Log-linear Guardedness and its Implications

Shauli Ravfogel[1,2] **Yoav Goldberg**[1,2] **Ryan Cotterell**[3]
[1]Bar-Ilan University [2]Allen Institute for Artificial Intelligence [3]ETH Zürich
{shauli.ravfogel, yoav.goldberg}@gmail.com ryan.cotterell@inf.ethz.ch

Abstract

Methods for erasing human-interpretable concepts from neural representations that assume linearity have been found to be tractable and useful. However, the impact of this removal on the behavior of downstream classifiers trained on the modified representations is not fully understood. In this work, we formally define the notion of log-linear guardedness as the inability of an adversary to predict the concept directly from the representation, and study its implications. We show that, in the binary case, under certain assumptions, a downstream log-linear model cannot recover the erased concept. However, we demonstrate that a multiclass log-linear model *can* be constructed that indirectly recovers the concept in some cases, pointing to the inherent limitations of log-linear guardedness as a downstream bias mitigation technique. These findings shed light on the theoretical limitations of linear erasure methods and highlight the need for further research on the connections between intrinsic and extrinsic bias in neural models.

https://github.com/rycolab/guardedness

1 Introduction

Neural models of text have been shown to represent human-interpretable concepts, e.g., those related to the linguistic notion of morphology (Vylomova et al., 2017), syntax (Linzen et al., 2016), semantics (Belinkov et al., 2017), as well as extra-linguistic notions, e.g., gender distinctions (Caliskan et al., 2017). Identifying and erasing such concepts from neural representations is known as **concept erasure**. Linear concept erasure in particular has gained popularity due to its potential for obtaining formal guarantees and its empirical effectiveness (Bolukbasi et al., 2016; Dev and Phillips, 2019; Ravfogel et al., 2020; Dev et al., 2021; Kaneko and Bollegala, 2021; Shao et al., 2023b,a; Kleindessner et al., 2023; Belrose et al., 2023).

A common instantiation of concept erasure is removing a concept (e.g., gender) from a representation (e.g., the last hidden representation of a transformer-based language model) such that it cannot be predicted by a log-linear model. Then, one fits a *secondary* log-linear model for a downstream task over the erased representations. For example, one may fit a log-linear sentiment analyzer to predict sentiment from gender-erased representations. The hope behind such a pipeline is that, because the concept of gender was erased from the representations, the predictions made by the log-linear sentiment analyzer are oblivious to gender. Previous work (Ravfogel et al., 2020; Elazar et al., 2021; Jacovi et al., 2021; Ravfogel et al., 2022a) has implicitly or explicitly relied on this assumption that erasing concepts from representations would also result in a downstream classifier that was oblivious to the target concept.

In this paper, we formally analyze the effect concept erasure has on a downstream classifier. We start by formalizing concept erasure using Xu et al.'s (2020) $\mathcal{V}$-information.[1] We then spell out the related notion of **guardedness** as the inability to predict a given concept from concept-erased representations using a specific family of classifiers. Formally, if $\mathcal{V}$ is the family of distributions realizable by a log-linear model, then we say that the representations are guarded against gender with respect to $\mathcal{V}$. The theoretical treatment in our paper specifically focuses on **log-linear guardedness**, which we take to mean the inability of a *log-linear* model to recover the erased concept from the representations. We are able to prove that when the downstream classifier is binary valued, such as a binary sentiment classifier, its prediction indeed cannot leak information about the erased concept (§ 3.2) under certain assumptions. On the contrary, in the case of multiclass classification with a log-linear model, we show that predictions *can* potentially leak a substantial amount of information about the erased concept, thereby recovering the guarded information completely.

The theoretical analysis is supported by experiments on commonly used linear erasure techniques

[1] We also consider a definition based on accuracy.

(§ 5). While previous authors (Goldfarb-Tarrant et al. 2021, Orgad et al. 2022, *inter alia*) have empirically studied concept erasure's effect on downstream classifiers, to the best of our knowledge, we are the first to study it theoretically. Taken together, these findings suggest that log-linear guardedness may have limitations when it comes to preventing information leakage about concepts and should be assessed with extreme care, even when the downstream classifier is merely a log-linear model.

2 Information-Theoretic Guardedness

In this section, we present an information-theoretic approach to guardedness, which we couch in terms of $\mathcal{V}$-information (Xu et al., 2020).

2.1 Preliminaries

We first explain the concept erasure paradigm (Ravfogel et al., 2022a), upon which our work is based. Let $\mathbf{X}$ be a representation-valued random variable. In our setup, we assume representations are real-valued, i.e., they live in $\mathbb{R}^D$. Next, let Z be a binary-valued random variable that denotes a protected attribute, e.g., binary gender.[2] We denote the two binary values of Z by $\mathcal{Z} \overset{\text{def}}{=} \{\bot, \top\}$. We assume the existence of a **guarding function** $h : \mathbb{R}^D \to \mathbb{R}^D$ that, when applied to the representations, removes the ability to predict a concept Z given concept by a specific family of models. Furthermore, we define the random variable $\widehat{\mathrm{Y}} = t(h(\mathbf{X}))$ where $t : \mathbb{R}^D \to \mathcal{Y} \overset{\text{def}}{=} \{0, \ldots, |\mathcal{Y}|\}$ is a function[3] that corresponds to a linear classifier for a downstream task. For instance, t may correspond to a linear classifier that predicts the sentiment of a representation.

Our discussion in this paper focuses on the case when the function t is derived from the argmax of a log-linear model, i.e., in the binary case we define $\widehat{\mathrm{Y}}$'s conditional distribution given $h(\mathbf{X})$ as

$$p(\widehat{\mathrm{Y}} = y \mid h(\mathbf{X}) = h(\boldsymbol{x})) = \begin{cases} 1, & \textbf{if } y = y^* \\ 0, & \textbf{else} \end{cases} \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter column vector, $\phi \in \mathbb{R}$ is a scalar bias term, and

$$y^* = \begin{cases} 1, & \textbf{if} \quad \boldsymbol{\theta}^\top h(\boldsymbol{x}) + \phi > 0 \\ 0, & \textbf{else} \end{cases} \tag{2}$$

And, in the multivariate case we define $\widehat{\mathrm{Y}}$'s conditional distribution given $h(\mathbf{X})$ as

$$p(\widehat{\mathrm{Y}} = y \mid h(\mathbf{X}) = h(\boldsymbol{x})) = \begin{cases} 1, & \textbf{if } y = y^* \\ 0, & \textbf{else} \end{cases} \tag{3}$$

where $y^* = \mathrm{argmax}_{y' \in \mathcal{Y}} \, (\boldsymbol{\Theta}^\top h(\boldsymbol{x}) + \boldsymbol{\phi})_{y'}$ and $\boldsymbol{\Theta}_y \in \mathbb{R}^D$ denotes the y^{th} column of $\boldsymbol{\Theta} \in \mathbb{R}^{D \times K}$, a parameter matrix, and $\boldsymbol{\phi} \in \mathbb{R}^K$ is the bias term. Note K is the number of classes.

2.2 $\mathcal{V}$-Information

Intuitively, a set of representations is guarded if it is not possible to predict a protected attribute $z \in \mathcal{Z}$ from a representation $\boldsymbol{x} \in \mathbb{R}^D$ using a specific predictive family. As a first attempt, we naturally formalize predictability in terms of mutual information. In this case, we say that Z is not predictable from $\mathbf{X}$ if and only if $\mathrm{I}(\mathbf{X}; \mathrm{Z}) = 0$. However, the focus of this paper is on *linear* guardedness, and, thus, we need a weaker condition than simply having the mutual information $\mathrm{I}(\mathbf{X}; \mathrm{Z}) = 0$. We fall back on Xu et al.'s (2020) framework of $\mathcal{V}$-information, which introduces a generalized version of mutual information. In their framework, they restrict the predictor to a family of functions $\mathcal{V}$, e.g., the set of all log-linear models.

We now develop the information-theoretic background to discuss $\mathcal{V}$-information. The **entropy** of a random variable is defined as

$$\mathrm{H}(\mathrm{Z}) \overset{\text{def}}{=} - \mathop{\mathbb{E}}_{z \sim p(\mathrm{Z})} \log p(z) \tag{4}$$

Xu et al. (2020) analogously define the **conditional $\mathcal{V}$-entropy** as follows

$$\mathrm{H}_{\mathcal{V}}(\mathrm{Z} \mid \mathbf{X}) \overset{\text{def}}{=} - \sup_{q \in \mathcal{V}} \mathop{\mathbb{E}}_{(\boldsymbol{x}, z) \sim p(\mathbf{X}, \mathrm{Z})} \log q(z \mid \boldsymbol{x}) \tag{5}$$

The **$\mathcal{V}$-entropy** is a special case of Eq. (5) without conditioning on another random variable, i.e.,

$$\mathrm{H}_{\mathcal{V}}(\mathrm{Z}) \overset{\text{def}}{=} - \sup_{q \in \mathcal{V}} \mathop{\mathbb{E}}_{z \sim p(\mathrm{Z})} \log q(z) \tag{6}$$

Xu et al. (2020) further define the **$\mathcal{V}$-information**, a generalization of mutual information, as follows

$$\mathrm{I}_{\mathcal{V}}(\mathbf{X} \to \mathrm{Z}) \overset{\text{def}}{=} \mathrm{H}_{\mathcal{V}}(\mathrm{Z}) - \mathrm{H}_{\mathcal{V}}(\mathrm{Z} \mid \mathbf{X}) \tag{7}$$

In words, Eq. (7) is the *best* approximation of the mutual information realizable by a classifier belonging to the predictive family $\mathcal{V}$. Furthermore,

[2] Not all concepts are binary, but our analysis in § 2 makes use of this simplifying assumption.

[3] The elements of $\mathcal{Y}$ are denoted y.

in the case of log-linear models, Eq. (7) can be approximated empirically by calculating the negative log-likelihood loss of the classifier on a given set of examples, as $\mathrm{H}_{\mathcal{V}}(\mathrm{Z})$ is the entropy of the label distribution, and $\mathrm{H}_{\mathcal{V}}(\mathrm{Z} \mid \mathbf{X})$ is the minimum achievable value of the cross-entropy loss.

2.3 Guardedness

Having defined $\mathcal{V}$-information, we can now formally define guardedness as the condition where the $\mathcal{V}$-information is small.

Definition 2.1 ($\mathcal{V}$-Guardedness)**.** *Let $\mathbf{X}$ be a representation-valued random variable and let Z be an attribute-valued random variable. Moreover, let $\mathcal{V}$ be a predictive family. A guarding function h ε-**guards** $\mathbf{X}$ with respect to Z over $\mathcal{V}$ if $\mathrm{I}_{\mathcal{V}}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$.*

Definition 2.2 (Empirical $\mathcal{V}$-Guardedness)**.** *Let $\mathcal{D} = \{(\boldsymbol{x}_n, z_n)\}_{n=1}^{N}$ where $(\boldsymbol{x}_n, z_n) \sim p(\mathbf{X}, \mathrm{Z})$. Let $\widetilde{\mathbf{X}}$ and $\widetilde{\mathrm{Z}}$ be random variables over $\mathbb{R}^D$ and $\mathcal{Z}$, respectively, whose distribution corresponds to the marginals of the empirical distribution over $\mathcal{D}$. We say that a function $h(\cdot)$ **empirically** ε-**guards** $\mathcal{D}$ with respect to the family $\mathcal{V}$ if $\mathrm{I}_{\mathcal{V}}(h(\widetilde{\mathbf{X}}) \to \widetilde{\mathrm{Z}}) < \varepsilon$.*

In words, according to Definition 2.2, a dataset is log-linearly guarded if no linear classifier can perform better than the trivial classifier that completely ignores X and always predicts Z according to the proportions of each label. The commonly used algorithms that have been proposed for linear subspace erasure can be seen as approximating the condition we call log-linear guardedness (Ravfogel et al., 2020, 2022a,b). Our experimental results focus on empirical guardedness, which pertains to practically measuring guardedness on a finite dataset. However, determining the precise bounds and guarantees of empirical guardedness is left as an avenue for future research.

3 Theoretical Analysis

In the following sections, we study the implications of guardedness on *subsequent* linear classifiers. Specifically, if we construct a third random variable $\widehat{\mathrm{Y}} = t(h(\mathrm{X}))$ where $t : \mathbb{R}^D \to \mathcal{Y}$ is a function, what is the degree to which $\widehat{\mathrm{Y}}$ can reveal information about Z? As a practical instance of this problem, suppose we impose ε-guardedness on the last hidden representations of a transformer model, i.e., $\mathbf{X}$ in our formulation, and then fit a linear classifier t over the guarded representations $h(\mathbf{X})$ to predict sentiment. Can the predictions of the sentiment classifier indirectly leak information on gender? For expressive $\mathcal{V}$, the data-processing inequality (Cover and Thomas, 2006, §2.8) applied to the Markov chain $\mathbf{X} \to \widehat{\mathrm{Y}} \to \mathrm{Z}$ tells us the answer is no. The reason is that, in this case, $\mathcal{V}$-information is equivalent to mutual information and the data processing inequality tells us such leakage is *not* possible. However, the data processing inequality does not generally apply to $\mathcal{V}$-information (Xu et al., 2020). Thus, it *is* possible to find such a predictor t for less expressive $\mathcal{V}$. Surprisingly, when $|\mathcal{Y}| = 2$, we are able to prove that constructing such a t that leaks information is impossible under a certain restriction on the family of log-linear models.

3.1 Problem Formulation

We first consider the case where $|\mathcal{Y}| = 2$.

3.2 A Binary Downstream Classifier

We begin by asking whether the predictions of a binary log-linear model trained over the guarded set of representations can leak information on the protected attribute. Our analysis relies on the following simplified family of log-linear models.

Definition 3.1 (Discretized Log-Linear Models)**.** *The family of **discretized binary log-linear models** with parameter $\delta \in (0,1)$ is defined as*

$$\mathcal{V}^{\delta} \overset{\text{def}}{=} \left\{ f \;\middle|\; \begin{array}{l} f(0) = \rho_{\delta}(\sigma(\boldsymbol{\alpha}^{\top}\boldsymbol{x} + \gamma)) \\ f(1) = \rho_{\delta}(1 - \sigma(\boldsymbol{\alpha}^{\top}\boldsymbol{x} + \gamma)) \end{array} \right\} \quad (8)$$

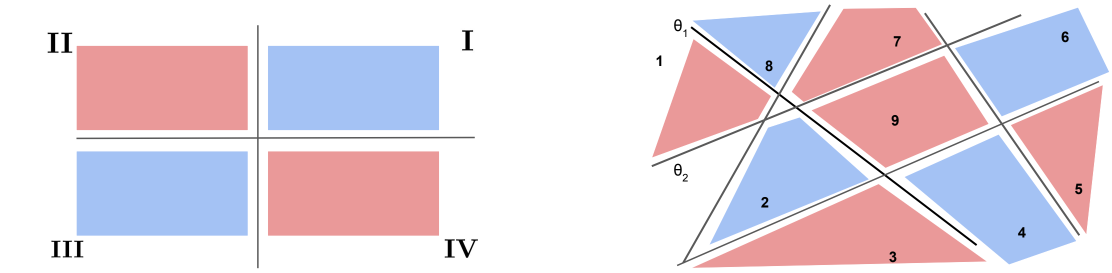
with $\boldsymbol{\alpha} \in \mathbb{R}^D$, $\gamma \in \mathbb{R}$, σ being the logistic function, and where we define the δ-discretization function as

$$\rho_{\delta}(p) \overset{\text{def}}{=} \begin{cases} \delta, & \textbf{\textit{if}}\ p \geq \frac{1}{2} \\ 1 - \delta, & \textbf{\textit{else}} \end{cases} \quad (9)$$

In words, ρ_{δ} is a function that maps the probability value to one of two possible values. Note that ties must be broken arbitrarily in the case that $p = \frac{1}{2}$ to ensure a valid probability distribution.

Our analysis is based on the following simple observation (see Lemma A.1 in the Appendix) that the composition of two δ-discretized log-linear models is itself a δ-discretized log-linear model. Using this fact, we show that when $|\mathcal{Y}| = |\mathcal{Z}| = 2$, and the predictive family is the set of δ-discretized binary log-linear models, ε-guarded representations $h(\mathbf{X})$ cannot leak information through a downstream classifier.

Theorem 3.2. *Let $\mathcal{V}^{\delta}$ be the family of δ-discretized log-linear models, and let $\mathbf{X}$ be a*



(a) Log-linearly guarded data in $\mathbb{R}^2$ with axis-aligned clusters.

(b) Log-inearly guarded data in $\mathbb{R}^2$ with clusters that are not axis-aligned.

Figure 1: Construction of a log-linear model that breaks log-linear guardedness.

representation-valued random variable. Define $\widehat{\mathrm{Y}}$ *as in Eq.* (1)*, then* $\mathrm{I}_{\mathcal{V}^\delta}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$ *implies* $\mathrm{I}_{\mathcal{V}^\delta}(\widehat{\mathrm{Y}} \to \mathrm{Z}) < \varepsilon$.

Proof. Define the hard thresholding function

$$\tau(x) \overset{\text{def}}{=} \begin{cases} 1, & \textbf{if} \quad x > 0 \\ 0, & \textbf{else} \end{cases} \tag{10}$$

We assume, by way of contradiction, that $\mathrm{I}_{\mathcal{V}^\delta}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$, but $\mathrm{I}_{\mathcal{V}^\delta}(\widehat{\mathrm{Y}} \to \mathrm{Z}) > \varepsilon$. We start by algebraically manipulating $\mathrm{I}_{\mathcal{V}^\delta}(\widehat{\mathrm{Y}} \to \mathrm{Z})$:

$$\begin{aligned} &\mathrm{I}_{\mathcal{V}^\delta}(\widehat{\mathrm{Y}} \to \mathrm{Z}) = \mathrm{H}_{\mathcal{V}^\delta}(\mathrm{Z}) - \mathrm{H}_{\mathcal{V}^\delta}(\mathrm{Z} \mid \widehat{\mathrm{Y}}) \\ &= \mathrm{H}_{\mathcal{V}^\delta}(\mathrm{Z}) + \sup_{q \in \mathcal{V}^\delta} \underset{(z,y)\sim p}{\mathbb{E}} \log q(z \mid y) \qquad (11) \\ &= \mathrm{H}_{\mathcal{V}^\delta}(\mathrm{Z}) + \sup_{q \in \mathcal{V}^\delta} \underset{(z,\boldsymbol{x})\sim p}{\mathbb{E}} \log q(z \mid \tau(\boldsymbol{\theta}^\top h(\boldsymbol{x}) + \phi)) \end{aligned}$$

for some $\boldsymbol{\theta}$ and ϕ as in the definition of t in Eq. (1). Now, by Lemma A.1, we note that, for all $q \in \mathcal{V}^\delta$, there exists a classifier $r \in \mathcal{V}^\delta$ such that $r(z \mid h(\boldsymbol{x})) = q(z \mid \tau(\boldsymbol{\theta}^\top h(\boldsymbol{x}) + \phi))$. This implies that $\mathrm{I}_{\mathcal{V}^\delta}(h(\mathbf{X}) \to \mathrm{Z}) \geq \mathrm{I}_{\mathcal{V}^\delta}(\widehat{\mathrm{Y}} \to \mathrm{Z}) > \varepsilon$,[4] contradicting the assumption that $\mathrm{I}_{\mathcal{V}^\delta}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$. Thus, $\mathrm{I}_{\mathcal{V}^\delta}(\widehat{\mathrm{Y}} \to \mathrm{Z}) < \varepsilon$, as desired. ■

3.3 A Multiclass Downstream Classifier

The above discussion shows that when both Z and Y are binary, ε-log-linear guardedness with respect to the family of discretized log-linear models (Definition 3.1) implies limited leakage of information about Z from $\widehat{\mathrm{Y}}$. It was previously implied (Ravfogel et al., 2020; Elazar et al., 2021) that linear concept erasure prevents information leakage about Z through the labeling of a log-linear classifier $\widehat{\mathrm{Y}}$, i.e., it was assumed that Theorem 3.2 in § 3.2 can be generalized to the multiclass case. Specifically, it was argued that a subsequent linear layer, such as the linear language-modeling head, would not be able to recover the information because it is linear. In this paper, however, we note a key flaw in this argument. If the data is log-linearly guarded, then it is easy to see that the *logits*, which are a linear transformation of the guarded representation, cannot encode the information. However, multiclass classification is usually performed by a softmax classifier, which adds a non-linearity. Note that the decision boundary of the softmax classifier for every pair of labels is linear since class i will have higher softmax probability than class j if, and only if, $(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^\top \boldsymbol{x} > 0$.

Next, we demonstrate that this is enough to break guardedness. We start with an example. Consider the data in $\mathbb{R}^2$ presented in Fig. 1(a), where the distribution $p(\mathbf{X}, \mathrm{Z})$ has 4 distinct clusters, each with a different label from $\mathcal{Z}$, corresponding to Voronoi regions (Voronoi, 1908) formed by the intersection of the axes. The red clusters correspond to $\mathrm{Z} = \top$ and the blue clusters correspond to $\mathrm{Z} = \bot$. The data is taken to be log-linearly guarded with respect to Z.[5] Importantly, we note that knowledge of the quadrant (i.e., the value of Y), renders Z recoverable by a 4-class log-linear model.

[4] Lemma A.1 only guarantees that a classifier of the form $q(z \mid \tau(\boldsymbol{\theta}^\top h(\boldsymbol{x}) + \phi)$, where $q \in \mathcal{V}^\delta$, can be converted into a classifier $r(z \mid h(\boldsymbol{x})) \in \mathcal{V}^\delta$. However, we have no proof of the opposite implication. Hence, we have only shown $\mathrm{I}_{\mathcal{V}^\delta}(h(\mathbf{X}) \to \mathrm{Z}) \geq \mathrm{I}_{\mathcal{V}^\delta}(\widehat{\mathrm{Y}} \to \mathrm{Z})$.

[5] Information-theoretic guardedness depends on the density over $p(\mathbf{X})$, which is *not* depicted in the illustrations in Fig. 1(a).

Assume the parameter matrix $\boldsymbol{\Theta} \in \mathbb{R}^{4\times 2}$ of this classifier is composed of four columns $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4$ such that $\boldsymbol{\theta}_1 = \alpha \cdot [1, 1]^\top, \boldsymbol{\theta}_2 = \alpha \cdot [-1, 1]^\top, \boldsymbol{\theta}_3 = \alpha \cdot [-1, -1]^\top, \boldsymbol{\theta}_4 = \alpha \cdot [1, -1]^\top$, for some $\alpha > 0$. These directions encode the quadrant of a point: When the norm of the parameter vectors is large enough, i.e., for a large enough α, the probability of class i under a log-linear model will be arbitrarily close to 1 if, and only if, the input is in the i^{th} quadrant and arbitrarily close to 0 otherwise. Given the information about the quadrant, the data is rendered *perfectly linearly separable*. Thus, the labels $\widehat{\mathrm{Y}}$ predicted by a multiclass softmax classifier can recover the linear separation according to Z.

This argument can be generalized to a separation that is not axis-aligned (Fig. 1(b)).

Definition 3.3. *Let $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ be column vectors orthogonal to corresponding linear subspaces, and let $R_1, \dots, R_M$ be the Voronoi regions formed by their intersection (Fig. 1(b)). Let $p(\mathbf{X}, \mathrm{Z})$ be any data distribution such that any two points in the same region have the same value of Z:*

$$\boldsymbol{x}_i \in R_k \wedge \boldsymbol{x}_j \in R_k \implies z_i = z_j \tag{12}$$

*for all $(\boldsymbol{x}_i, z_i), (\boldsymbol{x}_j, z_j) \sim p(\mathbf{X}, \mathrm{Z})$ and for all Voronoi regions R_k. We call such distribution a K-**Voronoi** distribution.*

Theorem 3.4. *Fix $\varepsilon > 0$. Let $p(\mathbf{X}, \mathrm{Z})$ be a K-Voronoi distribution, and let h linearly ε-guard $\mathbf{X}$ against Z with respect to the family $\mathcal{V}$ of log-linear models. Then, for every $\eta > 0$, there exists a K-class log-linear model such that $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}} \to \mathrm{Z}) > 1 - \eta$.*[6]

Proof. By assumption, the support of $p(\mathbf{X})$ is divided up into K Voronoi regions, each with a label from $\mathcal{Z}$. See Fig. 1 for an illustrative example.

Define the region identifier $\iota_k(i)$ for each region k as follows

$$\iota_k(i) \overset{\text{def}}{=} \begin{cases} 1, & \textbf{if } \boldsymbol{\theta}_i^\top \boldsymbol{x} > 0 \textbf{ for } \boldsymbol{x} \in R_k \\ -1, & \textbf{if } \boldsymbol{\theta}_i^\top \boldsymbol{x} < 0 \textbf{ for } \boldsymbol{x} \in R_k \end{cases} \tag{13}$$

We make the simplifying assumption that points $\boldsymbol{x}$ that lie on line $\boldsymbol{\theta}_i^\top \boldsymbol{x}$ for any i occur with probability zero. Consider a K-class log-linear model with a parameter matrix $\boldsymbol{\Theta}^\star \in \mathbb{R}^{D\times K}$ that contains, in its j^{th} column, the vector $\boldsymbol{\theta}_j^\star \overset{\text{def}}{=} \alpha \sum_{k=0}^{K} \iota_j(k) \boldsymbol{\theta}_k$, i.e., we sum over all $\boldsymbol{\theta}_k$ and give positive weight to a vector $\boldsymbol{\theta}_k$ if a positive dot product with it is a necessary condition for a point $\boldsymbol{x}$ to belong to the k^{th} Voronoi region. Additionally, we scale the parameter vector by some $\alpha > 0$. Let $\boldsymbol{x} \in R_j$ and let R_m be a Voronoi region such that $j \neq m$. We next inspect the ratio

$$\begin{aligned} r(\alpha) &\overset{\text{def}}{=} \frac{\text{softmax}(\boldsymbol{\Theta}^{\star\top}\boldsymbol{x})_j}{\text{softmax}(\boldsymbol{\Theta}^{\star\top}\boldsymbol{x})_m} && \text{(14a)} \\ &= e^{(\boldsymbol{\theta}^\star{}_j - \boldsymbol{\theta}^\star_m)^\top \boldsymbol{x}} && \text{(14b)} \\ &= e^{(\alpha \sum_{k=0}^{K} \iota_j(k)\boldsymbol{\theta}_k - \alpha \sum_{k=0}^{K} \iota_m(k)\boldsymbol{\theta}_k)^\top \boldsymbol{x}} \\ &= e^{\alpha(\sum_{k=0}^{K}(\iota_j(k) - \iota_m(k))\boldsymbol{\theta}_k)^\top \boldsymbol{x}} && \text{(14c)} \end{aligned}$$

We now show that $\alpha(\sum_{k=0}^{K}(\iota_j(k) - \iota_m(k))\boldsymbol{\theta}_k)^\top \boldsymbol{x} > 0$ through the consideration of the following three cases:

- **Case 1**: $\iota_j(k) = \iota_m(k)$. In this case, the subspace $\boldsymbol{\theta}_k$ is a necessary condition for belonging to both regions j and m. Thus, the summand is zero.
- **Case 2**: $\iota_j(k) = 1$, but $\iota_m(k) = -1$. In this case, $\iota_j(k) - \iota_m(k) = 2$. As $\boldsymbol{x} \in R_j$, we know that $\boldsymbol{\theta}_k{}^\top \boldsymbol{x} > 0$, and the summand is positive.
- **Case 3**: $\iota_j(k) = -1$, but $\iota_m(k) = 1$. In this case, $\iota_j(k) - \iota_m(k) = -2$. As $\boldsymbol{x} \in R_j$, we know that $\boldsymbol{\theta}_k{}^\top \boldsymbol{x} < 0$, and the summand is, again, positive.

Since $j \neq m$, a summand corresponding to cases 2 and 3 must occur. Thus, the sum is strictly positive. It follows that $\lim_{\alpha\to\infty} r(\alpha) = 1$. Finally, for $\widehat{\mathrm{Y}}$ defined as in Eq. (3), we have $p(\widehat{\mathrm{Y}} = j \mid \boldsymbol{x} \in R_j) = 1$ for α large. Now, because all points in each R_j have a distinct label from $\mathcal{Z}$, it is trivial to construct a binary log-linear model that places arbitrarily high probability on R_j's label, which gives us $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}} \to \mathrm{Z}) > 1 - \eta$ for all $\eta > 0$ small. This completes the proof. ■

This construction demonstrates that one should be cautious when arguing about the implications of log-linear guardedness when multiclass softmax classifiers are applied over the guarded representations. When log-linear guardedness with respect to a binary Z is imposed, there may still exist a set of $k > 2$ linear separators that separate Z.

[6] In base two, $\log_2(2) = 1$ is the maximum achievable $\mathcal{V}$-information value in the binary case.

4 Accuracy-Based Guardedness

We now define an accuracy-based notion of guardedness and discuss its implications. Note that the information-theoretic notion of guardedness described above does not directly imply that the accuracy of the log-linear model is damaged. To see this, consider a binary log-linear model on balanced data that always assigns a probability of $\frac{1}{2} + \Delta$ to the correct label and $\frac{1}{2} - \Delta$ to the incorrect label. For small enough Δ, the cross-entropy loss of such a classifier will be arbitrarily close to the entropy $\log(2)$, even though it has perfect accuracy. This disparity motivates an accuracy-based notion of guardedness.

We first define the **accuracy function** ℓ as

$$\ell(q, \boldsymbol{x}, z) = \begin{cases} 1, & \textbf{if} \operatorname*{argmax}_{z' \in \mathcal{Z}} q(z' \mid \boldsymbol{x}) = z \\ 0, & \textbf{else} \end{cases} \tag{15}$$

The **conditional $\mathcal{V}$-accuracy** is defined as

$$\mathrm{A}_{\mathcal{V}}\left(\mathrm{Z} \mid \mathbf{X}\right) \stackrel{\text{def}}{=} \sup_{q \in \mathcal{V}} \mathop{\mathbb{E}}_{(z, \boldsymbol{x}) \sim p(\mathrm{Z}, \mathbf{X})} \ell(q, \boldsymbol{x}, z) \tag{16}$$

The **$\mathcal{V}$-accuracy** is a special case of Eq. (16) when no random variable is conditioned on

$$\mathrm{A}_{\mathcal{V}}\left(\mathrm{Z}\right) \stackrel{\text{def}}{=} \sup_{q \in \mathcal{V}} \mathop{\mathbb{E}}_{z \sim p(\mathrm{Z})} \ell(q, z) \tag{17}$$

where we have overloaded ℓ to take only two arguments. We can now define an analogue of Xu et al.'s (2020) $\mathcal{V}$-information for accuracy as the difference between the unconditional $\mathcal{V}$-accuracy and the conditional $\mathcal{V}$-accuracy[7]

$$\mathrm{I}^{\mathrm{A}}_{\mathcal{V}}(\mathbf{X} \to \mathrm{Z}) \stackrel{\text{def}}{=} \mathrm{A}_{\mathcal{V}}\left(\mathrm{Z} \mid \mathbf{X}\right) - \mathrm{A}_{\mathcal{V}}\left(\mathrm{Z}\right) \tag{18}$$

Note that the $\mathcal{V}$-accuracy is bounded below by 0 and above by $\frac{1}{2}$.

Definition 4.1 (Accuracy-based $\mathcal{V}$-guardedness)**.** *Let $\mathbf{X}$ be a representation-valued random variable and let Z be an attribute-valued random variable. Moreover, let $\mathcal{V}$ be a predictive family. A guarding function h ε-guards $\mathbf{X}$ against Z with respect to $\mathcal{V}$ if $\mathrm{I}^{\mathrm{A}}_{\mathcal{V}}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$.*

Definition 4.2 (Accuracy-based Empirical $\mathcal{V}$-guardedness)**.** *Let $\mathcal{D} = \{(\boldsymbol{x}_n, z_n)\}_{n=1}^N$ where $(\boldsymbol{x}_n, z_n) \sim p(\mathbf{X}, \mathrm{Z})$. Let $\widetilde{\mathbf{X}}$ and $\widetilde{\mathrm{Z}}$ be random variables over $\mathbb{R}^D$ and $\mathcal{Z}$, respectively, whose distribution corresponds to the marginals of the empirical distribution over $\mathcal{D}$. A guarding function h empirically ε-guards $\mathcal{D}$ with respect to $\mathcal{V}$ if $\mathrm{I}^{\mathrm{A}}_{\mathcal{V}}(h(\widetilde{\mathbf{X}}) \to \widetilde{\mathrm{Z}}) < \varepsilon$.*

When focusing on accuracy in predicting Z, it is natural to consider the **independence** (also known as **demographic parity**) (Feldman et al., 2015) of the downstream classifiers that are trained over the representations.

Definition 4.3. *The L_1 **independence gap** measures the difference between the distribution of the model's predictions on the examples for which $\mathrm{Z} = \bot$, and the examples for which $\mathrm{Z} = \top$. It is formally defined as*

$$\begin{aligned} & GAP_{ind}(\widehat{\mathrm{Y}} \to \mathrm{Z} \mid \mathbf{X}) \qquad\qquad (19) \\ & \stackrel{\text{def}}{=} \sum_{y \in \mathcal{Y}} \Bigg| \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) \\ & \qquad - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) \Bigg| \end{aligned}$$

where $p(\mathbf{X} \mid \mathrm{Z})$ is the conditional distribution over representations given the protected attribute.

In Prop. 4.4, we prove that if the data is linearly ε-guarded and **globally balanced** with respect to Z, i.e., if $p(\mathrm{Z} = \bot) = p(\mathrm{Z} = \top) = \frac{1}{2}$, then the prediction of any linear binary downstream classifier is 4ε independent of Z. Note that is true *regardless* of any imbalance of the protected attribute Z within each class $y \in \mathcal{Y}$ in the downstream task: the data only needs to be globally balanced.

Proposition 4.4. *Let $\mathcal{V}$ be the family of binary log-linear models, and assume that $p(\mathbf{X}, \mathrm{Z})$ is globally balanced, i.e., $p(\mathrm{Z} = \bot) = p(\mathrm{Z} = \top) = \frac{1}{2}$. Furthermore, let h be a guarding function that ε-guards $\mathbf{X}$ against Z with respect to $\mathcal{V}$ in terms of accuracy (Definition 4.2), i.e., $\mathrm{I}^{\mathrm{A}}_{\mathcal{V}}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$. Let $\widehat{\mathrm{Y}}$ be defined as in Eq. (1). Then, the L_1 independence gap (Eq. (19)) satisfies $GAP_{ind}(\widehat{\mathrm{Y}} \to \mathrm{Z} \mid h(\mathbf{X})) \leq 4\varepsilon$.*

Proof. See App. A.2 for the proof. ■

5 Experimental Evaluation

In the empirical portion of our paper, we evaluate the extent to which our theory holds in practice.

Data. We perform experiments on gender bias mitigation on the Bias in Bios dataset (De-Arteaga et al., 2019), which is composed of short biographies annotated by both gender and profession.

[7] Note the order of the terms is reversed in $\mathcal{V}$-accuracy.

We represent each biography with the `[CLS]` representation in the final hidden representation of pre-trained BERT, which creates our representation random variable $\mathbf{X}$. We then try to guard against the protected attribute gender, which constitutes Z.

Approximating log-linear guardedness. To approximate the condition of log-linear guardedness, we use RLACE (Ravfogel et al., 2022a). The method is based on a minimax game between a log-linear predictor that aims to predict the concept of interest from the representation and an orthogonal projection matrix that aims to guard the representation against prediction. The process results in an orthogonal projection matrix $P \in \mathbb{R}^{D \times D}$, which, empirically, prevents log-linear models from predicting gender after the linear transformation P is applied to the representations. This process constitutes our guarding function h_{R}. Our theoretical result (Theorem 3.2) only holds for δ-discretized log-linear models. RLACE, however, guards against conventional log-linear models. Thus, we apply δ-discretization post hoc, i.e., after training.

5.1 Quantifying Empirical Guardedness

We test whether our theoretical analysis of leakage through binary and multiclass downstream classifiers holds in practice, on held-out data. Profession prediction serves as our downstream prediction task (i.e., our $\widehat{\mathrm{Y}}$), and we study binary and multiclass variants of this task. In both cases, we measure three $\mathcal{V}$-information estimates:

- **Evaluating** $\mathrm{I}_{\mathcal{V}}(\mathbf{X} \to \mathrm{Z})$. To compute an empirical upper bound on information about the protected attribute which is linearly extractable from the representations, we train a log-linear model to predict z_n from $\boldsymbol{x}_n$, i.e., from the unguarded representations. In other words, this is an upper bound on the information that could be leaked through the downstream classifier $\widehat{\mathrm{Y}}$.

- **Evaluating** $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{p}} \to \mathrm{Z})$. We quantify leakage through a downstream classifier $\widehat{\mathrm{Y}}$ by estimating $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}} \to \mathrm{Z})$, for binary and multiclass $\widehat{\mathrm{Y}}$, via two different approaches. The first of these, denoted $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{p}} \to \mathrm{Z})$, is computed by training two log-linear and stiching them together into a pipeline. First, we fit a log-linear model on top of the guarded representations $h_{\mathrm{R}}(\mathbf{X})$ to yield predictions for a downstream task $\widehat{\mathrm{Y}}_{\mathrm{p}} = t(h_{\mathrm{R}}(\mathbf{X}))$ where $t : \mathbb{R}^D \to \mathcal{Y}$ is the function induced by the trained classifier. Subsequently, we train a second log-linear model $\widehat{\mathrm{Z}} = r(\widehat{\mathrm{Y}}_{\mathrm{p}})$ with $r : \mathcal{Y} \to \{0, 1\}$ to predict Z from the output of $\widehat{\mathrm{Y}}$ In words, $\widehat{\mathrm{Y}}_{\mathrm{p}}$ represents the argmax from the distribution of profession labels (binary or multiclass). We approximate the $\mathcal{V}$-information $\widehat{\mathrm{Y}}_{\mathrm{p}}$ leaks about Z through the cross-entropy loss of a second classifier trained to predict the protected attribute from $\widehat{\mathrm{Y}}_{\mathrm{p}}$, i.e., we compute empirical guardedness (Definition 2.2) on held-out data.

- **Evaluating** $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{a}} \to \mathrm{Z})$. In addition to the standard scenario estimated by $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{p}} \to \mathrm{Z})$, we also ask: What is the maximum amount of information that a downstream classifier could leak about gender? $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{a}} \to \mathrm{Z})$ estimates this quantity, with a variant of the setup of $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{p}} \to \mathrm{Z})$. Namely, instead of training the two log-linear models separately, we train them together to find the $\widehat{\mathrm{Y}}_{\mathrm{a}}$ that is adversarially chosen to predict gender well. However, the argmax operation is not differentiable, so we remove it during training.

  In practice, this means $\widehat{\mathrm{Y}}_{\mathrm{a}}$ does not predict profession, but instead predicts a latent distribution which is adversarially chosen so as to best enable the prediction of gender.[8]

While high $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{a}} \to \mathrm{Z})$ indicates that there exists an adversarial log-linear model that leaks information about Z, it does not necessarily mean that common classifiers like those used to compute $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{p}} \to \mathrm{Z})$ would leak that information. Across all of 3 conditions, we explore how different values of the thresholds δ (applied after training) affect the $\mathcal{V}$-information. Refer to App. A.3 for comprehensive details regarding our experimental setup.

5.2 Binary Z and Y

We start by evaluating the case where both Z and Y take only two values each.

Experimental Setting. To create our set $\mathcal{Y}$ for the binary classification task, we randomly sample 15 pairs of professions from the Bias in Bios dataset; see App. A.3. We train a binary log-linear

[8] Only at inference time do we apply the argmax over the first log-linear model to get a prediction $\widehat{\mathrm{Y}}_{\mathrm{a}} = y$. We find that the loss of the composition model is not increased by the argmax operation.

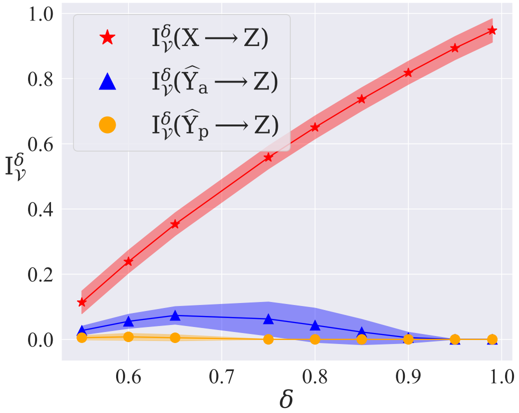


Figure 2: Results for § 5.2. Estimate of $\mathcal{V}$-information between the protected attribute and (1) the original representations (red); (2) the labels induced by the inner model within a composition of two log-linear models, trained to adversarially recover gender (blue); (3) labels for the downstream task (the predictions of profession classifiers; orange). The curve is the mean over different pairs of professions, and the shaded area representations 1 standard deviation. The x-axis presents results for different values of the threshold δ. Recall the thresholding is applied post hoc.

model to predict each profession from the representation after the application of the RLACE guarding function, $h_{\mathrm{R}}(\boldsymbol{x}_n)$. Empirically, we observe that our log-linear models achieve no more than 2% above majority accuracy on the protected data. For each pair of professions, we estimate three forms of $\mathcal{V}$-information.

Results. The results are presented in Fig. 2, for the 15 pairs of professions we experimented with (each curve is the mean over all pairs), the three quantities listed above, and different values of the threshold δ on the x-axis. Unsurprisingly, we observe that the $\mathcal{V}$-information estimated from the original representations (the red curve) has high values for some thresholds, indicating that BERT representations do encode gender distinctions. The blue curve, corresponding to $I_{\mathcal{V}}(\widehat{Y}_a \rightarrow Z)$, measures the ability of the adversarially constructed binary downstream classifier to recover the gender information. It is lower than the red curve but is nonzero, indicating that the solution found by RLACE does not generalize perfectly. Finally, the orange curve, corresponding to $I_{\mathcal{V}}(\widehat{Y}_p \rightarrow Z)$, measures the amount of leakage we get in practice from downstream classifiers that are trained on profession prediction. In that case, the numbers are significantly lower, showing that RLACE does provide decent guardedness in practice.

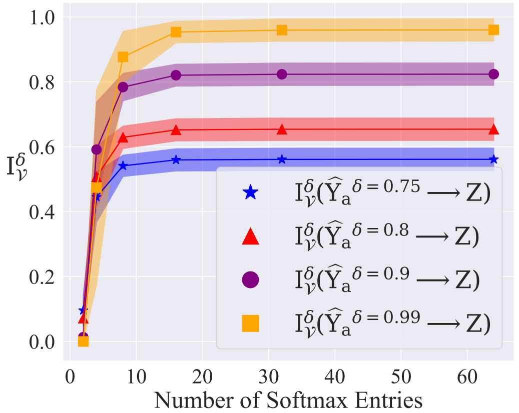


Figure 3: Results for § 5.3. Estimate of $\mathcal{V}$-information between the protected attribute and $\widehat{Y}_a$ with various δ.

5.3 Binary Z, Multiclass Y

Empirically, we have shown that RLACE provides good, albeit imperfect, protection against binary log-linear model adversaries. This finding is in line with the conclusions of Theorem 3.2. We now turn to experiments on multiclass classification, i.e., where $|\mathcal{Y}| > 2$. According to § 3.3, to the extent that the K-Voronoi assumption holds, we expect guardedness to be broken with a large enough $|\mathcal{Y}|$.

Experimental Setting. Note that, since $|\mathcal{Y}| > 2$, $\widehat{Y}$ is a multiclass log-linear classifier over $\mathcal{Y}$, but the logistic classifier that predicts gender from the argmax over these remains binary. We consider different values of $|\mathcal{Y}| = 2, 4, 8, 16, 32, 64$.

Results. The results are shown in Fig. 3. For all professions, we find a log-linear model whose predicted labels are highly predictive of the protected attribute. Indeed, softmax classifiers with 4 to 8 entries (corresponding to hidden neurons in the network which is the composition of two log-linear models) perfectly recover the gender information. This indicates that there are labeling schemes of the data using 4 or 8 labels that recover almost all information about Z.

5.4 Discussion

Even if a set of representations is log-linearly guarded, one can still adversarially construct a multiclass softmax classifier that recovers the information. These results stem from the disparity between the manifold in which the concept resides, and the expressivity of the (linear) intervention we perform: softmax classifiers can access information that is inaccessible to a purely linear classifier. Thus, interventions that are aimed at achieving guardedness should consider the specific adversary against which one aims to protect.

6 Related Work

Techniques for information removal are generally divided into adversarial methods and post-hoc linear methods. Adversarial methods (Edwards and Storkey, 2016; Xie et al., 2017; Chen et al., 2018; Elazar and Goldberg, 2018; Zhang et al., 2018) use a gradient-reversal layer during training to induce representations that do not encode the protected attribute. However, Elazar and Goldberg (2018) have shown that these methods fail to exhaustively remove all the information associated with the protected attribute. Linear methods have been proposed as a tractable alternative, where one identifies a linear subspace that captures the concept of interest, and neutralizes it using algebraic techniques. Different methods have been proposed for the identification of the subspace, e.g., PCA and variants thereof (Bolukbasi et al., 2016; Kleindessner et al., 2023), orthogonal rotation (Dev et al., 2021), classification-based (Ravfogel et al., 2020), spectral (Shao et al., 2023a,b) and adversarial approaches (Ravfogel et al., 2022a).

Different definitions have been proposed for fairness (Mehrabi et al., 2021), but they are mostly extrinsic—they concern themselves only with the predictions of the model, and not with its representation space. Intrinsic bias measures, which focus on the representation space of the model, have been also extensively studied. These measures quantify, for instance, the extent to which the word representation space encodes gender distinctions (Bolukbasi et al., 2016; Caliskan et al., 2017; Kurita et al., 2019; Zhao et al., 2019). The *relation* between extrinsic and intrinsic bias measures is understudied, but recent works have demonstrated empirically either a relatively weak or inconsistent correlation between the two (Goldfarb-Tarrant et al., 2021; Orgad et al., 2022; Cao et al., 2022; Orgad and Belinkov, 2022; Steed et al., 2022; Shen et al., 2022; Cabello et al., 2023).

7 Conclusion

We have formulated the notion of guardedness as the inability to *directly* predict a concept from the representation. We show that log-linear guardedness with respect to a binary protected attribute does not prevent a *subsequent* multiclass linear classifier trained over the guarded representations from leaking information on the protected attribute. In contrast, when the main task is binary, we can bound that leakage. Altogether, our analysis suggests that the deployment of linear erasure methods should carefully take into account the manner in which the modified representations are being used later on, e.g., in classification tasks.

Limitations

Our theoretical analysis targets a specific notion of information leakage, and it is likely that it does not apply to alternative ones. While the $\mathcal{V}$-information-based approach seems natural, future work should consider alternative extrinsic bias measures as well as alternative notions of guardedness. Additionally, our focus is on the linear case, which is tractable and important—but limits the generality of our conclusions. We hope to extend this analysis to other predictive families in future work.

Ethical Considerations

The empirical experiments in this work involve the removal of binary gender information from a pre-trained representation. Beyond the fact that gender is a non-binary concept, this task may have real-world applications, in particular such that relate to fairness. We would thus like to remind the readers to take the results with a grain of salt and be extra careful when attempting to deploy methods such as the one discussed here. Regardless of any theoretical result, care should be taken to measure the effectiveness of bias mitigation efforts in the context in which they are to be deployed, considering, among other things, the exact data to be used and the exact fairness metrics under consideration.

Acknowledgements

We thank Afra Amini, Clément Guerner, David Schneider-Joseph, Nora Belrose and Stella Biderman for their thoughtful comments and revision of this paper. This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No. 802774 (iEXTRACT). Shauli Ravfogel is grateful to be supported by the Bloomberg Data Science Ph.D Fellowship. Ryan Cotterell acknowledges the Google Research Scholar program for supporting the pro-

posal "Controlling and Understanding Representations through Concept Erasure."

References

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. *arXiv preprint arXiv:2306.03819*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29:4349–4357.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. *arXiv preprint arXiv:2304.10153*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2 edition. Wiley-Interscience.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 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*, pages 120–128.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR.

Harrison Edwards and Amos J. Storkey. 2016. Censoring representations with an adversary. In *4^{th} International Conference on Learning Representations*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. 2023. Efficient fair PCA for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5250–5270.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. *arXiv preprint arXiv:2210.11471*.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. *arXiv preprint arXiv:2204.06827*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022a. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. Kernelized concept erasure. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023a. Erasure of unaligned attributes from neural representations. *Transactions of the Association for Computational Linguistics*, 11:488–510.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023b. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is *Not* all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Georges Voronoi. 1908. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik*, 1908(134):198–287.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. 2017. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108, Copenhagen, Denmark. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 Composition of δ-discretized binary log-linear models

Lemma A.1. *Let $\mathcal{V}^\delta$ be the family of discretized binary log-linear models (Definition 3.1). Let $\tau(\boldsymbol{\theta}^\top \boldsymbol{x}+\phi)$ be a linear decision rule where τ is defined as in Eq. (10), and, furthermore, assume $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi \neq 0$ for all $\boldsymbol{x}$. Then, for any $\alpha, \beta \in \mathbb{R}$, there exists a function $r \in \mathcal{V}^\delta$ such that $r(0) = \rho_\delta(\sigma(\alpha \cdot \tau(\boldsymbol{\theta}^\top \boldsymbol{x}+\phi)+\beta))$ where ρ_δ is defined as in Definition 3.1.*

Proof. Consider the function composition $\sigma(\alpha \cdot \tau(\boldsymbol{\theta}^\top \boldsymbol{x}+\phi)+\beta)$. First, note that $\tau(\boldsymbol{\theta}^\top \boldsymbol{x}+\phi)$ is a step-function. And, thus, so, too, is $\sigma(\alpha \cdot \tau(\boldsymbol{\theta}^\top \boldsymbol{x}+\phi)+\beta)$, i.e.,

$$\widehat{y}(\boldsymbol{x}) \stackrel{\text{def}}{=} \sigma(\alpha \cdot \tau(\boldsymbol{\theta}^\top \boldsymbol{x}+\phi)+\beta) = \begin{cases} \frac{1}{1+e^{-\beta}} \stackrel{\text{def}}{=} a, & \textbf{if } \boldsymbol{\theta}^\top \boldsymbol{x}+\phi \leq 0 \\ \frac{1}{1+e^{-\alpha-\beta}} \stackrel{\text{def}}{=} b, & \textbf{else} \end{cases} \tag{20}$$

This results in a classifier with the following properties, if $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi \leq 0$, we have

$$p(\widehat{\mathrm{Y}} = 0 \mid \mathbf{X} = \boldsymbol{x}) = a \tag{21}$$

$$p(\widehat{\mathrm{Y}} = 1 \mid \mathbf{X} = \boldsymbol{x}) = 1 - a \tag{22}$$

Otherwise, if $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi > 0$, we have

$$p(\widehat{\mathrm{Y}} = 0 \mid \mathbf{X} = \boldsymbol{x}) = b \tag{23}$$

$$p(\widehat{\mathrm{Y}} = 1 \mid \mathbf{X} = \boldsymbol{x}) = 1 - b \tag{24}$$

By assumption, we have $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi \neq 0$. Now, observe that a binary δ-discretized classifier can represent any distribution of the form

$$r(0) = \begin{cases} \delta & \textbf{if } \boldsymbol{\theta}^\top \boldsymbol{x}+\phi > 0 \\ 1-\delta & \textbf{else} \end{cases} \tag{25}$$

$$r(1) = \begin{cases} 1-\delta & \textbf{if } \boldsymbol{\theta}^\top \boldsymbol{x}+\phi > 0 \\ \delta & \textbf{else} \end{cases} \tag{26}$$

We show how to represent r as a δ-discretized binary log-linear model in four cases:

- **Case 1**: $a > \frac{1}{2}$ and $b < \frac{1}{2}$. In this case, we require a classifier that places probability δ on 0 if $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi < 0$ and probability $1-\delta$ on 0 if $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi < 0$.
- **Case 2**: $a < \frac{1}{2}$ and $b > \frac{1}{2}$. In this case, we require a classifier that places probability δ on 0 if $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi < 0$ and probability $1-\delta$ on 0 if $\boldsymbol{\theta}^\top \boldsymbol{x}+\phi > 0$.
- **Case 3**: $a, b > \frac{1}{2}$. In this case, we set $\boldsymbol{\theta} = \mathbf{0}$ and $\phi > 0$.
- **Case 4**: $a, b < \frac{1}{2}$. In this case, we set $\boldsymbol{\theta} = \mathbf{0}$ and $\phi < 0$.

This proves the result. ■

A.2 Accuracy-based Guardedness: the Balanced Case

Proposition 4.4. *Let $\mathcal{V}$ be the family of binary log-linear models, and assume that $p(\mathbf{X}, \mathrm{Z})$ is globally balanced, i.e., $p(\mathrm{Z} = \bot) = p(\mathrm{Z} = \top) = \frac{1}{2}$. Furthermore, let h be a guarding function that ε-guards $\mathbf{X}$ against Z with respect to $\mathcal{V}$ in terms of accuracy (Definition 4.2), i.e., $\mathrm{I}^{\mathrm{A}}_{\mathcal{V}}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$. Let $\widehat{\mathrm{Y}}$ be defined as in Eq. (1). Then, the L_1 independence gap (Eq. (19)) satisfies GAP$_{ind}(\widehat{\mathrm{Y}} \to \mathrm{Z} \mid h(\mathbf{X})) \leq 4\varepsilon$.*

Proof. In the following proof, we use the notation $\mathbf{X} = \boldsymbol{x}$ for the guarded variable $h(\mathbf{X}) = h(\boldsymbol{x})$ to avoid notational cutter. Assume, by way of contradiction, that the L_1 independence gap (Eq. (19)), $\sum_y \left| \underset{\boldsymbol{x}\sim p(\mathbf{X}|\mathrm{Z}=\perp)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \perp, \mathbf{X} = \boldsymbol{x}) - \underset{\boldsymbol{x}\sim p(\mathbf{X}|\mathrm{Z}=\top)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) \right| > 4\varepsilon$. Then, there exists a $y \in \mathcal{Y}$ such that

$$\left| \underset{h(\boldsymbol{x})\sim p(h(\mathbf{X})|\mathrm{Z}=\perp)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \perp, \mathbf{X} = \boldsymbol{x}) - \underset{\boldsymbol{x}\sim p(\mathbf{X}|\mathrm{Z}=\top)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) \right| > 2\varepsilon. \tag{27}$$

We will show that we can build a classifier $q^\star \in \mathcal{V}$ that breaks the assumption $\mathrm{I}_{\mathcal{V}}^{\mathrm{A}}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$. Next, we define the random variable $\widehat{\mathrm{Z}}_q$ for convenience as

$$\widehat{\mathrm{Z}}_q(z) \overset{\text{def}}{=} \begin{cases} 1, & \textbf{if } z = \underset{z'}{\operatorname{argmax}}\, q(z' \mid \boldsymbol{x}) \\ 0, & \textbf{else} \end{cases} \tag{28}$$

In words, $\widehat{\mathrm{Z}}_q$ is a random variable that ranges over possible predictions, derived from the argmax, of the binary log-linear model q. Now, consider the following two cases.

- **Case 1**: There exits a y such that $\underset{\boldsymbol{x}\sim p(\mathbf{X}|\mathrm{Z}=\perp)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \perp, \mathbf{X} = \boldsymbol{x}) - \underset{\boldsymbol{x}\sim p(\mathbf{X}|\mathrm{Z})}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) > 2\varepsilon$. Let $\widehat{\mathrm{Y}}$ be defined as in Eq. (1). Next, consider a random variable $\widehat{\mathrm{Z}}_r$ defined as follows

$$p(\widehat{\mathrm{Z}}_r = \perp \mid \widehat{\mathrm{Y}} = y) \overset{\text{def}}{=} \begin{cases} 1, & \textbf{if } \widehat{\mathrm{Y}} = y \\ 0, & \textbf{else} \end{cases} \tag{29a}$$

$$p(\widehat{\mathrm{Z}}_r = \top \mid \widehat{\mathrm{Y}} = y) \overset{\text{def}}{=} \begin{cases} 1, & \textbf{if } \widehat{\mathrm{Y}} \neq y \\ 0, & \textbf{else} \end{cases} \tag{29b}$$

Now, note that we have

$$p(\widehat{\mathrm{Z}}_r = \perp \mid \mathbf{X} = \boldsymbol{x}) = \sum_{y\in\mathcal{Y}} p(\widehat{\mathrm{Z}}_r = \perp \mid \widehat{\mathrm{Y}} = y) p(\widehat{\mathrm{Y}} = y \mid \mathbf{X} = \boldsymbol{x}) \tag{30a}$$

$$= p(\widehat{\mathrm{Y}} = y \mid \mathbf{X} = \boldsymbol{x}) \tag{30b}$$

and

$$p(\widehat{\mathrm{Z}}_r = \top \mid \mathbf{X} = \boldsymbol{x}) = \sum_{y\in\mathcal{Y}} p(\widehat{\mathrm{Z}}_r = \top \mid \widehat{\mathrm{Y}} = y) p(\widehat{\mathrm{Y}} = y \mid \mathbf{X} = \boldsymbol{x}) \tag{31a}$$

$$= p(\widehat{\mathrm{Y}} \neq y \mid \mathbf{X} = \boldsymbol{x}) \tag{31b}$$

We perform the algebra below where the step from Eq. (37c) to Eq. (37d) follows because of the fact that, despite the nuisance variable $\widehat{\mathrm{Y}}$, the decision boundary of $p(\widehat{\mathrm{Z}}_r = \top \mid \mathbf{X} = \boldsymbol{x})$ is linear and, thus, there exists a binary log-linear model in $\mathcal{V}$ which realizes it. Now, consider the following steps

$$\mathrm{A}_{\mathcal{V}}\left(\mathrm{Z} \mid \mathbf{X}\right) \overset{\text{def}}{=} \sup_{q \in \mathcal{V}} \underset{(\boldsymbol{x}, z) \sim p(\mathrm{Z}, \mathbf{X})}{\mathbb{E}} \ell(q, \boldsymbol{x}, z) \tag{32a}$$

$$= \sup_{q \in \mathcal{V}} \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z})}{\mathbb{E}} \; \underset{z \sim p(\mathrm{Z})}{\mathbb{E}} \ell(q, \boldsymbol{x}, z) \tag{32b}$$

$$= \sup_{q \in \mathcal{V}} \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} p(\widehat{\mathrm{Z}}_q = \bot \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \bot) \tag{32c}$$

$$+ \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} p(\widehat{\mathrm{Z}}_q = \top \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \top)$$

$$\geq \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} p(\widehat{\mathrm{Z}}_r = \bot \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \bot) \tag{32d}$$

$$+ \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} p(\widehat{\mathrm{Z}}_r = \top \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \top)$$

$$= \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \bot) \tag{32e}$$

$$+ \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} p\left(\widehat{\mathrm{Y}} \neq y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}\right) p(\mathrm{Z} = \top) \tag{32f}$$

$$= \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \bot) \tag{32g}$$

$$+ \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} (1 - p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x})) p(\mathrm{Z} = \top) \tag{32h}$$

$$= \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} \frac{1}{2} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) \tag{32i}$$

$$+ \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} \frac{1}{2} (1 - p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x})) \tag{32j}$$

$$= \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} \frac{1}{2} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) \tag{32k}$$

$$+ \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} \frac{1}{2} - \frac{1}{2} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) \tag{32l}$$

$$= \frac{1}{2} \underbrace{\underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) - \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) + \frac{1}{2}}_{> 2\varepsilon \text{ by assumption}} \tag{32m}$$

$$> \frac{1}{2} (2\varepsilon) + \frac{1}{2} \tag{32n}$$

$$= \frac{1}{2} + \varepsilon \tag{32o}$$

- **Case 2**: There exits a y such that

$$\underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) - \underset{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)}{\mathbb{E}} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) > 2\varepsilon \tag{33}$$

Let $\widehat{\mathrm{Y}}$ be defined as in Eq. (1). Next, consider a random variable defined as follows

$$p(\widehat{\mathrm{Z}}_r = \bot \mid \widehat{\mathrm{Y}} = y) \overset{\text{def}}{=} \begin{cases} 1, & \textbf{if } \widehat{\mathrm{Y}} \neq y \\ 0, & \textbf{else} \end{cases} \tag{34a}$$

$$p(\widehat{\mathrm{Z}}_r = \top \mid \widehat{\mathrm{Y}} = y) \overset{\text{def}}{=} \begin{cases} 1, & \textbf{if } \widehat{\mathrm{Y}} = y \\ 0, & \textbf{else} \end{cases} \tag{34b}$$

Now, note that we have

$$
\begin{align}
p(\widehat{\mathrm{Z}}_r = \bot \mid \mathbf{X} = \boldsymbol{x}) &= \sum_{y \in \mathcal{Y}} p(\widehat{\mathrm{Z}}_r = \bot \mid \widehat{\mathrm{Y}} = y) p(\widehat{\mathrm{Y}} = y \mid \mathbf{X} = \boldsymbol{x}) \tag{35a} \\
&= p(\widehat{\mathrm{Y}} \neq y \mid \mathbf{X} = \boldsymbol{x}) \tag{35b}
\end{align}
$$

and

$$
\begin{align}
p(\widehat{\mathrm{Z}}_r = \top \mid \mathbf{X} = \boldsymbol{x}) &= \sum_{y \in \mathcal{Y}} p(\widehat{\mathrm{Z}}_r = \top \mid \widehat{\mathrm{Y}} = y) p(\widehat{\mathrm{Y}} = y \mid \mathbf{X} = \boldsymbol{x}) \tag{36a} \\
&= p(\widehat{\mathrm{Y}} = y \mid \mathbf{X} = \boldsymbol{x}) \tag{36b}
\end{align}
$$

We proceed by algebraic manipulation

$$
\begin{align}
\mathrm{A}_{\mathcal{V}}\left(\mathrm{Z} \mid \mathbf{X}\right) &\overset{\text{def}}{=} \sup_{q \in \mathcal{V}} \mathop{\mathbb{E}}_{(\boldsymbol{x}, z) \sim p(\mathrm{Z}, \mathbf{X})} \ell(q, \boldsymbol{x}, z) \tag{37a} \\
&= \sup_{q \in \mathcal{V}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z})} \mathop{\mathbb{E}}_{z \sim p(\mathrm{Z})} \ell(q, \boldsymbol{x}, z) \tag{37b} \\
&= \sup_{q \in \mathcal{V}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} p(\widehat{\mathrm{Z}}_q = \bot \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \bot) \tag{37c} \\
&\qquad + \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} p(\widehat{\mathrm{Z}}_q = \top \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \top) \notag \\
&\geq \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} p(\widehat{\mathrm{Z}}_r = \bot \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \bot) \tag{37d} \\
&\qquad + \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} p(\widehat{\mathrm{Z}}_r = \top \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \top) \notag \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} p(\widehat{\mathrm{Y}} \neq y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \bot) \tag{37e} \\
&\qquad + \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} p\left(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}\right) p(\mathrm{Z} = \top) \tag{37f} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} (1 - p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x})) p(\mathrm{Z} = \bot) \tag{37g} \\
&\qquad + \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) p(\mathrm{Z} = \top) \tag{37h} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} \frac{1}{2} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) \tag{37i} \\
&\qquad + \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} \frac{1}{2} (1 - p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x})) \tag{37j} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} \frac{1}{2} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) \tag{37k} \\
&\qquad + \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} \frac{1}{2} - \frac{1}{2} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x}) \tag{37l} \\
&= \frac{1}{2} \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \top)} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \top, \mathbf{X} = \boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\mathbf{X} \mid \mathrm{Z} = \bot)} p(\widehat{\mathrm{Y}} = y \mid \mathrm{Z} = \bot, \mathbf{X} = \boldsymbol{x})}_{> 2\varepsilon \text{ by assumption}} + \frac{1}{2} \tag{37m} \\
&> \frac{1}{2} \left(2\varepsilon\right) + \frac{1}{2} \tag{37n} \\
&= \frac{1}{2} + \varepsilon \tag{37o}
\end{align}
$$

In both cases, we have $\mathrm{A}_{\mathcal{V}}\left(\mathrm{Z} \mid \mathbf{X} = \boldsymbol{x}\right) > \frac{1}{2} + \varepsilon$. Thus, $\mathbb{E}_{\boldsymbol{x} \sim p(\mathbf{X})} \mathrm{A}_{\mathcal{V}}\left(\mathrm{Z} \mid \mathbf{X} = \boldsymbol{x}\right) = \mathrm{A}_{\mathcal{V}}\left(\mathrm{Z} \mid h(\mathbf{X})\right) \geq$

$\frac{1}{2} + \varepsilon$. Note that the distribution $p(\mathrm{Z}, \mathbf{X})$ is globally balanced, we have $\mathrm{A}_{\mathcal{V}}(\mathrm{Z}) = \frac{1}{2}$. Thus,

$$
\begin{aligned}
\mathrm{I}^{\mathrm{A}}_{\mathcal{V}}(h(\mathbf{X}) \to \mathrm{Z}) &= \mathrm{A}_{\mathcal{V}}(\mathrm{Z}) - \mathrm{A}_{\mathcal{V}}(\mathrm{Z} \mid h(\mathbf{X})) && \text{(38a)} \\
&= \mathrm{A}_{\mathcal{V}}(\mathrm{Z} \mid h(\mathbf{X})) - \frac{1}{2} && \text{(38b)} \\
&> \frac{1}{2} + \varepsilon - \frac{1}{2} && \text{(38c)} \\
&= \varepsilon && \text{(38d)}
\end{aligned}
$$

However, this contradicts the assumption that $\mathrm{I}^{\mathrm{A}}_{\mathcal{V}}(h(\mathbf{X}) \to \mathrm{Z}) < \varepsilon$. This completes the proof. ■

A.3 Experimental Setting

In this appendix, we give additional information necessary to replicate our experiments (§ 5).

Data. We use the same train–dev–test split of the biographies dataset used by Ravfogel et al. (2020), resulting in training, evaluation and test sets of sizes 255,710, 39,369, and 98,344, respectively. We reduce the dimensionality of the representations to 256 using PCA. The dataset is composed of short biographies, annotated with both gender and profession. We randomly sampled 15 pairs of professions from the dataset: (*professor*, *attorney*), (*journalist*, *surgeon*), (*physician*, *nurse*), (*professor*, *physician*), (*psychologist*, *teacher*), (*attorney*, *teacher*), (*physician*, *journalist*), (*professor*, *dentist*), (*teacher*, *surgeon*), (*psychologist*, *surgeon*), (*photographer*, *surgeon*), (*attorney*, *psychologist*), (*physician*, *teacher*), (*professor*, *teacher*), (*professor*, *psychologist*)

Optimization. We run RLACE (Ravfogel et al., 2022a) with a simple SGD optimization, with a learning rate of 0.005, a weight decay of $1e^{-5}$ and a momentum of 0.9, chosen by experimenting with the development set. We use a batch size of 128. The algorithm is based on an adversarial game between a predictor that aims to predict gender, and an orthogonal projection matrix adversary that aims to prevent gender classification. We choose the adversary which yielded *highest* classification loss. All training is done on a single NVIDIA GeForce GTX 1080 Ti GPU.

Estimating $\mathcal{V}$-information. After running RLACE, we get an approximately linearly-guarded representation by projecting $\boldsymbol{x}_n \leftarrow P\boldsymbol{x}_n$, where P is the orthogonal projection matrix returned by RLACE. We validate guardedness by training log-linear models over the projected representations; they achieve accuracy less than 2% above the majority accuracy. Then, to estimate $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{a}} \rightarrow \mathrm{Z})$, we fit a simple neural network of the form of a composition of two log-linear models. The inner model either has a single hidden neuron with a logistic activation (in the binary experiment), or $K = 2, 4, 8, 16, 32, 64$ hidden neurons with softmax activations, in the multiclass experiment (§ 5.3). The networks are trained end to end to recover binary gender for 25000 batches of size 2048. Optimization is done with Adam with the default parameters. We use the loss of the second log-linear model to estimate $\mathrm{I}_{\mathcal{V}}(\widehat{\mathrm{Y}}_{\mathrm{a}} \rightarrow \mathrm{Z})$, according to Definition 2.2.