Learning Uninformative Representations

Richard Zemel

Simons Workshop

Adversarial Approaches in Machine Learning

February, 2022

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Removing Information from Representations

Long-standing aim of representation learning: preserve information in data

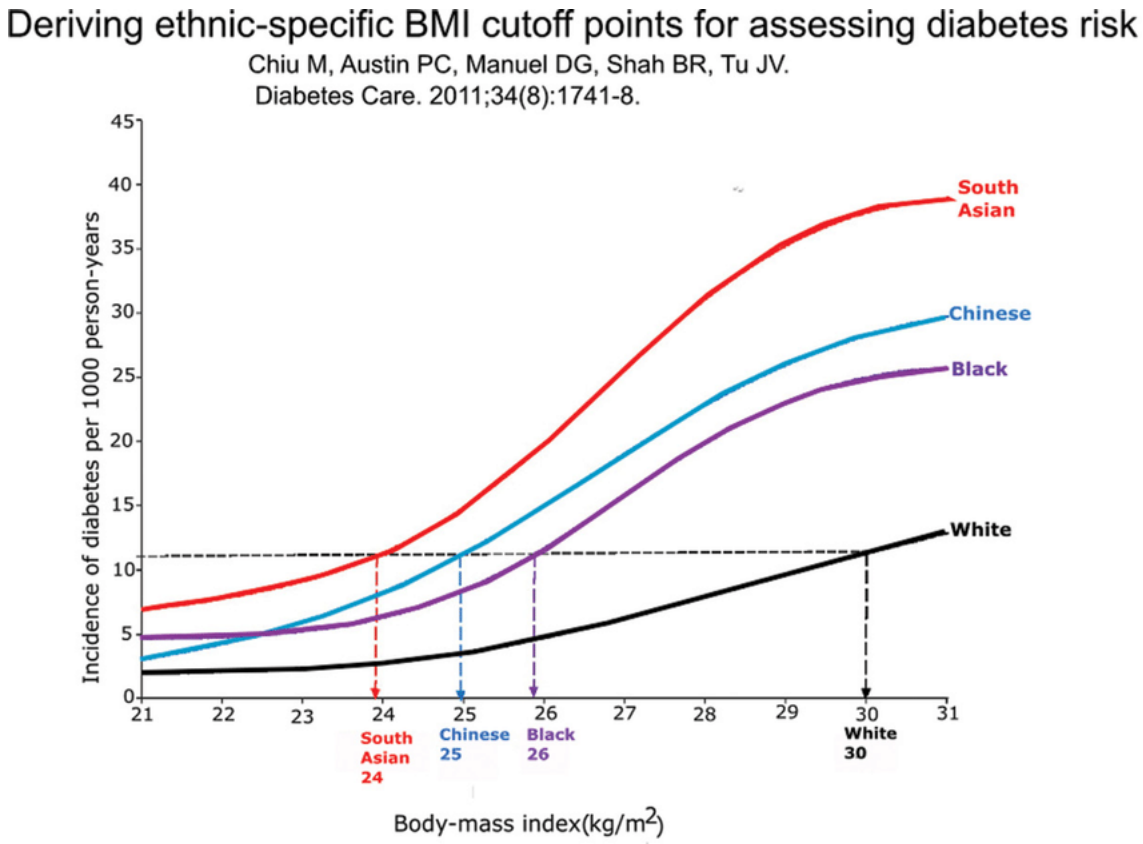
- **Infomax:** objective is to learn a representation *Z* of input *X* that maximizes the average Shannon mutual information between *Z* and *X* (Linsker 1988)
- related to the principle of redundancy reduction proposed for biological sensory processing by Barlow (1961)
- one application: Independent Components Analysis -- decompose input into non-Gaussian independent components (Comon, 1991; Bell & Sejnowski, 1997)

However, in many contexts, a key aim is instead to *remove* information about particular quantities

(1). Info Removal: Remove spurious features

Aim to learn predictor, factor out particular known spurious features

Example: diabetes predictors, uncovering factors beyond BMI



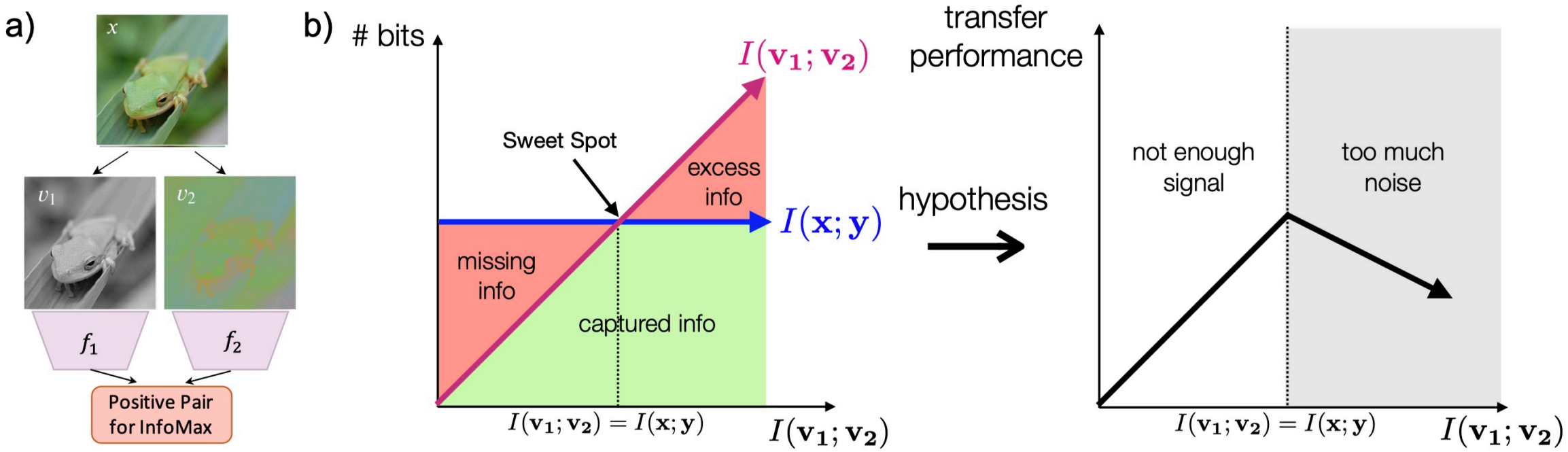
(2). Info Removal: Information Bottleneck

Form representation *T* that only retains information in input *X* that is needed to predict *Y*

$$\min_{p(t|x)} \; I(X;T) - \beta I(T;Y)$$

(Tishby, Pereira & Bialek, 1999)

(3). Info Removal: Self-Supervised Learning



InfoMin Principle: Form representation that retains information in multiple views of input relevant to target, loses all other information between views

maximize $I(\mathbf{v_1};\mathbf{y})$ and $I(\mathbf{v_2};\mathbf{y})$ minimize $I(\mathbf{v_1};\mathbf{v_2})$

(Tian, et al, 2020)

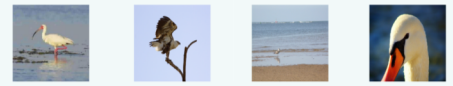
(4). Info Removal: Invariant learning

Training data: disjoint “domains”/”environments”

Assumes each example comes with side-information *c* indicating which environment data from

Environment-based loss:

$$\ell_c(f) = \frac{1}{n_c} \sum_{i=1}^{n} \ell(f(x_i), y_i) \mathbb{1}\{c_i = c\}$$

Dataset	Domains
Colored MNIST	+90% +80% -90%  *(degree of correlation between color and label)*
Rotated MNIST	0° 15° 30° 45° 60° 75° 
VLCS	Caltech101 LabelMe SUN09 VOC2007 
PACS	Art Cartoon Photo Sketch 
Office-Home	Art Clipart Product Photo 
Terra Incognita	L100 L38 L43 L46  *(camera trap location)*
DomainNet	Clipart Infographic Painting QuickDraw Photo Sketch 

(Gulrajani and Lopez-Paz 2020)

Invariant Learning

Invariant Learning: A form of domain generalization in which we generalize from training to test environments by **learning and predicting from invariant features** learned from environments seen in training

Aim of invariant learning: Discover features that reliably predict the class label regardless of the environment -- loses information about the environment

Invariant Risk Minimization: For input X^e and labels Y^e, find a transform Φ of input space such that, $P(Y^e|\Phi(X^e))$ is the same for all environments e (Arjovsky et al, 2019)

Illustration of Invariant Learning

Colored MNIST

Digits with misleading colors

	Y=0	Y=1
{0,1,2,3,4}	0.75	0.25
{5,6,7,8,9}	0.25	0.75

The optimal classification rate on the basis of the shape only is 75%.

Random guess is 50%.

	Red	Green
Y=0	e	$1-e$
Y=1	$1-e$	e

During the training $e \in \{0.8, 0.9\}$. The color is a better indicator than the shape, but not a stable one.

Then we test with $e = 0.1$.

Training with $e \in \{0.8, 0.9\}$	Testing with $e \in \{0.8, 0.9\}$	Testing with $e = 0.1$
Minimize empirical risk after mixing data from both environments	84.3%	10.1%
Minimize empirical risk with invariant regularization	70.0%	70.0%

- Network is a MLP with 256 hidden units on 14x14 images.
- Invariant regularization tuned high : regularization term is nearly zero.

(5). Info Removal: Fair representation learning

Fair classification is the most common setup, involving:

- X, some data
- Y, a label to predict
- $\hat{Y}$, the model prediction
- A, a **sensitive attribute** (race, gender, age, socio-economic status)

We want to learn a classifier which is:

1. accurate
2. fair with respect to A

- Fair classification: learn $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$
  - encoder f, classifier g
- Fair representation: learn $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$
- $Z = f(X)$ should:
  - Maintain **useful information** in X
  - **Yield fair downstream classification** for vendors g

(Zemel et al, 2014)

Info Removal Methods: Distribution Matching

Match moments of distributions:

- consider distance between empirical statistics of the distributions:

$$\left\| \frac{1}{N_0} \sum_{i=1}^{N_0} \psi(\mathbf{x}_i) - \frac{1}{N_1} \sum_{i=1}^{N_1} \psi(\mathbf{x}'_i) \right\|^2$$

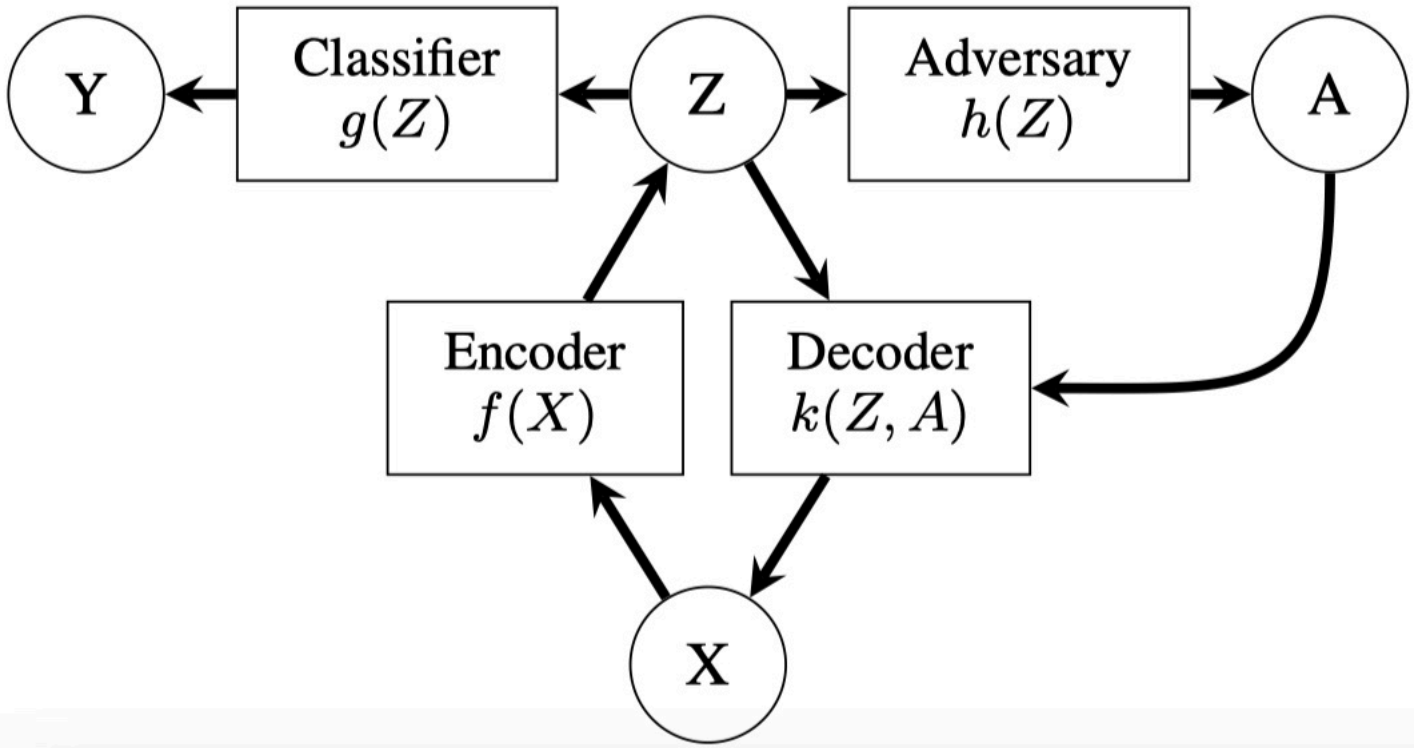
- estimate via kernel trick: Maximum Mean Discrepancy (Gretton, 2006)

$$\ell_{\mathrm{MMD}}(\mathbf{X}, \mathbf{X}') = \frac{1}{N_0^2} \sum_{n=1}^{N_0} \sum_{m=1}^{N_0} k(\mathbf{x}_n, \mathbf{x}_m) + \frac{1}{N_1^2} \sum_{n=1}^{N_1} \sum_{m=1}^{N_1} k(\mathbf{x}'_n, \mathbf{x}'_m) - \frac{2}{N_0 N_1} \sum_{n=1}^{N_0} \sum_{m=1}^{N_1} k(\mathbf{x}_n, \mathbf{x}'_m).$$

- formulate as regularizer in VAE (Louizos et al, 2015)

$$\ell_{\mathrm{MMD}}(\mathbf{Z}_{1\mathbf{s}=0}, \mathbf{Z}_{1\mathbf{s}=1}) = \| \mathbb{E}_{\tilde{p}(\mathbf{x}|\mathbf{s}=0)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=0)}[\psi(\mathbf{z}_1)]] - E_{\tilde{p}(\mathbf{x}|\mathbf{s}=1)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=1)}[\psi(\mathbf{z}_1)]] \|^2$$

Info Removal Methods: Adversarial



$$\mathcal{D}_i^j = \{(x, y, a) \in \mathcal{D} | a = i, y = j$$

$$L_{Adv}^{DP}(h) = 1 - \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i|} \sum_{(x,a) \in \mathcal{D}_i} |h(f(x, a)) - a|$$

$$L_{Adv}^{EO}(h) = 2 - \sum_{(i,j) \in \{0,1\}^2} \frac{1}{|\mathcal{D}_i^j|} \sum_{(x,a) \in \mathcal{D}_i^j} |h(f(x, a)) - a|$$

Invariant Learning with Unknown Environments

What if environment labels are not known?

- environment labels may not be known for all applications
- may be suboptimal when are known

Subgroup fairness without demographic labels: Multicalibration [Kim et al 2018], Fairness Gerrymandering [Kearns et al 2018]

Fairness for Unobserved Characteristics:
Insights from Technological Impacts on Queer Communities

Nenad Tomasev
nenadt@deepmind.com
DeepMind
London, United Kingdom

Kevin R. McKee
kevinrmckee@deepmind.com
DeepMind
London, United Kingdom

Jackie Kay*
kayj@deepmind.com
DeepMind
London, United Kingdom

Shakir Mohamed
shakir@deepmind.com
DeepMind
London, United Kingdom

How to define environments that will help identify those features?

How to identify what information we want to be removed from the representation?

Environment Inference for Invariant Learning

ICML 2021

arXiv:2010.07249

Elliot Creager

Joern Jacobsen

Notation & Definitions

input space $\mathcal{X}$, set of environments (a.k.a. "domains") $\mathcal{E}$, target space $\mathcal{Y}$, representation space $\mathcal{H}$

observational data $x, y, e \sim p^{obs}(x, y, e)$ with $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $e \in \mathcal{E}$, loss $\ell : \mathcal{H} \times \mathcal{Y} \to \mathbb{R}$

Predictor $w \circ \Phi$ comprises linear classifier $w : \mathcal{H} \to \mathcal{Y}$

applied to representation extractor ("model") $\Phi : \mathcal{X} \to \mathcal{H}$

$$C^{ERM}(\Phi) = \mathbb{E}_{p^{obs}(x,y,e)}[\ell(\Phi(x), y)]$$

- Domain Generalization: low error rates on samples $p(x, y|e_{test})$ from unseen $e_{test} \notin \mathcal{E}^{obs}$
- Domain Adaptation: model parameters can be adapted at test time using unlabelled samples

Invariant Learning ⟺ Fairness

Many parallels between invariant learning and algorithmic fairness

Consider the sensitive attribute in fairness analogous to environment indicator *e*

- IRM aims to minimize Environment Invariance Constraint:

$$\mathbb{E}[y|\Phi(x) = h, e_1] = \mathbb{E}[y|\Phi(x) = h, e_2]$$
$$\forall\, h \in \mathcal{H} \quad \forall\, e_1, e_2 \in \mathcal{E}^{obs}.$$

- Group-sufficiency [Chouldechova et al, 2017; Liu et al, 2018]:

$$\textbf{match}\ \mathbb{E}[y|S(x), e]\ \forall\, e$$

Environment Inference for Invariant Learning

Hypothesis: Learning systems tend to find shortcuts (Geirhos et al, 2020)

environments defined based on shortcuts → invariant learning will focus on other features

Example: shortcut classifier relies on color in Color-MNIST
assign E1=red; E2=green →
color features are not invariant across domains

Idea: Find “worst case” environments

Environment Inference for Invariant Learning

Recall the aim is to satisfy the Environment Invariance Constraint (EIC):

$$\mathbb{E}[y|\Phi(x) = h, e_1] = \mathbb{E}[y|\Phi(x) = h, e_2]$$
$$\forall\, h \in \mathcal{H} \quad \forall\, e_1, e_2 \in \mathcal{E}^{obs}.$$

Per-environment risk:

$$R^e = \mathbb{E}_{p^{obs}(x,y|e)}[\ell]$$

IRM regularizes ERM with a differentiable proxy to EIC:

$$C^{IRM}(\Phi) = \sum_{e \in \mathcal{E}^{obs}} R^e(\Phi) + \lambda ||\nabla_{\bar{w}} R^e(\bar{w} \circ \Phi)||$$

Worst case environment found by maximizing EIC, based on proxy regularizer

Summary of EIIL

- Requirement of hand-crafted environments replaced with reference model
- Reference model can be learned directly from observational data: maps X to Y, defines putative invariant features
- Experiments: environment assignment per example a Bernoulli probability **q**

1. Input *reference model* $\tilde{\Phi}$
2. Fix $\Phi \leftarrow \tilde{\Phi}$ and fully optimize the inner loop of (EIIL) to infer environments $\tilde{\mathbf{q}}_i(e) = \tilde{q}(e|x_i, y_i)$
3. Fix $\mathbf{q} \leftarrow \tilde{\mathbf{q}}$ and fully optimize the outer loop to yield the new model Φ.

EIIL Results: Color-MNIST

Method	Handcrafted Environments	Train	Test
ERM	✗	86.3 ± 0.1	13.8 ± 0.6
IRM	✓	71.1 ± 0.8	65.5 ± 2.3
EIIL	✗	73.7 ± 0.5	$\mathbf{68.4} \pm \mathbf{2.7}$

EIIL Results: Waterbirds

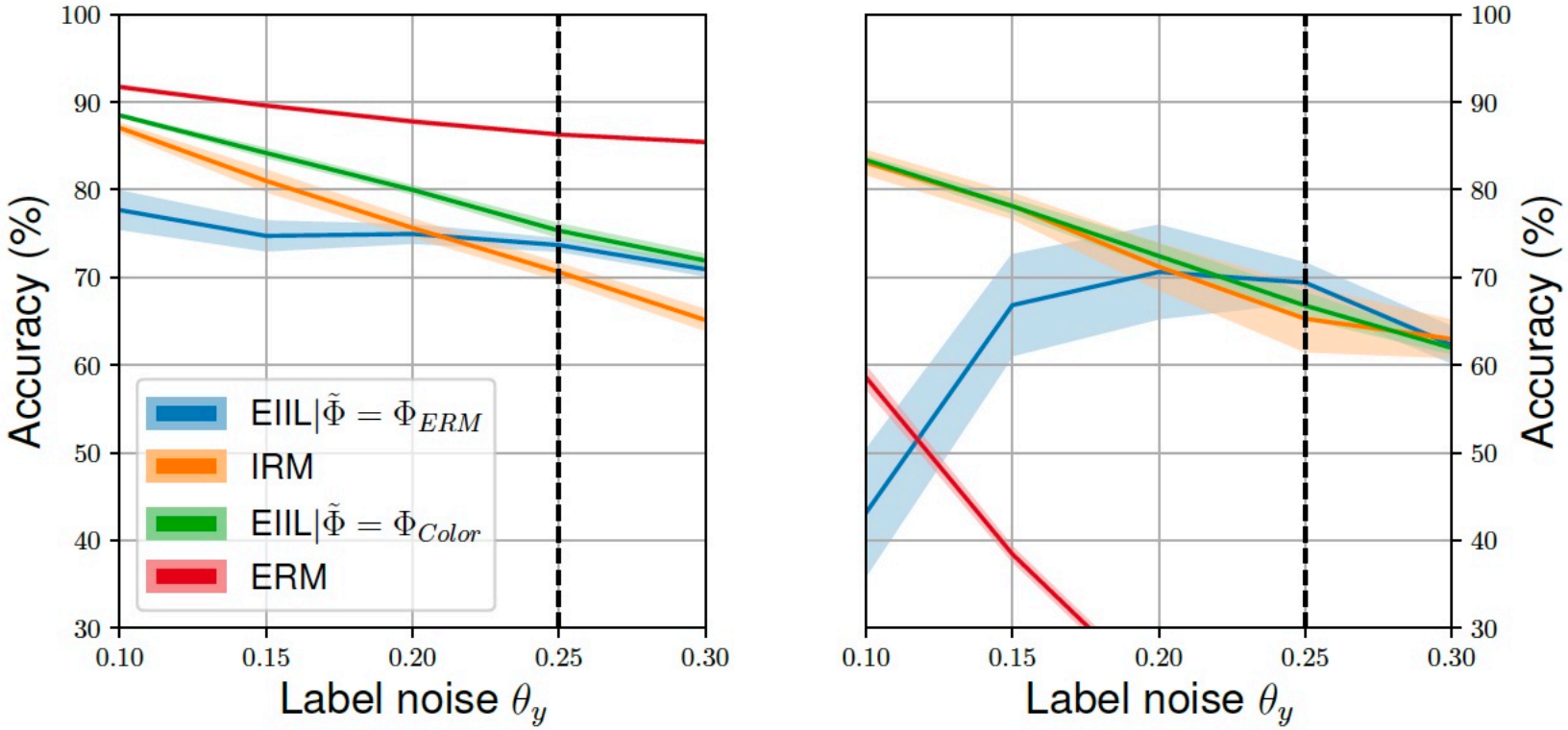
ERM: poor worst-group performance

GDRO (oracle) can mitigate [Sagawa et al, 2020]

EIIL: infer environments, then optimize GDRO based on those

Method	Train (avg)	Test (avg)	Test (worst group)
ERM	100.0	**97.3**	60.3
EIIL	99.6	96.9	**78.7**
GDRO (oracle)	99.1	96.6	84.6

EIIL: Dependence on reference model



(a) Train accuracy.

(b) Test accuracy

Discussion & Open Questions

EIIL

- Important aim is to discover environments (sensitive groups)
- Challenging -- dependent on reference model -- can we infer target new environments
- What kinds of distribution shift, out-of-context generalization are feasible, relevant?

Current methods for removing information from representations are insufficient

- distribution matching does not scale to high dimensional, continuous representation space
- adversarial methods present computational challenges
- differentiable proxies have unclear relationship to desired invariance properties