Embeddability of Analogies as Parallelograms in S^d

Narutatsu Ri
Department of Computer Science
Columbia University
`wl2787@columbia.edu`

Abstract

Static word embeddings possess the remarkable property of additive composition, facilitating the recovery of semantic analogies through basic vector arithmetic. However, previous studies have predominantly focused on analyzing these learned embeddings without considering the normalization of vector lengths, despite the common practice of normalizing word embeddings prior to their application in downstream tasks.

In this report, we explore the capabilities of length-normalized word embeddings in representing word analogies as parallelograms within the embedding space. We reduce the problem of embedding words on the d-sphere to solving a system of linear equations, and demonstrate the representation dimension of the embeddings necessarily scales with the number of analogy conditions.

1 Introduction

Dense vector representations such as *word2vec* [Mikolov et al., 2013a] and GloVe [Pennington et al., 2014], are built upon the premise that the inner product of two vectors should reflect the statistical co-occurrence of their corresponding words in the training corpus. Interestingly, these embeddings possess a property known as *additive compositionality* [Mikolov et al., 2013b], enabling elementary vector arithmetic to capture semantic analogies such as "man : woman = king : queen." Alongside empirical investigations, a surge of theoretical explanations aimed at unraveling how these models implicitly learn analogy relationships have emerged, prompted by the observation of additive compositionality in *word2vec*.

However, while empirical studies consistently employ *normalized* vectors, most studies do not take vector length into consideration, and studies that target the properties of normalized word vectors is scarce [Schakel and Wilson, 2015]. No study has systematically debunked the properties of normalized word vectors or considered their limitations in expressiveness. To address this gap, we investigate the conditions on embedding n points on the d-dimensional Euclidean ball while satisfying C analogy conditions.

2 Related Work

Word Embedding Models. Dense vector representations were popularized by *word2vec* and the GloVe model. Various alternate representations have been explored [Bojanowski et al., 2016, Seonwoo et al., 2019], but *word2vec* still remains as the most popular choice. In recent years, static word embeddings in non-Euclidean spaces have also garnered interest [Leimeister and Wilson, 2018, Nurmukhamedov et al., 2022, Dhingra et al., 2018, Meng et al., 2019, Nickel and Kiela, 2017, Tifrea et al., 2018] . Such studies show that embedding word vectors in non-Euclidean space provides empirical benefits such as the capability to capture hierarchical word similarities.

Theory. The first attempts to debunk the mechanism of word embedding models is by Levy and Goldberg [2014], who claim that *word2vec* is implicitly factorizing the shifted Pointwise Mutual Information (PMI)[1] matrix.

Follow-up work in Li et al. [2015] attempt to show that *word2vec* is explicitly factorizing the word co-occurrence matrix. Hashimoto et al. [2016] formulate learning dense word embeddings as metric recovery of a vector space over concepts

[1]For words i, j, PMI is defined as PMI(i, j) = $\log \frac{\#(i,j)}{\#(i)\cdot\#(j)}$, where $\#(i, j)$ denotes co-occurrence count between i and j.

where Euclidean distances between points are assumed to represent semantic similarities using word co-occurrence statistics.

The first attempts to understand how *word2vec* implicitly learns analogies are by Arora et al. [2015], who propose that analogies are recovered as parallel lines whe word embeddings recover PMI statistics with vector products under specific generative assumptions on the dataset. Follow-up work in Gittens et al. [2017] define the notion of a *paraphrase* and explain the formation of analogies with paraphrases. Recent studies take inspiration from Gittens et al. [2017] and improve upon the formulation of analogies as paraphrases by showing the condition in Gittens et al. [2017] holds with weaker assumptions [Allen and Hospedales, 2019, Allen et al., 2018, Ethayarajh et al., 2019].

3 Problem Construction

3.1 Notation

Denote the number of points to embed as n, the number of analogy relationships as C. Denote the dimension of the embedding space as d. Denote words as $a, b, \ldots$ and their dense vector representations as $\vec{a}, \vec{b}, \ldots$. d will be the number of dimensions and n will be the number of words to embed.

Definition 3.1 (Relation)**.** A **relation** is defined as a 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**.

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 iff the two relations are equivalent, and we say the analogy *expresses* the relation.

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

Definition 3.3 (Concept)**.** A **concept** c_i is a set of equivalent relations.

We say that the concept *expresses* the relation and denote it as r_{c_i}. We will represent the set of words participating in the concept with w_{c_i}. The set of words that participate in each relation as a source is denoted as a **source set** and the set of words that participate in each relation within each analogy as a sink is denoted as a **sink set**.

For example, for words a, b, c, d, e, f, if $r_{a,b} = r_{c,d}, r_{a,b} = r_{e,f}$ 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 common types of point 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}, 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.

3.1.1 Assumptions

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

The 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 are *consistent*. In other words, they are embeddable in some dimension d without contradiction.

Assumption 3. For two concepts c_1 and c_2, if $|w_{c_1} \cap w_{c_2}| > 2, |w_{c_1} \cap w_{c_2}| \equiv 0 \mod 2$ (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 $|w_{c_1} \cap w_{c_2}| > 1, |w_{c_1} \cap w_{c_2}| \equiv 1 \mod 2$ (i.e. they share an odd number of points greater than 1), then the shared points must all be either only sinks or 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.

3.2 Problem Construction

Here, we prove various properties for points on the surface of an L_2 ball residing in d dimensions. We will denote the unit L_2 ball as S^d. All proofs are included in the Appendix.

Lemma 3.1. *For any two distinct arbitrary points $x, y \in S^d, S^d \subset \mathbb{R}^d$, the vector $l = x - y$ satisfies the properties $x \cdot l = -y \cdot l$, $\|x - \text{proj}_l\, x\|_2 = \|y - \text{proj}_l\, y\|_2$.*

Lemma 3.1 implies that for two arbitrary points there always exists a direction in the space where the two points have equal L_2 distance from. If we translate the points in a way such that the axis is a scalar multiple of a basis vector, then the two points will have the same values for all entries besides one, which takes the negative value of the other point's corresponding entry. For simplicity, we will denote such an l as the *axis*.

Moreover, we show a property for four points residing on S^d in $\mathbb{R}^d$ that form a parallelogram must satisfy.

Lemma 3.2. *Consider two distinct arbitrary points $x, y \in S^d, S^d \subset \mathbb{R}^d$ and a corresponding axis l where the conditions of Lemma 3.1 are satisfied.*

For two other distinct points $a, b \in S^d$, if $x - y = b - a$, then $\|x - \text{proj}_l\, x\|_2^2 = \|y - \text{proj}_l\, y\|_2^2 = \|a - \text{proj}_l\, a\|_2 = \|b - \text{proj}_l\, b\|_2$.

Lemma 3.2 informs us the condition any four points need to satisfy to form an analogy geometrically. For all pairs of words in a concept, the word vectors must adhere to this condition. We can geometrically consider the region where the points can reside as the set of points that are of distance r from an axis l on the surface of S^d.

Now, we formalize the above into a mathematical expression for the set of points:

Lemma 3.3. *For a concept c, given an arbitrary unit vector l whose direction represents the axis and a radius $0 < r < 1$, the embeddings x of the words participating in c satisfy the property:*

$$\sum_{i=1}^{d} x_i l_i = \sqrt{1 - r^2}$$

where x_i, l_i denote the ith entry of the vectors x, l respectively.

Note that not all set of points x_i that satisfy the equality $\sum_{i=1}^{d} x_i l_i = \sqrt{1 - r^2}$ must be an embedding for a word. Instead, satisfying the equality is a necessary condition; if a word is participating in a concept, its embedding must satisfy the equality, but the converse is not necessarily true.

Now, consider two concepts c_1, c_2 that have overlapping points. By Lemma 3.3, the regions that embeddings can reside in when participating in each concept can be represented as:

$$\sum_{i=1}^{d} x_i^{(1)} l_i^{(1)} = \sqrt{1 - (r^{(1)})^2},$$
$$\sum_{i=1}^{d} x_i^{(2)} l_i^{(2)} = \sqrt{1 - (r^{(2)})^2},$$

where $x^{(1)}, x^{(2)}$ represents the regions for c_1, c_2 respectively.

Thus, for embeddings that participate in both concepts, we only need to solve for the set of points x that satisfy both equations. Finding the regions that overlapping points can reside in reduces to solving a system of linear equations in addition to the nonlinear equation that x must reside on S^d, namely $\|x\|_2 = 1$.

3.3 Dimensionality for Embedding Analogies

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

Theorem 3.4 (Embeddability of Disjoint Concepts). *If $\forall i, j \in [|C|], i \neq j : |w_{c_i} \cap w_{c_j}| = 0$, then for any $n \geq 4$, all n points can be embedded on S^2 for $d = 3$.*

Theorem 3.5 (Embeddability of Weak Overlaps). *If $\exists i, j \in [|C|], i \neq j : |w_{c_i} \cap w_{c_j}| = 1$, for any $n \geq 4$, all n points can be embedded on S^2 for $d = 3$.*

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

Now, we consider when there exists concepts that have strict overlap.

Theorem 3.6 (Embeddability of Strong Overlaps). *For any $n \geq 4$, if $k \leq |C|$ concepts strongly overlap, then the concepts can be embedded in S^{d-1} where $d \geq k + 2$.*

Before we consider strong overlaps between k concepts, we first establish a condition that the axes l_i need to satisfy for all k analogy conditions to be satisfied in the embeddings.

Lemma 3.7. *For an arbitrary set of four embeddings a, b, c, d and concepts c_1, c_2, if the analogies $(a : b) = (c : d), (a : c) = (b : d)$ hold and represent the concepts c_1, c_2 respectively, then the axes l_1, l_2 must be orthogonal for each concept.*

Lemma 3.7 can be trivially generalized to k concepts with strict overlap, in which case at least k orthogonal axes are required.

Theorem 3.8 (Embeddability of Strict Overlaps). *For any $n \geq 4$, if $k \leq |C|$ concepts strictly overlap, then C can be embedded in S^{d-1} for any $d \geq k + 2$.*

The contrast between Lemma 3.7 and Therorem 3.8 is whether the orthogonality of the axes. This has no effect on the system of equations, as the coefficients are independent.

Note that k strictly overlapping concepts *can* be embedded in $d = k + 1$ dimensions with a limitation on the number of words that are able to participate in the concepts. Namely, this construction will yield at most 2^{k+1} possible number of embeddable words, as the common region between k systems of linear equations with the unit norm constraint gives two possible values for each dimension.

4 Conclusion

We have explored the relationship between concept overlaps and the minimum number of dimensions necessary to preserve all analogy conditions. By considering the interplay between concept overlaps and dimensionality, our findings contribute to a deeper understanding of the fundamental aspects of word embeddings and their ability to capture semantic relationships. The insights gained from this study shed light on the possibilities and limitations of embedding techniques, providing valuable guidance for the design and application of word embeddings in various natural language processing tasks.

Note that while our approach focused on determining the possible regions where points can reside to satisfy the analogies rather than precisely pinpointing their positions in space, the arrangement of the points can be achieved by sequentially placing them in space while deliberately selecting appropriate words once the region is identified.

References

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`.

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.

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.

Adriaan M. J. Schakel and Benjamin J. Wilson. Measuring word significance using distributed representations of words, 2015. URL https://arxiv.org/abs/1508.02297.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

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.

Matthias Leimeister and Benjamin J. Wilson. Skip-gram word embeddings in hyperbolic space, 2018. URL https://arxiv.org/abs/1809.01498.

Sultan Nurmukhamedov, Thomas Mach, Arsen Sheverdin, and Zhenisbek Assylbekov. From hyperbolic geometry back to word embeddings, 2022. URL https://arxiv.org/abs/2204.12481.

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.

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.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations, 2017. URL https://arxiv.org/abs/1705.08039.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings, 2018. URL https://arxiv.org/abs/1810.06546.

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.

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.

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.

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.

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.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations, 2019. URL https://arxiv.org/abs/1908.06361.

A Proofs

A.1 Proof of Lemma 3.1

First, we show $x \cdot l = -y \cdot l$:

$$\begin{aligned}
x \cdot l &= x \cdot (x - y) = \|x\|_2^2 - xy \\
&= 1 - xy \quad (\because \text{All points are on } S^d) \\
&= \|y\|^2 - xy \\
&= -y \cdot (x - y) \\
&= -y \cdot l.
\end{aligned}$$

Now, we show $\|x - \mathrm{proj}_l\, x\|_2 = \|y - \mathrm{proj}_l\, y\|_2$. Using the fact $\|x\|_2 = \|y\|_2 = 1$,

$$\begin{aligned}
\|x - \mathrm{proj}_l\, x\|_2^2 = \|x - (x \cdot \hat{l})\hat{l}\|_2^2 &= \left\| x - \left(x \cdot \frac{x-y}{\|x-y\|_2} \right) \frac{x-y}{\|x-y\|_2} \right\|_2^2 \\
&= \|x\|_2^2 + 2\left(x \cdot \left(x \cdot \frac{x-y}{\|x-y\|_2} \right) \frac{x-y}{\|x-y\|_2} \right) + \left\| \left(x \cdot \frac{x-y}{\|x-y\|_2} \right) \frac{x-y}{\|x-y\|_2} \right\|_2^2 \\
&= 3 + \left(\frac{1-xy}{\|x-y\|_2} \right)^2
\end{aligned}$$

$$\begin{aligned}
\|y - \mathrm{proj}_l\, y\|_2^2 = \|y - (y \cdot \hat{l})\hat{l}\|_2^2 &= \left\| y - \left(y \cdot \frac{x-y}{\|x-y\|_2} \right) \frac{x-y}{\|x-y\|_2} \right\|_2^2 \\
&= \|y\|_2^2 + 2\left(y \cdot \left(y \cdot \frac{x-y}{\|x-y\|_2} \right) \frac{x-y}{\|x-y\|_2} \right) + \left\| \left(y \cdot \frac{x-y}{\|x-y\|_2} \right) \frac{x-y}{\|x-y\|_2} \right\|_2^2 \\
&= 3 + \left(\frac{xy-1}{\|x-y\|_2} \right)^2
\end{aligned}$$

Thus, $\|x - \mathrm{proj}_l\, x\|_2^2 = \|y - \mathrm{proj}_l\, y\|_2^2 \Leftrightarrow \|x - \mathrm{proj}_l\, x\|_2 = \|y - \mathrm{proj}_l\, y\|_2$. ∎

A.2 Proof of Lemma 3.2

Without loss of generality, consider a translation of the points x, y, a, b such that l is a basis and the entries of x, y are equal besides the dth entry where $x_d = -y_d$ and denote $\|x - \mathrm{proj}_l\, x\|_2^2 = \|y - \mathrm{proj}_l\, y\|_2^2 = R$. We first show that if $x - y = a - b$, then $\|a - \mathrm{proj}_l\, a\|_2 = R$. Note that $\|a\|_2 = \|b\|_2 = 1$.

$x - y = b - a \Leftrightarrow b = a + (x - y)$. As $\|b\|_2 = 1$, if we denote the ith entry of a vector v as v_i,

$$\begin{aligned}
\|a + (x - y)\|_2^2 &= (a_1 + (x_1 - y_1))^2 + (a_1 + (x_1 - y_1))^2 \cdots + (a_d + (x_d - y_d))^2 \\
&= \sum_{i=1}^{d} a_i^2 + 2\sum_{i=1}^{d}(a_i(x_i - y_i)) + \sum_{i=1}^{d}(x_i - y_i)^2 = 1
\end{aligned}$$

By construction, $\forall i \in [d-1] : x_i = y_i$, so

$$\begin{aligned}
&\sum_{i=1}^{d} a_i^2 + 2\sum_{i=1}^{d}(a_i(x_i - y_i)) + \sum_{i=1}^{d}(x_i - y_i)^2 = \sum_{i=1}^{d} a_i^2 + 2(a_d(x_d - y_d)) + (x_d - y_d)^2 = 1 \\
&\Leftrightarrow a_d = \frac{-(x_d - y_d)^2}{2(x_d - y_d)} = \frac{-4x_d^2}{4x_d} = -x_d
\end{aligned}$$

Note that as $\|x\|_2 = 1, \sum_{i=1}^{d-1} x_i^2 = R$ as $\|x - \mathrm{proj}_l\, x\|_2^2$ is equivalent to the distance between x and the axis. Then, $x_n^2 = 1 - R$.

As $\|a\|_2 = 1$,

$$\sum_{i=1}^{d-1} a_i^2 + a_d^2 = 1 \Leftrightarrow \sum_{i=1}^{d-1} a_i^2 = 1 - a_d^2 = 1 - (1 - R) = R$$

A similar argument can be made for b. ∎

A.3 Proof of Lemma 3.3

Consider a vector x that is distance r away from the axis. This is equivalent to

$$\|x - \mathrm{proj}_l\, x\| = r$$

Moreover, by construction, $(x \cdot l)^2 + r^2 = 1 \Leftrightarrow (x \cdot l)^2 = 1 - r^2$.

Utilizing this,

$$\begin{aligned}
\|x - \mathrm{proj}_l\, x\|^2 &= \|x - (x \cdot l)l\|^2 \\
&= \sum_{i=1}^{d}\left(x_i - \sum_{j=1}^{d}(x_j l_j) l_i\right)^2 \\
&= \sum_{i=1}^{d} x_i^2 - 2\sum_{i=1}^{d} x_i \sum_{j=1}^{d}(x_j l_j) l_i + \sum_{i=1}^{d}\left(\sum_{j=1}^{d}(x_j l_j) l_j\right)^2 \\
&= 1 - 2\sqrt{1 - r^2}\sum_{i=1}^{d} x_i l_i + 1 - r^2 \\
&= r^2
\end{aligned}$$

Thus, $\sum_{i=1}^{d} x_i l_i = \sqrt{1 - r^2}$. ■

A.4 Proof of Theorem 3.4

Consider an arbitrary concept $c \in C$. We show that if $|w_c| > 4$, then the words cannot be embedded on S^1 while preserving all analogies.

By Lemma 3.2, for word embeddings to form analogies, we consider an arbitrary axis l and a radius r of which the embeddings of the words satisfy the conditions:

$$\begin{aligned}
\|v_i\|_2 &= 1, \\
d(v_i, l) &= r,
\end{aligned}$$

where $d(v_i, l)$ denotes the distance from the word embedding for word i to the axis l. Note that for any l there exists at most 4 points that are of distance r from l on the surface of the unit ball (the unit circle in $d = 2$). Thus, there can be at most 4 words participating in the concept.

Conversely, when $d \geq 3$, the set of points that satisfy the above two conditions form a subspace S^{d-1} of which the cardinality is uncountably infinite. Thus, an arbitrary number of points can participate in c. ■

A.5 Proof of Theorem 3.5

We want to find the lowest d where the system of two linear equations

$$\begin{aligned}
\sum_{i=1}^{d} x_i^{(1)} l_i^{(1)} &= \sqrt{1 - (r^{(1)})^2} \\
\sum_{i=1}^{d} x_i^{(2)} l_i^{(2)} &= \sqrt{1 - (r^{(2)})^2}
\end{aligned}$$

can intersect while being able to assign as many words to each concept as possible.

When $d = 3$, if we want to find the embeddings x that satisfy both equations, the above becomes:

$$\begin{aligned}
x_1 l_1^{(1)} + x_2 l_2^{(1)} + x_3 l_3^{(1)} &= \sqrt{1 - (r^{(1)})^2}, \\
x_1 l_1^{(2)} + x_2 l_2^{(2)} + x_3 l_3^{(2)} &= \sqrt{1 - (r^{(2)})^2},
\end{aligned}$$

with the constraint $x_1^2 + x_2^2 + x_3^2 = 1$. With 3 equations and 3 unknowns, we are able to embed a finite set of points that satisfy the above equations. ■

A.6 Proof of Theorem 3.6

Given k overlapping concepts, we are required to solve for the linear system of k equations:

$$\sum_{i=1}^{d} x_i l_i^{(1)} = \sqrt{1-(r^{(1)})^2}$$
$$\sum_{i=1}^{d} x_i l_i^{(2)} = \sqrt{1-(r^{(2)})^2}$$
$$\cdots$$
$$\sum_{i=1}^{d} x_i l_i^{(k)} = \sqrt{1-(r^{(k)})^2},$$

with the constraint $\sum_{i=1}^{d} x_i^2 = 1$, therefore $k+1$ equations. To exhibit an arbitrary number of overlapping points, there must be at least $k+2$ unknown variables. This condition is only satisfied when $d \geq k+2$. ■

A.7 Proof of Lemma 3.7

By Lemma 3.1, we know that l_1, l_2 can be chosen such that $l_1 = \frac{1}{\|a-b\|}(a-b), l_2 = \frac{1}{\|a-c\|}(a-c)$. From the conditions on the concepts, the condition

$$a-b = c-d$$
$$a-c = b-d$$

holds. Therefore,

$$\begin{aligned}
&\|a-b\|^2 = \|c-d\|^2 \\
\Leftrightarrow &\|a\|^2 + \|b\|^2 - 2a\cdot b = \|c\|^2 + \|d\|^2 - 2c\cdot d \\
\Leftrightarrow &2 - 2a\cdot b = 2 - 2c\cdot d \\
\Leftrightarrow &a\cdot b = c\cdot d.
\end{aligned}$$

Now, consider the inner product $(a-b)\cdot(a-c)$:

$$\begin{aligned}
(a-b)\cdot(a-c) &= \|a\|^2 - a\cdot b - a\cdot c + b\cdot c \\
&= 1 + c(b-a) - a\cdot b \\
&= 1 + c(d-c) - a\cdot c \\
&= 1 - 1 + cd - ab \\
&= 0.
\end{aligned}$$

Therefore, $a-b \perp a-c \Rightarrow l_1 \perp l_2$. ■

A.8 Proof of Theorem 3.8

Generalizing Lemma 3.7, we know that k orthogonal axes are required to embed k strictly overlapping concepts, meaning $d \geq k$ must hold true. Given k linear equations where the vectors constructed by the coefficients l_i are all mutually orthogonal, we have $k+1$ equations to satisfy. To ensure an arbitrary number of points can reside in the region, we require $d \geq k+2$. ■