Lemma 2.4 (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.

We can now connect to the intervention notion:

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,*

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

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

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

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

In words: adding $\bar{\lambda}_W$ to the language model representation of the context changes the probability of the target concept (W), but not the probability of off-target concepts (Z).

3. Inner Product for Language Model Representations

Given linear representations, we would like to make use of them by doing things like measuring the similarity between different representations, or editing concepts by projecting onto a target direction. Similarity and projection are both notions that require an inner product. We now consider the question of which inner product is appropriate for understanding language model representations.

Preliminaries We define $\bar{\Gamma}$ to be the space of differences between elements of Γ. Then, $\bar{\Gamma}$ is a d-dimensional real vector space.[1] We consider defining inner products on $\bar{\Gamma}$. Unembedding representations are naturally directions (unique only up to scale). Once we have an inner product, we define the *canonical* unembedding representation $\bar{\gamma}_W$ to be the element of the cone with $\langle \bar{\gamma}_W, \bar{\gamma}_W \rangle = 1$. This lets us define inner products between unembedding representations.

Unidentifiability of the inner product We might hope that there is some natural inner product that is picked out (identified) by the model training. It turns out that this is not the case. To understand the challenge, consider transforming the unembedding and embedding spaces according to

$$g(y) \leftarrow A\gamma(y) + \beta, \quad l(x) \leftarrow A^{-\top}\lambda(x), \tag{3.1}$$

where $A \in \mathbb{R}^{d \times d}$ is some invertible linear transformation and $\beta \in \mathbb{R}^d$ is a constant. It's easy to see that this transformation preserves the softmax distribution $\mathbb{P}(y \mid x)$:

$$\frac{\exp(\lambda(x)^\top \gamma(y))}{\sum_{y'} \exp(\lambda(x)^\top \gamma(y'))} = \frac{\exp(l(x)^\top g(y))}{\sum_{y'} \exp(l(x)^\top g(y'))}, \quad \forall x, y.$$

However, the objective function used to train the model depends on the representations only through the softmax probabilities. Thus, the representation γ is identified (at best) only up to some invertible affine transformation.

This also means that the concept representations $\bar{\gamma}_W$ are identified only up to some invertible linear transformation A. The problem is that, given any fixed inner product,

$$\langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle \neq \langle A\bar{\gamma}_W, A\bar{\gamma}_Z \rangle,$$

in general. Accordingly, there is no obvious reason to expect that algebraic manipulations based on, e.g., the Euclidean inner product, should be semantically meaningful.

3.1. Causal Inner Products

We require some additional principles for choosing an inner product on the representation space. The intuition we follow here is that causally separable concepts should be represented as orthogonal vectors. For example, `English⇒French` and `Male⇒Female`, should be orthogonal. We define an inner product with this property:

Definition 3.1 (Causal Inner Product). A *causal inner product* $\langle \cdot, \cdot \rangle_\mathrm{C}$ on $\bar{\Gamma} \simeq \mathbb{R}^d$ is an inner product such that

$$\langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle_\mathrm{C} = 0,$$

for any pair of causally separable concepts W and Z.

This choice turns out to have the key property that it unifies the unembedding and embedding representations:

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 $\langle \cdot, \cdot \rangle_\mathrm{C}$ is a causal inner product, then the Riesz isomorphism $\bar{\gamma} \mapsto \langle \bar{\gamma}, \cdot \rangle_\mathrm{C}$, for $\bar{\gamma} \in \bar{\Gamma}$, maps the unembedding representation $\bar{\gamma}_W$ of each concept W to its embedding representation $\bar{\lambda}_W$:*

$$\langle \bar{\gamma}_W, \cdot \rangle_\mathrm{C} = \bar{\lambda}_W^\top.$$

To understand this result intuitively, notice we can represent embeddings as row vectors and unembeddings as column

[1] Note that the unembedding space Γ is only an affine space, since the softmax is invariant to adding a constant.

vectors. If the causal inner product were the Euclidean inner product, the isomorphism would simply be the transpose operation. The theorem is the (Riesz isomorphism) generalization of this idea: each linear map on $\bar{\Gamma}$ corresponds to some $\lambda \in \Lambda$ according to $\lambda^\top : \bar{\gamma} \mapsto \lambda^\top \bar{\gamma}$. So, we can map $\bar{\Gamma}$ to Λ by mapping each $\bar{\gamma}_W$ to a linear function according to $\bar{\gamma}_W \to \langle \bar{\gamma}_W, \cdot \rangle_\mathrm{C}$. The theorem says this map sends each unembedding representation of a concept to the embedding representation of the same concept.

In the experiments, we will make use of this result to construct embedding representations from unembedding representations. In particular, this allows us to find interventional representations of concepts. This is important because it is difficult in practice to find pairs of prompts that directly satisfy Definition 2.3.

3.2. An Explicit Form for Causal Inner Product

The next problem is: if a causal inner product exists, how can we find it? In principle, this could be done by finding the unembedding representations of a large number of concepts, and then finding an inner product that maps each pair of causally separable directions to zero. In practice, this is infeasible because of the number of concepts required to find the inner product, and the difficulty of estimating the representations of each concept.

We now turn to developing a more tractable approach based on the following insight: knowing the value of concept W expressed by a randomly chosen word tells us little about the value of a causally separable concept Z expressed by that word. For example, if we learn that a randomly sampled word is French (not English), this does not give us significant information about whether it refers to a man or woman.[2] We formalize this idea as follows:

Assumption 3.3. Suppose W, Z are causally separable concepts and that γ is an unembedding vector sampled uniformly from the vocabulary. Then, $\bar{\lambda}_W^\top \gamma$ and $\bar{\lambda}_Z^\top \gamma$ are independent[3] for any embedding representations $\bar{\lambda}_W$ and $\bar{\lambda}_Z$ for W and Z, respectively.

This assumption lets us connect causal separability with something we can actually measure: the statistical dependency between words. The next result makes this precise.

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

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

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

Notice that causal orthogonality only imposes $d(d-1)/2$ constraints on the inner product, but there are $d(d-1)/2+d$ degrees of freedom in identifying the positive definite matrix M (hence, an inner product)—thus, we expect d degrees of freedom in choosing a causal inner product. Theorem 3.4 gives a characterization of this class of inner products, in the form of (3.2). Here, D is a free parameter with d degrees of freedom. Each D defines the inner product. We do not have a principle for picking out a unique choice of D. In our experiments, we will work with the *choice* $D = I_d$, which gives us $M = \mathrm{Cov}(\gamma)^{-1}$. Then, we have a simple closed form for the corresponding inner product:

$$\langle \bar{\gamma}, \bar{\gamma}' \rangle_\mathrm{C} := \bar{\gamma}^\top \mathrm{Cov}(\gamma)^{-1} \bar{\gamma}', \quad \forall \bar{\gamma}, \bar{\gamma}' \in \bar{\Gamma}. \tag{3.3}$$

Note that although we don't have a unique inner product, we can rule out most inner products. E.g., the Euclidean inner product is not a causal inner product if $M = I_d$ does not satisfy (3.2) for any D.

Unified representations The choice of inner product can also be viewed as defining a choice of representations g and l in (3.1) (hence, $\bar{g} = A\bar{\gamma}$). With $A = M^{1/2}$, Theorem 3.2 further implies that a *causal* inner product makes the embedding and unembedding representations of concepts the same, that is, $\bar{g}_W = \bar{l}_W$. Moreover, in the transformed space, the Euclidean inner product *is* the causal inner product: $\langle \bar{\gamma}, \bar{\gamma}' \rangle_\mathrm{C} = \bar{g}^\top \bar{g}'$. In Figure 1, we illustrated this unification of unembedding and embedding representations. This is convenient for experiments, because it allows the use of standard Euclidean tools on the transformed space.

4. Experiments

We now turn to empirically validating the existence of linear representations, the estimated causal inner product, and the predicted relationships between the subspace, measurement, and intervention notions of linear representation. Code is available at github.com/KihoPark/linear_rep_geometry.

We use the LLaMA-2 model with 7 billion parameters (Touvron et al., 2023) as our testbed. This is a decoder-only Transformer LLM (Vaswani et al., 2017; Radford et al., 2018), trained using the forward LM objective and a 32K

[2] Note that this assumption is about words *sampled randomly from the vocabulary*, not words sampled randomly from natural language sources. In the latter, there may well be non-causal correlations between causally separable concepts.

[3] In fact, to prove our next result, we only require that $\bar{\lambda}_W^\top \gamma$ and $\bar{\lambda}_Z^\top \gamma$ are uncorrelated. In Appendix D.6, we verify that the causal inner product we find satisfies the uncorrelatedness condition.

token vocabulary. We include further details on all experiments in Appendix C.

Concepts are represented as directions in the unembedding space We start with the hypothesis that concepts are represented as directions in the unembedding representation space (Definition 2.1). This notion relies on counterfactual pairs of words that vary only in the value of the concept of interest. We consider 22 concepts defined in the Big Analogy Test Set (BATS 3.0) (Gladkova et al., 2016), which provides such counterfactual pairs.[4] We also consider 4 language concepts: `English⇒French`, `French⇒German`, `French⇒Spanish`, and `German⇒Spanish`, where we use words and their translations as counterfactual pairs. Additionally, we consider the concept `frequent⇒infrequent` capturing how common a word is—we use pairs of common/uncommon synonyms (e.g., "bad" and "terrible") as counterfactual pairs. We provide a table of all 27 concepts we consider in Appendix C.

If the subspace notion of the linear representation hypothesis holds, then all counterfactual token pairs should point to a common direction in the unembedding space. In practice, this will only hold approximately. However, if the linear representation hypothesis holds, we still expect that, e.g., $\gamma(\text{"queen"}) - \gamma(\text{"king"})$ will align with the `male⇒female` direction (more closely than the difference between random word pairs will). To validate this, for each concept W, we look at how the direction defined by each counterfactual pair, $\gamma(y_i(1)) - \gamma(y_i(0))$, is geometrically aligned with the unembedding representation $\bar{\gamma}_W$. We estimate $\bar{\gamma}_W$ as the (normalized) mean[5] among all counterfactual pairs: $\bar{\gamma}_W := \tilde{\gamma}_W / \sqrt{\langle \tilde{\gamma}_W, \tilde{\gamma}_W \rangle_{\mathrm{C}}}$, where

$$\tilde{\gamma}_W = \frac{1}{n_W} \sum_{i=1}^{n_W} \left[\gamma(y_i(1)) - \gamma(y_i(0))\right],$$

n_W denotes the number of counterfactual pairs for W, and $\langle \cdot, \cdot \rangle_{\mathrm{C}}$ denotes the causal inner product defined in (3.3).

Figure 2 presents histograms of each $\gamma(y_i(1)) - \gamma(y_i(0)))$ projected onto $\bar{\gamma}_W$ with respect to the causal inner product. Since $\bar{\gamma}_W$ is computed using $\gamma(y_i(1)) - \gamma(y_i(0))$, we compute each projection using a leave-one-out (LOO) estimate $\bar{\gamma}_{W,(-i)}$ of the concept direction that excludes $(y_i(0), y_i(1))$. Across the three concepts shown (and 23 others shown in Appendix D.1), the differences between counterfactual pairs are substantially more aligned with $\bar{\gamma}_W$ than those between random pairs. The sole exception is `thing⇒part`, which does not appear to have a linear representation.

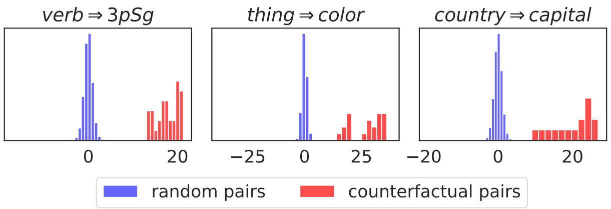

Figure 2. Projecting counterfactual pairs onto their corresponding concept direction shows a strong right skew, as we expect if the linear representation hypothesis holds. The projections of the counterfactual pairs, $\langle \bar{\gamma}_{W,(-i)}, \gamma(y_i(1)) - \gamma(y_i(0)) \rangle_{\mathrm{C}}$, are shown in red. For reference, we also project the differences between 100K randomly sampled word pairs onto the estimated concept direction, as shown in blue. See Table 2 for details about each concept W (the title of each plot).

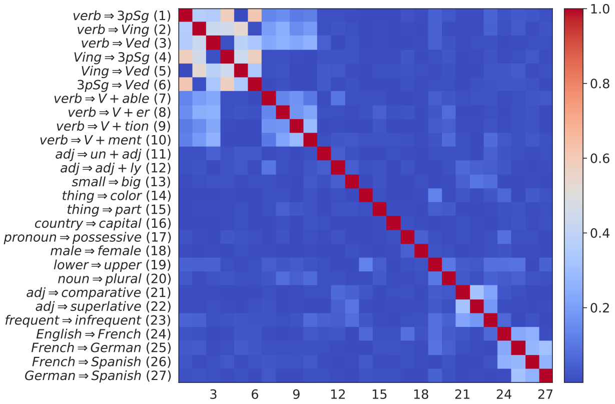

Figure 3. Causally separable concepts are represented approximately orthogonally under the estimated causal inner product based on (3.3). The heatmap shows $|\langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle_{\mathrm{C}}|$ for the estimated unembedding representations of each concept pair (W, Z). The detail for each concept is given in Table 2.

The results are consistent with the linear representation hypothesis: the differences computed by each counterfactual pair point to a common direction representing a linear subspace (up to some noise). Further, $\bar{\gamma}_W$ is a reasonable estimator for that direction.

The estimated inner product respects causal separability Next, we directly examine whether the estimated inner product (3.3) chosen from Theorem 3.4 is indeed approximately a causal inner product. In Figure 3, we plot a heatmap of the inner products between all pairs of the estimated unembedding representations for the 27 concepts. If the estimated inner product is a causal inner product, then we expect values near 0 between causally separable concepts.

The first observation is that most pairs of concepts are nearly orthogonal with respect to this inner product. Interestingly, there is also a clear block diagonal structure.

[4] We only utilize words that are single tokens in the LLaMA-2 model. See Appendix C for details.

[5] Previous work on word embeddings (Drozd et al., 2016; Fournier et al., 2020) motivate taking the mean to improve the consistency of the concept direction.