The Linear Representation Hypothesis and the Geometry of Large Language Models

Kiho Park [1] **Yo Joong Choe** [1] **Victor Veitch** [1]

Abstract

Informally, the "linear representation hypothesis" is the idea that high-level concepts are represented linearly as directions in some representation space. In this paper, we address two closely related questions: What does "linear representation" actually mean? And, how do we make sense of geometric notions (e.g., cosine similarity and projection) in the representation space? To answer these, we use the language of counterfactuals to give two formalizations of linear representation, one in the output (word) representation space, and one in the input (context) space. We then prove that these connect to linear probing and model steering, respectively. To make sense of geometric notions, we use the formalization to identify a particular (non-Euclidean) inner product that respects language structure in a sense we make precise. Using this *causal inner product*, we show how to unify all notions of linear representation. In particular, this allows the construction of probes and steering vectors using counterfactual pairs. Experiments with LLaMA-2 demonstrate the existence of linear representations of concepts, the connection to interpretation and control, and the fundamental role of the choice of inner product. Code is available at github.com/KihoPark/linear_rep_geometry.

1. Introduction

In the context of language models, the "Linear Representation Hypothesis" is the idea that high-level concepts are represented linearly in the representation space of a model (e.g., Mikolov et al., 2013c; Arora et al., 2016; Elhage et al., 2022). High-level concepts might include: is the text in French or English? Is it in the present or past tense? If the text is about a person, are they male or female? The appeal of the *linear representation hypothesis* is that—were it true—the tasks of interpreting and controlling model behavior could exploit linear algebraic operations on the representation space. Our goal is to formalize the linear representation hypothesis, and clarify how it relates to interpretation and control.

The first challenge is that it is not clear what "linear representation" means. There are (at least) three interpretations:

1. **Subspace:** (e.g., Mikolov et al., 2013c; Pennington et al., 2014) The first idea is that each concept is represented as a (1-dimensional) subspace. For example, in the context of word embeddings, it has been argued empirically that $\mathrm{Rep}(\text{“woman”}) - \mathrm{Rep}(\text{“man”})$, $\mathrm{Rep}(\text{“queen”}) - \mathrm{Rep}(\text{“king”})$, and all similar pairs belong to a common subspace (Mikolov et al., 2013c). Then, it is natural to take this subspace to be a representation of the concept of `Male/Female`.
2. **Measurement:** (e.g., Nanda et al., 2023; Gurnee & Tegmark, 2023) Next is the idea that the probability of a concept value can be measured with a linear probe. For example, the probability that the output language is French is logit-linear in the representation of the input. In this case, we can take the linear map to be a representation of the concept of `English/French`.
3. **Intervention:** (e.g., Wang et al., 2023; Turner et al., 2023) The final idea is that the value a concept takes on can be changed, without changing other concepts, by adding a suitable steering vector—e.g., we change the output from English to French by adding an `English/French` vector. In this case, we take this added vector to be a representation of the concept.

It is not clear a priori how these ideas relate to each other, nor which is the "right" notion of linear representation.

Next, suppose we have somehow found the linear representations of various concepts. We can then use linear algebraic operations on the representation space for interpretation and control. For example, we might compute the cosine similarity between a representation and known concept directions, or edit representations projected onto target directions. However, similarity and projection are geometric notions: they require an inner product on the representation space. The second challenge is that it is not clear which inner product

[1]University of Chicago, Illinois, USA.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

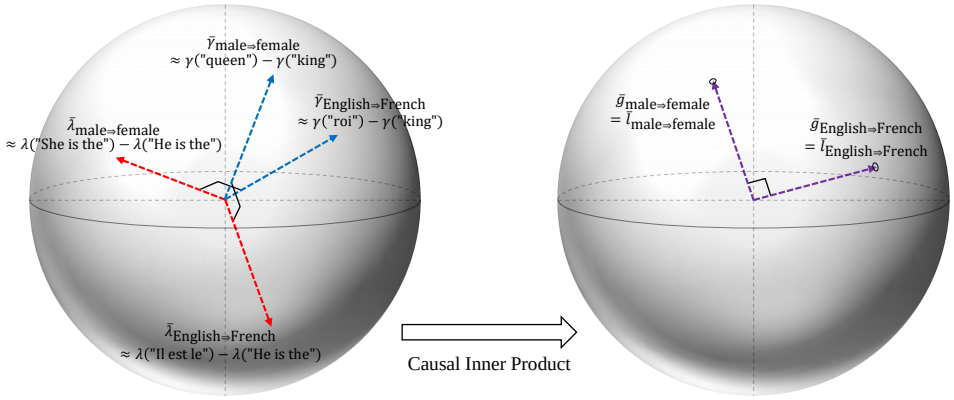

Figure 1. The geometry of linear representations can be understood in terms of a *causal inner product* that respects the semantic structure of concepts. In a language model, each concept has two separate linear representations, $\bar{\lambda}$ (red) in the embedding (input context) space and $\bar{\gamma}$ (blue) in the unembedding (output word) space, as drawn on the left. The causal inner product induces a linear transformation for the representation spaces such that the transformed linear representations coincide (purple), as drawn on the right. In this unified representation space, causally separable concepts are represented by orthogonal vectors.

is appropriate for understanding model representations.

To address these, we make the following contributions:

1. First, we formalize the subspace notion of linear representation in terms of counterfactual pairs, in both "embedding" (input context) and "unembedding" (output word) spaces. Using this formalization, we prove that the unembedding notion connects to measurement, and the embedding notion to intervention.
2. Next, we introduce the notion of a *causal inner product*: an inner product with the property that concepts that can vary freely of each other are represented as orthogonal vectors. We show that such an inner product has the special property that it unifies the embedding and unembedding representations, as illustrated in Figure 1. Additionally, we show how to estimate the inner product using the LLM unembedding matrix.
3. Finally, we study the linear representation hypothesis empirically using LLaMA-2 (Touvron et al., 2023). We find the subspace notion of linear representations for a variety of concepts. Using these, we give evidence that the causal inner product respects semantic structure, and that subspace representations can be used to construct measurement and intervention representations.

Background on Language Models We will require some minimal background on (large) language models. Formally, a language model takes in context text x and samples output text. This sampling is done word by word (or token by token). Accordingly, we'll view the outputs as single words. To define a probability distribution over outputs, the language model first maps each context x to a vector $\lambda(x)$ in a representation space $\Lambda \simeq \mathbb{R}^d$. We will call these *embedding vectors*. The model also represents each word y as an *unembedding vector* $\gamma(y)$ in a separate representation space $\Gamma \simeq \mathbb{R}^d$. The probability distribution over the next words is then given by the softmax distribution:

$$\mathbb{P}(y \mid x) \propto \exp(\lambda(x)^\top \gamma(y)).$$

2. The Linear Representation Hypothesis

We begin by formalizing the subspace notion of linear representation, one in each of the unembedding and embedding spaces of language models, and then tie the subspace notions to the measurement and intervention notions.

2.1. Concepts

The first step is to formalize the notion of a concept. Intuitively, a concept is any factor of variation that can be changed in isolation. For example, we can change the output from French to English without changing its meaning, or change the output from being about a man to about a woman without changing the language it is written in.

Following Wang et al. (2023), we formalize this idea by taking a *concept variable* W to be a latent variable that is caused by the context X, and that acts as a cause of the output Y. For simplicity of exposition, we will restrict attention to binary concepts. Anticipating the representation of concepts by vectors, we introduce an ordering on each binary concept—e.g., `male⇒female`. This ordering makes the sign of a representation meaningful (e.g., the representation of `female⇒male` will have the opposite sign).

Each concept variable W defines a set of counterfactual outputs $\{Y(W = w)\}$ that differ only in the value of W. For example, for the concept `male⇒female`, $(Y(0), Y(1))$ is a random element of the set {("man", "woman"), ("king",

"queen"), ...}. In this paper, we assume the value of concepts can be read off deterministically from the sampled output (e.g., the output "king" implies $W = 0$). Then, we can specify concepts by specifying their corresponding counterfactual outputs.

We will eventually need to reason about the relationships between multiple concepts. We say that two concepts W and Z are *causally separable* if $Y(W = w, Z = z)$ is well-defined for each w, z. That is, causally separable concepts are those that can be varied freely and in isolation. For example, the concepts `English⇒French` and `male⇒female` are causally separable—consider {"king", "queen", "roi", "reine"}. However, the concepts `English⇒French` and `English⇒Russian` are not because they cannot vary freely.

We'll write $Y(W = w, Z = z)$ as $Y(w, z)$ when the concepts are clear from context.

2.2. Unembedding Representations and Measurement

We now turn to formalizing linear representations of a concept. The first observation is that there are two distinct representation spaces in play—the embedding space Λ and the unembedding space Γ. A concept could be linearly represented in either space. We begin with the unembedding space. Defining the cone of vector v as $\mathrm{Cone}(v) = \{\alpha v : \alpha > 0\}$,

Definition 2.1 (Unembedding Representation)**.** We say that $\bar{\gamma}_W$ is an *unembedding representation* of a concept W if $\gamma(Y(1)) - \gamma(Y(0)) \in \mathrm{Cone}(\bar{\gamma}_W)$ almost surely.

This definition captures the subspace notion in the unembedding space, e.g., that $\gamma(\text{"queen"}) - \gamma(\text{"king"})$ is parallel to $\gamma(\text{"woman"}) - \gamma(\text{"man"})$. We use a cone instead of subspace because the sign of the difference is significant—i.e., the difference between "king" and "queen" is in the opposite direction as the difference between "woman" and "man". The unembedding representation (if it exists) is unique up to positive scaling, consistent with the linear subspace hypothesis that concepts are represented as directions.

Connection to Measurement The first result is that the unembedding representation is closely tied to the measurement notion of linear representation:

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 $\lambda \in \Lambda$,*

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

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

All proofs are given in Appendix B.

In words: if we know the output token is either "king" or "queen" (say, the context was about a monarch), then the probability that the output is "king" is logit-linear in the language model representation with regression coefficients $\bar{\gamma}_W$. The random scalar α is a function of the particular counterfactual pair $\{Y(0), Y(1)\}$—e.g., it may be different for {"king", "queen"} and {"roi", "reine"}. However, the direction used for prediction is the same for all counterfactual pairs demonstrating the concept.

Theorem 2.2 shows a connection between the subspace representation and the linear representation learned by fitting a linear probe to predict the concept. Namely, in both cases, we get a predictor that is linear on the logit scale. However, the unembedding representation differs from a probe-based representation in that it does not incorporate any information about correlated but off-target concepts. For example, if French text were disproportionately about men, a probe could learn this information (and include it in the representation), but the unembedding representation would not. In this sense, the unembedding representation might be viewed as an ideal probing representation.

2.3. Embedding Representations and Intervention

The next step is to define a linear subspace representation in the embedding space Λ. We'll again go with a notion anchored in demonstrative pairs. In the embedding space, each $\lambda(x)$ defines a distribution over concepts. We consider pairs of sentences such as $\lambda_0 = \lambda(\text{"He is the monarch of England, "})$ and $\lambda_1 = \lambda(\text{"She is the monarch of England, "})$ that induce different distributions on the target concept, but the same distribution on all off-target concepts. A concept is embedding-represented if the differences between all such pairs belong to a common subspace. Formally,

Definition 2.3 (Embedding Representation)**.** We say that $\bar{\lambda}_W$ is an *embedding representation* of a concept W if we have $\lambda_1 - \lambda_0 \in \mathrm{Cone}(\bar{\lambda}_W)$ for any context embeddings $\lambda_0, \lambda_1 \in \Lambda$ that satisfy

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

for each concept Z that is causally separable with W.

The first condition ensures that the direction is relevant to the target concept, and the second condition ensures that the direction is not relevant to off-target concepts.

Connection to Intervention It turns out the embedding representation is closely tied to the intervention notion of linear representation. For this, we need the following lemma relating embedding and unembedding representations.