

A. Summary of Main Results

In Figure 6, we give a high-level summary of our main results. In Section 2, we have given the definitions of unembedding and embedding representations and how they also yield measurement and intervention representations, respectively. In Section 3, we have defined the causal inner product and show how it unifies the unembedding and embedding representations via the induced Riesz isomorphism.

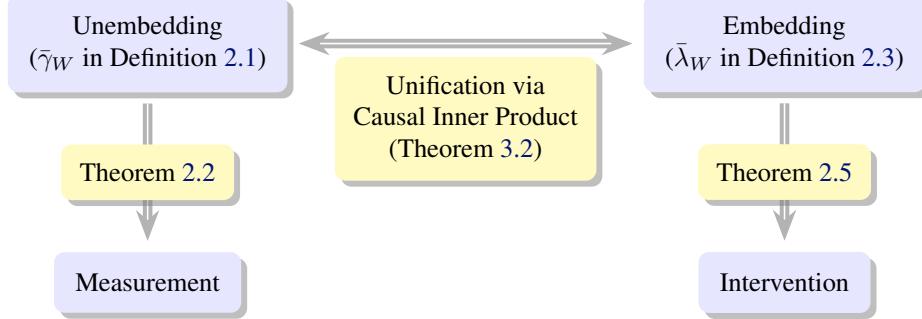


Figure 6. A high-level summary of our main results, illustrating the connections between the different notions of linear representations.

B. Proofs

B.1. Proof of Theorem 2.2

Theorem 2.2 (Measurement Representation). *Let W be a concept, and let $\bar{\gamma}_W$ be the unembedding representation of W . Then, given any context embedding $\lambda \in \Lambda$,*

$$\text{logit } \mathbb{P}(Y = Y(1) | Y \in \{Y(0), Y(1)\}, \lambda) = \alpha \lambda^\top \bar{\gamma}_W,$$

where $\alpha > 0$ (a.s.) is a function of $\{Y(0), Y(1)\}$.

Proof. The proof involves writing out the softmax sampling distribution and invoking Definition 2.1.

$$\text{logit } \mathbb{P}(Y = Y(1) | Y \in \{Y(0), Y(1)\}, \lambda) \tag{B.1}$$

$$= \log \frac{\mathbb{P}(Y = Y(1) | Y \in \{Y(0), Y(1)\}, \lambda)}{\mathbb{P}(Y = Y(0) | Y \in \{Y(0), Y(1)\}, \lambda)} \tag{B.2}$$

$$= \lambda^\top \{\gamma(Y(1)) - \gamma(Y(0))\} \tag{B.3}$$

$$= \alpha \cdot \lambda^\top \bar{\gamma}_W. \tag{B.4}$$

In (B.3), we simply write out the softmax distribution, allowing us to cancel out the normalizing constants for the two probabilities. Equation (B.4) follows directly from Definition 2.1; note that the randomness of α comes from the randomness of $\{Y(0), Y(1)\}$. \square

B.2. Proof of Lemma 2.4

Lemma B.1 (Unembedding-Embedding Relationship). *Let $\bar{\lambda}_W$ be the embedding representation of a concept W , and let $\bar{\gamma}_W$ and $\bar{\gamma}_Z$ be the unembedding representations for W and any concept Z that is causally separable with W . Then,*

$$\bar{\lambda}_W^\top \bar{\gamma}_W > 0 \quad \text{and} \quad \bar{\lambda}_W^\top \bar{\gamma}_Z = 0. \tag{2.1}$$

Conversely, if a representation $\bar{\lambda}_W$ satisfies (2.1), and if there exist concepts $\{Z_i\}_{i=1}^{d-1}$, such that each Z_i is causally separable with W and $\{\bar{\gamma}_W\} \cup \{\bar{\gamma}_{Z_i}\}_{i=1}^{d-1}$ is the basis of \mathbb{R}^d , then $\bar{\lambda}_W$ is the embedding representation for W .

Proof. Let λ_0, λ_1 be a pair of embeddings such that

$$\frac{\mathbb{P}(W = 1 | \lambda_1)}{\mathbb{P}(W = 1 | \lambda_0)} > 1 \quad \text{and} \quad \frac{\mathbb{P}(W, Z | \lambda_1)}{\mathbb{P}(W, Z | \lambda_0)} = \frac{\mathbb{P}(W | \lambda_1)}{\mathbb{P}(W | \lambda_0)}, \tag{B.5}$$

for any concept Z that is causally separable with W . Then, by Definition 2.3,

$$\lambda_1 - \lambda_0 \in \text{Cone}(\bar{\lambda}_W). \quad (\text{B.6})$$

The condition (B.5) is equivalent to

$$\frac{\mathbb{P}(W = 1 | \lambda_1)}{\mathbb{P}(W = 1 | \lambda_0)} > 1 \quad \text{and} \quad \frac{\mathbb{P}(Z = 1 | W, \lambda_1)}{\mathbb{P}(Z = 1 | W, \lambda_0)} = 1. \quad (\text{B.7})$$

These two conditions are also equivalent to the following pair of conditions, respectively:

$$\frac{\mathbb{P}(Y = Y(1) | Y \in \{Y(0), Y(1)\}, \lambda_1)}{\mathbb{P}(Y = Y(1) | Y \in \{Y(0), Y(1)\}, \lambda_0)} > 1 \quad (\text{B.8})$$

and

$$\frac{\mathbb{P}(Y = Y(W, 1) | Y \in \{Y(W, 0), Y(W, 1)\}, \lambda_1)}{\mathbb{P}(Y = Y(W, 1) | Y \in \{Y(W, 0), Y(W, 1)\}, \lambda_0)} = 1 \quad (\text{B.9})$$

The reason is that, conditional on $Y \in \{Y(0, 0), Y(0, 1), Y(1, 0), Y(1, 1)\}$, conditioning on W is equivalent to conditioning on $Y \in \{Y(W, 0), Y(W, 1)\}$. And, the event $Z = 1$ is equivalent to the event $Y = Y(W, 1)$. (In words: if we know the output is one of “king”, “queen”, “roi”, “reine” then conditioning on $W = 1$ is equivalent to conditioning on the output being “king” or “roi”. Then, predicting whether the word is in English is equivalent to predicting whether the word is “king”.)

By Theorem 2.2, the two conditions (B.8) and (B.9) are respectively equivalent to

$$\alpha(Y(0), Y(1))(\lambda_1 - \lambda_0)^\top \bar{\gamma}_W > 0 \quad \text{and} \quad \alpha(Y(W, 0), Y(W, 1))(\lambda_1 - \lambda_0)^\top \bar{\gamma}_Z = 0, \quad (\text{B.10})$$

where α 's are positive a.s. These are in turn respectively equivalent to

$$\bar{\lambda}_W^\top \bar{\gamma}_W > 0 \quad \text{and} \quad \bar{\lambda}_W^\top \bar{\gamma}_Z = 0. \quad (\text{B.11})$$

Conversely, if a representation $\bar{\lambda}_W$ satisfies (B.11) and there exist concepts $\{Z_i\}_{i=1}^{d-1}$ such that each concept is causally separable with W and $\{\bar{\gamma}_W\} \cup \{\bar{\gamma}_{Z_i}\}_{i=1}^{d-1}$ is the basis of \mathbb{R}^d , then $\bar{\lambda}_W$ is unique up to positive scaling. If there exists λ_0 and λ_1 satisfying (B.5), then the equivalence between (B.5) and (B.10) says that

$$(\lambda_1 - \lambda_0)^\top \bar{\gamma}_W > 0 \quad \text{and} \quad (\lambda_1 - \lambda_0)^\top \bar{\gamma}_Z = 0. \quad (\text{B.12})$$

In other words, $\lambda_1 - \lambda_0$ also satisfies (B.11), implying that it must be the same as $\bar{\lambda}_W$ up to positive scaling. Therefore, for any λ_0 and λ_1 satisfying (B.5), $\lambda_1 - \lambda_0 \in \text{Cone}(\bar{\lambda}_W)$. \square

B.3. Proof of Theorem 2.5

Theorem 2.5 (Intervention Representation). *Let $\bar{\lambda}_W$ be the embedding representation of a concept W . Then, for any concept Z that is causally separable with W ,*

$$\mathbb{P}(Y = Y(W, 1) | Y \in \{Y(W, 0), Y(W, 1)\}, \lambda + c\bar{\lambda}_W)$$

is constant in $c \in \mathbb{R}$, and

$$\mathbb{P}(Y = Y(1, Z) | Y \in \{Y(0, Z), Y(1, Z)\}, \lambda + c\bar{\lambda}_W)$$

is increasing in $c \in \mathbb{R}$.

Proof. By Theorem 2.2,

$$\text{logit } \mathbb{P}(Y = Y(W, 1) | Y \in \{Y(W, 0), Y(W, 1)\}, \lambda + c\bar{\lambda}_W) \quad (\text{B.13})$$

$$= \alpha \cdot (\lambda + c\bar{\lambda}_W)^\top \bar{\gamma}_Z \quad (\text{B.14})$$

$$= \alpha \cdot \lambda^\top \bar{\gamma}_Z + \alpha c \cdot \bar{\lambda}_W^\top \bar{\gamma}_Z \quad (\text{B.15})$$

Therefore, the first probability is constant since $\bar{\lambda}_W^\top \bar{\gamma}_Z = 0$ by Lemma 2.4.

Also, by Theorem 2.2,

$$\text{logit } \mathbb{P}(Y = Y(1, Z) \mid Y \in \{Y(0, Z), Y(1, Z)\}, \lambda + c\bar{\lambda}_W) \quad (\text{B.16})$$

$$= \alpha \cdot (\lambda + c\bar{\lambda}_W)^\top \bar{\gamma}_W \quad (\text{B.17})$$

$$= \alpha \cdot \lambda^\top \bar{\gamma}_Z + \alpha c \cdot \bar{\lambda}_W^\top \bar{\gamma}_W \quad (\text{B.18})$$

Therefore, the second probability is increasing since $\bar{\lambda}_W^\top \bar{\gamma}_W > 0$ by Lemma 2.4. \square

B.4. Proof of Theorem 3.2

Theorem 3.2 (Unification of Representations). *Suppose that, for any concept W , there exist concepts $\{Z_i\}_{i=1}^{d-1}$ such that each Z_i is causally separable with W and $\{\bar{\gamma}_W\} \cup \{\bar{\gamma}_{Z_i}\}_{i=1}^{d-1}$ is a basis of \mathbb{R}^d . If $\langle \cdot, \cdot \rangle_C$ is a causal inner product, then the Riesz isomorphism $\bar{\gamma} \mapsto \langle \bar{\gamma}, \cdot \rangle_C$, for $\bar{\gamma} \in \bar{\Gamma}$, maps the unembedding representation $\bar{\gamma}_W$ of each concept W to its embedding representation $\bar{\lambda}_W$:*

$$\langle \bar{\gamma}_W, \cdot \rangle_C = \bar{\lambda}_W^\top.$$

Proof. The causal inner product defines the Riesz isomorphism ϕ such that $\phi(\bar{\gamma}) = \langle \bar{\gamma}, \cdot \rangle_C$. Then, we have

$$\phi(\bar{\gamma}_W)(\bar{\gamma}_W) = \langle \bar{\gamma}_W, \bar{\gamma}_W \rangle_C > 0 \quad \text{and} \quad \phi(\bar{\gamma}_W)(\bar{\gamma}_Z) = \langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle_C = 0, \quad (\text{B.19})$$

where the second equality follows from Definition 3.1. By Lemma 2.4, $\phi(\bar{\gamma}_W)$ expresses the unique unembedding representation $\bar{\lambda}_W$ (up to positive scaling); specifically, $\phi(\bar{\gamma}_W) = \bar{\lambda}_W^\top$ where $\bar{\lambda}_W^\top : \bar{\gamma} \mapsto \bar{\lambda}_W^\top \bar{\gamma}$. \square

B.5. Proof of Theorem 3.4

Theorem 3.4 (Explicit Form of Causal Inner Product). *Suppose there exists a causal inner product, represented as $\langle \bar{\gamma}, \bar{\gamma}' \rangle_C = \bar{\gamma}^\top M \bar{\gamma}'$ for some symmetric positive definite matrix M . If there are mutually causally separable concepts $\{W_k\}_{k=1}^d$, such that their canonical representations $G = [\bar{\gamma}_{W_1}, \dots, \bar{\gamma}_{W_d}]$ form a basis for $\bar{\Gamma} \simeq \mathbb{R}^d$, then under Assumption 3.3,*

$$M^{-1} = GG^\top \text{ and } G^\top \text{Cov}(\gamma)^{-1}G = D, \quad (\text{3.2})$$

for some diagonal matrix D with positive entries, where γ is the unembedding vector of a word sampled uniformly at random from the vocabulary.

Proof. Since $\langle \cdot, \cdot \rangle_C$ is a causal inner product,

$$0 = \bar{\gamma}_W^\top M \bar{\gamma}_Z \quad (\text{B.20})$$

for any causally separable concepts W and Z . By applying (B.20) to the canonical representations $G = [\bar{\gamma}_{W_1}, \dots, \bar{\gamma}_{W_d}]$, we obtain

$$I = G^\top MG. \quad (\text{B.21})$$

This shows that $M = G^{-\top} G^{-1}$, proving the first half of (3.2).

Next, observe that $M \bar{\gamma}_{W_i}$ is an embedding representation for each concept W_i for $i = 1, \dots, d$ by the proof of Lemma 2.4 and Theorem 3.2. Then, by Assumption 3.3,

$$0 = \text{Cov}(\bar{\gamma}_{W_i}^\top M \gamma, \bar{\gamma}_{W_j}^\top M \gamma) \quad (\text{B.22})$$

$$= \bar{\gamma}_{W_i}^\top M \text{Cov}(\gamma) M \bar{\gamma}_{W_j}. \quad (\text{B.23})$$

for $i \neq j$. Thus,

$$D^{-1} = G^\top M \text{Cov}(\gamma) MG, \quad (\text{B.24})$$

for some diagonal matrix D with positive entries. Substituting in $M = G^{-\top} G^{-1}$, we get

$$\text{Cov}(\gamma) = GD^{-1}G^\top, \quad (\text{B.25})$$

proving the second half of (3.2). \square