Learning Invariant Representations using Inverse Contrastive Loss (ICL)

Aditya Kumar Akash†, Vishnu Lokhande†, Sathya Ravi‡, Vikas Singh†

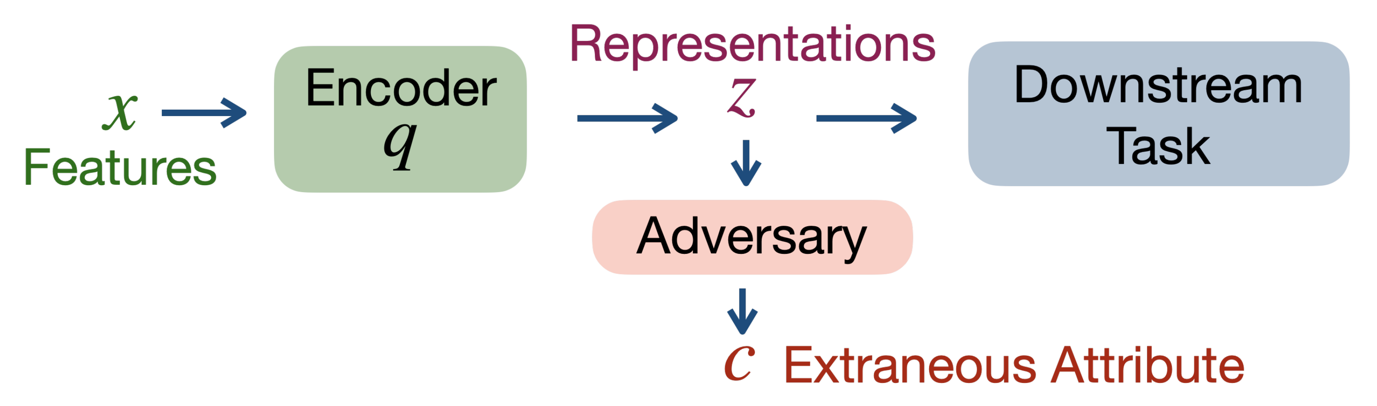
† University of Wisconsin Madison, USA ‡ University of Illinois at Chicago, USA

The AAAI Conference on Artificial Intelligence

Overview

- ✧ In this paper we propose **Inverse Contrastive Loss (ICL)**, a computationally efficient way to learn Invariant Representations
- ✧ We interpret ICL by drawing **relation** with well studied **Maximum Mean Discrepancy** (MMD)
- ✧ ICL provides **invariant representations** for not only **discrete** extraneous variables but also in the difficult case of **continuous** ones

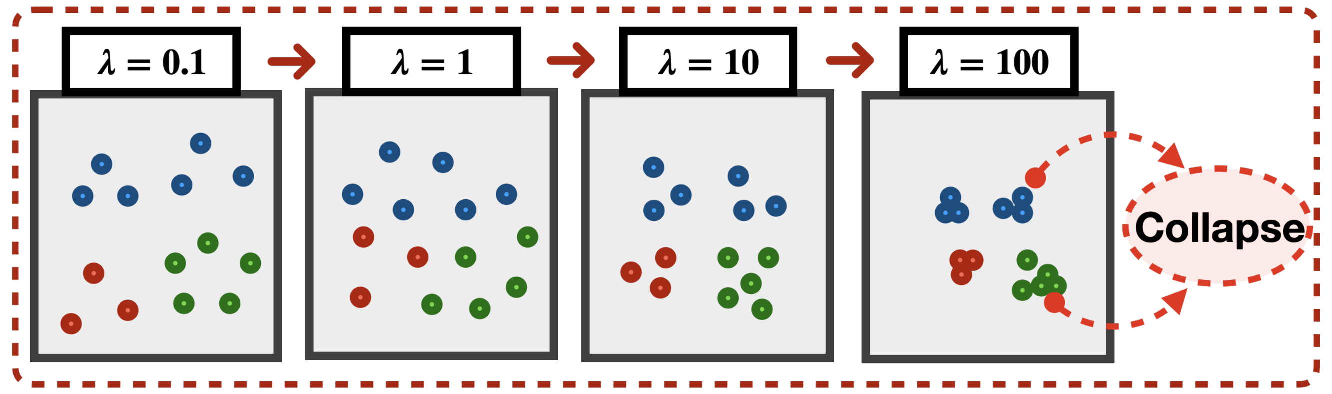
Basic Setting



Motivation

- Existing methods impose invariance by enforcing **statistical independence** $z \perp c$ through mutual information **approx** which **compresses** c from z
- **Increasing the compression** weight parameter λ to get more invariance **may lead to a collapsed** latent space

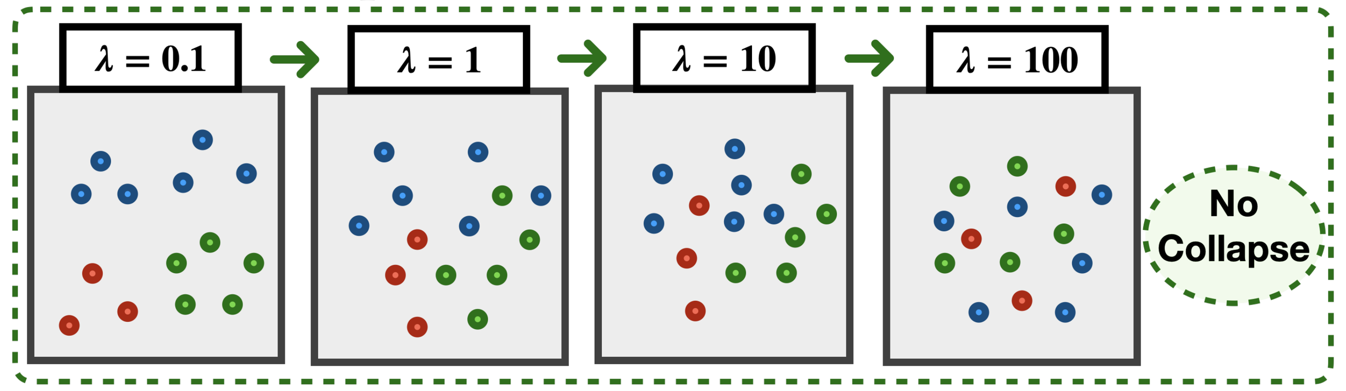
$$\underbrace{\lambda}_{\text{Weight}} \underbrace{\mathbb{E}[KL[q(z|x)||q(z)]]}_{\text{Compression Regularizer}}$$



- We propose ICL to explicitly use c in the compression term for preventing collapse

ICL and Benefits

- ✧ **Intuitively** we propose to
- **Intermix** representations for **different** c's
- **Spread out** representations for **same** c's



- ✧ **Contrastive losses** express high intraclass similarity $z^T z^+$ and low interclass similarity $z^T z^-$. **We invert this class** to impose invariance through ICL
- Switch roles of $z^T z^+$ and $z^T z^-$ by sign flip
- Apply increasing function on $z^T z^-$
- Apply decreasing function on $z^T z^+$

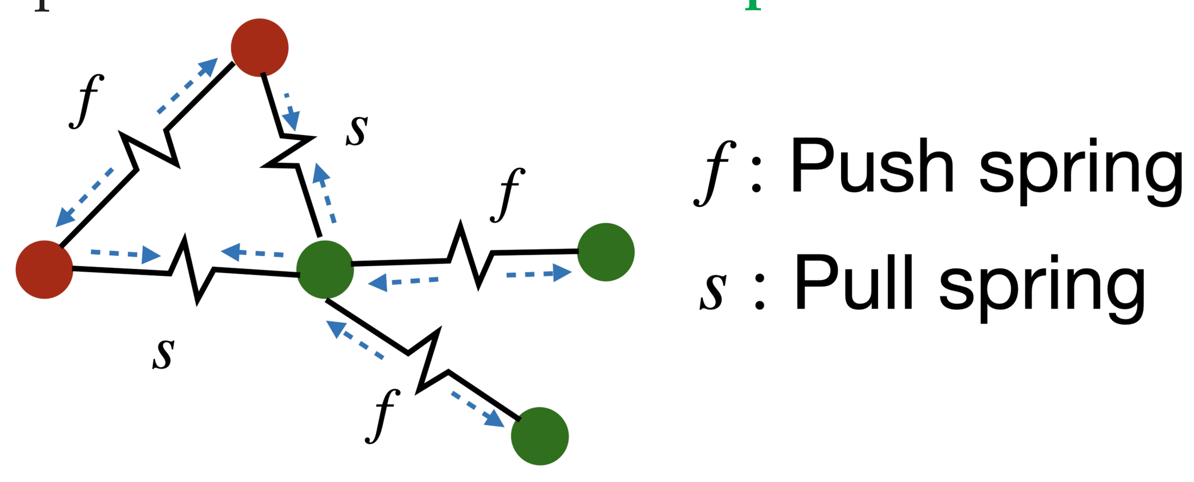
Similar C **Dissimilar C**

$$\text{ICL}(z,c) = \mathbb{E}_{\substack{(z,c)\sim p(z,c)\\(z',c')\sim p(z',c')}} \left[\mathbf{1}(c' \in \mathcal{N}_\delta(c))\, f(z,z') + \mathbf{1}(c' \notin \mathcal{N}_\delta(c))\, s(z,z')\right]$$

$f(z,z') = \exp(\alpha - \beta d_Z(z,z'))$
Decreasing loss, Spreads out

$s(z,z') = d_Z^2(z,z')$
Increasing loss, Intermixes

- ✧ **Optimizing ICL** is equivalent to driving **spring system to equilibrium** where
- Samples with **same** c's have **push** connection
- Samples with **different** c's have **pull** connection

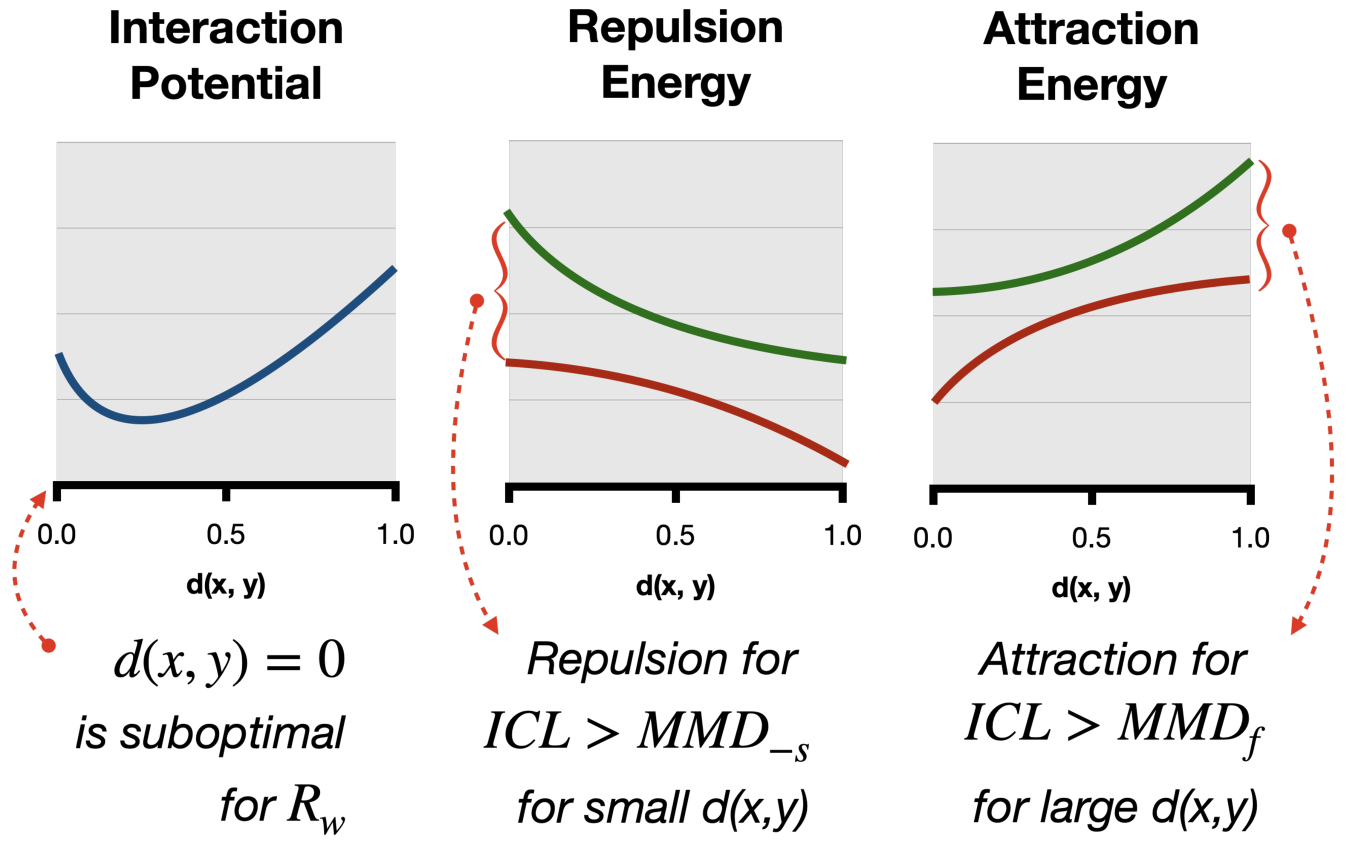


Optimizing ICL = System Equilibrium

- ✧ **Lemma 1** - For $c \in \{0,1\}$, $p(c=0)=1/2$ $\exists$ conditionally positive definite kernel g and an interaction energy functional R_w such that

$$\text{ICL}(Z,C) = \text{MMD}_g(p_0,p_1) + R_w(p_0,p_1)$$

- **Significance of R_w** - **Prevents collapse** of latent space since $d(x,y)=0$ is now suboptimal



- ✧ **Benefit 1** - **ICL is well suited for first order methods**
- When $d(x,y)$ is large, attraction for ICL is larger than MMD_f. Hence farther particles come closer faster.
- ✧ **Benefit 2** - **ICL prevents particles from collapsing**
- Repulsion for ICL is larger than MMD_{-s} when $d(x,y)$ is small. This pushes intraclass partices apart when they are close hence prevents collapse.
- ✧ **Lemma 2** - For c continuous, L-Lipschitz adversary b, $\rho = P_{c,c'}(|c-c'| > \delta)$, $\exists \alpha,\ \epsilon < \delta^2\rho^2/L^2$ such that for $\text{ICL}(z,c) < \epsilon$

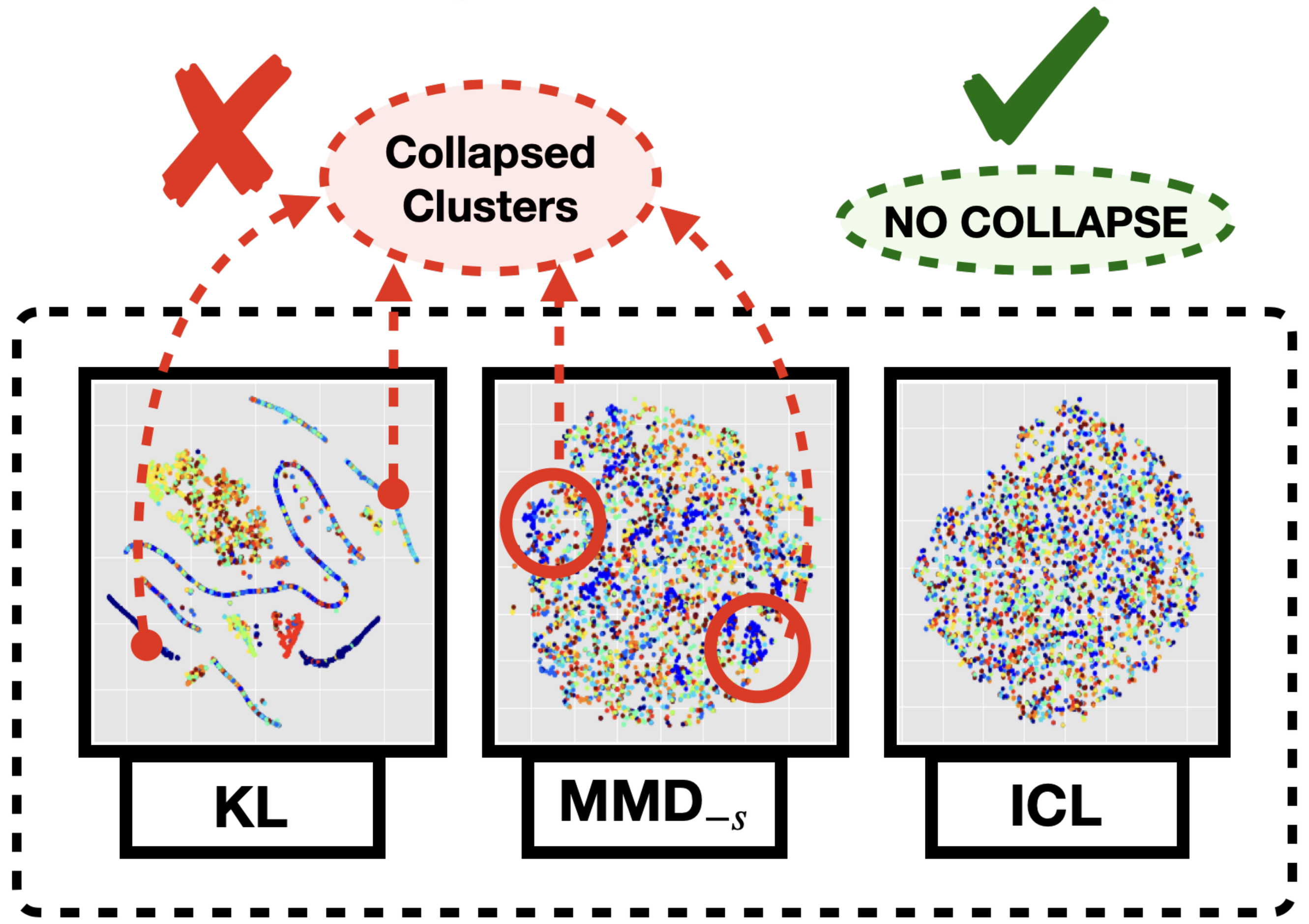
$$\mathbb{E}_z[(b(z)-c)^2] \geq (\delta\rho - L\sqrt{\epsilon})^2/4$$

- For sufficient small ICL, **no Lipschitz adversary can have an arbitrarily small MSE**

This research was supported in part by grant NIH R01 AG062336, NSF CAREER RI#1252725, NSF CCF #1918211.

Experimental Results

- ✧ We apply ICL in both **generative and discriminative** model settings such as -
- **Learning style information** in MNIST dataset
- Invariant representations for **Fairness datasets** (Adult and German)
- Invariance wrt **continuous extraneous attribute** for Adult dataset
- **Rotation Invariance** for MNIST-ROT
- **Predicting Alzheimer's disease** status while controlling for scanner confounds **(ADNI dataset)**
- ✧ t-SNE shows ICL produces uniform latent space



- ✧ ICL provides best invariance while providing good predictive accuracy

	MNIST		Adult		German		MNIST-Rot		ADNI	
	R↓	A↓	P↑	A↓	P↑	A↓	P↑	A↓	P↑	A↓
Unregularized	12.1	46	84	84	73	78	96	42	83	55
MI	13.2	50	84	78	70	76	96	38	-	-
MMD_{-s}	15.8	55	84	82	73	75	96	35	85	49
MMD_f	15.8	50	83	80	74	78	96	34	86	57
OT	14.4	61	83	78	72	75	-	-	-	-
CAI	11.8	48	84	81	73	75	96	38	85	51
UAI	-	-	84	83	73	75	98	34	84	49
ICL (Ours)	16.6	**32**	83	**75**	75	**75**	96	**33**	84	**46**

P: Prediction Accuracy, **R**: Reconstruction Error, **A**: Adversarial Invariance