w_1 = 0$ the trajectories suggest that both ML-ARL and MaxEnt-ARL converge to the local optima, ($w_1 = w_2 = w_3 = 0$). When $w_2 = 0$, the MaxEnt-ARL trajectories converge to the local optima. The ML-ARL trajectories converge to the optima only when they start far away from 0 along w_3. The trajectories starting closer to $w_3 = 0$, however, do not converge to $w_1 = 0$. When $w_3 = 0$, the game reduces to a two-player adversarial game (akin to a GAN[7]), where ML-ARL shows non-convergent cyclic behavior while MaxEnt-ARL converges.



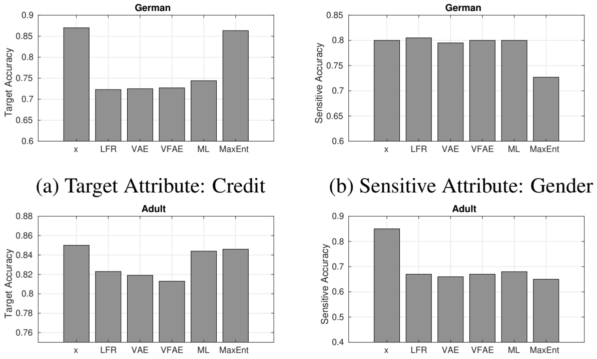
(a) Input (b) ML-ARL (c) MaxEnt-ARL

Figure 4: Samples from four Gaussians with target (shape) and sensitive attributes (color). (a) input space, (b) learned embedding z for ML-ARL, (c) learned embedding for MaxEnt-ARL. We can now notice that ML-ARL has some isolated samples with different colors (sensitive label), while MaxEnt-ARL results in slightly better mixing of the colors.

man data, MaxEnt-ARL's prediction accuracy is 86.33% which is close to that of the original data (87%). Other models such as, LFR, VAE, VFAE and ML-ARL have target accuracies of 72.3%, 72.5%, 72.7% and 74.4% respectively. On the other hand, for the sensitive attribute, the MaxEnt-ARL adversary's accuracy is 72.7%. Other models reveal much more information with adversary accuracies of 80%, 80.5%, 79.5%, 79.7% and 80.2% for the original data, LFR, VAE, VFAE and ML-ARL, respectively. For the adult income dataset, the target accuracy for original data, ML-ARL and MaxEnt-ARL is 85%, 84.4% and 84.6%, respectively, while the adversary's performance on the sensitive attribute is 67.7% and 65.5% for ML-ARL and MaxEnt-ARL, respectively.

5.4. Illumination Invariant Face Classification

We consider the task of face classification under different illumination conditions. We used the Extended Yale B dataset [6] comprising of face images of 38 people under different lighting conditions (directions of the light source) : upper right, lower right, lower left, upper left, or the front. Our target task is to identify one of the 38 people in the



(a) Target Attribute: Credit

(b) Sensitive Attribute: Gender

(c) Target Attribute: Income

(d) Sensitive Attribute: Gender

Figure 5: Representation Learning for Fair Classification

Table 1: Illumination Invariant Face Classification (%)

Method	s (lighting)	t (identity)
LR	96	78
NN + MMD [13]	-	82
VFAE [14]	57	85
ML-ARL [18]	57	**89**
Maxent-ARL	**40**	**89**

dataset with the direction of the light source being the sensitive attribute. We follow the experimental setup of Xie et al. [18] and Louizos et al. [14] using the same train/test split strategy and no validation set. $38 \times 5 = 190$ samples are used for training and the rest of the 1,096 data samples are used for testing. Following the model setup in [18], the encoder is a one layer neural network, target predictor is a linear layer and the discriminator has two hidden layers where each hidden layer consists of 100 units. The parameters are trained using Adam [10] with a learning rate of 10^{-4} and weight decay of 5×10^{-2}.

We report baseline [13, 14, 18] results for this experiment in Table 1 and compare with the proposed MaxEnt-ARL framework. Louizos et al. [14] regularize their neural networks via Maximum Mean Discrepancy to remove lighting conditions from data whereas Xie et al. [18] use the ML-ARL framework. The MaxEnt-ARL achieves an accuracy of 89% for identity classification (same as ML-ARL) while outperforming MMD (82%) and VFAE (85%). In terms of protecting sensitive attribute i..e, illumination direction, adversary's classification accuracy reduces from 57% for ML-ARL to 40.2% for MaxEnt-ARL. It is clear from the table that, MaxEnt-ARL is able to remove more information from the image compared to the baselines.

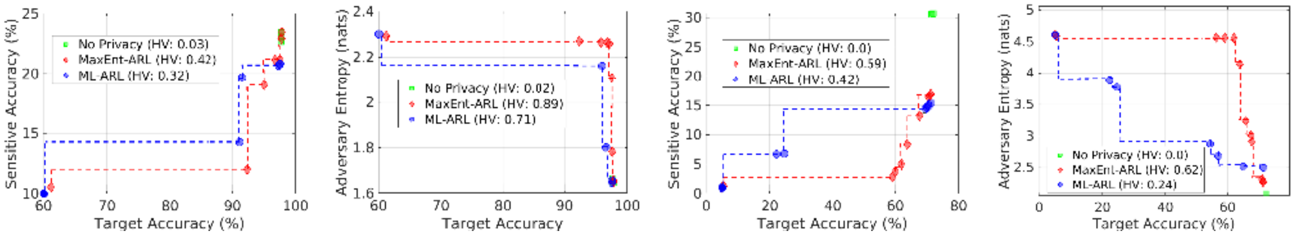
5.5. CIFAR-10

We create a new binary target classification problem on the CIFAR-10 dataset[12]. The CIFAR-10 dataset consists of 10 basic classes, namely, ('airplane', 'automobile', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', 'truck'). We divide the classes into two groups: living and non-living objects. We expect the living objects to have visually discriminative properties like smooth shapes compared to regular geometric shapes of non-living objects. The target task is binary classification of an image into these two supersets with the underlying class label being the sensitive attribute. For example, the task of classifying an object as living ('dog' or 'cat') or non-living ('ship' or 'truck') should not reveal any information about its underlying identity ('dog', 'cat', 'truck' or 'ship'). But as we will see, this is a challenging problem and the image representation might not be able to prevent leakage of the sensitive label.

Implementation Details: We adopt the ResNet-18 [8] architecture as the encoder, and the discriminator and adversary are 2-layered neural networks with 256 and 64 neurons, respectively. The encoder and the target predictor are trained using SGD with momentum of 0.9, learning rate of 10^{-3} and weight-decay of 10^{-3} for the prediction task. Both the discriminator and the adversary, however, are trained using Adam with a learning rate of 10^{-4} and weight-decay of 10^{-3} for 300 epochs.

Experimental Results: We evaluate performance of the predictor and adversary as we vary the trade-off parameter α. We first note that, ideally, the desired predictor accuracy is 100%, adversary accuracy is 10% (random chance for 10 classes) and adversary entropy is 2.3 nats (uniform distribution for 10 classes). Figure 6 (a)-(b) shows the trade-off achieved between predictor and adversary along with the corresponding normalized hyper-volume (HV). For the predictor and adversary accuracy, the HV corresponds to area above the trade-off curve, while for the predictor accuracy and adversary entropy the HV is the area under the curve.

We obtain these results by repeating all the experiments five times and retaining the non-dominated solutions i.e., a solution that is no worse than any other solution in both the objectives. From these results, we observe that without privacy considerations, the representation achieves the best target accuracy but also leaks significant information. In contrast adversarial learning of the representation achieves a better trade-off between utility and information leakage. Among ARL approaches, we observe that MaxEnt-ARL is able to obtain a better trade-off compared to ML-ARL. Furthermore, among all possible solutions, MaxEnt-ARL achieves the solution closest to the ideal desired point.



(a) (b) (c) (d)

Figure 6: **Adversary Representation Learning on CIFAR Datasets:** Trade-off fronts for two different ARL approaches, ML-ARL and MaxEnt-ARL, in comparison to standard no privacy representation learning. Plots (a)-(b) and (c)-(d) correspond to CIFAR-10 and CIFAR-100 experiments, respectively. In (a) and (c) the ideal desired solution is the bottom right corner, while in (b) and (d) it is the top right corner. HV in the legend corresponds to normalized hyper-volume. Exact numerical values are available in the supplementary material.

Table 2: Main classes and Superclasses in CIFAR-100

Superclass	Main Class
aquatic mammals	beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food containers	bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household furniture	bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

5.6. CIFAR-100

We formulate a new privacy problem on the CIFAR-100 dataset. The dataset consists of 100 classes and are grouped into 20 superclasses (Table 2). Each image has a “fine” (the class to which it belongs) and a “coarse” (the superclass to which it belongs) label. We treat the “coarse” (superclass) and “fine” (class) labels as the target and sensitive attribute, respectively. So the encoder is tasked to learn features of the super-classes while not revealing the information of the underlying classes. We adopt ResNet-18 as the encoder while the predictor, discriminator and adversary are all 2-layered fully connected networks. The adversarial game is trained for 150 epochs, followed by training the adversary for 100 epochs while the parameters of the encoder are frozen.

Just as in the case of CIFAR-10, we report the trade-off achieved between predictor and adversary along with the corresponding normalized hyper-volume (HV) in Fig. 6 (c)-(d). Here we note that, ideally, we desire predictor accuracy of 100%, adversary accuracy of 1% (random chance for 100 classes) and adversary entropy of $\ln 100 = 4.61$ nats (uniform distribution for 100 classes). We make the following observations from the results. Firstly, the performance of the different approaches suggest that this task is significantly harder than the CIFAR-10 task, with much lower achievable target accuracy and much higher adversary accuracy. Secondly, representation learning without privacy considerations leaks significant amount of information. Thirdly, MaxEnt-ARL is able to significantly outperform ML-ARL on this task, achieving trade-off solutions that are far better, both in terms of adversary accuracy and entropy of adversary.

6. Conclusion

This paper introduced a new formulation of *Adversarial Representation Learning* called *Maximum Entropy Adversarial Representation Learning* (MaxEnt-ARL) for mitigating information leakage from learned representations under an adversarial setting. In this model, the encoder is optimized to maximize the entropy of the adversary’s distribution of a sensitive attribute as opposed to minimizing the likelihood (ML-ARL) of the true sensitive label. We analyzed the equilibrium and convergence properties of the ML-ARL and MaxEnt-ARL. Numerical experiments on multiple datasets suggests that MaxEnt-ARL is a promising framework for preventing information leakage from image representations, outperforming the baseline minimum likelihood objective.

References

[1] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017. 2

[2] D. Dua and C. Graff. UCI machine learning repository, 2017. 5

[3] H. Edwards and A. J. Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*, 2016. 1, 2

[4] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015. 2

[5] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2

[6] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):643–660, 2001. 6

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 6

[8] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016. 7

[9] H. K. Khalil. Nonlinear systems. *Printice-Hall Inc*, 1996. 2, 4, 5

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[12] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 7

[13] Y. Li, K. Swersky, and R. Zemel. Learning unbiased features. *arXiv preprint arXiv:1412.5244*, 2014. 7

[14] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR)*, 2016. 1, 2, 5, 7

[15] L. Mescheder, S. Nowozin, and A. Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[16] V. Nagarajan and J. Z. Kolter. Gradient descent gan optimization is locally stable. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5585–5595, 2017. 2

[17] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[18] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 3, 4, 5, 7

[19] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, 2013. 2, 5

[20] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018. 2