Embeddability of Analogies as Parallelograms in S^{d-1}

Narutatsu (Edward) Ri

Department of Computer Science Columbia University wl2787@columbia.edu

December 13, 2022

Abstract

Static word embeddings exhibit *additive composition* properties that facilitate the recovery of linguistic analogies through vector arithmetic. However, existing analyses have predominantly neglected the impact of vector length normalization, despite its common application in downstream tasks. This paper investigates the capabilities of length-normalized word embeddings in representing word analogies as parallelograms within the embedding space. We formulate the embedding of words on the unit sphere in \mathbb{R}^d as a system of linear equations and demonstrate that the required representation dimension scales with the number of analogy conditions. Our findings provide insights into the geometric structure of normalized embeddings and their capacity to capture different analogy relationship configurations.

1 Introduction

Word embeddings, such as those generated by Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014), are trained so that inner products between word vectors approximate the co-occurrence statistics of their corresponding words. A well-studied property of these embeddings is *additive compositionality* (Mikolov et al., 2013b), which enables simple vector arithmetic to recover linguistic analogies (e.g., "man : woman :: king : queen") as parallel structures. Motivated by empirical observations in Word2Vec and related embeddings, various theoretical studies have emerged to explain how these models implicitly learn to encode analogies as parallelograms.

Despite the common practice of normalizing word embeddings for downstream tasks, the role of vector length remains underexplored (Schakel and Wilson, 2015). Few studies have systematically examined the geometric properties of normalized word vectors or assessed their limitations in representing analogy constraints. To address this gap, we investigate conditions for embedding n word vectors in a d-dimensional Euclidean space while satisfying C analogy constraints.

Our analysis shows that certain geometric configurations, specifically those representing analogy relationships as parallelograms (or rhomboids), cannot be realized on the unit sphere under mild assumptions. From this finding, we derive lower bounds on the dimensionality required to support multiple analogy constraints, demonstrating that the necessary embedding dimension scales linearly with the number of analogies that share words. Since words typically participate in numerous analogy relationships, this result suggests that a high-dimensional space is required to represent these constraints perfectly. As our findings rely on minimal assumptions regarding word similarity or distributional properties, they establish a fundamental baseline on the dimension necessary for representing analogy structures.

2 Related Work

Word Embedding Models. Dense, low-dimensional vector representations were popularized by Word2Vec and the GloVe models. Various alternative representations have been explored (Bojanowski et al., 2016; Seonwoo et al., 2019); however, Word2Vec remains the most widely adopted approach. In recent years, there has been growing interest in static word embeddings within non-Euclidean spaces (Leimeister and Wilson, 2018; Nurmukhamedov et al., 2022; Dhingra et al., 2018; Meng et al., 2019; Nickel and Kiela, 2017; Tifrea et al., 2018). These studies demonstrate that embedding word vectors in non-Euclidean geometries offers empirical advantages, such as the ability to capture hierarchical word similarities.

Theory of Word Embeddings. Initial efforts to elucidate the mechanisms underlying word embedding models were undertaken by Levy and Goldberg (2014), who proposed that Word2Vec implicitly factorizes the shifted *Pointwise Mutual Information* (PMI)¹ matrix. Subsequent work by Li et al. (2015) attempted to demonstrate that Word2Vec explicitly factorizes the word co-occurrence matrix. Additionally, Hashimoto et al. (2016) framed the learning of dense word embeddings as a metric recovery problem within a vector space over concepts, wherein Euclidean distances between vectors are assumed to represent semantic similarities based on word co-occurrence statistics.

Early investigations into how Word2Vec implicitly captures analogical relationships were conducted by Arora et al. (2015), who suggested that arbitrary word pairs form parallel lines when word embeddings recover PMI statistics through vector products under specific generative assumptions about the training corpora. Follow-up work by Gittens et al. (2017) introduced the concept of a *paraphrase* and provided an explanation for the formation of analogies based on paraphrases. Recent studies have built upon the framework established by Gittens et al. (2017), refining the formulation of analogies as paraphrases and demonstrating that the conditions proposed by Gittens et al. (2017) hold under less restrictive assumptions (Allen and Hospedales, 2019; Allen et al., 2018; Ethayarajh et al., 2019).

3 Problem Construction

3.1 Notation

Let *n* denote the number of words to embed, *C* denote the number of analogy relationships, and *d* denote the dimension of the embedding space. Denote the words as a, b, \ldots , with their corresponding vector representations given by $\mathbf{a}, \mathbf{b}, \ldots$ in \mathbb{R}^d . Vectors in general will be denoted by \mathbf{x} . We first present preliminary definitions to formalize the notion of analogy relationships.

Definition 3.1 (Relation). A relation is defined as an ordered semantic relationship between any two words a, b, expressed as $r_{a,b} = (a:b)$. Denote a as the source and b as the sink.

¹For words *i* and *j*, PMI is defined as PMI(*i*, *j*) = log $\frac{\Pr[i,j]}{\Pr[i]\Pr[j]}$, where $\Pr[i,j]$ denotes the co-occurrence probability of words *i* and *j* and $\Pr[i]$ denotes the occurrence probability of word *i*.

Relations are not symmetric $(r_{a,b} \neq r_{b,a})$. An example is " $r_{\text{man,woman}} = (\text{man} : \text{woman})$," where $r_{\text{man,woman}}$ represents the relation of change of gender from male to female, which is not equivalent to $r_{\text{woman,man}}$. Note that the explicit representation of a relation need not be known.

Definition 3.2 (Analogy). An *analogy* is a relationship that exists between two relations if and only if the two relations are equivalent, and we say the analogy *expresses* the relation.

For example, for relations $r_{a,b} = (a : b)$ and $r_{c,d} = (c : d)$, the relations $r_{a,b}$ and $r_{c,d}$ form an analogy if and only if $r_{a,b} = r_{c,d}$. Then, the words a, b, c, d satisfy (a : b) = (c : d).

Definition 3.3 (Concept). A concept c_i is defined as a set of equivalent relations. Each concept expresses a relation, denoted by r_{c_i} . The set of words participating in the concept is represented by w_{c_i} . For each relation, the set of words that act as sources is referred to as the source set, while the set of words that act as sinks within each analogy is referred to as the sink set.

For example, for words a, b, c, d, e, f, if $r_{a,b} = r_{c,d}$ and $r_{a,b} = r_{e,f}$, then there exists a concept for the set of points participating in the relations, expressed as $c = \{(a:b), (c:d), (e:f)\}$. The source set is $\{a, c, e\}$ and the sink set is $\{b, d, f\}$, and $r_c = r_{a,b}(=r_{c,d} = r_{e,f})$. Additionally, $w_c = \{a, b, c, d, e, f\}$. Note that concepts must contain an even number of words, multiple concepts can contain the same word, and the relation a concept expresses must be unique.

Next, we define three types of word overlap between concepts.

Definition 3.4 (Weak Overlap). Two concepts c_1, c_2 are weakly overlapping when $|w_{c_1} \cap w_{c_2}| = 1$.

Definition 3.5 (Strong Overlap). k concepts c_1, c_2, \ldots, c_k are strongly overlapping when c_1, \ldots, c_k either have the same source set or sink set.

Definition 3.6 (Strict Overlap). Two concepts c_1, c_2 are strictly overlapping when $w_{c_1} = w_{c_2}$.

Note that by Assumption 3, as $|w_{c_1}|, |w_{c_2}| \equiv 0 \mod 2$ holds true, all points in w_{c_1} and w_{c_2} each participate in exactly 2 relations expressing each concept when c_1, c_2 strictly overlap.

Definition 3.7 (Embeddability). n words and concepts C are *embeddable* if the vector representations for all n words can be placed on a specified subspace in dimension d while preserving all analogy conditions for all concepts.

While other arrangements of overlap between concepts are possible, we will focus on the embeddability of strong and strict overlaps as they are the most common and interesting forms of analogies that occur in natural language.

3.2 Assumptions

Assumption 1. All *n* words participate in at least one concept.

This assumption implies that the embeddings for any of the n words cannot be freely chosen. Otherwise, the embedding of the word can be arbitrarily placed in the embedding space.

Assumption 2. The set of concepts is mutually consistent; that is, they can be embedded in some dimension d without contradiction.

Assumption 3. For two concepts c_1 and c_2 , if $|w_{c_1} \cap w_{c_2}| > 2$ and $|w_{c_1} \cap w_{c_2}| \equiv 0 \mod 2$ (i.e., they share an even number of points greater than 2), then each shared point must form both the relation r_{c_1} with another shared point and the relation r_{c_2} with another different shared point.

Assumption 4. For two concepts c_1 and c_2 , if $|w_{c_1} \cap w_{c_2}| > 1$ and $|w_{c_1} \cap w_{c_2}| \equiv 1 \mod 2$ (i.e., they share an odd number of points greater than 1), then the shared points must all be either only sinks or only sources.

For example, given words a, b, c, d, e, f, g, h, if we assume concept c_1 expresses the relation $r_1 = r_{a,b} = r_{c,d} = r_{e,f} = r_{g,h} = (a:b): (c:d): (e:f): (g:h)$, then the second concept denoted c_2 can only express a relation equivalent to $r_2 = r_{a,c}: r_{b,d}: r_{e,g}: r_{f,h}$ and cannot be any arbitrarily chosen set of relations.

Assumptions 3 and 4 are attributed to how analogies are generally observed to form in natural language. For example, for the words "man, woman, king, queen, boy, girl, prince, princess," there exists a relation expressing the concept of change of gender from masculine to feminine where the analogies are "(man : woman) = (king : queen) = (boy : girl) = (prince : princess)." There can also exist a relation expressing the concept of royalty where the analogies are "(man : king) = (woman : queen) = (boy : prince) = (girl : princess)," but analogies that attempt to represent the relationship between "(man : queen)" are drastically rarer.

4 Embeddability of Analogies as Parallelograms on S^{d-1}

Here, we establish various properties of points on the surface of a unit ℓ_2 ball in d dimensions. The unit ℓ_2 ball is denoted as S^d . All proofs are provided in the Appendix.

Lemma 4.1. For any two distinct points $\mathbf{x}, \mathbf{y} \in S^d$, where $S^d \subset \mathbb{R}^d$, the vector $\mathbf{l} = \mathbf{x} - \mathbf{y}$ satisfies the following properties:

- $\mathbf{x} \cdot \mathbf{l} = -\mathbf{y} \cdot \mathbf{l}$, and
- $\|\mathbf{x} \operatorname{proj}_{\mathbf{l}}(\mathbf{x})\|_2 = \|\mathbf{y} \operatorname{proj}_{\mathbf{l}}(\mathbf{y})\|_2$,

where $\operatorname{proj}_{\mathbf{l}}(\mathbf{x})$ and $\operatorname{proj}_{\mathbf{l}}(\mathbf{y})$ denote the projection of \mathbf{x} and \mathbf{y} onto \mathbf{l} , respectively.

Lemma 4.1 implies that for any two arbitrary points, there always exists a direction in the space such that the two points have equal ℓ_2 distance from that direction. If the points are translated so that this direction aligns with a scalar multiple of a basis vector, then the two points will share identical values for all entries except one, where the value of one point is the negation of the corresponding entry of the other point. For simplicity, we refer to such a direction as the *axis*.

Moreover, we establish a property that must be satisfied by four points on S^d in \mathbb{R}^d that form a parallelogram.

Lemma 4.2. Consider two distinct points $\mathbf{x}, \mathbf{y} \in S^d$, where $S^d \subset \mathbb{R}^d$, and a corresponding axis \mathbf{l} such that the conditions of Lemma 4.1 are satisfied.

For two additional distinct points $\mathbf{a}, \mathbf{b} \in S^d$, if $\mathbf{x} - \mathbf{y} = \mathbf{b} - \mathbf{a}$, then the following holds:

$$\|\mathbf{x} - \text{proj}_{\mathbf{l}}(\mathbf{x})\|_{2}^{2} = \|\mathbf{y} - \text{proj}_{\mathbf{l}}(\mathbf{y})\|_{2}^{2} = \|\mathbf{a} - \text{proj}_{\mathbf{l}}(\mathbf{a})\|_{2}^{2} = \|\mathbf{b} - \text{proj}_{\mathbf{l}}(\mathbf{b})\|_{2}^{2}$$

Lemma 4.2 establishes the condition that any four points must satisfy to form an analogy geometrically. For all pairs of words within a concept, the corresponding word vectors must adhere to this condition. Geometrically, the region where such points can reside can be described as the set of points on the surface of S^d that are at a fixed distance r from a given axis **l**.

We now formalize the above into a mathematical expression for the set of points:

Lemma 4.3. For a concept $c \in C$, given an arbitrary unit vector \mathbf{l} whose direction represents the axis and a radius 0 < r < 1, the embeddings \mathbf{x} of the words participating in c satisfy the following property:

$$\sum_{i=1}^{d} \mathbf{x}_i \mathbf{l}_i = \sqrt{1 - r^2},$$

where \mathbf{x}_i and \mathbf{l}_i denote the *i*th entries of the vectors \mathbf{x} and \mathbf{l} , respectively.

Note that not every set of points \mathbf{x} satisfying the equality $\sum_{i=1}^{d} \mathbf{x}_i \mathbf{l}_i = \sqrt{1 - r^2}$ corresponds to a valid embedding for a word. Rather, satisfying this equality is a necessary condition: if a word participates in a concept, its embedding must satisfy the equality. However, the converse does not necessarily hold; satisfying the equality does not guarantee that the point corresponds to a word embedding.

Now, consider two concepts $c_1, c_2 \in C$ that have overlapping points. By Lemma 4.3, the regions where embeddings can reside while participating in each concept can be represented as follows:

$$\sum_{i=1}^{d} \mathbf{x}_{i}^{(1)} \mathbf{l}_{i}^{(1)} = \sqrt{1 - (r^{(1)})^{2}},$$
$$\sum_{i=1}^{d} \mathbf{x}_{i}^{(2)} \mathbf{l}_{i}^{(2)} = \sqrt{1 - (r^{(2)})^{2}},$$

where $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ represent the embeddings for the regions corresponding to concepts c_1 and c_2 , respectively, and $\mathbf{l}^{(1)}$ and $\mathbf{l}^{(2)}$ denote the axes of the respective regions. Thus, for embeddings that participate in both concepts, we need to solve for the set of points \mathbf{x} that satisfy both equations. Determining the regions where overlapping points can reside reduces to solving the system of linear equations along with the nonlinear constraint that \mathbf{x} must lie on S^d , given by $\|\mathbf{x}\|_2 = 1$.

4.1 Dimensionality for Embedding Analogies

With the above constructions, we analyze the relationship between the dimensionality d and the analogy relationships C for embedding n points on S^d .

Theorem 4.4 (Embeddability of Disjoint Concepts). If $\forall i, j \in [|C|], i \neq j$, it holds that $|w_{c_i} \cap w_{c_j}| = 0$, then for any $n \geq 4$, all n points can be embedded on S^2 .

Theorem 4.5 (Embeddability of Weak Overlaps). If $\exists i, j \in [|C|], i \neq j$, such that $|w_{c_i} \cap w_{c_j}| = 1$, then for any $n \geq 4$, all n points can be embedded on S^2 .

Note that the statement still holds even if there are more than two overlapping concepts involving the same word, as we can choose an arbitrary number of linear equations that include a particular point in their region.

Now, we consider the case where there exist concepts with strict overlap.

Theorem 4.6 (Embeddability of Strong Overlaps). For any $n \ge 4$, if $k \le |C|$ concepts strongly overlap, then the concepts can be embedded in S^{d-1} , where $d = \Omega(k)$.

Before we consider strong overlaps between k concepts, we first establish a condition that the axes \mathbf{l}_i need to satisfy for all k analogy conditions to hold in the embeddings.

Lemma 4.7. For an arbitrary set of four words a, b, c, d and concepts c_1, c_2 , if the analogies (a : b) = (c : d) and (a : c) = (b : d) hold and correspond to the concepts c_1 and c_2 , respectively, then the axes \mathbf{l}_1 and \mathbf{l}_2 associated with each concept must be orthogonal.

Lemma 4.7 can be trivially generalized to k concepts with strict overlap, in which case at least k orthogonal axes are required.

Theorem 4.8 (Embeddability of Strict Overlaps). For any $n \ge 4$, if $k \le |C|$ concepts strictly overlap, then C can be embedded in S^{d-1} for any $d = \Omega(k)$.

The contrast between Lemma 4.7 and Theorem 4.8 lies in the orthogonality of the axes. This distinction has no effect on the system of equations, as the coefficients remain independent.

Note that k strictly overlapping concepts can be embedded in d = k + 1 dimensions, but with a limitation on the number of words that can participate in the concepts. Specifically, this construction allows at most 2^{k+1} embeddable words, as the common region determined by k systems of linear equations, combined with the unit norm constraint, results in two possible values for each dimension.

5 Conclusion

We have investigated the relationship between concept overlaps and the minimum dimensionality required to preserve all analogy conditions. By analyzing the interaction between analogy overlaps and dimensionality, our findings provide insights into the dimensionality requirements for word embeddings and their capacity to encode linguistic analogies.

It is important to note that our approach focused on identifying the feasible regions where points can reside to satisfy analogy constraints, rather than precisely determining their exact positions in space. Once these regions are identified, the arrangement of points can be achieved by sequentially positioning them while strategically selecting appropriate words to ensure the embedding satisfies all required conditions.

References

- Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings, 2019. URL https://arxiv.org/abs/1901.09813.
- Carl Allen, Ivana Balažević, and Timothy Hospedales. What the vec? towards probabilistically grounded embeddings, 2018. URL https://arxiv.org/abs/1805.12164.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings, 2015. URL https://arxiv.org/abs/1502.03520.

- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606, 2016.
- Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi:10.18653/v1/W18-1708. URL https://aclanthology.org/W18-1708.
- Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations, 2019. URL https://arxiv.org/abs/1908.06361.
- Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. Skip-gram Zipf + uniform = vector additivity. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 69–76, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi:10.18653/v1/P17-1007. URL https://aclanthology.org/ P17-1007.
- Tatsunori B. Hashimoto, David Alvarez-Melis, and Tommi S. Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4: 273–286, 2016. doi:10.1162/tacl_a_00098. URL https://aclanthology.org/Q16-1020.
- Matthias Leimeister and Benjamin J. Wilson. Skip-gram word embeddings in hyperbolic space, 2018. URL https://arxiv.org/abs/1809.01498.
- Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. Advances in Neural Information Processing Systems, 3(January):2177–2185, 2014. ISSN 1049-5258. null; Conference date: 08-12-2014 Through 13-12-2014.
- Yitan Li, Linli Xu, Fei Tian, Liang Jiang, Xiaowei Zhong, and Enhong Chen. Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3650–3656. AAAI Press, 2015. ISBN 9781577357384.
- Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. Spherical text embedding, 2019. URL https://arxiv.org/abs/1911.01196.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013a. URL http://arxiv.org/abs/1301.3781.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013b. Curran Associates Inc.
- Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations, 2017. URL https://arxiv.org/abs/1705.08039.

- Sultan Nurmukhamedov, Thomas Mach, Arsen Sheverdin, and Zhenisbek Assylbekov. From hyperbolic geometry back to word embeddings, 2022. URL https://arxiv.org/abs/2204.12481.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.
- Adriaan M. J. Schakel and Benjamin J. Wilson. Measuring word significance using distributed representations of words, 2015. URL https://arxiv.org/abs/1508.02297.
- Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. Additive compositionality of word vectors. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pages 387–396, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-5551. URL https://aclanthology.org/D19-5551.
- Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings, 2018. URL https://arxiv.org/abs/1810.06546.

A Proofs

Proof of Lemma 4.1. We begin by proving that $\mathbf{x} \cdot \mathbf{l} = -\mathbf{y} \cdot \mathbf{l}$. Using the definition of \mathbf{l} , we write:

$$\mathbf{x} \cdot \mathbf{l} = \mathbf{x} \cdot (\mathbf{x} - \mathbf{y})$$
$$= \|\mathbf{x}\|^2 - \mathbf{x} \cdot \mathbf{y},$$

where the second equality follows from the distributive property of the dot product. Since $\mathbf{x}, \mathbf{y} \in S^d$, we have $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, leading to:

$$\mathbf{x} \cdot \mathbf{l} = 1 - \mathbf{x} \cdot \mathbf{y}.$$

Similarly, for $\mathbf{y} \cdot \mathbf{l}$:

$$\mathbf{y} \cdot \mathbf{l} = \mathbf{y} \cdot (\mathbf{x} - \mathbf{y})$$
$$= \mathbf{y} \cdot \mathbf{x} - \|\mathbf{y}\|^2$$
$$= \mathbf{x} \cdot \mathbf{y} - 1.$$

Combining these results, we find:

$$\mathbf{x} \cdot \mathbf{l} = -\mathbf{y} \cdot \mathbf{l}.$$

Next, we show that $\|\mathbf{x} - \text{proj}_{\mathbf{l}}(\mathbf{x})\|_2 = \|\mathbf{y} - \text{proj}_{\mathbf{l}}(\mathbf{y})\|_2$. Using the fact that $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, we expand:

$$\begin{aligned} \|\mathbf{x} - \operatorname{proj}_{\mathbf{l}}(\mathbf{x})\|_{2}^{2} &= \|\mathbf{x} - (\mathbf{x} \cdot \hat{\mathbf{l}})\hat{\mathbf{l}}\|_{2}^{2} \\ &= \|\mathbf{x}\|^{2} - 2(\mathbf{x} \cdot \hat{\mathbf{l}})^{2} + (\mathbf{x} \cdot \hat{\mathbf{l}})^{2} \\ &= 1 - (\mathbf{x} \cdot \hat{\mathbf{l}})^{2}, \end{aligned}$$

where $\hat{\mathbf{l}} = \frac{1}{\|\mathbf{l}\|}$. Repeating this calculation for **y**:

$$\|\mathbf{y} - \operatorname{proj}_{\mathbf{l}}(\mathbf{y})\|_{2}^{2} = 1 - (\mathbf{y} \cdot \hat{\mathbf{l}})^{2}.$$

Since $\mathbf{x} \cdot \hat{\mathbf{l}} = -\mathbf{y} \cdot \hat{\mathbf{l}}$, the equality $\|\mathbf{x} - \text{proj}_1(\mathbf{x})\|_2 = \|\mathbf{y} - \text{proj}_1(\mathbf{y})\|_2$ follows.

Proof of Lemma 4.2. Without loss of generality, let **l** form a basis and assume that all entries of **x** and **y** are equal except for the *d*-th entry, where $\mathbf{x}_d = -\mathbf{y}_d$. Define $R = \|\mathbf{x} - \text{proj}_1(\mathbf{x})\|_2^2 = \|\mathbf{y} - \text{proj}_1(\mathbf{y})\|_2^2$.

Suppose $\mathbf{x} - \mathbf{y} = \mathbf{a} - \mathbf{b}$, and consider $\mathbf{b} = \mathbf{a} + (\mathbf{x} - \mathbf{y})$. Since $\|\mathbf{b}\| = 1$, expanding $\|\mathbf{b}\|^2$ gives:

$$\|\mathbf{a} + (\mathbf{x} - \mathbf{y})\|^2 = \|\mathbf{a}\|^2 + 2\mathbf{a} \cdot (\mathbf{x} - \mathbf{y}) + \|\mathbf{x} - \mathbf{y}\|^2 = 1.$$

By construction, the entries of \mathbf{x} and \mathbf{y} in dimensions $1, \ldots, d-1$ are identical. Consequently, the above simplifies to:

$$\|\mathbf{a}\|^2 + 2\mathbf{a}_d(\mathbf{x}_d - \mathbf{y}_d) + (\mathbf{x}_d - \mathbf{y}_d)^2 = 1.$$

Solving for \mathbf{a}_d , we find:

 $\mathbf{a}_d = -\mathbf{x}_d.$

Since $\|\mathbf{x}\| = 1$, the squared sum of the first d-1 entries of \mathbf{x} is R. Similarly, for \mathbf{a} , we have:

$$\|\mathbf{a}\|^2 = \sum_{i=1}^{d-1} \mathbf{a}_i^2 + \mathbf{a}_d^2 = R + (1 - R) = 1.$$

A similar argument holds for **b**, completing the proof. \blacksquare

Proof of Lemma 4.3. Let \mathbf{x} be a vector with distance r from the axis defined by \mathbf{l} . By definition, this implies:

$$\|\mathbf{x} - \operatorname{proj}_{\mathbf{l}}(\mathbf{x})\| = r$$

From the Pythagorean theorem, we have:

$$(\mathbf{x} \cdot \mathbf{l})^2 + r^2 = 1,$$

leading to:

$$(\mathbf{x} \cdot \mathbf{l})^2 = 1 - r^2$$

Using the definition of the projection, we expand:

$$\|\mathbf{x} - \operatorname{proj}_{\mathbf{l}}(\mathbf{x})\|^{2} = \|\mathbf{x} - (\mathbf{x} \cdot \hat{\mathbf{l}})\hat{\mathbf{l}}\|^{2}$$
$$= \|\mathbf{x}\|^{2} - 2(\mathbf{x} \cdot \hat{\mathbf{l}})^{2} + (\mathbf{x} \cdot \hat{\mathbf{l}})^{2}$$
$$= 1 - (\mathbf{x} \cdot \hat{\mathbf{l}})^{2} = r^{2}.$$

Thus, $(\mathbf{x} \cdot \mathbf{l})^2 = 1 - r^2$, as required.

Proof of Theorem 4.4. Let $\mathbf{c} \in C$ be an arbitrary concept. We show that if $|w_{\mathbf{c}}| > 4$, it is impossible to embed the corresponding words on S^1 while preserving all analogies.

By Lemma 4.2, word embeddings that preserve analogies must satisfy the following two conditions for an axis l and a radius r:

$$\|\mathbf{v}_i\|_2 = 1,$$

$$d(\mathbf{v}_i, \mathbf{l}) = r,$$

where $d(\mathbf{v}_i, \mathbf{l})$ denotes the distance from the embedding \mathbf{v}_i to the axis \mathbf{l} .

On the unit circle (S^1) in \mathbb{R}^2 , it is geometrically evident that at most four points can lie at the same distance r from a given axis l. Hence, no more than four embeddings can simultaneously satisfy the conditions above on S^1 .

Conversely, for $d \ge 3$, the set of points satisfying $\|\mathbf{v}_i\|_2 = 1$ and $d(\mathbf{v}_i, \mathbf{l}) = r$ forms a sphere S^{d-1} of dimension d-1, which is uncountably infinite. Therefore, an arbitrary number of embeddings can satisfy the conditions when $d \ge 3$. This completes the proof.

Proof of Theorem 4.5. We aim to determine the minimum d such that the system of two linear equations:

$$\sum_{i=1}^{d} \mathbf{x}_{i}^{(1)} \mathbf{l}_{i}^{(1)} = \sqrt{1 - (r^{(1)})^{2}},$$
$$\sum_{i=1}^{d} \mathbf{x}_{i}^{(2)} \mathbf{l}_{i}^{(2)} = \sqrt{1 - (r^{(2)})^{2}},$$

has a non-trivial solution while allowing embeddings to assign as many words as possible to each concept.

When d = 3, we express the embeddings **x** as follows:

$$\begin{aligned} \mathbf{x}_1 \mathbf{l}_1^{(1)} + \mathbf{x}_2 \mathbf{l}_2^{(1)} + \mathbf{x}_3 \mathbf{l}_3^{(1)} &= \sqrt{1 - (r^{(1)})^2}, \\ \mathbf{x}_1 \mathbf{l}_1^{(2)} + \mathbf{x}_2 \mathbf{l}_2^{(2)} + \mathbf{x}_3 \mathbf{l}_3^{(2)} &= \sqrt{1 - (r^{(2)})^2}, \end{aligned}$$

subject to the constraint:

$$\mathbf{x}_1^2 + \mathbf{x}_2^2 + \mathbf{x}_3^2 = 1.$$

With three equations and three unknowns $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, the system admits a finite set of solutions. Thus, for d = 3, it is possible to find embeddings satisfying the conditions of both equations. This establishes d = 3 as the minimum dimension.

Proof of Theorem 4.6. Suppose there are k overlapping concepts. To encode these concepts while preserving analogies, the word embeddings must satisfy the following system of k linear equations:

$$\sum_{i=1}^{d} \mathbf{x}_{i} \mathbf{l}_{i}^{(1)} = \sqrt{1 - (r^{(1)})^{2}},$$
$$\sum_{i=1}^{d} \mathbf{x}_{i} \mathbf{l}_{i}^{(2)} = \sqrt{1 - (r^{(2)})^{2}},$$
$$\vdots$$
$$\sum_{i=1}^{d} \mathbf{x}_{i} \mathbf{l}_{i}^{(k)} = \sqrt{1 - (r^{(k)})^{2}},$$

along with the constraint:

$$\sum_{i=1}^{d} \mathbf{x}_i^2 = 1.$$

These equations impose k + 1 constraints on the embeddings **x** (the k equations plus the norm constraint). To have sufficient degrees of freedom for solutions, we require at least k + 2 unknown variables. Consequently, the dimension d must satisfy:

$$d \ge k + 2.$$

Therefore, a minimum of k + 2 dimensions is necessary to encode k overlapping concepts. This completes the proof.

Proof of Lemma 4.7. By Lemma 4.1, the vectors l_1 and l_2 can be chosen such that:

$$\mathbf{l}_1 = \frac{\mathbf{a} - \mathbf{b}}{\|\mathbf{a} - \mathbf{b}\|}, \quad \mathbf{l}_2 = \frac{\mathbf{a} - \mathbf{c}}{\|\mathbf{a} - \mathbf{c}\|}.$$

From the conditions on the concepts, the following relationships hold:

$$\mathbf{a} - \mathbf{b} = \mathbf{c} - \mathbf{d},$$

 $\mathbf{a} - \mathbf{c} = \mathbf{b} - \mathbf{d}.$

We begin by verifying the equality $\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{c} - \mathbf{d}\|^2$:

$$\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a} \cdot \mathbf{b},$$
$$\|\mathbf{c} - \mathbf{d}\|^2 = \|\mathbf{c}\|^2 + \|\mathbf{d}\|^2 - 2\mathbf{c} \cdot \mathbf{d}.$$

Using the fact that $\|\mathbf{a}\| = \|\mathbf{b}\| = \|\mathbf{c}\| = \|\mathbf{d}\| = 1$ (since all points lie on S^d), the above simplifies to:

$$2 - 2\mathbf{a} \cdot \mathbf{b} = 2 - 2\mathbf{c} \cdot \mathbf{d}.$$

Hence, we conclude:

 $\mathbf{a} \cdot \mathbf{b} = \mathbf{c} \cdot \mathbf{d}.$

Next, consider the inner product $(\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{c})$. Expanding this expression, we have:

$$(\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{c}) = \|\mathbf{a}\|^2 - \mathbf{a} \cdot \mathbf{b} - \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c},$$

= 1 - $\mathbf{a} \cdot \mathbf{b} - \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c}.$

Substituting the conditions on $\mathbf{a} - \mathbf{b}$ and $\mathbf{a} - \mathbf{c}$, we rewrite $\mathbf{b} - \mathbf{c}$ as $\mathbf{d} - \mathbf{c}$ and simplify:

$$(\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{c}) = 1 - \mathbf{a} \cdot \mathbf{b} - \mathbf{a} \cdot \mathbf{c} + \mathbf{c} \cdot (\mathbf{d} - \mathbf{c}),$$
$$= 1 - \mathbf{a} \cdot \mathbf{b} - \mathbf{a} \cdot \mathbf{c} + \mathbf{c} \cdot \mathbf{d} - \|\mathbf{c}\|^2,$$
$$= 1 - \mathbf{a} \cdot \mathbf{b} - \mathbf{a} \cdot \mathbf{c} + \mathbf{c} \cdot \mathbf{d} - 1,$$
$$= \mathbf{c} \cdot \mathbf{d} - \mathbf{a} \cdot \mathbf{b}.$$

Since $\mathbf{a} \cdot \mathbf{b} = \mathbf{c} \cdot \mathbf{d}$, it follows that:

$$(\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{c}) = 0.$$

Thus, $\mathbf{a} - \mathbf{b} \perp \mathbf{a} - \mathbf{c}$, which implies $\mathbf{l}_1 \perp \mathbf{l}_2$, as required.

Proof of Theorem 4.8. Generalizing Lemma 4.7, we conclude that k orthogonal axes are necessary to encode k strictly overlapping concepts. Specifically, for each concept, the axes $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_k$ are mutually orthogonal.

To encode these k concepts, the embedding must satisfy k + 1 equations: k equations corresponding to the orthogonal axes and one equation enforcing the unit norm constraint:

$$\sum_{i=1}^{d} \mathbf{x}_i^2 = 1.$$

For the system to admit a solution, the dimension **d** must be sufficient to provide k + 2 degrees of freedom. Therefore, the minimum dimension required to embed k strictly overlapping concepts is $d \ge k + 2$. This completes the proof.