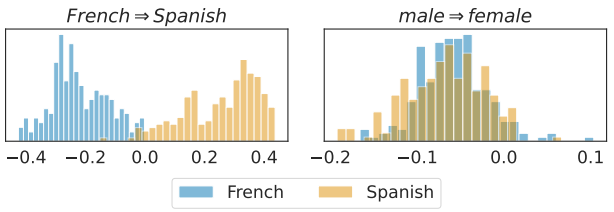

Figure 4. The subspace representation $\bar{\gamma}_W$ acts as a linear probe for W. The histograms show $\bar{\gamma}_W^\top \lambda(x_j^{\text{fr}})$ vs. $\bar{\gamma}_W^\top \lambda(x_j^{\text{es}})$ (left) and $\bar{\gamma}_Z^\top \lambda(x_j^{\text{fr}})$ vs. $\bar{\gamma}_Z^\top \lambda(x_j^{\text{es}})$ (right) for $W = \texttt{French}\Rightarrow\texttt{Spanish}$ and $Z = \texttt{male}\Rightarrow\texttt{female}$, where $\{x_j^{\text{fr}}\}$ and $\{x_j^{\text{es}}\}$ are random contexts from French and Spanish Wikipedia, respectively. We also see that $\bar{\gamma}_Z$ does *not* act as a linear probe for W, as expected.

This arises because the concepts are grouped by semantic similarity. For example, the first 10 concepts relate to verbs, and the last 4 concepts are language pairs. The additional non-zero structure also generally makes sense. For example, `lower⇒upper` (capitalization, concept 19) has non-trivial inner product with the language pairs *other than* `French⇒Spanish`. This may be because French and Spanish obey similar capitalization rules, while English and German each have different conventions (e.g., German capitalizes all nouns, but English only capitalizes proper nouns). In Appendix D.2, we compare the Euclidean inner product to the causal inner product for both the LLaMA-2 model and a more recent Gemma large language model (Mesnard et al., 2024).

Concept directions act as linear probes Next, we check the connection to the measurement notion of linear representation. We consider the concept $W = \texttt{French}\Rightarrow\texttt{Spanish}$. To construct a dataset of French and Spanish contexts, we sample contexts of random lengths from Wikipedia pages in each language. Note that these are *not* counterfactual pairs. Following Theorem 2.2, we expect $\bar{\gamma}_W^\top \lambda(x_j^{\text{fr}}) < 0$ and $\bar{\gamma}_W^\top \lambda(x_j^{\text{es}}) > 0$. Figure 4 confirms this expectation, showing that $\bar{\gamma}_W$ is a linear probe for the concept W in Λ (left). Also, the representation of an off-target concept $Z = \texttt{male}\Rightarrow\texttt{female}$ does not have any predictive power for this task (right). Appendix D.3 includes analogous results using all 27 concepts.

Concept directions map to intervention representations Theorem 2.5 says that we can construct an intervention representation by constructing an embedding representation. Doing this directly requires finding pairs of prompts that vary only on the distribution they induce on the target concept, which can be difficult to find in practice.

Here, we will instead use the isomorphism between embedding and unembedding representations (Theorem 3.2) to construct intervention representations from unembedding

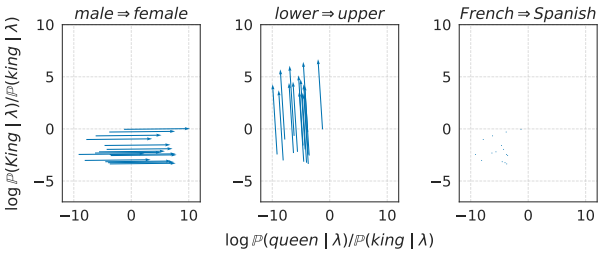

Figure 5. Adding $\alpha\bar{\lambda}_C$ to λ changes the target concept C without changing off-target concepts. The plots illustrate change in $\log(\mathbb{P}(\text{“queen”} \mid x)/\mathbb{P}(\text{“king”} \mid x))$ and $\log(\mathbb{P}(\text{“King”} \mid x)/\mathbb{P}(\text{“king”} \mid x))$, after changing $\lambda(x_j)$ to $\lambda_{C,\alpha}(x_j)$ as α increases from 0 to 0.4, for $C = \texttt{male}\Rightarrow\texttt{female}$ (left), $\texttt{lower}\Rightarrow\texttt{upper}$ (center), $\texttt{French}\Rightarrow\texttt{Spanish}$ (right). The two ends of the arrow are $\lambda(x_j)$ and $\lambda_{C,0.4}(x_j)$, respectively. Each context x_j is presented in Table 4.

representations. We take

$$\bar{\lambda}_W := \text{Cov}(\gamma)^{-1}\bar{\gamma}_W. \tag{4.1}$$

Theorem 2.5 predicts that adding $\bar{\lambda}_W$ to a context representation should increase the probability of W, while leaving the probability of all causally separable concepts unaltered.

To test this for a given pair of causally separable concepts W and Z, we first choose a quadruple $\{Y(w,z)\}_{w,z\in\{0,1\}}$, and then generate contexts $\{x_j\}$ such that the next word should be $Y(0,0)$. For example, if $W = \texttt{male}\Rightarrow\texttt{female}$ and $Z = \texttt{lower}\Rightarrow\texttt{upper}$, then we choose the quadruple (“king”, “queen”, “King”, “Queen”), and generate contexts using ChatGPT-4 (e.g., “Long live the”). We then intervene on $\lambda(x_j)$ using $\bar{\lambda}_C$ via

$$\lambda_{C,\alpha}(x_j) = \lambda(x_j) + \alpha\bar{\lambda}_C, \tag{4.2}$$

where $\alpha > 0$ and C can be W, Z, or some other causally separable concept (e.g., $\texttt{French}\Rightarrow\texttt{Spanish}$). For different choices of C, we plot the changes in $\text{logit}\,\mathbb{P}(W = 1 \mid Z, \lambda)$ and $\text{logit}\,\mathbb{P}(Z = 1 \mid W, \lambda)$, as we increase α. We expect to see that, if we intervene in the W direction, then the intervention should linearly increase $\text{logit}\,\mathbb{P}(W = 1 \mid Z, \lambda)$, while the other logit should stay constant; if we intervene in a direction C that is causally separable with both W and Z, then we expect both logits to stay constant.

Figure 5 shows the results of one such experiment shown for three target concepts (24 others shown in Appendix D.4), confirming our expectations. We see, for example, that intervening in the $\texttt{male}\Rightarrow\texttt{female}$ direction raises the logit for choosing “queen” over “king” as the next word, but does not change the logit for “King” over “king”.

A natural follow-up question is to see if the intervention in a concept direction (for W) pushes the probability of $Y(W = 1)$ being the next word to be the largest among all

Table 1. Adding the intervention representation $\alpha\bar{\lambda}_W$ pushes the probability over completions to reflect the concept W. As the scale of intervention increases, the probability of seeing $Y(W = 1)$ ("**queen**") increases while the probability of seeing $Y(W = 0)$ ("*king*") decreases. We show the top-5 most probable words over the entire vocabulary following the intervention (4.2) in the $W =$ `male⇒female` direction, i.e., $\lambda_{W,\alpha}(x) = \lambda(x) + \alpha\bar{\lambda}_W$, for $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4\}$. The original context $x =$ "Long live the " is a sentence fragment that ends with the word $Y(W = 0)$ ("*king*"). The most likely words reflect the concept, with "**queen**" being top-1. In Appendix D.5, we provide more examples.

Rank	$\alpha = 0$	0.1	0.2	0.3	0.4
1	*king*	Queen	**queen**	**queen**	**queen**
2	King	**queen**	Queen	Queen	Queen
3	Queen	*king*	_	lady	lady
4	**queen**	King	lady	woman	woman
5	_	_	*king*	women	women

tokens. We expect to see that, as we increase the value of α, the target concept should eventually be reflected in the most likely output words according to the LM.

In Table 1, we show an illustrative example in which W is the concept `male⇒female` and the context x is a sentence fragment that can end with the word $Y(W = 0)$ ("*king*"). For $x =$ "Long live the ", as we increase the scale α on the intervention, we see that the target word $Y(W = 1)$ ("**queen**") becomes the most likely next word, while the original word $Y(W = 0)$ drops below the top-5 list. This illustrates how the intervention can push the probability of the target word high enough to make it the most likely word while decreasing the probability of the original word.

5. Discussion and Related Work

The idea that high-level concepts are encoded *linearly* is appealing because—if it is true—it may open up simple methods for interpretation and control of LLMs. In this paper, we have formalized 'linear representation', and shown that all natural variants of this notion can be unified.[6] This equivalence already suggests some approaches for interpretation and control—e.g., we show how to use collections of pairs of words to define concept directions, and then use these directions to predict what the model's output will be, and to change the output in a controlled fashion. A major theme is the role played by the choice of inner product.

Linear subspaces in language representations The linear subspace hypotheses was originally observed empirically in the context of word embeddings (e.g., Mikolov et al., 2013b;c; Levy & Goldberg, 2014; Goldberg & Levy, 2014; Vylomova et al., 2016; Gladkova et al., 2016; Chiang et al., 2020; Fournier et al., 2020). Similar structure has been observed in cross-lingual word embeddings (Mikolov et al., 2013a; Lample et al., 2018; Ruder et al., 2019; Peng et al., 2022), sentence embeddings (Bowman et al., 2016; Zhu & de Melo, 2020; Li et al., 2020; Ushio et al., 2021), representation spaces of Transformer LLMs (Meng et al., 2022; Merullo et al., 2023; Hernandez et al., 2023), and vision-language models (Wang et al., 2023; Trager et al., 2023; Perera et al., 2023). These observations motivate Definition 2.1. The key idea in the present paper is providing formalization in terms of counterfactual pairs—this is what allows us to connect to other notions of linear representation, and to identify the inner product structure.

Measurement, intervention, and mechanistic interpretability There is a significant body of work on linear representations for interpreting (probing) (e.g., Alain & Bengio, 2017; Kim et al., 2018; nostalgebraist, 2020; Rogers et al., 2021; Belinkov, 2022; Li et al., 2022; Geva et al., 2022; Nanda et al., 2023) and controlling (steering) (e.g., Wang et al., 2023; Turner et al., 2023; Merullo et al., 2023; Trager et al., 2023) models. This is particularly prominent in *mechanistic interpretability* (Elhage et al., 2021; Meng et al., 2022; Hernandez et al., 2023; Turner et al., 2023; Zou et al., 2023; Todd et al., 2023; Hendel et al., 2023). With respect to this body of work, the main contribution of the present paper is to clarify the linear representation hypothesis, and the critical role of the inner product. However, we do not address interpretability of either model parameters, nor the activations of intermediate layers. These are main focuses of existing work. It is an exciting direction for future work to understand how ideas here—particularly, the causal inner product—translate to these settings.

Geometry of representations There is a line of work that studies the geometry of word and sentence representations (e.g., Arora et al., 2016; Mimno & Thompson, 2017; Ethayarajh, 2019; Reif et al., 2019; Li et al., 2020; Hewitt & Manning, 2019; Chen et al., 2021; Chang et al., 2022; Jiang et al., 2023). This work considers, e.g., visualizing and modeling how the learned embeddings are distributed, or how hierarchical structure is encoded. Our work is largely orthogonal to these, since we are attempting to define a suitable inner product (and thus, notions of similarity and projection) that respects the semantic structure of language.

Causal representation learning Finally, the ideas here connect to causal representation learning (e.g., Higgins et al., 2016; Hyvarinen & Morioka, 2016; Higgins et al., 2018; Khemakhem et al., 2020; Zimmermann et al., 2021; Schölkopf et al., 2021; Moran et al., 2021; Wang et al., 2023). Most obviously, our causal formalization of concepts is inspired by Wang et al. (2023), who establish a characterization of latent concepts and vector algebra in dif-

[6] In Appendix A, we summarize these results in a figure.

fusion models. Separately, a major theme in this literature is the identifiability of learned representations—i.e., to what extent they capture underlying real-world structure. Our causal inner product results may be viewed in this theme, showing that an inner product respecting semantic closeness is not identified by the usual training procedure, but that it can be picked out with a suitable assumption.

Acknowledgements

Thanks to Gemma Moran for comments on an earlier draft. This work is supported by ONR grant N00014-23-1-2591 and Open Philanthropy.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL https://aclanthology.org/K16-1002.

Chang, T., Tu, Z., and Bergen, B. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, 2022.

Chen, B., Fu, Y., Xu, G., Xie, P., Tan, C., Chen, M., and Jing, L. Probing BERT in hyperbolic spaces. In *International Conference on Learning Representations*, 2021.

Chiang, H.-Y., Camacho-Collados, J., and Pardos, Z. Understanding the source of semantic regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 119–131, 2020.

Choe, Y. J., Park, K., and Kim, D. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3036–3045, 2020.

Drozd, A., Gladkova, A., and Matsuoka, S. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers*, pp. 3519–3530, 2016.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Fournier, L., Dupoux, E., and Dunbar, E. Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 365–375, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.29. URL https://aclanthology.org/2020.conll-1.29.

Geva, M., Caciularu, A., Wang, K., and Goldberg, Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, 2022.

Gladkova, A., Drozd, A., and Matsuoka, S. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pp. 8–15, 2016.

Goldberg, Y. and Levy, O. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, art. arXiv:2310.02207, October 2023. doi: 10.48550/arXiv.2310.02207.

Hendel, R., Geva, M., and Globerson, A. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*, 2023.