Optimal Representations for Covariate Shift

Yangjun Ruan

Joint with Yann Dubois, Chris J. Maddison

ICLR 2022
Overview
ML experiences **distribution shifts** from train (source) to test (target)
ML experiences **distribution shifts** from train (source) to test (target)

**Goal:** learn robust representations $Z$ of data $X$ from which source ($d_s$) predictors perform well on target ($d_t$)
ML experiences **distribution shifts** from train (source) to test (target)

**Goal:** learn robust representations $Z$ of data $X$ from which source ($d_s$) predictors perform well on target ($d_t$)

**Optimal $Z^*$:** all source optimal predictors **minimize** target risk
We characterize the optimally robust $Z^*$ to covariate shift

😊 prove **sufficient and necessary** condition for optimal $Z^*$
We characterize the optimally robust $Z^*$ to covariate shift

😊 prove **sufficient and necessary** condition for optimal $Z^*$

😊 derive practical **self-supervised** objectives for learning $Z^*$
We characterize the optimally robust $Z^*$ to covariate shift

😊 prove **sufficient and necessary** condition for optimal $Z^*$

😊 derive practical **self-supervised** objectives for learning $Z^*$

😊 show why CLIP [4] is more robust over other SSL methods
We characterize the optimally robust $Z^*$ to covariate shift

- prove sufficient and necessary condition for optimal $Z^*$
- derive practical self-supervised objectives for learning $Z^*$
- show why CLIP [4] is more robust over other SSL methods
- improve CLIP’s robustness with our objectives
Theory: Characterizing $\mathbb{Z}^*$
**Desiderata:** reduce to typical ML setup in $Z$ space
**Desiderata:** reduce to typical ML setup in $Z$ space

- **Sufficient condition** (...most previous work hinted towards)
Desiderata: reduce to typical ML setup in $Z$ space

- ✓ Sufficient condition (...most previous work hinted towards)
- ❌ Necessary? Achievable?
Intuition

**Minimal sufficiency:** $Z^*$ should

- remain discriminative about $Y$
- have invariant support

![Discriminative & Support Diagram](image.png)
Intuition

Minimal sufficiency: $Z^*$ should

- remain **discriminative** about $Y$
- have **invariant support**

![Diagram showing discriminative and support](image_url)
Minimal sufficiency: $Z^*$ should

- remain discriminative about $Y$
- have invariant support
Formalization with domain generalization (DG) language:

1. Given
   - A set of domains $\mathcal{D}$
   - Domain-specific $\{p_{X,Y|d}\}_{d \in \mathcal{D}}$
   - Loss $\ell : \mathcal{Y} \times \Gamma \rightarrow \mathbb{R}_{\geq 0}$

[Asm: discrete finite]
[Asm: gen. covariate shift]
Formalization with domain generalization (DG) language:

1. Given
   - A set of domains \( \mathcal{D} \)  
   - Domain-specific \( \{ p_{X, Y} \mid d \} \) \( d \in \mathcal{D} \)  
   - Loss \( \ell : \mathcal{Y} \times \Gamma \rightarrow \mathbb{R}_{\geq 0} \)  

2. Learn an encoder \( p_{Z \mid X} \)
Formalization with domain generalization (DG) language:

1. Given
   - A set of domains $\mathcal{D}$ [Asm: discrete finite]
   - Domain-specific $\{p_{X,Y} \mid d\}_{d \in \mathcal{D}}$ [Asm: gen. covariate shift]
   - Loss $\ell : \mathcal{Y} \times \Gamma \rightarrow \mathbb{R}_{\geq 0}$

2. Learn an encoder $p_Z \mid X$

3. Measure DG risk:
   - Select a random source $D_s$ and target $D_t$
   - Train a source predictor: $h \in \mathcal{H}_{D_s}^* := \arg \min_h R_{h}^{D_s}[Y \mid Z]$
   - Measure target risk $R_{h}^{D_t}[Y \mid Z]$
     where $R_{h}^{d}[Y \mid Z] := \mathbb{E}_{p_{Z,Y} \mid d}[\ell(Y, h(Z))]$
Problem Setup

Goal: minimize the idealized domain generalization (IDG) risk w.r.t. $Z$

$$R_{\text{IDG}}[Y|Z] := \mathbb{E}_{P_{D_s,D_t}} \sup_{h \in \mathcal{H}_{D_s}^*} R_h^{D_t}[Y|Z]$$

Uniform guarantees:

- random domains
- worst-case source predictor
**Problem Setup**

**Goal:** minimize the *idealized domain generalization* (IDG) risk w.r.t. $Z$

$$R_{IDG} [Y | Z] := \mathbb{E}_{P_{D_s, D_t}} \sup_{h \in \mathcal{H}^*_D} \mathbb{R}^D_{h} [Y|Z]$$

**Uniform guarantees:**

- random domains
- **worst-case** source predictor

**Idealized setup** for simplicity:

- population risk used for source predictor selection
- universal hypothesis class
Characterization of $Z^*$

Theorem (Optimality conditions, informal)

Under generalized covariate shift and some mild assumptions, $Z^*$ is optimal for IDG if and only if it

- remains discriminative: $\mathbb{R}[Y \mid Z^*] = \mathbb{R}[Y \mid X]$
- has invariant support: $\text{supp}(p_{Z^*} \mid d_s) = \text{supp}(p_{Z^*} \mid d_t)$, $\forall d_s, d_t \in \mathcal{D}$
Characterization of \( Z^* \)

**Theorem (Optimality conditions, informal)**

Under generalized covariate shift and some mild assumptions, \( Z^* \) is optimal for IDG if and only if it

- remains **discriminative**: \( \mathbb{R} \left[ Y \mid Z^* \right] = \mathbb{R} \left[ Y \mid X \right] \)
- has **invariant support**: \( \text{supp}(p_{Z^* \mid d_s}) = \text{supp}(p_{Z^* \mid d_t}), \quad \forall d_s, d_t \in \mathcal{D} \)

😊 achievable sufficient and necessary condition
Characterization of $Z^*$

Theorem (Optimality conditions, informal)

Under generalized covariate shift and some mild assumptions, $Z^*$ is optimal for IDG if and only if it

- remains **discriminative**: $\mathbb{R} [Y \mid Z^*] = \mathbb{R} [Y \mid X]$
- has **invariant support**: $\text{supp}(p_{Z^* \mid d_s}) = \text{supp}(p_{Z^* \mid d_t}), \forall d_s, d_t \in \mathcal{D}$

😉 achievable sufficient and necessary condition
😊 requires access to labeled target domain
Proposition (No free lunch for IDG, informal)

Let $Z_{d_s}$ be any rep. chosen on some source $d_s$ and $C$ a constant rep.

Under mild assumptions, if $Z_{d_s}$ outperforms $C$ on some “good” targets outside the source’s support, there are many “bad” targets on which $Z_{d_s}$ is strictly worse than $C$. 
Proposition (No free lunch for IDG, informal)

Let $Z_{ds}$ be any rep. chosen on some source $d_s$ and $C$ a constant rep. Under mild assumptions, if $Z_{ds}$ outperforms $C$ on some “good” targets outside the source’s support, there are many “bad” targets on which $Z_{ds}$ is strictly worse than $C$.

✅ implies the failure of current DG methods

⊆ unable to outperform ERM on a unified benchmark [3]
⊆ insufficient access to or strong asmp. on targets
**Proposition (No free lunch for IDG, informal)**

Let $Z_{d_s}$ be any rep. chosen on some source $d_s$ and $C$ a constant rep.

Under mild assumptions, if $Z_{d_s}$ outperforms $C$ on some “good” targets outside the source’s support, there are many “bad” targets on which $Z_{d_s}$ is strictly worse than $C$.

- ✓ implies the failure of current DG methods
  - ⊗ unable to outperform ERM on a unified benchmark [3]
  - ⊗ insufficient access to or strong asmp. on targets
- ✗ how to deal with necessary (but unrealistic) access to targets?
Method: Learning $Z^*$ with SSL
Recent SSL methods learn transferable and **robust** reps.:

- train on large-scale **unlabelled** data (≫ = ImageNet)
- use **augmentations** as surrogate information for $Y$

**SimCLR [1]: image aug.**

**CLIP [4]: text caption as aug.**
Recent SSL methods learn transferable and robust reps.:

- train on large-scale unlabelled data ($\gg \approx$ ImageNet)
- use augmentations as surrogate information for $Y$

Robustness of different SSL methods varies:

- CLIP achieves incredible robustness to distribution shifts
Augmentation $A$ for learning $Z^*$:

- Label-perserving: retain information about $Y$
Learning $Z^*$ with SSL

Augmentation $A$ for learning $Z^*$:

- Label-perserving: retain information about $Y$
- Domain-agnostic: no correlation with domain
Learning $Z^*$ with SSL

**Augmentation $A$ for learning $Z^*$:**

- Label-perserving: retain information about $Y$
- **Domain-agnostic:** no correlation with domain

**Domain-agnostic $A$**

- ✔ Example: image-text aug. (e.g., CLIP [4])
- ✗ Counterexample: standard image aug. (e.g., SimCLR [1])

“A dog with floppy ears.”  “A pointy-eared dog.”

CLIP aug. $\Rightarrow$ domain-agnostic rep.  SimCLR aug. $\Rightarrow$ domain-correlated rep.
Learning $Z^*$ with SSL

Augmentation $A$ for learning $Z^*$:

- Label-perserving: retain information about $Y$
- **Domain-agnostic:** no correlation with domain

**Domain-agnostic $A$**

✔ Example: image-text aug. (e.g., CLIP [4])
✘ Counterexample: standard image aug. (e.g., SimCLR [1])

“A dog with floppy ears.” “A pointy-eared dog.”

CLIP aug. $\Rightarrow$ domain-agnostic rep. SimCLR aug. $\Rightarrow$ domain-correlated rep.

😊 implies the incredible robustness of CLIP over other SSL models
Proposition (Learning $Z^*$ with domain-agnostic $A$)

Let $p_{A|X}$ be a domain-agnostic augmenter. Then any optimal solution $p_{Z^*|X}$ of the following objective is optimal for IDG:

$$\max_{p_{Z|X}} I[A; Z]$$

s.t. \quad \text{supp}(p_{Z|d}) = \text{supp}(p_Z), \quad \forall d \in \mathcal{D}$$
Proposition (Learning $Z^*$ with domain-agnostic $A$)

Let $p_A | X$ be a domain-agnostic augmenter. Then any optimal solution $p_{Z^*} | X$ of the following objective is optimal for IDG:

$$\max_{p_Z | X} I[A; Z]$$

s.t. $\text{supp}(p_Z|d) = \text{supp}(p_Z), \forall d \in D$

😊 No $Y$ anymore!

😭 support invariance constraint
Learning $Z^*$ with SSL

Practical objectives:

$$\arg \min_{p_{Z \mid X}} \left\{ -I[A; Z] + \lambda B[Z, D] \right\}$$

- Maximize $I[A; Z]$: MI lower bound (e.g., InfoNCE)
- Domain bottleneck $B[Z, D]$: enforce support invariance

- Contrastive adversarial domain (CAD) bottleneck $I[Z; D] \dashv$ Requires no explicit trainable domain classifier
- Constructs an implicit domain classifier from contrastive var. dist.

- Entropy (Ent) bottleneck $H[Z] \dashv$ Requires no access to domain information
Learning $Z^*$ with SSL

**Practical objectives:**

\[
\arg \min_{p_{z|x}} \left\{ -I[A; Z] + \lambda B[Z, D] \right\}
\]

- Maximize $I[A; Z]$: MI lower bound (e.g., InfoNCE)
- Domain bottleneck $B[Z, D]$: enforce support invariance
Learning $Z^*$ with SSL

Practical objectives:

$$\arg \min_{p_{Z \mid X}} -I[A; Z] + \lambda \max \text{ MI} + \lambda B[Z, D]$$

- Maximize $I[A; Z]$: MI lower bound (e.g., InfoNCE)
- Domain bottleneck $B[Z, D]$: enforce support invariance

Domain bottleneck: previous DG methods (e.g., DANN [2]) can apply

- Contrastive adversarial domain (CAD) bottleneck $I[Z; D]$
  - Requires no explicit trainable domain classifier
  - Constructs an implicit domain classifier from contrastive var. dist.
- Entropy (Ent) bottleneck $H[Z]$
  - Requires no access to domain information
Summary: one can learn optimal $Z^*$ with SSL using:

- large-scale unlabeled data
- contrastive learning with domain-agnostic augmentations
- domain bottlenecks
Experiments
Motivation: CLIP was trained

- ✓ with 400M image-text augmentations
- ✗ without explicit domain bottlenecks

Idea:

- Finetune CLIP with bottlenecks on available data
- Evaluate with linear probe on DomainBed [3]
### Exploiting Pretrained CLIP for $Z^*$

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>VLCS</th>
<th>PACS</th>
<th>OfficeHome</th>
<th>DomainNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM</td>
<td>77.6 ± 0.3</td>
<td>86.7 ± 0.3</td>
<td>66.4 ± 0.5</td>
<td>41.3 ± 0.1</td>
</tr>
<tr>
<td>DomainBed SOTA</td>
<td>79.9 ± 0.2</td>
<td>87.2 ± 0.1</td>
<td>68.4 ± 0.2</td>
<td>41.8 ± 0.1</td>
</tr>
<tr>
<td>DINO + CAD</td>
<td>69.6 ± 0.6</td>
<td>76.1 ± 0.1</td>
<td>56.9 ± 0.5</td>
<td>33.6 ± 0.1</td>
</tr>
<tr>
<td>CLIP</td>
<td>80.7 ± 0.4</td>
<td>93.7 ± 0.8</td>
<td>79.6 ± 0.1</td>
<td>52.8 ± 0.1</td>
</tr>
<tr>
<td>CLIP + CAD</td>
<td>81.6 ± 0.1</td>
<td>94.9 ± 0.3</td>
<td>80.0 ± 0.2</td>
<td>53.7 ± 0.1</td>
</tr>
</tbody>
</table>

😊 **SOTA result with domain-agnostic aug. and bottlenecks!**
Idea: learn task- and domain-agnostic robust reps.

- Task: use LAION-400M [5] with text-image contrastive loss
- Domain: finetune CLIP with Ent bottleneck
Towards Generic Robust Representations with SSL

Idea: learn task- and domain-agnostic robust reps.

- Task: use LAION-400M [5] with text-image contrastive loss
- Domain: finetune CLIP with Ent bottleneck

Evaluate: natural distribution shift [6]

<table>
<thead>
<tr>
<th></th>
<th>IN</th>
<th>IN-V2</th>
<th>IN-S</th>
<th>YT-BB</th>
<th>IN-Vid</th>
<th>ObjNet</th>
<th>IN-A</th>
<th>IN-R</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretrained</td>
<td>75.2</td>
<td>64.2</td>
<td>41.0</td>
<td>58.4</td>
<td>71.6</td>
<td>42.8</td>
<td>27.5</td>
<td>62.9</td>
<td>52.6</td>
</tr>
<tr>
<td>Tuned w/o Ent</td>
<td>73.8</td>
<td>62.1</td>
<td>37.0</td>
<td>56.9</td>
<td>68.8</td>
<td>41.3</td>
<td>26.0</td>
<td>58.1</td>
<td>50.0</td>
</tr>
<tr>
<td>Tuned w/ Ent</td>
<td>74.2</td>
<td>62.7</td>
<td>38.9</td>
<td>58.1</td>
<td>70.1</td>
<td>42.1</td>
<td>26.2</td>
<td>60.8</td>
<td>51.3</td>
</tr>
</tbody>
</table>

😊 Consistently improved robustness with bottlenecks!
😊 Gains could be larger if end-to-end trained with bottlenecks!
Future Directions

- Non-idealized setups: finite sample case, constrained hypothesis?
- Approx. optimality: relaxed constraints?
- More practical methods for learning $Z^*$?
- Implicit regularization effect for learning $Z^*$?
- ...
Thank you!

Amazing co-authors:

Yann Dubois

Chris J. Maddison


