Learning Invariant Representations using Inverse Contrastive Loss

Aditya Kumar Akash[1], **Vishnu Suresh Lokhande**[1], **Sathya N. Ravi**[2], **Vikas Singh**[1]
[1] University of Wisconsin-Madison,
[2] University of Illinois at Chicago
aakash@wisc.edu, lokhande@cs.wisc.edu, sathya@uic.edu, vsingh@biostat.wisc.edu

Abstract

Learning invariant representations is a critical first step in a number of machine learning tasks. A common approach corresponds to the so-called information bottleneck principle in which an application dependent function of mutual information is carefully chosen and optimized. Unfortunately, in practice, these functions are not suitable for optimization purposes since these losses are agnostic of the metric structure of the parameters of the model. We introduce a class of losses for learning representations that are invariant to some extraneous variable of interest by inverting the class of contrastive losses, i.e., inverse contrastive loss (ICL). We show that if the extraneous variable is binary, then optimizing ICL is equivalent to optimizing a regularized MMD divergence. More generally, we also show that if we are provided a metric on the sample space, our formulation of ICL can be decomposed into a sum of convex functions of the given distance metric. Our experimental results indicate that models obtained by optimizing ICL achieve significantly better invariance to the extraneous variable for a fixed desired level of accuracy. In a variety of experimental settings, we show applicability of ICL for learning invariant representations for both continuous and discrete extraneous variables. The project page with code is available at *https://github.com/adityakumarakash/ICL*

1 Introduction

Removing or controlling for the influence of certain observed or unobserved extraneous variables, that may have an unintended effect on a learning task, is often a critical step in model estimation (Xie et al. 2017). Often, we want to explicitly control for their influence on the response variable, and estimate model coefficients that are, roughly speaking, immune to one or more confounding factors (Wasserman 2013). These tasks involve understanding invariance properties of data representations and/or parameters of the model we wish to learn. While mechanisms to control for extraneous variables are not strictly necessary in typical supervised learning tasks, where one focuses on predictive accuracy, over the last few years, many results have indicated how it can be quite useful (Lokhande et al. 2020). For instance, controlling the influence of a protected attribute such as race or gender on a response variable such as credit worthiness enables the design of fair machine learning models (Donini et al. 2018). Invariance is also relevant in domain adaptation when analyzing data from multiple sources or sites. Representations that are invariant to the categorical variable (e.g., which identifies the site) leads to models that are more immune (or less biased) to site-specific artifacts (Zhou et al. 2018). While invariance is prominent in a number of other settings (Baktashmotlagh et al. 2013; Moyer et al. 2019), we will focus on learning representations that are minimally informative of such extraneous variables yet preserve enough information to reliably predict the response/target variable or label.

Related works. Classical regression analysis techniques for handling extraneous variables based on residual scores and ANOVA (Girden 1992) are not easily applicable for deep neural networks. Instead, one approaches the question in one of two ways. A common approach is to use a standalone adversarial module (Xie et al. 2017) tasked with using the latent representations of the data to predict the extraneous variable whose influence (on the representations) we wish to remove. If the adversary succeeds, we have not yet fully controlled for the extraneous variable and so, the representations must be modified. This necessitates the design of an adversary tailored to the form of the downstream task. Further, the evaluation of sample complexity, convergence behavior of the training procedure, and the degree to which the representations remain invariant when the datasets are scaled or if an additional confound must be controlled for, require careful treatment and remain an active area of research (Jaiswal et al. 2019) (Jaiswal et al. 2018). An alternative strategy is to ask for statistical independence of the latent representations learned by the network and the extraneous variable. For example, one may approximately measure mutual information (Cover 1999) between the latent representations and the extraneous variable Moyer et al. (2018). This idea as well as the use of alternative distance and divergence measures is popular (Li, Swersky, and Zemel 2014; Louizos et al. 2016), and in most cases, perfectly models the innate requirements of the task. In practice, however, their viability depends on a variety of computational and implementation considerations, where design/approximation choices may frequently lead to representations where a modest adversary can successfully recover information about the extraneous variable fairly reliably.

Moving from theory to practice. To operationalize the statistical independence criterion described above, a sensible

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

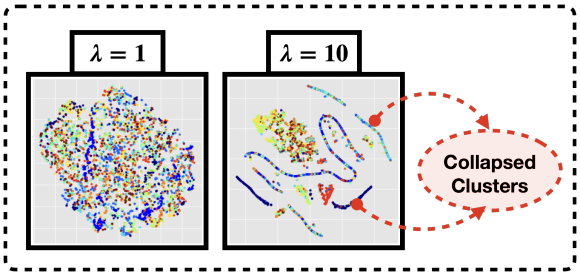

Figure 1: t-SNE plots for MNIST style experiment where the digit label is the extraneous variable c. For existing compression regularizers, increasing the regularization weight λ results in the collapse of latent space as indicated by the plot.

modeling choice is to use mutual information (Moyer et al. 2018) and then choose a good approximation. We describe this setting briefly to identify some practical issues that affect the overall behavior: instead of minimizing the mutual information $I(z, c)$, where z denotes the latent representation of the data x and c denotes the extraneous variable, we may minimize a suitable upper bound instead as shown below

$$I(z,c) \leq \underbrace{\mathbb{E}[KL[q(z|x)||q(z)]]}_{a)\ \text{Compression}} - \underbrace{\mathbb{E}[\log(p(x|z,c))]}_{b)\ \text{Reconstruction}} \quad (1)$$

The bound in (1) considers contributions from two terms: **(a)** one that *compresses* x into z via an encoder modeled using conditional likelihood $q(z|x)$ (whose marginal is $q(z)$), and **(b)** the second which *reconstructs* x from z and c via a decoder $p(x|z, c)$. Clearly, if c is available for free during decoding, there is no reason for the model not to aggressively compress x, while keeping just enough information content to reliably reconstruct it during decoding. When both terms function as intended, the balance will lead to representations that are invariant to c, as desired.

Let us temporarily set aside the reconstruction term and evaluate the compression term in (1) which ideally will remove from z the information regarding c. It is used as an invariance regularizer and controlled using a weight parameter λ as follows

$$\underbrace{\lambda}_{\text{Weight}} \underbrace{\mathbb{E}[KL[q(z|x)||q(z)]]}_{\text{Compression Regularizer}} \quad (2)$$

To minimize (2) in a computationally tractable way, one may model $q(z|x)$ as a Gaussian which allows (2) to be approximated using pairwise distances $KL(q(z|x_i), q(z|x_j))$, where x_i, x_j are different input samples. With this assumption, $KL(q(z|x_i), q(z|x_j))$ admits a closed form and roughly translates to the difference between the means of these two Gaussians scaled by the covariance (Wasserman 2013). For a reasonable weighting λ, we obtain some invariance to c in z but an adversary can still recover c from z. Increasing λ – to improve the invariance behavior – leads to the means of the conditionals coming closer to each other. Since the mean of the conditional $q(z|x)$ is used as the encoded representation of x, this also brings the representations closer together. Note that this compression of the means is agnostic of the extraneous variable c. In practice, this leads to a collapse of the latent space and formation of clusters when the strength of the regularizer is increased, making it easier for an adversary to recover c from z.

An example. We illustrate the above behavior experimentally using the setup of (Moyer et al. 2018) for unsupervised representation learning in MNIST in Figure 1. We wish to learn representations which are only informative of the style of digits but uninformative of the digit label. We gradually increase the strength of the compression term, via the weight parameter λ, and evaluate its effect. Since images of the same digits are similar to begin with, they map to representations which are in close proximity. This means that the latent space already has a rough grouping of representations based on the digits. A modest increase of the compression strength causes the inter group distances to decrease. This makes it more difficult to distinguish one group from another and can be seen as improving invariance – but not yet enough that an adversary cannot recover the digit label from the representation. However, when the regularizer is increased further, we observe that the latent representations start to form smaller clusters associated with the variable c or collapse completely – degrading invariance – in fact, making it easier for the adversary to identify the digit class. The foregoing behavior (see Figure 2) is not an artifact of approximation choices.

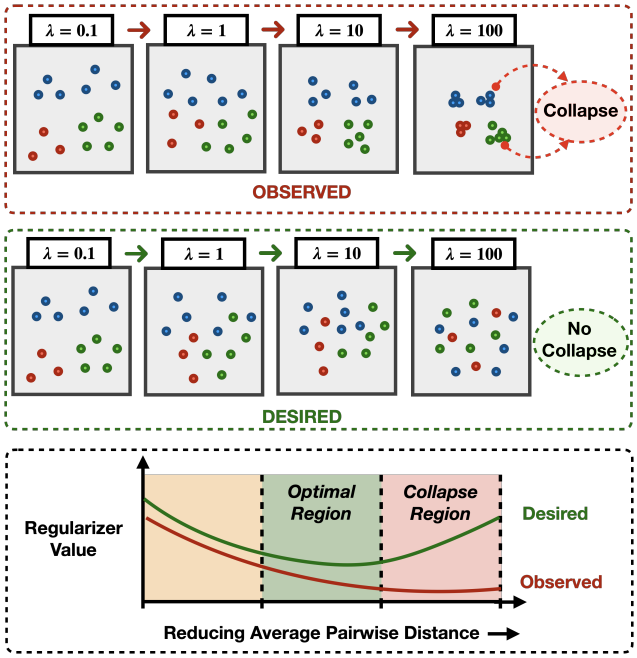

Figure 2: (Top): Representations generated by existing regularizers have some invariance for moderate weights λ, but form clusters for large λ values. (Center): The desired behavior is to spread intraclass samples and mix interclass samples giving rise to high invariance. The proposed ICL regularizer intuitively captures this notion. (Bottom): Existing compression regularizers are observed to let average distance between samples decrease and do not discourage cluster formation. A desired regularizer would assign high penalty in the collapse region and prevent clustering.

Consider a latent variable z and a binary variable $c \in \{0, 1\}$ we wish to control for. Here, we are concerned with the conditional distributions $p(z|c = 0)$ and $p(z|c = 1)$. Let us assume that we use a divergence $\mathcal{D}$, and statistical independence between z and c is imposed by a soft-version of the constraint $\mathcal{D}(p(z|c = 0), p(z|c = 1)) = 0$. For each value of c, if the latent space has clusters to begin with and one optimizes *both* $p(z|c = 0)$ and $p(z|c = 1)$ together, with no mechanism to spread/inter-mix the representations, the latent space may remain clustered with respect to c when we increase the weight of the invariance term. The above issue has less to do with how the distributional overlap is measured and can instead be attributed to not discouraging the formation of clusters. It seems that an explicit use of the extraneous variable during the encoding step may provide an effective workaround.

The basic intuition expressed above, is that x's that pertain to different values of c's should map to representations z's which are "mixed" yet contain enough information to keep the reconstruction error low. At the same time, representations for a specific value of c should be spread out, and not locally collapse to a point even when the weight parameter λ is increased.

The **main contributions** of this paper include (a) We propose *Inverse* Contrastive Loss (ICL) for learning invariant representations inspired from the class of contrastive losses (LeCun and Huang 2005). Our proposed loss is computationally efficient as it does not require specialized solvers or additional training through adversarial modules. (b) We interpret ICL by drawing a relation with the well studied Maximum Mean Discrepancy (MMD) as well as energy functionals used in dynamical systems analysis. (c) We demonstrate that ICL provides invariant representations for not only discrete extraneous variables but also continuous ones.

2 From Contrastive Models to Inverse Contrastive Representation Learners

We will now briefly review concepts from the recently proposed framework of *Contrastive Loss (CL)* functions. We will denote our input data using tuple of random variables $(x, x^-) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ where x^- is a *negative* sample, that is, if x can semantically be classified as y, then x^- is closer to a different class $y^- \neq y$. As usual, in unsupervised learning, y, y^- are not available during training. Let z (and similarly, z^-) denote the latent representation of x that may be obtained using a feature extraction scheme like ResNet, DenseNet or others (He et al. 2016). Finally, a CL function is defined by $\ell\left(z^T\left(z^+ - z^-\right)\right)$ where z^+ is the representation of a sample from the same class as x and ℓ can be any classification loss function such as hinge, softmax etc., see Definition 2.3 in (Saunshi et al. 2019). In essence, the definition of CL function captures the simple notion of **contrastiveness** that *semantically* similar points should have *geometrically* similar representations (Hadsell, Chopra, and LeCun 2006). To see this, assume that ℓ is the logistic loss, then it is easy to see that $\ell\left(z^T\left(z^+ - z^-\right)\right)$ is small for a high intraclass similarity $z^T z^+$ and a low interclass similarity $z^T z^-$. We say that a model is *contrastive* if it satisfies the contrastiveness property. We will now list some basic mathematical notations that we will use throughout the rest of the paper.

Basic Notations. For any pair of random variables (x_1, x_2), we will use $p(x_1, x_2), p(x_1|x_2)$ to denote the joint and conditional distribution respectively. $\delta(x)$ represents the dirac delta measure at $x \in \mathbb{R}^d$, and the indicator function $\mathbb{1}(\cdot)$ evaluates to 1 if the argument is true, and 0 otherwise. For a positive definite kernel $k(x, y)$, MMD divergence (Gretton et al. 2006) between distributions p, q is defined as,

$$\text{MMD}_k = \mathop{\mathbb{E}}_{\substack{x\sim p\\ x'\sim p}} k(x, x') + \mathop{\mathbb{E}}_{\substack{y\sim q\\ y'\sim q}} k(y, y') - 2 \mathop{\mathbb{E}}_{\substack{x\sim p\\ y\sim q}} k(x, y) \quad (3)$$

For $z, z' \in \mathbb{R}^d$, we will use $d(z, z')$ to be the Euclidean norm $\|z - z'\|_2$ unless otherwise stated, and $\mathcal{N}_\delta(z)$ denotes the Euclidean ball of radius δ centered at z. For a subset $X \subseteq \mathbb{R}^d$, we will use $\mathcal{P}(X)$ to denote the space of probability distributions over X.

2.1 How to Invert a CL Function to Learn Invariant Representations?

In this section, we will define our Inverse Contrastive Loss (ICL) that can be used to learn representations that are *invariant* to an extraneous (random) variable c. At a high level, our procedure consists of the following two steps:

1. **Formal Inversion (FI):** invert the contrastiveness property to reflect low intraclass and high interclass similarity by switching the role of $z^T z^-$ and $z^T z^+$ via sign flip;
2. **Addition of Weighted Neighborhood Kinks (AWNK):** apply an increasing function on interclass similarity $z^T z^-$ and a decreasing function on the intraclass similarity $z^T z^+$

While the two step procedure mentioned above implicitly defines an Inverse CL (ICL) function, note that it is well defined as long as the CL function is. Before we present a precise definition of ICL, it is meaningful to see why FI+AWNK can improve invariance to an extraneous variable.

Sufficiency of FI+AWNK. As discussed earlier in Section 1, for learning invariant representations, it is desirable that features with similar c be spread apart in the latent space while the features with dissimilar c be closer to each other. FI explicitly formalizes the idea that invariance should be better for high interclass similarity $z^T z^-$ and a low intraclass similarity $z^T z^+$. AWNK can be thought of as a disentanglement step that allows us to handle interclass and intraclass similarities appropriately. The interclass similarity $z^T z^-$ is expressed with a quadratic function similar to (Hadsell, Chopra, and LeCun 2006). For intraclass similarity $z^T z^+$, (Hadsell, Chopra, and LeCun 2006) suggests using a clipped quadratic function which is inefficient for gradient based methods because the gradient in the clipped region of the function is always zero. In contrast, we propose to use exponential loss which provides non-zero gradient values. While other alternative functions are applicable here, we will see shortly in Section 2.3 that the exponential loss provides a means to draw an interesting connection between ICL and well-studied and mature ideas like MMD divergence. To sum up, AWNK's role is to prevent the intraclass representations from locally collapsing **even** for a wide range of values of the regularization parameter.

2.2 ICL – A Probabilistic Definition

From now on, we will use distance/metric $d(z, z')$ to measure similarity – closer points are similar. Intuitively this can be expressed by saying that on average features which share similar values for extraneous variables have representations that are further from each other, while features that have dissimilar values for extraneous variables have closer representation. We operationalize this intuition by inverting the class of contrastive losses.

Definition 1. *Let $p(z, c)$ be the joint distribution for representation variable $z \in Z$ and extraneous variable $c \in \mathcal{C}$. Let $d_Z(z, z')$ be the distance metric on Z and $\mathcal{N}_\delta(c)$ denote the δ-neighbourhood centered at c. For $s(z, z') = d_Z^2(z, z')$ and $f(z, z') = \exp(\alpha - \beta d_Z(z, z'))$, $\beta > 0$, we define $ICL^\delta_{\alpha,\beta}(z, c) : \mathcal{P}(Z \times \mathcal{C}) \mapsto \mathbb{R}_+$ as*

$$ICL^\delta_{\alpha,\beta}(z, c) = \mathop{\mathbb{E}}_{\substack{(z,c)\sim p(z,c)\\(z',c')\sim p(z',c')}} \Big[\mathbb{1}(c' \in \mathcal{N}_\delta(c)) f(z, z') + \mathbb{1}(c' \notin \mathcal{N}_\delta(c)) s(z, z') \Big] \quad (4)$$

In Definition 1, $\mathcal{N}_\delta(c)$ encodes the similarity aspect of extraneous variables using its (underlying) geometry. A simple calculation shows that ICL functions in (4) immediately possess two (desirable) geometrical properties by definition: **(a)** whenever samples have similar extraneous value, our loss function is specified by $f(z, z')$ – a *decreasing* function of $d_Z(z, z')$; and **(b)** for samples with dissimilar extraneous value, the loss is specified by $s(z, z')$ – an *increasing* function of $d_Z(z, z')$. For the remainder of the paper, we will hide the AWNK parameters α, β and radius δ in ICL functions (4) whenever appropriate.

Remark 1. *It turns out that optimizing ICL is equivalent to driving a spring system to equilibrium in which samples with similar extraneous values are connected by a push spring while samples with dissimilar extraneous values are connected by a pull spring, see (Hadsell, Chopra, and LeCun 2006). In particular, the neighborhood radius δ in our ICL functions* (4) *determines the level of control exerted by these connections in the system – a large δ forces the latent representations to come closer while a smaller δ drives the representation to be a bit more spread.*

Handling Discrete Extraneous Variables using ICL. The following Lemma states that definition of ICL function in (4) is closely related to the standard MMD distance in (3).

Lemma 1 (ICL is equivalent to R-MMD)**.** *Assume that the extraneous variable c is binary with $p(c = 0) = 1/2$. Then there exist a conditionally positive definite kernel g and an interaction energy functional R_w (see equation 1.1 in (Carrillo, Lisini, and Mainini 2014)) such that the following equality holds:*

$$ICL(Z, C) = MMD_g(p_0, p_1) + R_w(p_0, p_1), \quad (5)$$

where p_0 and p_1 denote the conditional distributions $p(z|c = 0)$ and $p(z|c = 1)$ respectively.

The proof of Lemma 1 is included in the appendix. In essence, Lemma 1 states that if c is binary, then optimizing ICL is equivalent to optimizing a Regularized-MMD (R-MMD) divergence between conditional distributions $p(z|c = 0)$ and $p(z|c = 1)$. Recall from Section 1 that $\text{MMD}(p_0, p_1) = 0$ is a sufficient condition for statistical independence between z and c. Hence, for the special case considered here, we see that ICL ensures statistical independence constraint using R-MMD divergence. To see that MMD_g is a valid divergence, note that the kernel g is conditionally positive definite since it is a composition of a laplacian kernel and a euclidean distance matrix. Please see appendix for details on how to generalize Lemma 1 to multiclass setting, when c is (discrete) uniformly distributed.

In practice, we are often *only* given access to empirical samples of z and c. This becomes problematic for optimization purposes since we can only evaluate the divergences **approximately** – approximate zeroth order oracle. In the next section, we study the finite sample optimization properties of R-MMD (5) using control theoretic constructions.

2.3 Exploring the Landscape of ICL Functions using Spring Forces

The following observation establishes a link between the R_w term in (5) and *distributional interaction energy functionals* used in analyzing dynamical systems (Carrillo, Lisini, and Mainini 2014).

Observation 1 (Significance of R_w)**.** *The regularizer $R_w(p, q)$ is composed of pairwise energy functional $w(x, y) \sim f(x, y) + s(x, y)$ between particles of the system (Hadsell, Chopra, and LeCun 2006). Intuitively, when input distributions p and q are decision variables of an optimization problem, MMD_g admits a trivial solution, $p = q = \delta(0)$, that is, p and q collapse to a single point mass. However, this trivial solution is almost surely* suboptimal *for R_w (see Figure 3a.), thus decreasing the chances of such a collapse. Indeed, since R_w forces representations to stay apart even when the regularizer weight is arbitrarily increased, which suggests that R_w may be reasonable for learning invariant representations.*

Plugging in the definition of R_w (see appendix) in equation (5) and rearranging, we have that,

$$\text{ICL}(z, c) = \overbrace{\mathop{\mathbb{E}}_{\substack{x\sim p_0\\x'\sim p_0}} \frac{f(x,x')}{4} + \mathop{\mathbb{E}}_{\substack{y\sim p_1\\y'\sim p_1}} \frac{f(y,y')}{4}}^{\text{Repulsion}} + \overbrace{\mathop{\mathbb{E}}_{\substack{x\sim p_0\\y\sim p_1}} \frac{s(x,y)}{2}}^{\text{Attraction}} \quad (6)$$

Intuitively, (6) shows that ICL can be decomposed into two terms: 1. **Attraction** $s(\cdot, \cdot)$ between interclass particles; and 2. **Repulsion** $f(\cdot, \cdot)$ between intraclass particles. That is, ICL can be interpreted as modeling interclass and intraclass connection between particles (representations) using two types of springs f, s. Indeed, a similar decomposition is also possible for MMD by setting $s = -k, f = k$. For optimization purposes, our choice of f and s in R-MMD immediately yields **two** crucial benefits that is absent in MMD:

Benefit 1 – ICL is well suited for First Order Methods. By definition, gradient of spring energy with respect to the

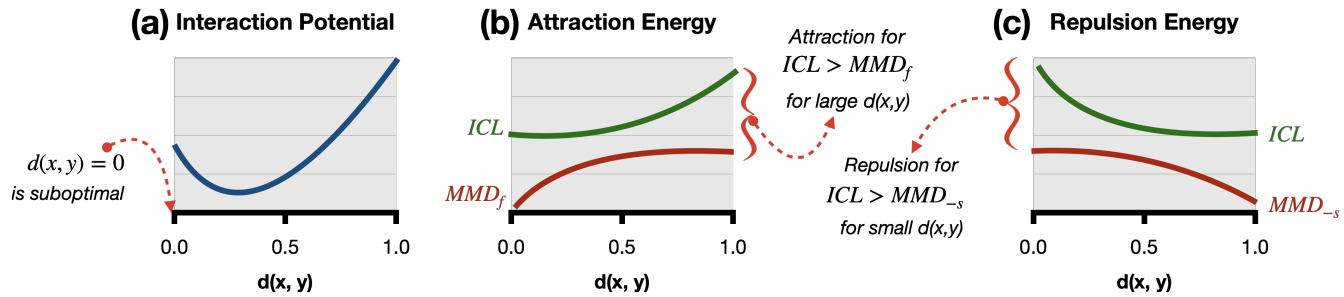

Figure 3: (a) We plot the interaction potential for the functional R_w. The functional R_w prevents the collapse of representation space by shifting the minima away from the trivial solution $d(x, y) = 0$. (b) We compare the attraction energy between ICL and MMD_f. The attraction for ICL is larger than for MMD_f when the particles are farther apart. (c) We plot the repulsion energy of ICL and MMD_{-s}. The repulsion for ICL is larger than MMD_{-s} when the particles are in close neighborhood.

distance $d(x, y)$ is the sum of attraction and repulsion connecting two particles. ICL and MMD_f differ in the attraction spring between interclass samples. When the distance between samples $d(x, y)$ is large, the attraction under ICL given by $\|\nabla_d s(x, y)\|_2$ is larger than the attraction under MMD_f given by $\|\nabla_d f(x, y)\|_2$ (Figure 3b). Furthermore attraction $\|\nabla_d s(x, y)\|_2$ increases with $d(x, y)$ while $\|\nabla_d f(x, y)\|_2$ deceases. Hence, while using first order methods like gradient descent, farther particles come closer *faster* while using ICL.

Benefit 2 – ICL prevents particles from collapsing. In the context of learning invariant representations, ICL and MMD_{-s} differ in repulse-only springs between intraclass samples. For ICL, the repulsive forces $\|\nabla_d f(x, y)\|_2$ between samples increases as the particles come close together while for MMD_{-s} the force $\|\nabla_d s(x, y)\|_2$ decreases (Figure 3c). Hence, whenever gradient based methods are used for training, ICL may be beneficial since the intraclass particles are pushed apart *strongly* when they are in the same neighborhood, as desired.

ICL Optimization provides adversarially invariant representations. It turns out that the above two benefits can be used to prove that models obtained by optimizing ICL derived loss can confuse adversaries. Formally, consider an adversary b that uses representation z to predict a continuous extraneous variable c. We will use the mean squared error (MSE) $\mathbb{E}_z[(b(z) - c)^2]$ to measure invariance, that is, a high value of MSE implies high invariance (desired). The following Lemma provides a lower bound on the MSE as a function of ICL under standard assumptions on b.

Lemma 2. *Assume that the extraneous variable c is continuous and b is L-lipschitz, and let $\rho = P_{c,c'}(|c - c'| > \delta)$. Then there exists α, and $\epsilon < \delta^2\rho^2/L^2$ such that for $\text{ICL}^{\delta}_{\alpha,\beta}(z, c) < \epsilon$, the MSE of adversary b is lower bounded i.e, $\mathbb{E}_z[(b(z) - c)^2] \geq (\delta\rho - L\sqrt{\epsilon})^2/4$.*

The proof of Lemma 2 is included in the appendix. Basically, Lemma 2 states that if ICL is made sufficiently small, then no Lipschitz adversary can have an arbitrarily small MSE as expected. We will now demonstrate the utility of Lemma 2 for analyzing datasets used in real world applications.

3 Applications of Inverse Contrastive Loss

Many representation learning schemes are built on Variational Auto-Encoder (VAE) based models (Kingma and Welling 2013). Recently (Cemgil et al. 2020) showed that one effective mechanism to improve adversarial robustness of representations obtained using VAE is via data augmentation: creating "fictive" data points. This can be thought of as providing invariance w.r.t. adversarial perturbations. However, obtaining such perturbations might not always be possible. While rotations, flips and crops work for natural images, this is problematic for brain imaging data where either a cropped brain or an image-flip that switches the asymmetrical relationship between the two hemispheres is meaningless. Applying a deformation to generate an augmented sample is defensible, but requires a great deal of care and user involvement. Similarly, deploying augmentation strategies for electronic health records (EHR) or audio data is not straightforward. Section 2 provides us the necessary guidance to explore the use of ICL regularizer for VAE based representation learners.

Setup. We use the setup based on Conditional VAE and Variational Information Bottleneck (VIB) (Alemi et al. 2017) for learning invariant representations in unsupervised and supervised setting respectively. These frameworks have been considered in the context of a mutual information based regularizer by (Moyer et al. 2018). Briefly, in the unsupervised setting one learns representations z using an encoder $q(z|x)$, that maps data x to conditional distribution $q(z|x)$, and a decoder $p(x|z, c)$ that reconstructs x from z and c. Gaussian reparameterization trick allows the encoder $q(z|x)$ to be written as $\mathcal{N}(\mu = h(x), \sigma(x))$, where $h(x)$ is the representation learner of interest. We augment this setup with ICL regularizer and propose optimizing the following objective,

$$\min_{p,q} \mathbb{E}_{x,c}\Big[\mathbb{E}_{z}[-\log p(x|z, c)] + \beta \text{KL}[q(z|x)||p(z)]\Big] + \lambda\ \text{ICL}(z, c) \tag{7}$$

where $p(z)$ is standard isotropic Gaussian prior.

For the supervised setting of predicting y from x we augment the VIB framework from (Alemi et al. 2017) with ICL regularizer and propose optimizing the following objective

$$\min_{p,q} -\mathbb{E}_{\substack{x\\c}}\Big[\mathbb{E}_{\substack{z\\y}} \log p(y|z) + \beta \mathbb{E}_{z} \log p(x|z, c)\Big] + \lambda\ \text{ICL} \tag{8}$$

where $p(y|z)$ is the learned prediction model.

Next, we show ICL's wide applicability by using it with discriminative encoders that are not based on VAE. Consider the task of predicting y from x in presence of extraneous c. To learn representations uninformative of c, the task is broken down into learning an encoder $h : x \mapsto z$ and a predicter $f : z \mapsto y$. We add the ICL regularizer to the loss objective ℓ and propose optimizing

$$\min_{f,h} \mathbb{E}_{x,y} [\ell(f(h(x), y))] + \lambda \text{ICL}(h(x), c) \tag{9}$$

h, f are generally parameterized using deep networks.

Baselines. As discussed in Section 1 invariance can be enforced using statistical independence or using adversarial modules. Our proposed ICL loss is compared with the following frameworks from both these categories: (a) Unregularized model, (b) MI regularizer (Moyer et al. 2018), (c) OT regularizer, where KL term in (b) is replaced with Wasserstein distance, (d) MMD_{-s} (Section 2.3), (e) MMD_f (Section 2.3), based on MMD (Li, Swersky, and Zemel 2014) and (f) CAI, Controllable invariance through adversarial feature learning (Xie et al. 2017), (g) UAI, Unsupervised Adversarial Invariance (Jaiswal et al. 2018).

Quantifying invariance. We follow (Xie et al. 2017) and train a three layered FC network as an adversary to predict the extraneous variable c from latent representations z. We report the accuracy of this adversary for discrete c and MSE for continuous c as the adversarial invariance measure (A).

We evaluate the frameworks in terms of task accuracy/reconstruction error and adversarial invariance on an unseen test set. The hyperparameter selection is done on a validation split such that best adversarial invariance is achieved for task accuracy within 5% of unregularized model for supervised tasks and reconstruction MSE within 5 points of unregularized model for unsupervised tasks. Mean and standard deviation are reported on ten runs, except when mentioned otherwise or quoting results from previous work. We use Adam optimizer for model training. More details on training and hyperparameters are provided in the appendix. Next, we present our results grouped by the nature of model (generative/discriminative) and the dataset.

3.1 Generative Model Families

First, we apply ICL to the family of generative models based on VAEs. Primarily we work with the setups (7) and (8).

Learning style information in MNIST Dataset. We consider the problem of learning representations that preserve only the style information of the digit (e.g., slant of digit, thickness of stroke etc.) while being invariant to the digit label. We use the VAE setup from (7). ***Results.*** Table 1 shows that ICL provides the best adversarial invariance amongst all the methods. Except for CAI, the invariance provided by other methods are significantly worse in comparison to the unregularized case. We reviewed this behavior in Section 1 and suspect that it is due to a high similarity between input examples of the same digit. We also show the effect of large regularizer weight to explain this behavior. t-SNE plots in Figure 4 show clusters and collapse of the latent space for KL and MMD_{-s}. In comparison, ICL has a uniform latent space, which partly explains why it provides better invariance.

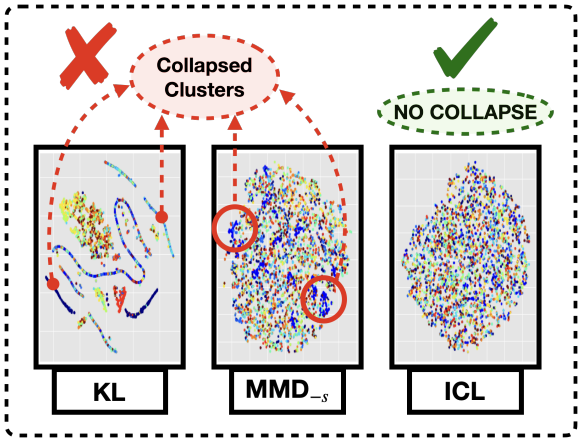

Figure 4: We plot t-SNE for latent representations of KL, MMD_{-s} and ICL for MNIST style experiment. Collapsed clusters are observed in the plots of KL, MMD_{-s}, whereas ICL generates a uniform latent space favoring invariance.

Learning invariant representation for Fairness Datasets. Next, we consider the problem of learning representations that are invariant to the extraneous variable which may be "protected" in fair classification models. The intuition is that such invariant features should help downstream fair algorithms that depend on these representations. We use the Adult and German datasets (Dua and Graff 2017) for this task. In Adult, the task is to predict if a person has over $\$50,000$ in savings, and the extraneous variable is Gender. In German, the task is to predict if a person has a good credit score and the extraneous variable is Age (binarized). We use the preprocessing from (Moyer et al. 2018), and follow the VIB (8) setup. ***Results.*** For Adult (Table 1), all methods show comparable prediction accuracy and ICL gives the best invariance. For German, ICL is amongst the methods with best adversarial invariance (Table 1) and provides best predictive accuracy. Accuracy higher than unregularized case suggests that removal of Age assists the downstream task.

3.2 Discriminative Model Families

Next we apply ICL to discriminative models (9) which are parameterized using a deep neural network such as ResNet18 (He et al. 2016). We seek to make representations at an internal layer of the network invariant, and so some VAE based baselines are not directly applicable.

Invariance w.r.t. continuous extraneous attribute for Adult Dataset. For Adult dataset, we evaluate ICL in the context of *age*, a continuous c variable. ***Results.*** In Table 2, we see that ICL provides a significantly better invariance in comparison to the baselines. Since continuous attributes are common in the fairness literature as well as in the context of applications in scientific disciplines, we believe this experiment shows the viability of ICL's use in this setting.

Rotation invariance for MNIST-ROT. This is a variant on MNIST dataset from (Jaiswal et al. 2018) where each digit is randomly rotated by an angle $\in \{0, \pm 22.5°, \pm 45°\}$. The task is to achieve invariance wrt rotation for predicting

R: Reconstruction Error, **P:** Prediction Accuracy, **A:** Adversarial Invariance Measure
↑: Higher Value is preferred, ↓: Lower Value is preferred

	MNIST		Adult		German		MNIST-ROT		ADNI	
	R ↓	A ↓	P ↑	A ↓	P ↑	A ↓	P ↑	A ↓	P ↑	A ↓
Unregularized	12.1 ± 0.5	46 ± 4	84 ± 0	84 ± 0	73 ± 2	78 ± 2	96 ± 0	42 ± 1	83 ± 3	55 ± 5
MI	13.2 ± 0.4	50 ± 3	84 ± 0	78 ± 2	70 ± 0	76 ± 3	96	38 ± 1	–	–
MMD_{-s}	15.8 ± 0.5	55 ± 5	84 ± 0	82 ± 0	73 ± 1	75 ± 2	96 ± 0	35 ± 2	85 ± 3	49 ± 3
MMD_f	15.8 ± 0.5	50 ± 5	83 ± 0	80 ± 0	74 ± 1	78 ± 2	96 ± 0	34 ± 1	86 ± 1	57 ± 6
OT	14.4 ± 0.4	61 ± 3	83 ± 0	78 ± 1	72 ± 2	75 ± 3	–	–	–	–
CAI	11.8 ± 0.3	48 ± 9	84 ± 0	81 ± 3	73 ± 1	75 ± 2	96	38	85 ± 2	51 ± 4
UAI	–	–	84 ± 0	83 ± 0	73 ± 2	75 ± 3	98	34	84 ± 3	49 ± 7
ICL (Ours)	16.6 ± 0.1	$\mathbf{32 \pm 0}$	83 ± 0	$\mathbf{75 \pm 2}$	75 ± 2	$\mathbf{75 \pm 2}$	96 ± 0	$\mathbf{33 \pm 1}$	84 ± 3	$\mathbf{46 \pm 7}$

Table 1: ICL achieves a better Adversarial Invariance Measure (A) relative to the baselines as indicated in bold. The Prediction Accuracy (P) / Reconstruction Error (R) for all the methods are comparable. We include the following baselines: (a) Unregularized setup, (b) MI (Moyer et al. 2018), (c) MMD_{-s}, (d) MMD_f, based on (Li, Swersky, and Zemel 2014), (e) OT based regularizer (f) CAI (Xie et al. 2017) (g) UAI (Jaiswal et al. 2018). The symbol $(-)$ indicates that the baseline was not applicable for the dataset.

the digit label. ***Results.*** In Table 1, we see that while being comparable in predictive accuracy, ICL provides the best adversarial invariance against rotation.

Predicting disease status while controlling for scanner confounds (ADNI dataset (adni.loni.usc.edu)). We finally show the effectiveness of ICL for predicting, using brain imaging data, whether an individual has Alzheimer's disease (AD) or is a healthy control subject (CN). Our pre-processed dataset consists of about 449 brain MRI scans of patients – of note here is that because the acquisitions are performed at different sites, the scanner manufacturers are different (e.g., Siemens, GE) (Giannelli et al. 2010). While the pulse sequences for the scans are standardized, because of differences in the magnetic coils and other factors, it is not realistic for the images to be completely harmonized. If a handful of coarse region of interest (ROI) summaries are obtained from the images via some pre-processing methods (such as Freesurfer), one may expect some immunity to scanner specific artifacts. But if the goal is to maximize performance using whole brain images, it becomes difficult to discourage an off-the-shelf CNN model from picking up scanner specific artifacts, especially if the demographics of the subjects are not perfectly matched across sites. Here, we use the imaging protocol (site/scanner) as the categorical variable we wish to control for. While more specialized models can be used if desired to further improve performance, we trained a simple ResNet-18 based model and use the output of the last hidden layer as the latent representation. The response variable was disease status: AD or CN. Since the dataset is small, the results are reported over five random training validation split. ***Results.*** We find that for this challenging setting, ICL gives the best adversarial invariance (Table 1) while also providing better predictive accuracy than the unregularized model.

Dataset: Adult with Age	P ↑	A_{MSE} ↑
Unregularized	83 ± 0	112 ± 1
CAI (Xie et al. 2017)	82 ± 2	129 ± 10
UAI (Jaiswal et al. 2018)	84 ± 0	114 ± 2
ICL (Ours)	83 ± 0	$\mathbf{161 \pm 15}$

Table 2: We study the continuous extraneous variable setting with the Adult dataset and *Age* as the extraneous attribute. We find that ICL attains a better Adversarial Invariance Measure (A_{MSE}↑) compared to the baselines applicable in this setting.

Discussion on ICL's use for downstream tasks: Our experiments show that ICL is effective in preventing an adversarial module from identifying the extraneous attribute from the latent representations. This would prevent the downstream models from using these extraneous features for prediction. These representations appear to be beneficial for use within fair algorithms. For some of our experiments, we observe that invariance leads to improved prediction accuracy of the downstream task. We also provide a real world application where invariant representations help in pooling data from multiple sites, relevant in scientific studies.

4 Conclusions

Whether for compliance with legislative policies that forbid preferential treatment (positive or negative) based on protected attributes or to derive some level of immunity to systematic variations when pooling data in a large observational study spanning participating institutions, it is clear that the need for invariant representations within a sub-class of problems in machine learning will continue to grow and be broadly adopted. The form of ICL described here exhibits a number of desirable properties and empirical behavior in scenarios/datasets that have been described in the literature. While contrastive losses are not new, recent results shed light on when one may be able to characterize their performance provably. As this literature continues to grow, at least some of the findings will translate to and help inform additional invariance properties afforded by ICL and its variants.

Acknowledgments

The authors are grateful to Eric Huang for help and suggestions. Research supported by NIH R01 AG062336, NSF CAREER RI#1252725, NSF CCF #1918211, NIH RF1 AG059312, NIH RF1 AG05986901 and UW CPCP (U54 AI117924). Sathya Ravi was also supported by UIC-ICR start-up funds. Correspondence should be directed to Ravi or Singh.

Broader Impact

The general idea of invariant representations is closely tied to ongoing research on fair algorithms. In that sense, ICL and other measures for invariance can enable the design of methods with a more desirable behavior, if the protected variables are appropriately controlled for. Such strategies can also facilitate pooling of data from multiple sites, and help answer important scientific questions that may not be possible to answer with small sized datasets. Controlling for undesirable observed variables will be an important consideration in a number of biomedical applications where deep learning models are getting increasingly adopted.

Appendix

Definition 2. *Let p, q be two distributions and $w(x, y)$ be the interaction energy potential. Then the distributional interaction energy functional between distributions p, q is defined as*

$$R_w(p,q) = \mathop{\mathbb{E}}_{\substack{z\sim p\\ z'\sim p}} w(z,z') + \mathop{\mathbb{E}}_{\substack{z\sim q\\ z'\sim q}} w(z,z') + 2\mathop{\mathbb{E}}_{\substack{z\sim p\\ z'\sim q}} w(z,z') \tag{10}$$

where the potential $w(x, y)$ is chosen suitably for different applications.

See equation 1.1 in (Carrillo, Lisini, and Mainini 2014)) for reference to interaction energy functional.

.1 Proof of Lemma 1

Proof. Recall the definition of $\mathrm{ICL}(z, c)$ from (4). For binary extraneous variable c, we have $\mathcal{N}_\delta(c) = \{c\}$. Using this to simplify $\mathbb{1}(c' \in \mathcal{N}_\delta(c)) = \mathbb{1}_{(c=c')}$ and $\mathbb{1}(c' \notin \mathcal{N}_\delta(c)) = \mathbb{1}_{(c\neq c')}$ and plugging in (4), we obtain

$$\mathrm{ICL}(z,c) = \mathop{\mathbb{E}}_{\substack{(z,c)\sim p(z,c)\\ (z',c')\sim p(z',c')}} \mathbb{1}_{(c=c')} f(z,z') + \mathbb{1}_{(c\neq c')} s(z,z') \tag{11}$$

Next we introduce the functions g and w used in the Lemma

$$g(z,z') = (f(z,z') - s(z,z'))/8 \tag{12}$$

$$w(z,z') = (f(z,z') + s(z,z'))/8 \tag{13}$$

Using (12) and (13) in (11) gives us

$$\frac{\mathrm{ICL}(z,c)}{4} = \mathop{\mathbb{E}}_{\substack{(z,c)\sim p(z,c)\\ (z',c')\sim p(z',c')}} \left[\left(\mathbb{1}_{(c=c')} - \mathbb{1}_{(c\neq c')}\right) g(z,z')\right] + \mathop{\mathbb{E}}_{\substack{(z,c)\sim p(z,c)\\ (z',c')\sim p(z',c')}} [w(z,z')] \tag{14}$$

Using law of total expectation we write (14) as

$$\frac{\mathrm{ICL}(z,c)}{4} = \mathop{\mathbb{E}}_{\substack{c\sim p(c)\\ c'\sim p(c)}} \left[\left(\mathbb{1}_{(c=c')} - \mathbb{1}_{(c\neq c')}\right) \mathop{\mathbb{E}}_{\substack{z\sim p(z|c)\\ z'\sim p(z|c')}} [g(z,z')]\right] + \mathop{\mathbb{E}}_{c,c'\sim p(c)} \left[\mathop{\mathbb{E}}_{\substack{z\sim p(z|c)\\ z'\sim p(z|c')}} [w(z,z')]\right]$$

Since $p(c = 0) = 1/2$, and using p_0 and p_1 to denote the conditional distributions $p(z|c = 0)$ and $p(z|c = 1)$ respectively, the expectation is expanded to get

$$\begin{aligned}\mathrm{ICL}(z,c) &= \underbrace{\mathop{\mathbb{E}}_{\substack{z\sim p_0\\ z'\sim p_0}} g(z,z') + \mathop{\mathbb{E}}_{\substack{z\sim p_1\\ z'\sim p_1}} g(z,z') - 2\mathop{\mathbb{E}}_{\substack{z\sim p_0\\ z'\sim p_1}} g(z,z')}_{\mathrm{MMD}_g}\\ &+ \underbrace{\mathop{\mathbb{E}}_{\substack{z\sim p_0\\ z'\sim p_0}} w(z,z') + \mathop{\mathbb{E}}_{\substack{z\sim p_1\\ z'\sim p_1}} w(z,z') + 2\mathop{\mathbb{E}}_{\substack{z\sim p_0\\ z'\sim p_1}} w(z,z')}_{R_w}\\ &= \mathrm{MMD}_g(p_0,p_1) + R_w(p_0,p_1)\end{aligned}$$

where $R_w(p, q)$ is defined in (10).

□

.2 Generalization of Lemma 1

We next show that Lemma 1 can be generalized to multi-class setting when c is (discrete) uniformly distributed.

Lemma 3. *Assume that the extraneous variable c is discrete with $c \in \{1, ..., m\}$ and is uniformly distributed, $p(c = i) = 1/m$. Then there exist a positive definite kernel g and an interaction energy functional R_w (see* (10)*) such that the following equality holds:*

$$ICL(z,c) = \sum_{\substack{i,j\in\{1,\dots,m\},\\ i<j}} MMD_g(p_i,p_j) + R_w(p_i,p_j),$$

where p_i denote the conditional distributions $p(z|c = i)$.

Proof. The proof proceeds on similar lines as the proof of Lemma 1. We introduce new functions g and w for the multi-class setting as

$$g(z,z') = \frac{1}{m^2}\left(\frac{f(z,z')}{m-1} - s(z,z')\right) \tag{15}$$

$$w(z,z') = \frac{1}{m^2}\left(\frac{f(z,z')}{m-1} + s(z,z')\right) \tag{16}$$

Using law of total expectation we write (11)

$$\begin{aligned}\mathrm{ICL}(z,c) &= \mathop{\mathbb{E}}_{c,c'\sim p(c)} \left[\mathbb{1}_{(c=c')} \mathop{\mathbb{E}}_{\substack{z\sim p(z|c)\\ z'\sim p(z|c')}} [f(z,z')] + \mathbb{1}_{(c\neq c')} \mathop{\mathbb{E}}_{\substack{z\sim p(z|c)\\ z'\sim p(z|c')}} [s(z,z')]\right]\\ &= \frac{1}{m^2} \sum_{i,j\in\{1,\dots,m\}} \mathbb{1}_{(i=j)} \mathop{\mathbb{E}}_{\substack{z\sim p_i\\ z'\sim p_j}} f(z,z') + \mathbb{1}_{(i\neq j)} \mathop{\mathbb{E}}_{\substack{z\sim p_i\\ z'\sim p_j}} s(z,z')\end{aligned}$$

Using (15), (16) in above and rearranging gives

$$\text{ICL}(z,c) = \sum_{\substack{i,j\in\{1,..,m\}\\ i<j}} \Bigg(\underbrace{\mathop{\mathbb{E}}_{\substack{z\sim p_i\\ z'\sim p_i}} g(z,z') + \mathop{\mathbb{E}}_{\substack{z\sim p_j\\ z'\sim p_j}} g(z,z') - 2\mathop{\mathbb{E}}_{\substack{z\sim p_i\\ z'\sim p_i}} g(z,z')}_{\text{MMD}_g} + \underbrace{\mathop{\mathbb{E}}_{\substack{z\sim p_i\\ z'\sim p_i}} w(z,z') + \mathop{\mathbb{E}}_{\substack{z\sim p_j\\ z'\sim p_j}} w(z,z') + 2\mathop{\mathbb{E}}_{\substack{z\sim p_i\\ z'\sim p_i}} w(z,z')}_{R_w} \Bigg)$$

$$= \sum_{\substack{i,j\in\{1,...,m\}\\ i<j}} \text{MMD}_g(p_i,p_j) + R_w(p_i,p_j)$$

□

.3 Proof of Lemma 2

Proof. Define $\mathcal{L}(f) := \mathbb{E}_{(z,c)}[(b(z)-c)^2]$ as the MSE of adversary b. Next we introduce imaginary samples $(z',c') \sim p(z,c)$ and have following

$$2\mathcal{L}(f) = 2\mathop{\mathbb{E}}_{(z,c)\sim p(z,c)}[(b(z)-c)^2] = \mathop{\mathbb{E}}_{\substack{(z,c)\sim p(z,c)\\ (z',c')\sim p(z,c)}}[(b(z)-c)^2 + (b(z')-c')^2]$$

We hide the subscript in expectation to simplify the notation.

$$\begin{aligned}
4\mathcal{L}(f) &= 2\mathbb{E}[(b(z)-c)^2 + (b(z')-c')^2] \\
&\geq \mathbb{E}\big[\big(b(z)-c-b(z')+c'\big)^2\big] \quad \{(a^2+b^2) \geq (a-b)^2/2\} \\
&= \mathbb{E}\Bigg[\Bigg((c'-c) - \big(b(z')-b(z)\big)\Bigg)^2\Bigg] \quad \{\text{Rearranging}\} \\
&\geq \mathbb{E}\big[\big(|c'-c| - |b(z')-b(z)|\big)^2\big]\{(a-b)^2 \geq (|a|-|b|)^2\} \\
&= \mathbb{E}\Bigg[\big(\mathbb{1}_{(c'\in\mathcal{N}_\delta(c))} + \mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}\big)\times \\
&\qquad \big(|c'-c| - |b(z')-b(z)|\big)^2\Bigg] \\
&\geq \mathbb{E}\big[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}\big(|c'-c| - |b(z')-b(z)|\big)^2\big] \\
&= \mathbb{E}\big[\big(\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|c'-c| - \mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|b(z')-b(z)|\big)^2\big] \\
&\geq \mathbb{E}\big[\big(\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|c'-c| - \mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|b(z')-b(z)|\big)^2\big] \\
&\{\text{Using } \mathbb{E}[X^2] \geq \mathbb{E}[X]^2, \mathbb{E}[X-Y] = \mathbb{E}[X] - \mathbb{E}[Y]\} \\
&\geq \mathbb{E}\Bigg[\Bigg(\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|c'-c| - \mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|b(z')-b(z)|\Bigg)\Bigg]^2 \\
&\geq \big(\underbrace{\mathbb{E}[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|c'-c|]}_{\text{I}} - \underbrace{\mathbb{E}[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|b(z')-b(z)|]}_{\text{II}}\big)^2
\end{aligned} \tag{17}$$

We lower bound I using the encoding $\mathcal{N}_\delta(c) = \{c' : |c-c'| \leq \delta\}$ as follows,

$$\begin{aligned}
\text{I} &= \mathbb{E}[\mathbb{1}(c' \notin \mathcal{N}_\delta(c))|c'-c|] \\
&\geq \delta\mathbb{E}[\mathbb{1}(c' \notin \mathcal{N}_\delta(c))] = \delta P_{c,c'}(|c-c'| > \delta) = \delta\rho
\end{aligned} \tag{18}$$

Next we use the fact that $\text{ICL}(z,c) < \epsilon$ to obtain following

$$\begin{aligned}
&\text{ICL}(z,c) < \epsilon \\
&\implies \mathbb{E}[\mathbb{1}_{(c'\in\mathcal{N}_\delta(c))}f(z,z') + \mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}s(z,z')] < \epsilon \\
&\implies \mathbb{E}[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}s(z,z')] < \epsilon \\
&\implies \mathbb{E}[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}d^2(z,z')] < \epsilon \;\; \{s(z,z') = d^2(z,z')\} \\
&\implies \mathbb{E}[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}d(z,z')] < \sqrt{\epsilon} \;\; \{\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}\}
\end{aligned} \tag{19}$$

Since b is L-lipschitz, $|b(z)-b(z')| \leq L\ d(z,z')$, which allows us to upper bound II as

$$\begin{aligned}
\text{II} &= \mathbb{E}[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}|b(z')-b(z)|] \\
&\leq L\ \mathbb{E}[\mathbb{1}_{(c'\notin\mathcal{N}_\delta(c))}d(z,z')] \quad \{\text{Using lipschitz definition}\} \\
&< L\sqrt{\epsilon} \quad \{\text{Using (19)}\}
\end{aligned} \tag{20}$$

We choose ϵ such that $\epsilon < \delta^2\rho^2/L^2$. Note that there exists a α such that $\text{ICL}^{\delta}_{\alpha,\beta}(z,c) < \epsilon$ for this choice. This allows us to use the lower and upper bounds of I and II respectively from (18) and (20) in (17) to give

$$\mathcal{L}(f) \geq (\delta\rho - L\sqrt{\epsilon})^2/4$$

□

.4 Detailed Setup for Applications

(a) Details on adversary. We follow (Xie et al. 2017) for training the adversary used for reporting invariance. We use a three-layered FC network with batch normalization and train it using Adam. For the MNIST-ROT experiment, we follow the setup of (Jaiswal et al. 2018).

(b) Evaluation methodology. We evaluate the frameworks in terms of task accuracy/reconstruction error and adversarial invariance on an unseen test set. The ADNI dataset is very small and hence for this dataset we use five fold random training validation splits to report the mean and standard deviation. For all other experiments, the mean and standard deviation are reported on an unseen test set for ten random runs, except when quoting results from previous works.

(c) Hyperparameter selection. The hyperparameter selection is done on a separate validation split such that on this set the model achieves the best adversarial invariance while the task accuracy remains within 5% of the unregularized model for supervised tasks and within 5 points of the unregularized model for unsupervised tasks. For the baselines, we grid search the best regularization weight in powers of ten and select the one with best invariance on validation set. For some of the experiments, we found it useful to initialize the regularization weight to a smaller value (0.01 times the regularizer weight) and multiplicatively update it (with factor 1.5) every epoch till it reaches the best found regularization weight. The same update rule is used for all the baselines.

(d) ICL parameters. For identifying ICL parameters α, β and δ, we perform simple grid search in powers of ten and its multiples of two and five. The δ hyperparameter is only relevant for the case of continuous extraneous attribute. For the continuous case, we normalize the extraneous variable to be in $[0,1]$ and search the δ parameter from multiples of 0.05.

References

Alemi, A.; Fischer, I.; Dillon, J.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *ICLR*. URL https://arxiv.org/abs/1612.00410.

Baktashmotlagh, M.; Harandi, M. T.; Lovell, B. C.; and Salzmann, M. 2013. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, 769–776.

Carrillo, J.; Lisini, S.; and Mainini, E. 2014. Gradient flows for non-smooth interaction potentials. *Nonlinear Analysis: Theory, Methods & Applications* 100: 122–147. ISSN 0362-546X. doi:https://doi.org/10.1016/j.na.2014.01.010. URL https://www.sciencedirect.com/science/article/pii/S0362546X14000236.

Cemgil, T.; Ghaisas, S.; Dvijotham, K. D.; and Kohli, P. 2020. Adversarially Robust Representations with Smooth Encoders. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=H1gfFaEYDS.

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

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2791–2801.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Giannelli, M.; Cosottini, M.; Michelassi, M. C.; Lazzarotti, G.; Belmonte, G.; Bartolozzi, C.; and Lazzeri, M. 2010. Dependence of brain DTI maps of fractional anisotropy and mean diffusivity on the number of diffusion weighting directions. *Journal of applied clinical medical physics* 11(1): 176–190.

Girden, E. R. 1992. *ANOVA: Repeated measures*. 84. Sage.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2006. A Kernel Method for the Two-Sample-Problem. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, 513–520. Cambridge, MA, USA: MIT Press.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 1735–1742.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*, 630–645. Springer.

Jaiswal, A.; Wu, R. Y.; Abd-Almageed, W.; and Natarajan, P. 2018. Unsupervised Adversarial Invariance. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 5092–5102. Curran Associates, Inc. URL http://papers.nips.cc/paper/7756-unsupervised-adversarial-invariance.pdf.

Jaiswal, A.; Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. Unified adversarial invariance. *arXiv preprint arXiv:1905.03629* .

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

LeCun, Y.; and Huang, F. J. 2005. Loss Functions for Discriminative Training of Energy-Based Models. In *AIStats*, volume 6, 34. Citeseer.

Li, Y.; Swersky, K.; and Zemel, R. 2014. Learning unbiased features. *arXiv preprint arXiv:1412.5244* .

Lokhande, V. S.; Akash, A. K.; Ravi, S. N.; and Singh, V. 2020. FairALM: Augmented Lagrangian Method for Training Fair Models with Little Regret. *arXiv preprint arXiv:2004.01355* .

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2016. The Variational Fair Autoencoder. *CoRR* abs/1511.00830.

Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; and Ver Steeg, G. 2018. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, 9084–9093.

Moyer, D.; Steeg, G. V.; Tax, C. M.; and Thompson, P. M. 2019. Scanner Invariant Representations for Diffusion MRI Harmonization. *arXiv preprint arXiv:1904.05375* .

Saunshi, N.; Plevrakis, O.; Arora, S.; Khodak, M.; and Khandeparkar, H. 2019. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *International Conference on Machine Learning*, 5628–5637.

Wasserman, L. 2013. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Xie, Q.; Dai, Z.; Du, Y.; Hovy, E.; and Neubig, G. 2017. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, 585–596.

Zhou, H. H.; Singh, V.; Johnson, S. C.; and Wahba, G. 2018. Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proceedings of the National Academy of Sciences* 115(7): 1481–1486. ISSN 0027-8424. doi:10.1073/pnas.1719747115. URL https://www.pnas.org/content/115/7/1481.