Addressing Divergent Representations from Causal Interventions on Neural Networks

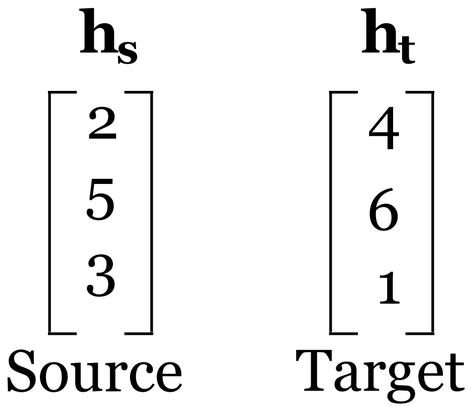
Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, Christopher Potts

Introduction

- Many interpretability methods use causal interventions on representations to understand what they encode.

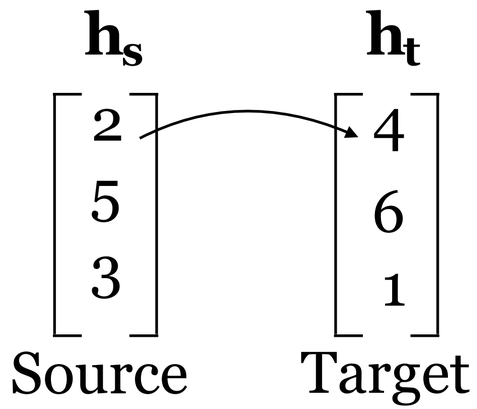
Introduction

- Many interpretability methods use causal interventions on representations to understand what they encode.



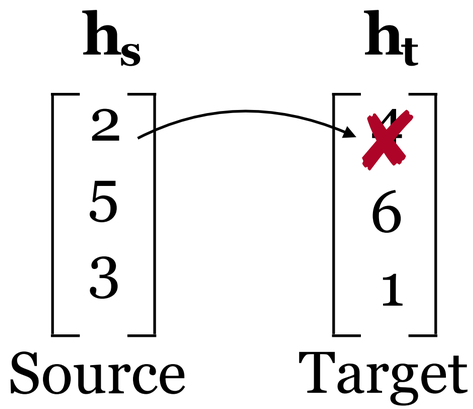
Introduction

- Many interpretability methods use causal interventions on representations to understand what they encode.



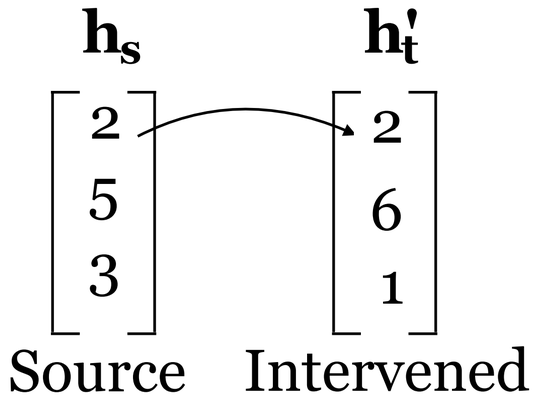
Introduction

- Many interpretability methods use causal interventions on representations to understand what they encode.



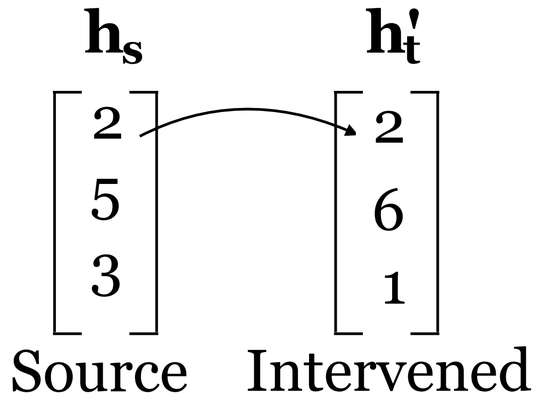
Introduction

- Many interpretability methods use causal interventions on representations to understand what they encode.



Introduction

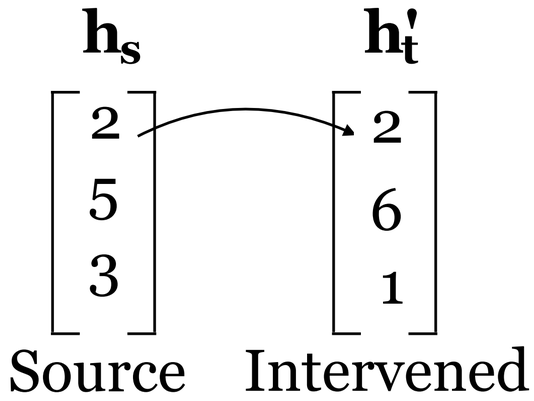
- Many interpretability methods use causal interventions on representations to understand what they encode.



- It is common to make claims about a network's inner mechanisms depending on the resulting computations from an intervention.

Introduction

- Many interpretability methods use causal interventions on representations to understand what they encode.

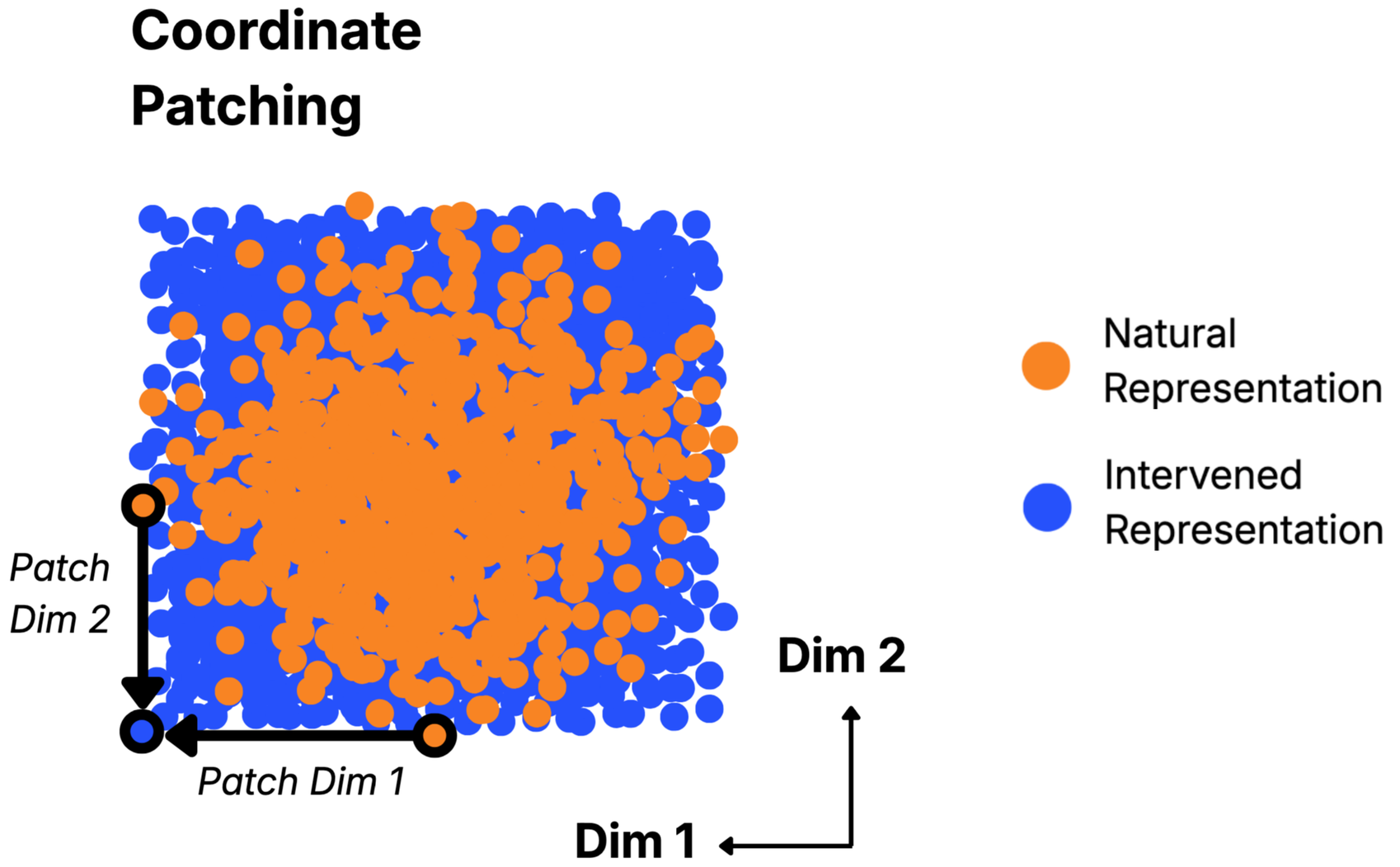


- It is common to make claims about a network's inner mechanisms depending on the resulting computations from an intervention.

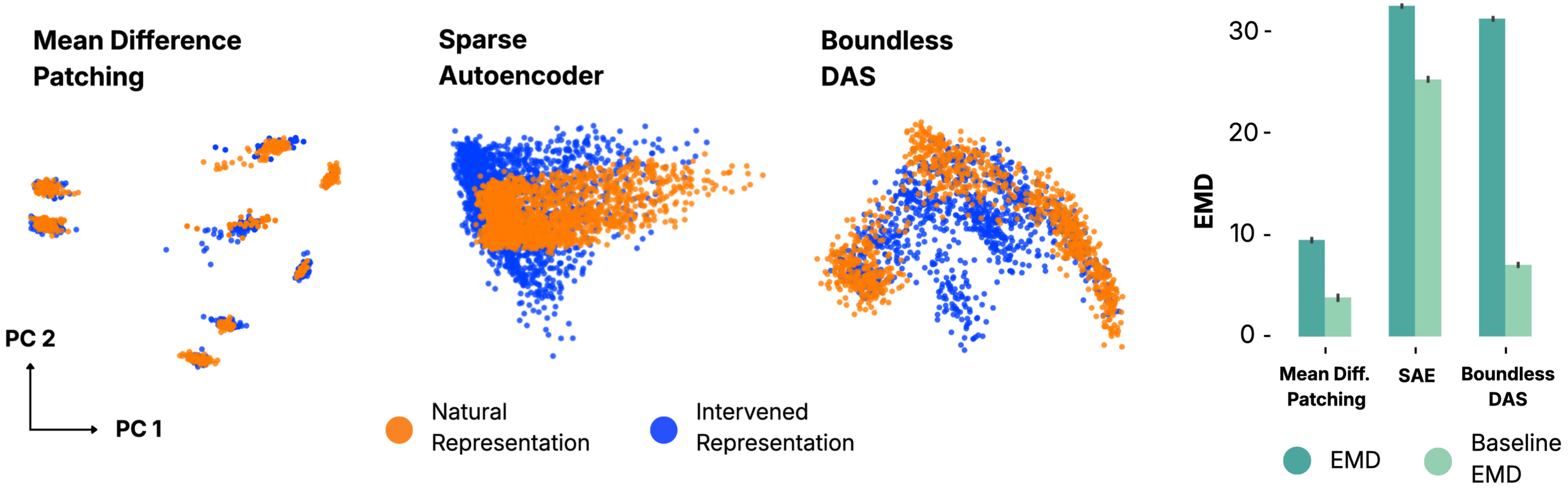
- What if causally intervened representations deviate from the natural distribution of representations?

Part 1: Do intervened representations diverge from the natural distribution?

Yes: Theoretical Representational Divergence



Yes: Empirical Representational Divergence



Part 2: Is this okay?

Harmfulness is Claim Dependent

- Intuitive example:

Harmfulness is Claim Dependent

- Intuitive example:
    - Claim 1: We have found ***a*** computational pathway that does X

Harmfulness is Claim Dependent

- Intuitive example:
  - Claim 1: We have found ***a*** computational pathway that does X
  - Claim 2: We have found ***the*** computational pathway that does X

Harmless Divergence Intuitions

- Matrix multiplication example:

Harmless Divergence Intuitions

- Matrix multiplication example:

$$Wh = z \qquad \begin{matrix} h \in \mathbb{R}^n \\ W \in \mathbb{R}^{m \times n} \end{matrix}$$

Harmless Divergence Intuitions

- Matrix multiplication example:

$$Wh = z \qquad h \in \mathbb{R}^n \qquad W \in \mathbb{R}^{m \times n}$$

- Intuitively, adding zero to a representation should be okay:

Harmless Divergence Intuitions

- Matrix multiplication example:

$$Wh = z \qquad \begin{matrix} h \in \mathbb{R}^n \\ W \in \mathbb{R}^{m \times n} \end{matrix}$$

- Intuitively, adding zero to a representation should be okay:

$$W(h + 0) = Wh = z$$

Harmless Divergence Intuitions

- Matrix multiplication example:

$$Wh = z \qquad \begin{matrix} h \in \mathbb{R}^n \\ W \in \mathbb{R}^{m \times n} \end{matrix}$$

- Intuitively, adding zero to a representation should be okay:

$$W(h + 0) = Wh = z$$

- Thus, any vector that is equivalent to adding zero should be okay:

Harmless Divergence Intuitions

- Matrix multiplication example:

$$Wh = z \qquad \begin{matrix} h \in \mathbb{R}^n \\ W \in \mathbb{R}^{m \times n} \end{matrix}$$

- Intuitively, adding zero to a representation should be okay:

$$W(h + 0) = Wh = z$$

- Thus, any vector that is equivalent to adding zero should be okay:

$$\mathcal{N}(W) = \{v \in \mathbb{R}^d \mid Wv = 0\}$$

Harmless Divergence Intuitions

- Matrix multiplication example:

$$Wh = z \qquad \begin{matrix} h \in \mathbb{R}^n \\ W \in \mathbb{R}^{m \times n} \end{matrix}$$

- Intuitively, adding zero to a representation should be okay:

$$W(h + 0) = Wh = z$$

- Thus, any vector that is equivalent to adding zero should be okay:

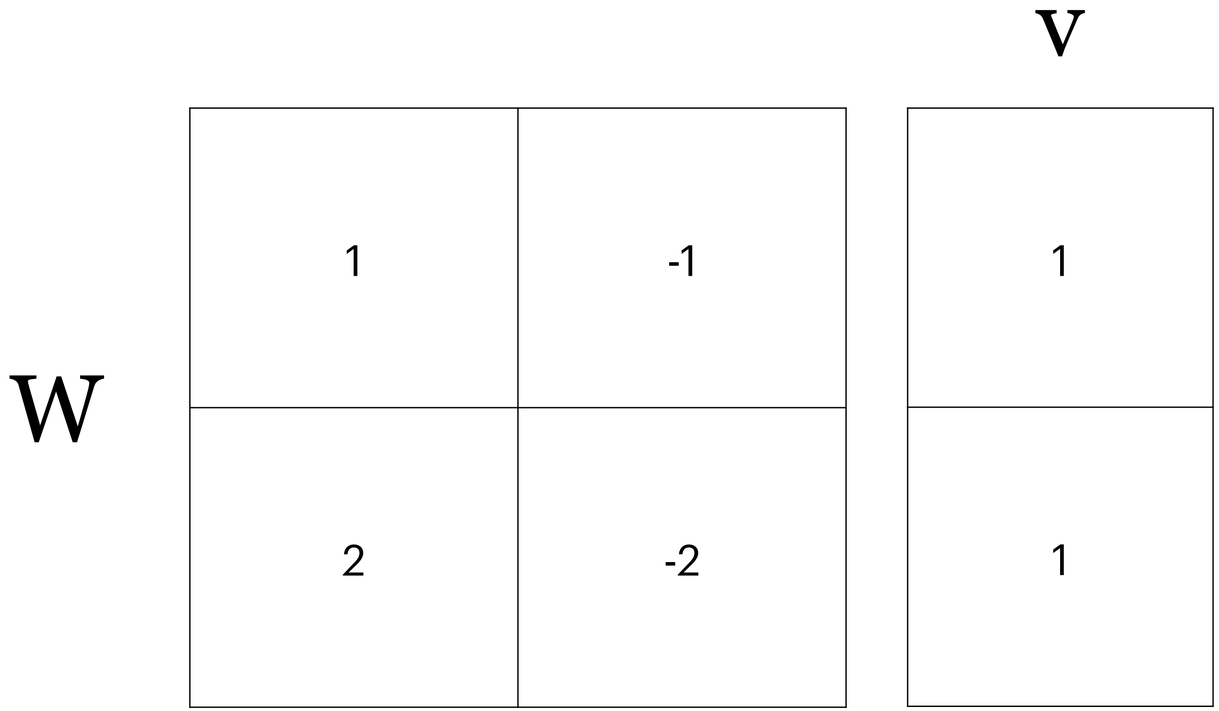
$$\mathcal{N}(W) = \{v \in \mathbb{R}^d \mid Wv = 0\}$$

$$\mathbf{v} \in \mathcal{N}(W)$$

$$W(h + \mathbf{v}) = Wh + W\mathbf{v} = Wh + 0 = W(h + 0) = z$$

Intuitions from Matrix Multiplication

v is in $\mathcal{N}$(W)



Intuitions from Matrix Multiplication

v^T

1	1

W

1	-1
2	-2

v is in $\mathcal{N}(W)$

Intuitions from Matrix Multiplication

v is in $\mathcal{N}(W)$

v^T

1		1	

W

1∗1	+	-1∗1	= 0
2∗1	+	-2∗1	= 0

Intuitions from Matrix Multiplication

v is in $\mathcal{N}(W)$

v^T

1	1

W

1∗1	+	-1∗1	
2∗1	+	-2∗1	

=

z

0
0

Intuitions from Matrix Multiplication

v is in $\mathcal{N}(W)$

v^T

1	1

W

1∗1	+	-1∗1
2∗1	+	-2∗1

$=$

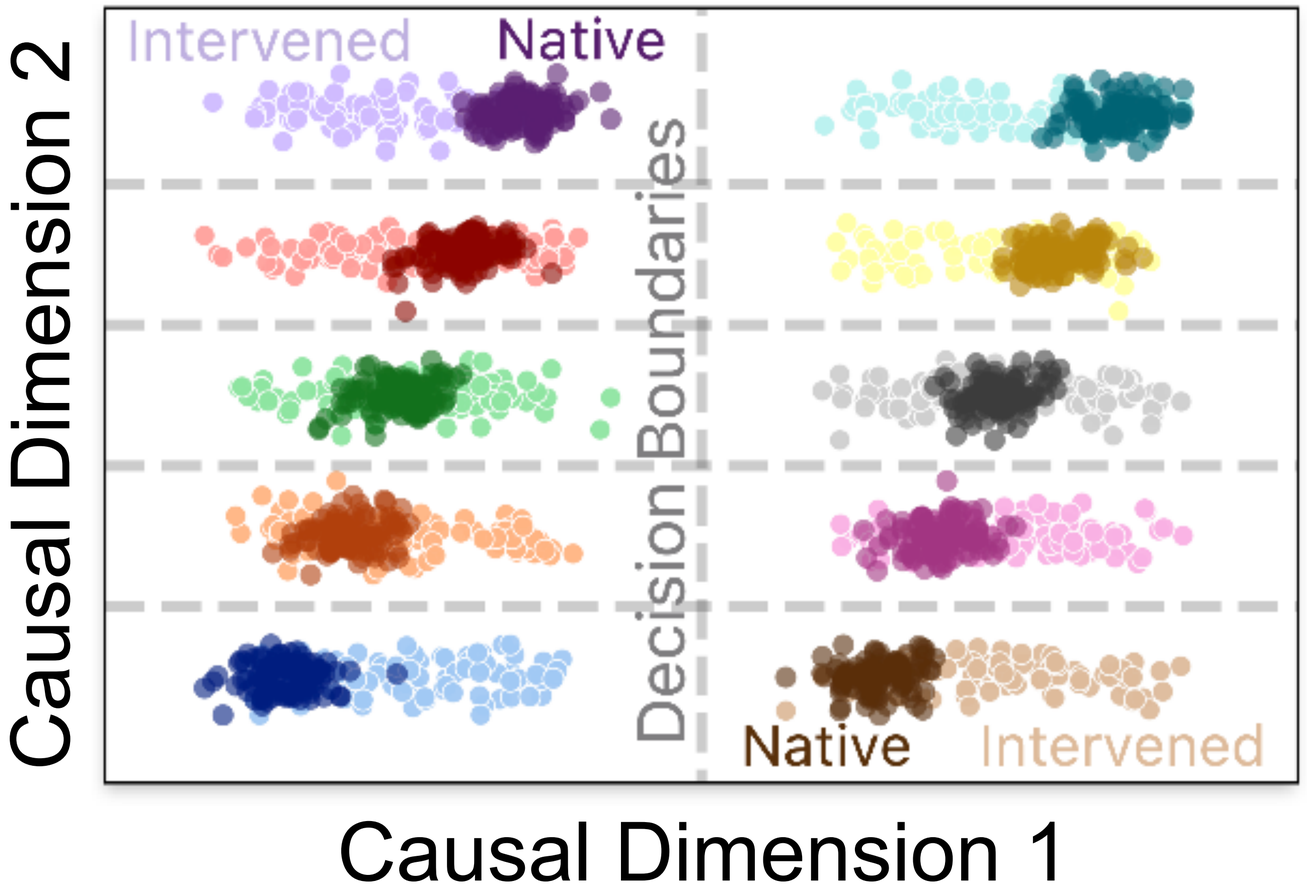
z

0
0

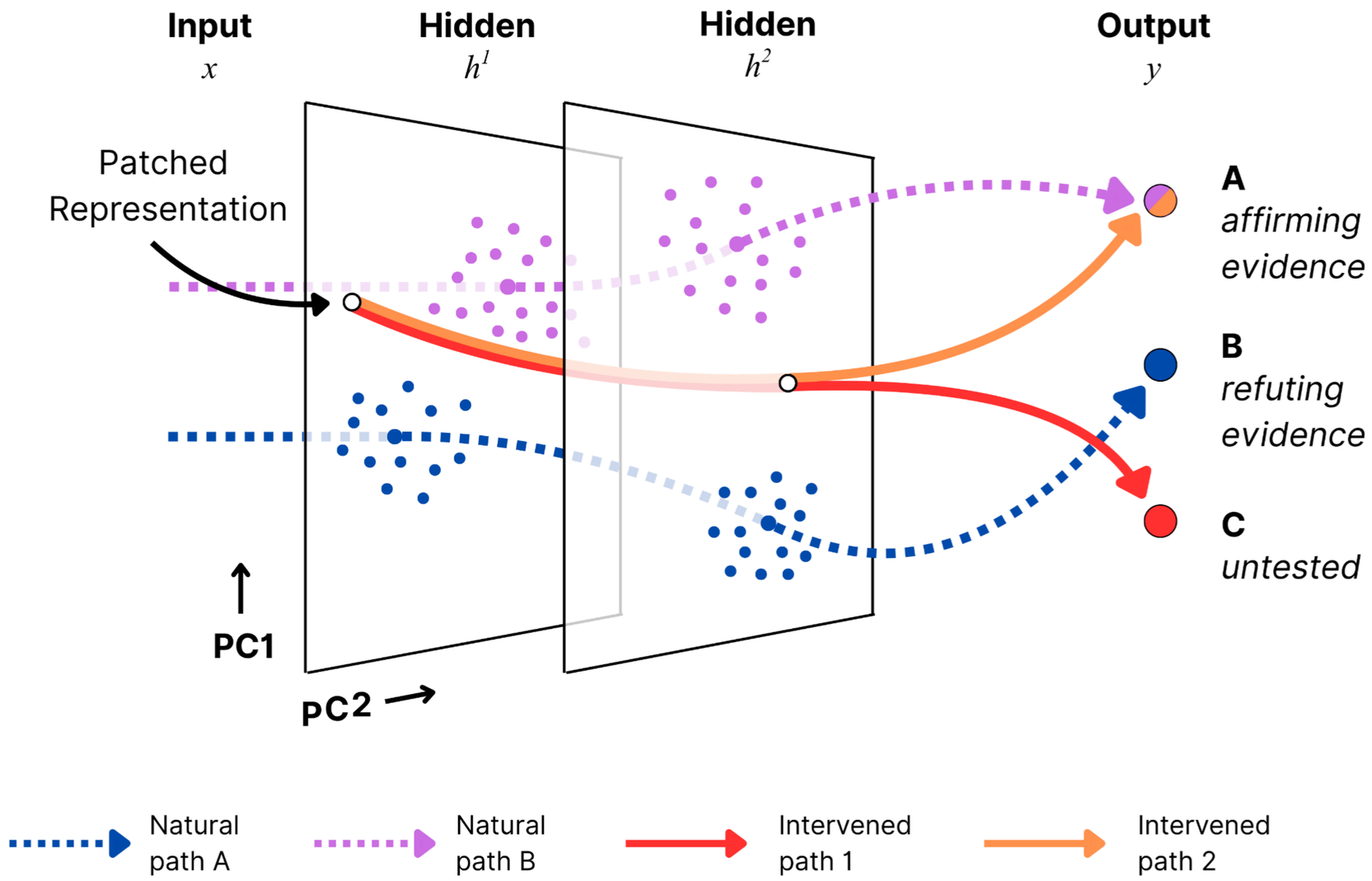
$$W_{i,j}v_j \neq W_{i,j}0$$

Divergence that is harmless for claims about the aggregate computation may be pernicious to claims about the subcomputations.

Harmless Divergence



Pernicious Divergence

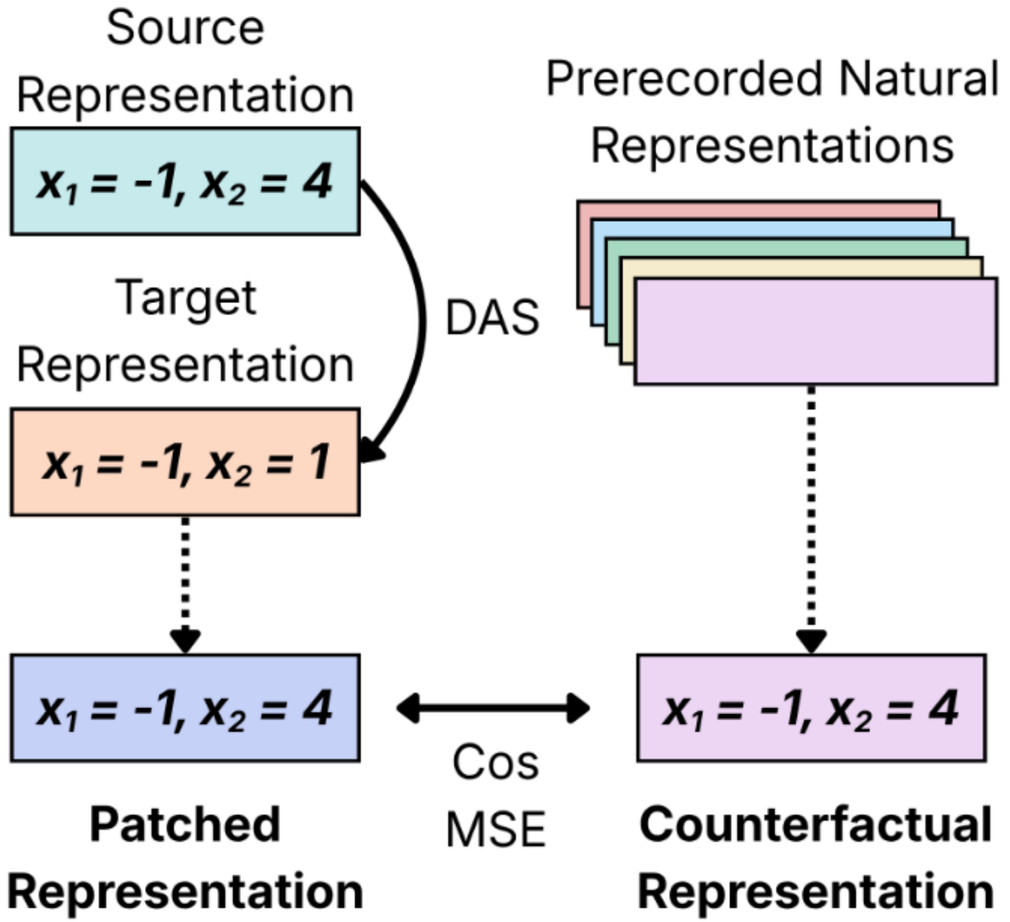


Part 3: How do we mitigate representational divergence?

Counterfactual Latent (CL) Loss

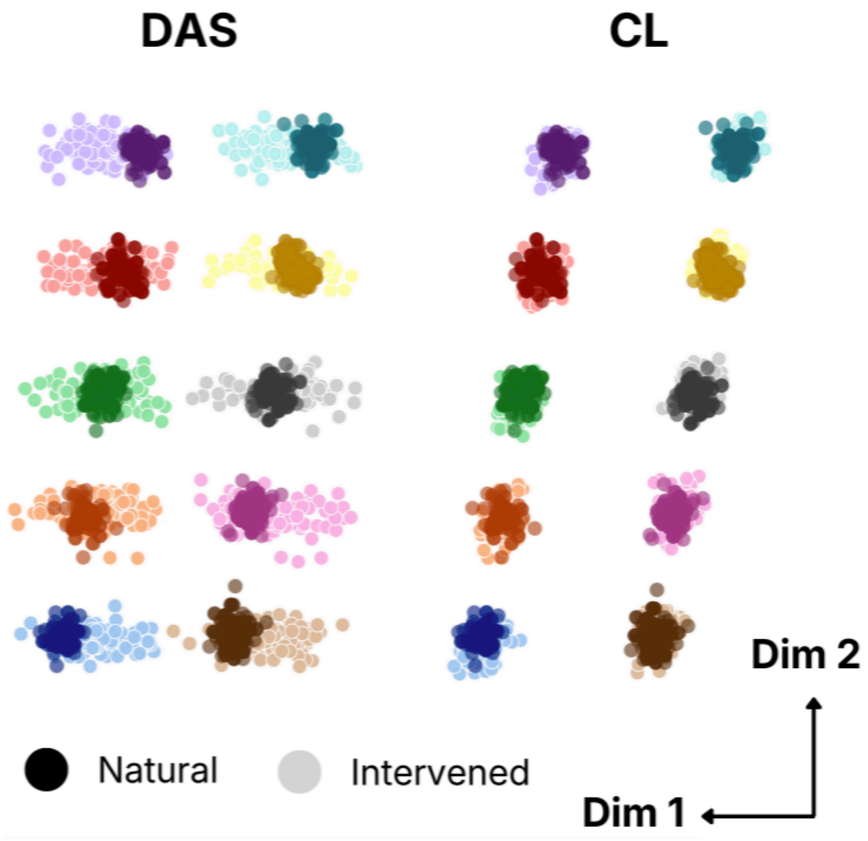
Instead of using the ***expected behavior*** of the causal abstraction as a training objective,

we can use the ***expected natural representation*** as a latent training objective.

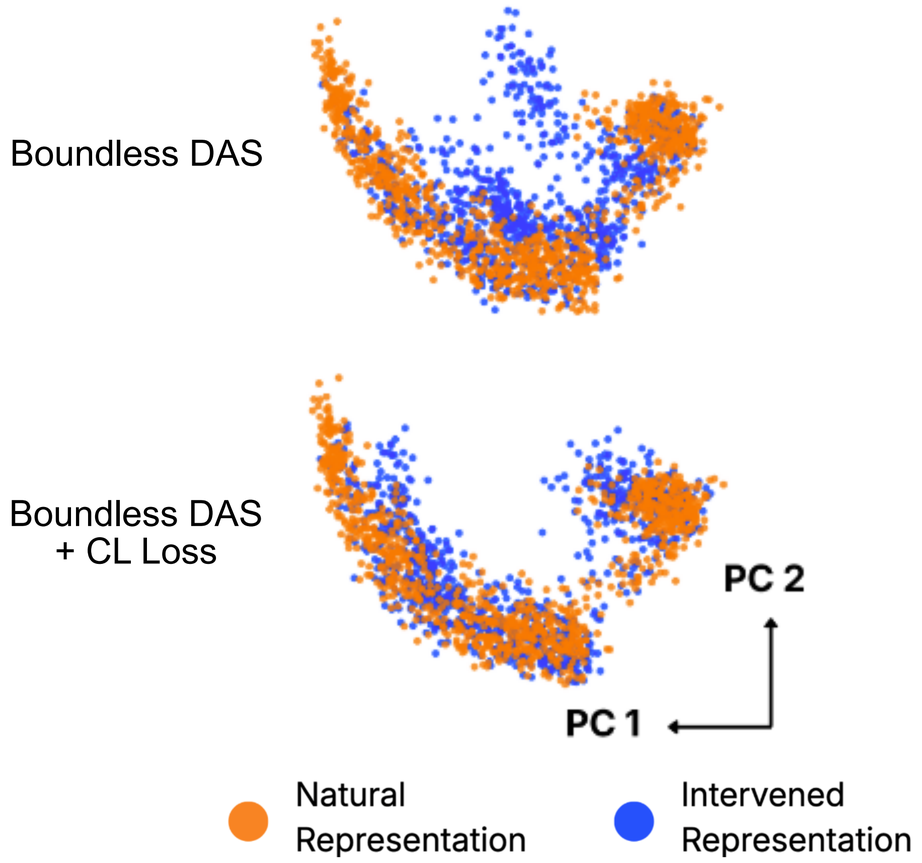


CL Loss Reduces Divergence

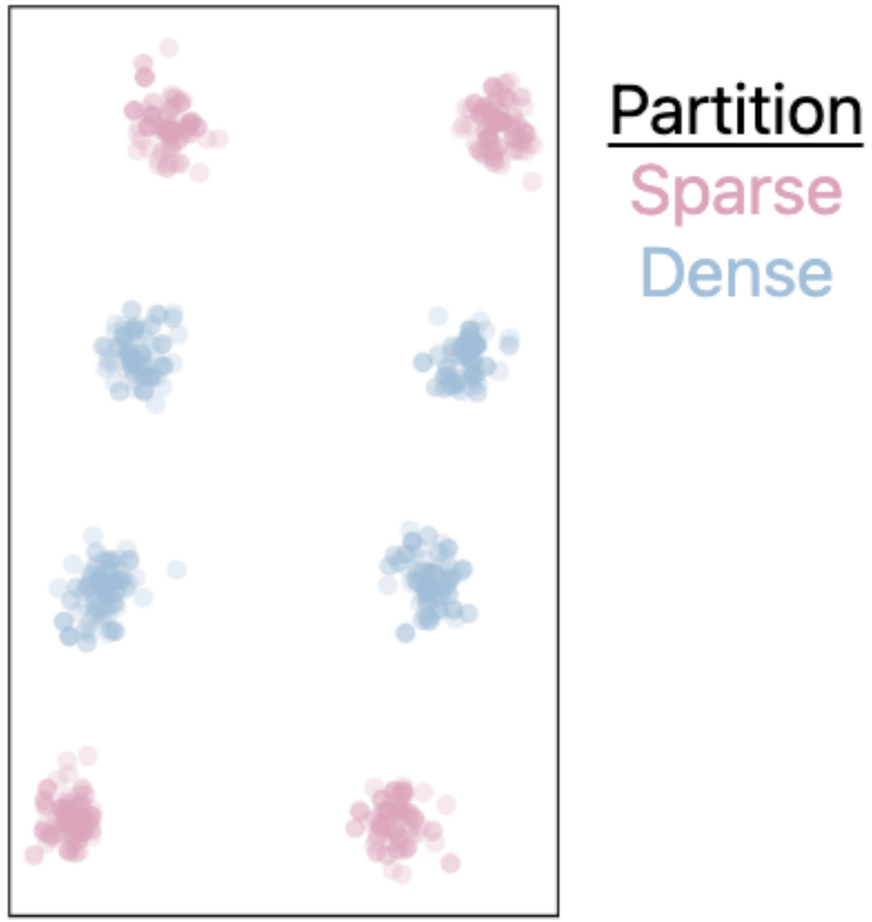
Synthetic Setting

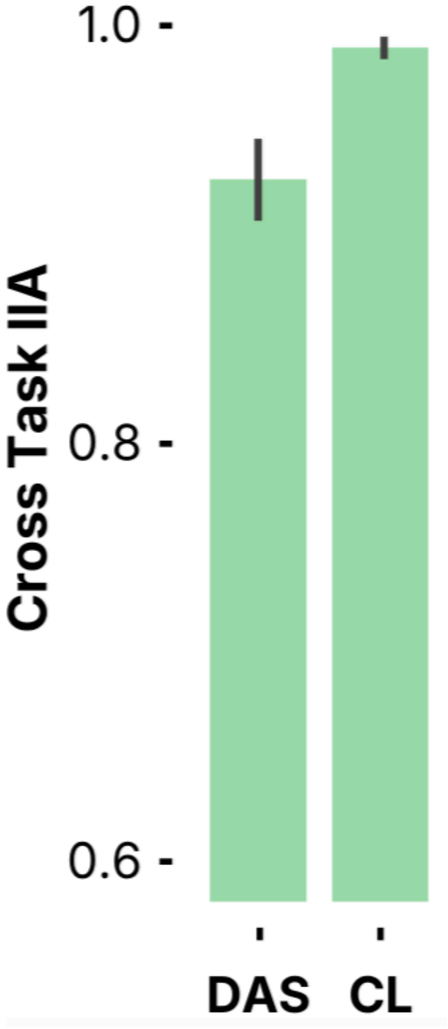


LLM Price Tagging Setting



CL Loss Can Improve OOD Interventions





Thank you!

Jerome Han

Alexa Tartaglini

Christopher Potts

Email: grantsrb@stanford.edu

Twitter: @satchelgrant