Motivation

- Current AI systems are still limited in terms of planning and reasoning
- Humans plan & reason using abstract concepts (e.g., objects & their properties)
- Causal models present a natural framework to represent such abstract concepts — latent causal variables and reason about interventions on them
- How to train representation learners that extract causal variables from high dimensional data (e.g., images) with minimal supervision?

Problem Statement

- True latent variables:
  - Observational distribution: \( z \sim \mathcal{P}_z \) with support \( \mathcal{Z} \)
  - Interventional distribution: \( z \sim \mathcal{P}_z^{(u)} \) with support \( \mathcal{Z}^{(u)} \)
- Mixing function: \( g: \mathbb{R}^d \rightarrow \mathbb{R}^n \), which is injective
- Observations: \( x \leftarrow g(z) \) with supports \( \mathcal{X}, \mathcal{X}^{(u)} \) in observational and interventional distribution
- Learn an auto encoder:
  - Reconstruction identity: \( \hat{f}(x) = x \), \( \forall x \in \mathcal{X} \cup \mathcal{X}^{(u)} \)
  - \( \hat{z} \triangleq f(x) \)
- Affine Identification: \( \hat{z} = A z + c \)
- Permutation and scaling identification: \( \tilde{z} = \Pi A z + c \)

Identification under Hard do Interventions

- **Assumption 1**: \( g \) is an injective polynomial, \( \mathcal{X} \) has a non-empty interior
- **Assumption 2**: \( \mathcal{P}_z^{(u)} \) hard intervention on \( z_i \)
- Do intervention constraint: \( f_k(x) = z_i, \forall x \in \mathcal{X}^{(u)} \)

Theorem (Informal): If Assumption 1 and 2 hold, then the solution to the reconstruction identity with \( h \) as a polynomial and do intervention constraint satisfies \( \hat{z}_i = c \hat{z}_i + h, \forall z \in \mathcal{X} \cup \mathcal{X}^{(u)} \)

- For general diffeomorphisms, we show approximate component-wise identification under multiple do interventions per component

Identification under Imperfect Interventions

- **Assumption 3**: For \( \mathcal{P}_z^{(u)} \bigcirc \mathcal{S} \), support of \( z_j \) is independent of latents in \( \mathcal{S} \)
- **IS constraint**: For a set \( \mathcal{S} \) support of \( \hat{z}_k \) is independent of latents in \( \mathcal{S} \)

Theorem (Informal): If Assumption 1 and 3 hold, then the solution to the reconstruction identity with \( h \) as a polynomial and support independence constraint achieves block-affine identification

\[
\hat{z}_i = a_i z_i + c_i, \quad \hat{z}_m = a_m z_m + c_m, \quad \forall m \in \mathcal{S} \setminus \mathcal{K}_i
\]

\( a_i \) and \( c_i \) do not share non-zero components.

Independent Support & Imperfect Interventions

- Independent Support (IS): \( \mathcal{X}^{(i)} = \mathcal{X}_1 \times \mathcal{X}_2 \)
- Statistical Independence \( \implies \) IS

Experiments

<table>
<thead>
<tr>
<th># Interv Outs</th>
<th>Uniform</th>
<th>SCM-Linear</th>
<th>SCM Non-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33.2</td>
<td>42.7</td>
<td>34.9</td>
</tr>
<tr>
<td>3</td>
<td>72.2</td>
<td>73.9</td>
<td>65.2</td>
</tr>
<tr>
<td>5</td>
<td>88.3</td>
<td>83.6</td>
<td>77.2</td>
</tr>
<tr>
<td>7</td>
<td>88.1</td>
<td>85.5</td>
<td>81.9</td>
</tr>
<tr>
<td>9</td>
<td>87.5</td>
<td>84.8</td>
<td>81.1</td>
</tr>
</tbody>
</table>