What the Vec? Towards Probabilistically Grounded Embeddings

Carl Allen **Ivana Balažević** **Timothy Hospedales**
School of Informatics
University of Edinburgh
{carl.allen, ivana.balazevic, t.hospedales}@ed.ac.uk

Abstract

Vector representation, or embedding, of words is commonly achieved with neural network methods, in particular *word2vec* (**W2V**). It has been shown that certain statistics of word co-occurrences are implicitly captured by properties of *W2V* vectors, but much remains unknown of them, e.g. any meaning of length, or more generally how it is that statistics can be reliably framed as vectors at all. By deriving a mathematical link between probabilities and vectors, we justify why *W2V* works and are able to create embeddings with probabilistically interpretable properties.

1 Introduction

Many machine learning tasks benefit from using vector representations of input data that is naturally discrete, e.g. users and products in recommender systems. This paradigm, of embedding each discrete datum as a vector encapsulating its information, is particularly crucial in natural language processing.

Word embedding methods typically rely on statistics of word observations, e.g. Latent Semantic Analysis [5]. Recently, neural networks have been used to effectively capture these statistics, as epitomised by *word2vec* (**W2V**) [14].[1] Vectors derived by *W2V* have far fewer dimensions than the number of words represented and are thus computationally efficient, whilst frequently outperforming non-neural network methods on downstream tasks. However, this improvement comes at the cost of transparency, i.e. of understanding what the embeddings *mean*. Such embeddings form the bedrock of many natural language processing tasks, from predictive language models and machine translation to image annotation and question answering, where they are largely 'plugged in' to a larger neural network architecture. Since the properties of the embeddings themselves are not mathematically interpretable, there is little hope of meaningfully interpreting a model using them, or building them into a probabilistic framework. An understanding of their properties is of interest as it may allow the development of better performing embeddings; improved interpretability, in turn, of models using them; or for their use in more principled probabilistic models; pushing back on the 'black box' nature of word embeddings and their use. Our work takes steps in this direction.

Many works explore properties of the *W2V* algorithm and its embeddings (e.g. [10, 2, 12, 8, 18, 15]), providing various degrees of insight. However, none fully explains *why W2V* works. Here, we develop a mathematical relationship between the probabilities of word occurrences and vectors that demonstrates how statistics can provably be coerced into vector properties and, as a consequence, justifies why models, with the architecture of *W2V* and a similar implicit loss function, *work*, in the sense of generating embeddings that seemingly capture word meaning. We our the derived relationship to understand properties of the resulting vectors and their arrangement with respect to one another. Our key contributions are:

[1] Throughout, we refer exclusively to the more commonly used *skipgram* implementation of W2V.

- to establish a mathematically principled framework for moving from probabilities to vectors by deriving a relationship between co-occurrence *statistics* and vector properties such as *lengths* and *angles*, thereby justifing why *W2V* and models with a sufficiently similar *de facto* loss function e.g. *Glove* [16], work (Section 3.2);
- to show how the two vectors for each word within the *W2V* architecture (one per layer) arise as complex conjugates and interpret their real and imaginary components (Section 3.5); and
- to train with a loss function based on our derivation that improves training in higher dimensions and generates embeddings with angles and lengths that more directly correspond to statistics of word occurrences, thus embedding probabilistic attributes into the vector space (Section 4.2).

We believe our main theorem, the encoding of joint and marginal probabilities into vector properties, may extend to other domains in which statistics are matrix factorised, e.g. recommender systems and graph embeddings, and may further extend to tensor factorisation such as for knowledge bases.

2 Background

2.1 W2V and its interpretations

The *W2V* algorithm is trained with pairs of words extracted from a text corpus. Each pair contains a *target word*, w_i, which is assigned to each word in the corpus in turn; and a *context word*, w'_j, which for a given target word is assigned in turn to each word within a defined *context window* (e.g. 10 words either side of the target word). *W2V* is often referred to as a 'neural network' approach, however, its architecture comprises two layers with no non-linear activation and it can therefore be more simply considered matrix multiplication, subject to a non-linear output function. We denote the matrix closest to the input layer, i.e. corresponding to the target words, **W** and the other matrix **C**, corresponding to each context word. Both matrices are $d \times n$ where d is the chosen dimensionality of the embeddings (e.g. 100) and n the number of distinct words in the corpus. The i^{th} column of **W**, v_i, therefore corresponds to target word w_i and similarly, v'_j, the j^{th} column of **C**, corresponds to context word w'_j. These v_i and v'_j are the embeddings, thus we have two for each word. Different authors have suggested different approaches to using or combining them, e.g. [16, 11].[2]

The authors of *W2V* [14] initially suggest training to achieve, for all word pairs w_i, w'_j:

$$p(w'_j \mid w_i) \quad = \quad \operatorname{softmax}(v_i{}^\top v'_j) \quad = \quad \frac{\exp\{v_i{}^\top v'_j\}}{\Sigma_k \exp\{v_i{}^\top v'_k\}} \ . \tag{1}$$

It can be observed that the training objective is met when $v_i{}^\top v'_j \propto \log p(w_i, w'_j)$, i.e. **W** and **C** together factorise an $n \times n$ matrix of log joint probabilities (subject to an offset due to the proportionality constant). However, to avoid the expense of computing the softmax denominator, the training objective is adapted based on the principle of noise contrastive estimation [7]. A logistic sigmoid layer, $\sigma(\cdot)$, replaces the softmax; and $\sigma(v_i{}^\top v_j)$ is trained under supervision to classify whether a word pair comes from the corpus or is an artificially generated *negative sample*. It has since been shown [10] that the adapted training objective has a new maximum:

$$\mathbf{W}^\top\mathbf{C} \quad = \quad [\log \tfrac{p(w_i, w'_j)}{p(w_i)p(w'_j)} - \log k\,] \quad = \quad [\,\mathrm{PMI}(w_i, w'_j) - \log k\,] \ , \tag{2}$$

where PMI refers to the information theoretic property *Pointwise Mutual Information* and k is a chosen ratio of negative samples to positive samples (resulting, again, in a constant offset or *shift*). Elegance aside, this result is of particular interest given that the resultant vectors prove particularly useful and that PMI has a long history in word embeddings and linguistic analysis. More recent observations [4, 9] show that *W2V* can also be interpreted as *exponential PCA* [3] of co-occurrence statistics, which follows from *W2V* employing matrix factorisation with a sigmoid non-linearity.

Whilst these interpretations concisely summarise the workings of *W2V*, they provide no particular insight into the vectors themselves, namely a probabilistic interpretation of their properties. For instance, we do not know what vector norms and the angles between them relate to, if anything; we

[2] Notation: Subscripts i, j indicate the index of the corresponding word within a dictionary of all distinct words observed in a corpus. Apostrophes distinguish context words and their corresponding vectors from their target word counterparts. For example v'_k represents the vector for word k when observed as the context word.

do not understand the relationship between v_i and v'_i, the two vectors corresponding to the same word – should they be the same? More fundamentally, we do not know if there is consistency in the relationship ${v_i}^\top v_j = \text{PMI}(w_i, w'_j)$. That is, having mapped probabilities to vectors, might a logical consequence within the vector space, when mapped back to the domain of probabilities, violate a fundamental law? Resolving these unknowns is of interest as it may provide a more fundamental understanding of embeddings and may enable them to be deployed in more principled ways, e.g. within a probabilistic framework, similarly, perhaps, to how PPCA [19] gives probabilistic meaning to PCA. We resolve the first two unknowns and make strides towards the broader final one.

2.2 Comparison to *Glove*

Another popular embedding model, *Glove* [16], with the same architecture as $W2V$, has loss function:

$$J = \textstyle\sum_{i,j} f(X_{ij})(w_i^\top w'_j + b_i + b_j - \log X_{ij})^2 \ \ ,$$

for $X_{ij} \propto p(w_i, w'_j)$, the observation count of word pair $\langle w_i, w'_j \rangle$, biases b_i, b_j and weighting $f(X_{ij})$. Component-wise, the loss function is optimised when $w_i^\top w'_j = \log p(w_i, w'_j) - b_i - b_j + \log Z$ (for normalising constant Z). This expression bears strong similarity to (2), with *Glove*, in principle, having greater flexibility from its biases. *Glove* is of particular relevance due to this similarity of loss function with *W2V*, which is known to capture word meaning and which we seek to explain.

2.3 Interpreting *W2V*'s vectors

Recent strides have been made towards ascribing probabilistic properties to vector embeddings so as to justify (2) and, in turn, $W2V$. For example, [2] draw the correspondence:

$$\log p(w) \approx \tfrac{\|v_w\|}{2d}^2 - \log Z \qquad \text{and} \qquad \log p(w, w') \approx \tfrac{\|v_w + v_{w'}\|}{2d}^2 - 2\log Z, \tag{3}$$

giving the appealing result:

$$\begin{aligned} \text{PMI}(w, w') &= \log p(w, w') - \log p(w) - \log p(w') \\ &\approx \tfrac{1}{2d}\Big(\|v_w + v_{w'}\|^2 - \|v_w\|^2 - \|v_{w'}\|^2\Big) \quad \propto \quad v_w^\top v_{w'}. \end{aligned}$$

However, as the authors themselves note, the positive relationship between the marginal word probabilities and corresponding vector norms suggested by (3) is not observed empirically for *W2V* or similar embeddings, suggesting these relationships may not hold. We note that [2] and a related work [8] are the closest to ours of which we are aware and from which we take inspiration.

3 Towards a consistent correspondence between probabilities and vectors

With relationship (2) in mind, we generalise the task of embedding words in a vector space to one of translating probabilities into vectors such that their respective laws are maintained, or at least not grossly violated within the region of values of interest. Thus for any word pair w_i, w'_j we consider the problem of encoding three probabilities, $p(w_i)$, $p(w'_j)$ and $p(w_i, w'_j)$ in two vectors, v_i and v'_j.

In the realm of probabilities $\{p(w_i)\}$, we have concepts such as marginal, joint, independent and conditionally independent; and properties such as $0 \leq p(w_i) \leq 1$; $\Sigma_i p(w_i) = 1$; $p(w_i, w_j) \leq p(w_i),\ p(w_j)$. For vectors $\{v_i\}$ in a Euclidean vector space, with orthonormal basis $\{b_k\}$, we have concepts such as origin, parallel, orthogonal and dot product; and properties such as $\|v_i\| > 0$; $\Sigma_k \|{v_i}^\top b_k\|^2 = \|v_i\|^2$; $\|v_i + v_j\| \leq \|v_i\| + \|v_j\|$. Mapping one regime to the other consistently is non-trivial and cannot be made by arbitrarily linking properties, without the risk of discovering that a logical consequence in the vector space, when translated back to probabilities, violates some law. We proceed by considering the desired properties of a *semantic* vector space so as to have a goal in mind. We then start from laws of probability to consider how they can be achieved.

3.1 Desired properties of a vector space

We consider the vector space in polar co-ordinates, parameterising each point by its *direction* and *distance* to origin. In terms of direction, we seek to capture the common notion whereby words with similar meaning have similar directions.[3] It seems intuitive for orthogonality to then correspond to unrelatedness – or, in probabilistic terms, *independence*, i.e. observing one word makes no difference to the probability of observing the other. Thus we want the angle between vectors of words w_i, w'_j to reflect a notion of 'similarity' between w_i, w'_j (however that may be captured) and be orthogonal if, and only if, $p(w_i, w'_j) = p(w_i)p(w'_j)$. This leaves distance from the origin, or vector length, as a free parameter. Having associated $p(w_i, w'_j)$ with the angle between word vectors it may seem intuitive to relate their lengths to $p(w_i)$ and $p(w'_j)$, each being specific to one word. However, we do not wish to *assign* probabilities to vector properties, rather to let a consistent relationship materialise.

3.2 Transforming probabilities into vectors

To map probabilities to vectors, we start with a probabilistic expression and aim to re-parameterise until what remains can be directly interpreted in terms of vectors that have the desired properties outlined in Section 3.1. We thus begin by linking the three probabilities of interest with:

$$p(w_i, w'_j) \quad = \quad p(w_i) \times \left\{ \frac{p(w_i, w'_j)}{p(w_i)p(w'_j)} \right\} \times p(w'_j), \tag{4}$$

$$\text{letting:} \qquad \mathrm{PR}_{i,j} \quad = \quad \frac{p(w_i, w'_j)}{p(w_i)p(w'_j)} \quad = \quad \frac{p(w_i|w'_j)}{p(w_i)} \quad = \quad \frac{p(w'_j|w_i)}{p(w'_j)} \tag{5}$$

denote the *probability ratio*, i.e the factor by which w_i becomes more likely (PR$_{i,j} > 1$), less likely (PR$_{i,j} < 1$), or indeed remains unchanged (PR$_{i,j} = 1$), in the context of w'_j, relative to its marginal probability. By the definition of PMI(w_i, w'_j) (shortened to PMI$_{i,j}$), we have PMI$_{i,j} = \log \mathrm{PR}_{i,j}$. Now, by considering PR$_{i,j}$ for different values and relationships between the joint and marginal probabilities of w_i and w'_j, we find that we can re-parameterise PR$_{i,j}$ to conclude:

Lemma 1. *For all word pairs w_i, w'_j, we can find parameters $\theta_{i,j} \in [0, \pi]$, $\mathcal{L}_{i,j} \geq 0$ such that:*

$$\mathrm{PMI}(w_i, w'_j) = \cos\theta_{i,j}\, \mathcal{L}_{i,j} \ \ ,$$

$$\theta_{i,j} \in [0, \tfrac{\pi}{2}] \quad \textit{if} \quad \mathrm{PR}_{i,j} = \max_{i \textit{ or } j} \mathrm{PR}_{i,j} \qquad \textit{and} \qquad \theta_{i,j} = \tfrac{\pi}{2} \quad \textit{iff} \quad p(w_i, w'_j) = p(w_i)p(w'_j) \tag{6}$$

Proof. We consider the value of PR$_{i,j}$ in three cases: (1) PR$_{i,j} = \max\{{}^{1}\!/_{p(w_i)}, {}^{1}\!/_{p(w'_j)}\}$ at notional maximality where $p(w_i|w'_j) = 1$ or $p(w'_j|w_i) = 1$; (2) PR$_{i,j} = 1$ when $p(w_i)$ and $p(w'_j)$ are independent; and (3) PR$_{i,j} = 0$ where $p(w_i, w'_j) = 0$. We say *notional* in case 1, since no word-context pair has such property, e.g. the whole context window would need to contain w_i. Thus, if we define PR$^{max}_{i,j}$ as the maximum value of PR$_{i,j}$ as either w_i or w'_j are varied (not both), we see that PR$^{max}_{i,j}$ is finitely upper bounded for all i, j. To be clear, for given w_i, w'_j, we can find $\max_k\{\mathrm{PR}_{k,j}\}$, holding the context fixed and varying the target word, and an equivalent holding the target word and cycling through all contexts. By PR$^{max}_{i,j}$ we refer to the higher of these. We can similarly define PR$^{min}_{i,j}$ as the lowest value PR$_{i,j}$ attains as i, j vary, which, for *observed* word pairs, must also be finite. Note that we have PR$^{max}_{i,j} > 1$ if the weak assumption holds that there exists any context w'_k for which $p(w_i|w'_k) > p(w_i)$ *or* a target word w_k for which $p(w_k|w'_j) > p(w_k)$. [4] Taking logs, we can define PMI$^{max}_{i,j} = \log \mathrm{PR}^{max}_{i,j}$ and PMI$^{min}_{i,j} = \log \mathrm{PR}^{min}_{i,j}$.

[*Intuition*: Notionally, we gradually transition from statistics to lengths and angles. We have considered extremes of PR$_{i,j}$ as we will find that they correspond to a property of maximum length.]

For each word i we now introduce $\mathcal{L}_{i,i} > 0$ such that $\mathcal{L}_{i,i} \geq \max\{\mathrm{PMI}^{max}_{i,i}, \text{-}\mathrm{PMI}^{min}_{i,i}\}$. Thus, we have -$\mathcal{L}_{i,i} \leq \mathrm{PMI}_{i,k} \leq \mathcal{L}_{i,i}$ and -$\mathcal{L}_{i,i} \leq \mathrm{PMI}_{k,i} \leq \mathcal{L}_{i,i}$, $\forall k$. We can then define $\mathcal{L}_{i,j}$ for $i \neq j$ by:

$$\mathcal{L}_{i,j} = \sqrt{\mathcal{L}_{i,i}}\sqrt{\mathcal{L}_{j,j}} \ \ , \tag{7}$$

from which it is straightforward to show that $\mathcal{L}_{i,j} \geq \mathrm{PMI}_{i,j}, \ \forall\, i, j$. This makes it valid to re-parameterise PMI$_{i,j} = \beta_{i,j}\mathcal{L}_{i,j}$ for new parameter $\beta_{i,j} \in [-1, 1]$; and $\beta_{i,j} = 0$ for independent w_i, w'_j (since PR$_{i,j} = 1$). We redefine case 1 as PR$_{i,j}$ = PR$^{max}_{i,j}$ and case 3 as PR$_{i,j}$ = PR$^{min}_{i,j}$, the observed extremes, and track the re-parameterisation steps in Table 1.

[*Intuition* Having chosen length-related $\mathcal{L}_{i,j}$, we see that β now takes the range of an angle cosine.]

[3] This follows intuitively in the case of semantics, but we note that syntactic properties may be relevant also.

[4] Empirically we find, for > 99% of words, that PR$_{i,i} > 1$, i.e. seeing a word once makes it more likely to be seen again for a sufficient context window.

Table 1: Summary of parameterisations steps for cases of $p(w_i, w'_j)$

Case	$p(w_i, w'_j)$	PR$_{i,j}$	PMI$_{i,j}$	$\beta_{i,j}$	$\theta_{i,j}$	Condition
(1)	?	PR$_{i,j}^{max}$	PMI$_{i,j}^{max}$	$[0, 1]$	$[\frac{\pi}{2}, 0]$	maximal observation
(2)	$p(w_i)p(w'_j)$	1	0	0	$\frac{\pi}{2}$	independence
(3)	?	PR$_{i,j}^{min}$	PMI$_{i,j}^{min}$	$[-1, 0]$	$[\pi, \frac{\pi}{2}]$	minimal observation

With this property of $\beta_{i,j}$, we make the final re-parameterisation: $\theta_{i,j} = \arccos \beta_{i,j}$. Note that we have made no explicit claim that $\theta_{i,j}$ *is* an angle, simply a parameter, which completes the proof:

$$\text{PMI}(w_i, w'_j) = \beta_{i,j}\mathcal{L}_{i,j} = \cos\theta_{i,j}\mathcal{L}_{i,j}.$$

□

By Lemma 1, we have linked PMI(w_i, w'_j) to a vector dot product if $\mathcal{L}_{i,j}$ can be expressed as the product of the norms of vectors v_i and v'_j, having angle $\theta_{i,j}$. With a few further steps, we have:

Theorem 1. *For all word pairs w_i, w'_j, we can find vectors v_i, v'_j such that ${v_i}^\top v'_j$ = PMI(w_i, w'_j).*

Proof. By Lemma 1 and the specific construction of $\mathcal{L}_{i,j}$ in (7), for any word pair w_i, w'_j, we can find parameters $\mathcal{L}_{i,j}$ and $\theta_{i,j}$ such that:

$$\text{PMI}(w_i, w'_j) = \cos\theta_{i,j}\,\mathcal{L}_{i,j} = \cos\theta_{i,j}\sqrt{\mathcal{L}_{i,i}}\sqrt{\mathcal{L}_{j,j}} \tag{8}$$

Thus if we choose vectors v_i, v'_j to have $\|v_i\| = \sqrt{\mathcal{L}_{i,i}}$ and $\|v'_j\| = \sqrt{\mathcal{L}_{j,j}}$ at angle $\theta_{i,j}$, then:

$$\text{PMI}(w_i, w'_j) = \cos\theta_{i,j}\,\|v_i\|\|v'_j\| = {v_i}^\top v'_j \tag{9}$$

□

Thus we have proved that PMI of word occurrences, can be encoded in the properties of vectors from first principles. Subject to considerations of arrangement, which we come to, we know that vectors can be found such that the dot product of any pair gives the appropriate PMI value. Also, by (6) we have provably encapsulated the desired properties of orthogonality for independent words and closer alignment of vectors for words that are similar, in the sense that one improves the probability of seeing the other relative to its marginal, i.e. the *probability ratio* (5).

We see that probabilities are not captured explicitly, which is perhaps key to the correspondence being made. The issues of maintaining consistency with the laws of probability are largely side-stepped, e.g. rather than being constrained to [0,1], PMI maps to the full real line allowing all points of the vector space to be *valid*. Expressing PMI as a dot product is not new; our contribution is the derivation of the relationship from first principles such that statistical properties are specifically assigned to attributes of vectors, allowing probabilistic interpretation of properties of the vector space.

We consider a few further details of the proof in Appendix B. We can now rearrange (9) to show:

Corollary 1.1.

$$\log p(w_i) \;=\; -\frac{{v_i}^\top v'_i}{2} + \frac{\log p(w_i, w'_i)}{2} \tag{10}$$

$$\log p(w_i, w'_j) \;=\; -\frac{(v_i - v_j)^\top (v'_i - v'_j)}{2} + \frac{\log p(w_i, w'_i)p(w_j, w'_j)}{2}. \tag{11}$$

We draw comparison between Corollary 1.1 and previous work [2, 8] in Appendix B.3.

3.3 The relationship between the context and target vectors of the same word

We assume vectors follow the construction in the proof of Theorem 1 and consider the relationship between $\mathcal{L}_{i,i}$ and $\theta_{i,i}$, properties of v_i and v'_i, target and context vectors of the i^{th} word. Setting $i = j$, by (8) and (9) we have:

$${v_i}^\top v'_i = \cos\theta_{i,i}\mathcal{L}_{i,i} = \text{PMI}(w_i, w'_i)\ , \tag{12}$$

an explicit relationship between $\mathcal{L}_{i,i}$ and $\theta_{i,i}$, so between properties of length and the angle between our two vectors (the 'internal angle'), related by PMI$_{i,i}$, a fixed value specific to word i. It shows how

$\mathcal{L}_{i,i}$ affects the vectors: higher values mean higher internal angles and vice versa. Overall, we see that (12) binds the internal angle and norms of the two vectors for a word, subject to $\text{PMI}_{i,i}$; and (9) binds vector norms of different words to the angles between them, subject to $\text{PMI}_{i,j}$. Lastly, *if* $\theta_{i,i} = 0$ (i.e. v_i, v'_i align) then $\mathcal{L}_{i,i} = \text{PMI}_{i,i}$ and $\|v_i\| = \|v'_i\| = \sqrt{\text{PMI}}_{i,i}$ (we term this its 'minimum length', d_i), and vector lengths relate explicitly to probabilities, that can be estimated empirically.

3.4 Considering vector arrangement

Having derived a system of equations that encode PMI as vector properties, we have a firm mathematical basis from which to proceed, enabling probabilistic and geometric implications to be drawn. Here we consider how the 'competing forces' of (9), restricted to $i \neq j$ (assumed here throughout), and (12) interact, to better understand the role of terms *internal angle* and *minimum length*.

By induction, we consider placing each vector pair v_i, v'_i corresponding to the i^{th} word, in turn, into the vector space V, assumed to have unconstrained dimensionality. We initially assume $v_i = v'_i$, hence by (12) we have $\|v_i\| = \|v'_i\| = \sqrt{\text{PMI}}_{i,i} = d_i$. By (9), v_i, v'_i must also satisfy a system of simultaneous constraints $\mathcal{C}^{:i}$, with respect to previously placed vectors, whose span we denote $U^{:i} \subseteq V$. By $S^{:i} \subseteq U^{:i}$ we denote those vectors in $U^{:i}$ that maximally satisfy $\mathcal{C}^{:i}$. We now consider each vector pair in turn, finding either (i) some $s \in S^{:i}$ with $\|s\| = d_v$; (ii) $\|s\| < d_v \ \forall s \in S^{:i}$; or (iii) $\|s\| > d_v \ \forall s \in S^{:i}$. Loosely speaking, we can either find a vector in $U^{:i}$ best satisfying (9) that happens to have length d_i and so satisfying (12) also; or it must be that all such vectors in $U^{:i}$, best satisfying (9), are shorter or longer than d_i (See Appendix C for illustration).

In case (i) a solution exists without need for dimensionality of V beyond $U^{:i}$. In (ii), a solution can be found by choosing any $s \in S$ and adding a sufficient orthogonal component $t \in V, t \notin U^{:i}$, such that $v = s + t$ has the required *minimum length*. Loosely speaking, since our minimum length vector is 'too long to fit in S', we 'prop it up' in a new dimension such that its (now shorter) projection in $U^{:i}$ satisfies $\mathcal{C}^{:i}$. In case (iii) the opposite problem occurs, any minimum length vector is shorter than any vector in $S^{:i}$. However, we have a *pair* of vectors v_i, v'_i and it is their dot product constrained by (12). Therefore their norms can be made arbitrarily greater than the minimum length as the internal angle ranges from 0 to $\frac{\pi}{2}$, whilst their dot product remains constant, satisfying (12), in effect they are 'split in a new dimension' (see Appendix C). Thus we can find v_i, v'_i of equal length with opposite components in the new dimension, whose (now longer) projection in $U^{:i}$ falls in S, satisfying (9).

Thus each vector pair must satisfy a constraint (12) specific to the word it represents, governing its length and, *exclusively*, its internal angle; whilst finding the most fitting length and direction such that all other vectors interact (dot product) with it satisfying their mutual PMI relationships (9, for $i \neq j$).

3.5 Re-interpreting the relationship between vectors of the same word

This induction in fact allows the relationship between v_i and v'_i to be specified more quantitatively.

Theorem 2. *The target and context vectors v_i, v'_i of word i are complex conjugates of one another.*

Proof. Cases (ii) and (iii) in Section 3.4 can be shown to be equivalent to finding quadratic root(s) $h = \sqrt{{d_w}^2 - \|s\|^2}$, where h takes the length of the component in the 'new dimension' (unit vector $\hat{t}$) for the solution vector $v = s + h\hat{t}$. In case (ii), h is a single positive value since $d_w > \|s\|$. In case (iii), $d_w < \|s\|$ implies ${d_w}^2 - \|s\|^2 < 0$ and we have a pair of complex conjugates: $h = \pm b\, i$. Thus by construction of v, we have $v_i = s + b\, i\, \hat{t}$ and $v'_i = s - b\, i\, \hat{t}$, thus v_i and v'_i are themselves complex conjugates. We can then see that ${v_i}^\top v'_i = {v_i}^\top \overline{v}_i = \|s\|^2 - (b\ \|\hat{t}\|)^2 = \|s\|^2 - b^2 = \|s\|^2 + h^2 = \|s\|^2 + ({d_i}^2 - \|s\|^2) = {d_i}^2 = \text{PMI}_{i,i}$, as expected (where $\overline{v}$ denotes the complex conjugate of v). □

Thus the internal angle reflects an imaginary component of the embeddings, induced by the competing requirements of (9) and (12). Furthermore, it is the real component that relates to interactions *between* vectors to satisfy (9). More broadly, Theorem 2 suggests that $\mathbf{W}$ and $\mathbf{C}$ are also complex conjugates: $\mathbf{W} = \mathbf{A} + \mathbf{B}i$; $\mathbf{C} = \mathbf{A} - \mathbf{B}i$ for matrices $\mathbf{A}$, $\mathbf{B}$ respectively containing the *real* and imaginary components of all embeddings and thus spanning the real and imaginary subspaces of V. We can also use this to rearrange and restate Corollary 1.1:

$$\exp\left\{-\tfrac{{v_i}^\top \overline{v}_i}{2}\right\} = \tfrac{p(w_i)}{\sqrt{\log p(w_i, w'_i)}} \qquad \exp\left\{-\tfrac{(v_i - v_j)^\top (\overline{v_i - v_j})}{2}\right\} = \tfrac{p(w_i, w'_j)}{\sqrt{\log p(w_i, w'_i) p(w_j, w'_j)}}, \quad (13)$$

exhibiting an interesting quasi-spherical property over the vector space.

Given the similarities between the result of Theorem 1 and the loss functions of *W2V* and *Glove*, we might expect a similar phenomena to apply to those models also. This would then justify the summing of target and context vectors (i.e computing $2{\times}\mathbf{A}$) performed in *Glove*, also subsequently shown to materially benefit *W2V* on similarity tasks [11]. From $\mathbf{A} = \frac{\mathbf{W}+\mathbf{C}}{2}$ and $\mathbf{B} = \frac{\mathbf{W}-\mathbf{C}}{2}$, we see that this corresponds to using only real components, removing any increased angular elements in imaginary dimensions which might otherwise cause two similar words to appear less so.

4 Experiments

Models: To test if our theoretical insight can improve training of word embeddings, we implement a series of *Prob2Vec* (**P2V**) models with the architecture of *W2V*, trained to minimise a respective MSE loss function, shown below ($\alpha_i = 0.5$). As a baseline, we run the *W2V* algorithm as implemented in the *Gensim* python library [17], using the same parameters on all models. We pre-compute PMI values over the corpus (similar to [16]), substituting $p(w_i, w'_i) = {}^2\!/\!_3\, p_{min}$ for missing joint probabilities *with respect to the same word*, where p_{min} is the minimum of such values we do observe (see Section B.2). We generate 5 random negative word pairs for each positive pair, as is common for dealing with sparse pairwise data. We vary dimensionality (d: 200, 500) and corpus size ($\mathcal{D}$: 3.2m and 17m tokens) taken from the *text8* data set[5] (sourced from the English Wikipedia dump on Mar. 3, 2006). We filter words that appear less than 5 times and apply down-sampling (similar to [14]). We find it sufficient to train for 100 epochs (full passes over the PMI matrix) with 3.2m words and for 50 epochs with 17m. The loss functions of our models are as follows:

$$\begin{aligned}
&P2V\text{-}D: && \textstyle\sum_{i,j} (\, {v_i}^\top v'_j - \mathrm{PMI}_{i,j} \,)^2 \\
&P2V\text{-}L: && \textstyle\sum_{i,j} (\, {v_i}^\top v'_j - \mathrm{PMI}_{i,j} \,)^2 \quad + \quad \alpha_1(\, \|v_i\| - \sqrt{\mathrm{PMI}}_{i,i} \,)^2 \quad + \quad \alpha_2(\, \|v'_j\| - \sqrt{\mathrm{PMI}}_{j,j} \,)^2 \\
&P2V\text{-}P: && \textstyle\sum_{i,j} (\, {v_i}^\top v'_j - \mathrm{PMI}_{i,j} \,)^2 \quad + \quad \alpha_1(\, {v_i}^\top v'_i - \mathrm{PMI}_{i,i} \,)^2 \quad + \quad \alpha_2(\, \|v_i\| - \|v'_i\| \,)^2
\end{aligned}$$

P2V-D trains according to Theorem 1, effectively removing the 'log k' term or *debiasing* the implicit loss function of *W2V*. We anticipate that this term may significantly impact the geometry of *W2V* vectors, since dot products are driven more negative, causing angles $\theta_{i,j}$ to increase.

P2V-L regularises vectors to *minimum length*, assigning each to a hypersphere and reducing degrees of freedom. *Internal angles* and thus imaginary components are implicitly encouraged towards zero.

P2V-P prioritises terms for two vectors of the same word in the loss function, without restricting the *internal angle*; and regularises for $\|v_i\| = \|v'_i\|$ to reduce what the model needs to learn.

4.1 Evaluation

For each model, we evaluate vectors of **W** and **A** (the real component of **W** and **C** from our theory) on two tasks: *similarity* and *analogy*. Similarity is tested using *WordSim353* [6] and its subsets for similarity ('SIM') and relatedness ('REL') [1], containing word pairs with human-assigned similarity scores. Each word pair is ranked by cosine similarity and the evaluation is the correlation (Spearman's ρ) between those rankings and human ratings. Analogies are tested using Google's analogy data set [13] of c. 20k questions 'a is to b as c is to ..?' for a mix of semantic and syntactic analogies. We filter questions with out-of-vocabulary words, as standard [11]. Accuracy is computed by comparing $\operatorname{argmin}_d \|v_a - v_b - v_c + v_d\|$ to the labelled answer.

4.2 Results

Table 2 shows the results for all experiments. The key observations we highlight are:

- *P2V-P* outperforms all models on all tasks, with **A** vectors consistently outperforming **W** – showing our loss function improves vectors, in particular by embedding information in *internal angles*;
- Vectors from **A** outperform those of **W** in almost every case for models able to make use of the internal angle, i.e. *P2V-L* shows minimal benefit – supporting our theory that real components **A** reflect *inter-word* relationships, which the imaginary components may obscure;

[5] `http://mattmahoney.net/dc/textdata.html`

Table 2: Accuracy across *Similarity* and *Analogy* tasks of vectors from $\mathbf{W}$ and $\mathbf{A}$ (with difference δ). Corpus size ($\mathcal{D}$) and dimensionality (d) are indicated with best cases per setting and task in bold, Best vector choice ($\mathbf{W}$ vs $\mathbf{A}$) are underlined and best overall per task in blue.

	$\mathcal{D}$		Similarity-SIM			Similarity-REL			Similarity-All			Analogy		
Model	(m)	d	W	A	δ	W	A	δ	W	A	δ	W	A	δ
W2V	3.2	500	.439	.476	+.037	.415	.415	+.000	.410	.424	+.014	.035	.014	-.021
P2V-D			.662	.650	-.012	.627	.617	-.010	**.656**	.646	-.010	.217	.212	-.005
P2V-L			**.678**	**.686**	+.008	**.631**	**.633**	+.002	**.655**	**.659**	+.004	**.266**	**.271**	+.005
P2V-P			.628	.661	+.033	.598	.631	+.033	.630	.651	+.021	.221	.258	+.037
W2V	17	200	.744	.628	-.116	.687	.589	-.098	.705	.594	-.111	.421	.109	-.312
P2V-D			**.784**	**.788**	+.004	**.714**	**.715**	+.001	**.737**	**.740**	+.003	**.435**	.433	-.002
P2V-L			.670	.678	+.008	.553	.551	-.002	.598	.601	+.003	.260	.265	+.005
P2V-P			.752	.784	+.032	.689	.699	+.010	.702	.724	+.022	.411	**.441**	+.030
W2V	17	500	.412	.508	+.096	.256	.397	+.141	.311	.458	+.137	.045	.059	+.014
P2V-D			.735	.764	+.029	.693	.704	+.011	.692	.714	+.022	.393	.430	+.037
P2V-L			.758	.766	+.008	.680	.676	-.004	.703	.703	+.000	**.451**	.447	-.004
P2V-P			**.769**	**.796**	+.027	**.703**	**.719**	+.016	**.725**	**.744**	+.019	.411	**.453**	+.042

- When data is limited ($\mathcal{D}$=3.2m), *P2V-L* is the best performing model over all tasks, showing similar performance from $\mathbf{A}$ and $\mathbf{W}$ vectors – indicating that the direct regularisation of norms helps in lower data-to-parameter settings, we presume by restricting degrees of freedom;
- When dimensionality is restricted (d=200), *P2V-D* performs best (Similarity) or near best (Analogy) with similar performance from $\mathbf{A}$ and $\mathbf{W}$ – indicating that sufficient dimensionality is required for vectors to benefit from the increased use of internal angles; and
- Overall, with relatively limited data ($\mathcal{D}$=17m), our proposed models show performance, particularly on similarity tasks, comparable to that shown using vastly larger data sets ($\mathcal{D}$=1.5b) [11].

5 Conclusion & discussion

From first principles, we have shown that co-occurrence statistics (PMI) can be encoded in properties of vector embeddings such that vectors of similar words are aligned and those of unrelated words are orthogonal. This connection leads to a new loss function for *W2V* architecture training, which has a principled justification and produces improved word embeddings by directly corresponding probabilities to vector properties. The relationship also rationalises existing heuristic loss functions by their similarity to ours, justifying why popular embedding models such as *W2V* and *Glove* are able to capture word similarity within their embeddings. We have shown how this relationship sheds light on vector properties and, in particular, that the two vectors of the same word are complex conjugates, justifying the sometimes used heuristic of summing them. We show that improved vectors can be trained to achieve markedly improved performance, far outperforming $W2V$ with relatively little data in high dimensions, whilst also improving on a more comparable baseline (*P2V-D*).

In essence, our proposed method reduces the problem of positioning vectors *anywhere* in a high-dimensional space to one of locating each within a defined region by identifying a system of angles between vectors. By regularising based on our theory, we can also show that the generated vectors are endowed with more accurate probabilistic properties (see Appendix D), which may enable them to contribute more meaningfully in downstream tasks. We do not claim to establish a complete one-to-one correspondence between the properties of probabilities and vectors, indeed, we are able to overlook limits at which the relationship breaks down, e.g. $p(w_i, w'_j) = 0$. We leave to future work a full understanding of the limits of the correspondence.

Lastly, we note that we rely on no assumptions specific to word occurrences *per se*, hence we believe a similar approach may be applicable to other domains in which co-occurrence statistics are factored, such as recommender systems, graph embeddings and knowledge bases.

Acknowledgements

We thank Iain Murray for useful discussions. Carl Allen and Ivana Balažević were supported by the Centre for Doctoral Training in Data Science, funded by EPSRC (grant EP/L016427/1) and the University of Edinburgh.

References

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009.

[2] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 2016.

[3] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, 2002.

[4] Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. Explaining and generalizing skip-gram through exponential family principal component analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017.

[5] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[6] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, 2001.

[7] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[8] 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*, 2016.

[9] Andrew J Landgraf and Jeremy Bellay. word2vec skip-gram with negative sampling is a weighted logistic pca. *arXiv preprint arXiv:1705.09755*, 2017.

[10] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, 2014.

[11] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 2015.

[12] Oren Melamud and Jacob Goldberger. Information-theory interpretation of the skip-gram negative-sampling objective function. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

[15] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[16] 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*, 2014.

[17] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[18] Adriaan MJ Schakel and Benjamin J Wilson. Measuring word significance using distributed representations of words. *arXiv preprint arXiv:1508.02297*, 2015.

[19] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Appendix A Summary of results and assumptions of Theorem 1

A.1 Summary of results and identities

$$\text{PMI}(w_i, w'_i) \overset{(9)}{=} {v_i}^\top v'_i = \|v_i\|\|v'_i\| \cos\theta_{i,i} = \|v_i\|^2 \cos\theta_{i,i} \tag{14}$$

$$\|v_i\| \overset{(14)}{=} \sqrt{\frac{\text{PMI}_{i,i}}{\cos\theta_{i,i}}} \tag{15}$$

$$\text{PMI}(w_i, w'_j) \overset{(9)}{=} {v_i}^\top v'_j \overset{(15)}{=} \sqrt{\frac{\text{PMI}_{i,i}}{\cos\theta_{i,i}}} \sqrt{\frac{\text{PMI}_{j,j}}{\cos\theta_{j,j}}} \cos\theta_{i,j} \tag{16}$$

A.2 Assumptions of Theorem 1

The proof of Theorem 1 and its precursor Lemma 1 requires the following assumptions:

1. the vector space has sufficient dimension for all pairwise vector relationships to be achieved;
2. PMI(w_i, w'_j) for unobserved word pair w_i, w'_j, where the empirical estimate of $p(w_i, w'_j)$ is zero, is not explicitly captured in the arrangement of their vectors, hence, where PMI(w_i, w'_j) is captured, $\log p(w_i, w'_j)$ is assumed to be finite. Note: such word pairs arise during training as *negative samples*; their corresponding vectors are 'pushed apart' but are never able to achieve an inner product of $-\infty$, the empirical estimate of PMI(w_i, w'_j); and
3. $\|v_i\| = \|v'_i\|$ based on the symmetric treatment of target and context words (see Appendix B.1).

Appendix B Further considerations of Theorem 1

B.1 Considering the respective lengths of $\mathbf{v_i}$ and $\mathbf{v'_i}$

In the proof of Lemma 1 we introduced a square root term when setting $\mathcal{L}_{i,j}$ equal to $\sqrt{\mathcal{L}}_{i,i}\sqrt{\mathcal{L}}_{j,j}$, which ultimately gives the property $\|v\| = \|v'\|$. We note that our results would hold equally if we chose $\mathcal{L}_{i,j} = [\mathcal{L}_{i,i}]^p[\mathcal{L}_{j,j}]^{1-p}$ for any $p \in [0, 1]$. However, the resulting embeddings would differ as p ranges over $[0, 1]$: lower values of p give shorter target word vectors v_i and longer context word vectors v'_j and vice versa. Importantly, the angles $\theta_{i,j}$ would adjust, altering the distribution of meanings across the vector space.

Here, due to the symmetric treatment of the target and context words and of **W** and **C** more generally, e.g. each word is, in expectation, considered equally as target and context, vectors interact via the dot product; it seems intuitive to set $p = 0.5$, such that equal length is assigned to each vector. We note that in other potential applications of the theory this may differ. We also point out that this relationship is generally observed in word embeddings from *W2V* architectures.

B.2 Computing PMI ($\boldsymbol{w_i}$, $\boldsymbol{w_i}$)

Equations 12 and 9 require the computation of PMI(w_i, w'_i) and PMI(w_j, w'_j). However, this term is undefined for any words that do not co-occur with themselves thus having a empirical estimate of the joint probability of zero, which is increasingly prevalent for rare words. Three options to overcome this: (i) ignore cases where this value cannot be computed from the training process; (ii) force the value to be computable by adding each target word to its context window; and (iii) fill in missing values of the joint probability, e.g. with the minimal value observed, or one slightly lower.

B.3 Comparison of our results to other works [2, 8]

Comparing our results in Corollary 1.1 to those of [2], as shown in (3), we can see a strong resemblance in overall form, in particular an exponential radial effect. However, importantly, we see that the signs are opposite: both dot products are negated and it is the difference between vectors not their sum that relates to the joint probability. However, it is not a case that relationships are simply *opposite*, since we see that the rightmost terms of (10) and (11) are not normalisation constants, as are the equivalent term in (3). Rather these terms corresponds to the joint probability of seeing the same word as both target and context. This probability understandably bares a strong relationship with the marginal $p(w_i)$ and is certainly not constant across words. Finally, (13) shows that simply taking vector norms, as contemplated in (3), is potentially meaningless as it assumes $v_i = v'_i$ and that all *internal angles* are zero.

Interestingly, we see in [8] that by a route very different to our own, a relationship is drawn between the expectation of $C_{i,j}$, the observation count of pair w_i, w'_j proportional to $p(w_i, w'_j)$, and the difference between word vectors, whereby:

$$\mathbb{E}[C_{i,j}] = \exp\{-\frac{\|v_i - v_j\|^2}{2} + a_i + b_j\}\ ,$$

for 'unigram normalisers' a_i, b_j. This bears stronger similarity to (11) in Corollary 1.1, in particular the signs match giving a more similar overall relationship. However, we differ in the use of vector norms compared to dot product (as above) and we have shown that we require log joint probabilities in place of the monogram biases, although there is a strong relationship between the two.

Appendix C Illustrating vector placement

Figures 1 and 2 respectively illustrate cases (ii) and (iii) in the induction step of placing each vector in turn. The subspace spanned by previously placed vectors U^i is indicated by the blue plane: we show two dimensions, u_x, u_y but there could be many more. The pink region indicates S^i (of which s is a particular member) the set of vectors in U^i that best satisfy constraint set $\mathcal{C}$ with respect to previously placed vectors, according to (9). The solution vectors v_i, v_i' are shown in blue, having component of length h in the 'new dimension' of unit vector $\hat{t}$. The black arc is of the minimum length d_i and thus indicates vectors with a *internal angle* of zero that satisfy (12). In the first case, we can find a solution that has the minimum length and projects onto S^i, in the latter case, that is impossible and the internal angle must be non-zero and the vector lengths increased from their minimum length.

Figure 1: Positioning a vector pair v_i, v_i' when *minimum length* d_i is 'too long', i.e. $d_i \geq \|s\| \; \forall s \in S^i$, by 'propping up the vector pair in a new dimension' (of unit vector $\hat{t}$).

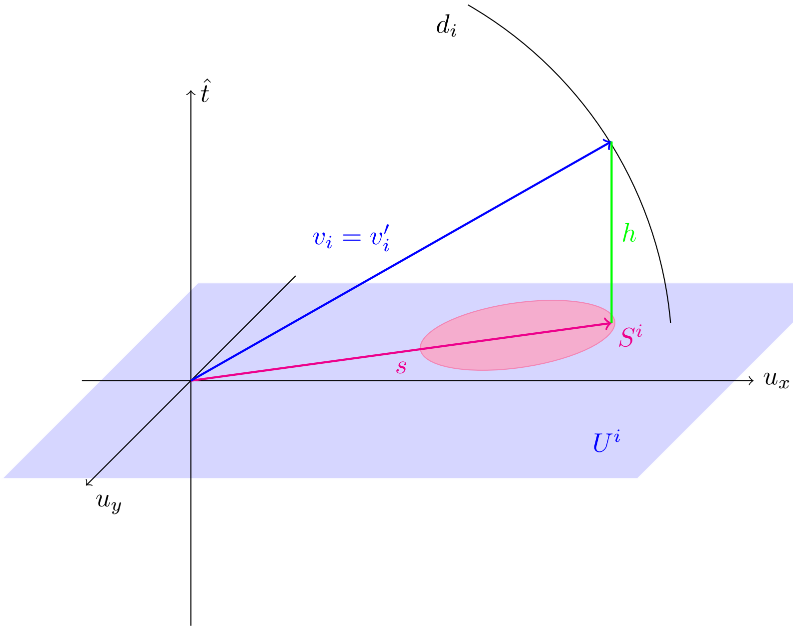


Figure 1 shows that if the *minimum length*, d_i, determined by $\text{PMI}_{i,i}$, is longer than all vectors in S^i, then no vector in U^i satisfies both (9) and (12), thus one must be found in V 'outside' U^i. Hence we can consider the vector pair 'propped up' in a new dimension so that their projection in U^i falls in S^i (as shown by s). We do not consider separating the vector pair as that would require them to be longer to satisfy their own constraint and would only move in the wrong direction in terms of satisfying (9).

Figure 2: Positioning a vector pair v_i, v'_i when *minimum length* d_i is 'too short', i.e. $d_i \leq \|s\| \; \forall s \in S^i$, by 'splitting the vector pair in a new dimension' (of unit vector $\hat{t}$).

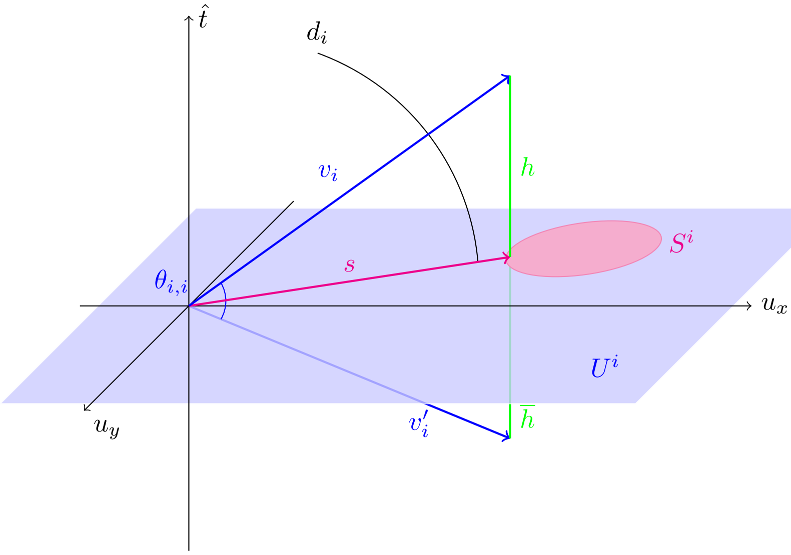


Figure 2 shows that if d_i (black arc) is shorter than any vector in S^i, then the vector pair v_i, v'_i can be 'split apart' in a new dimension whilst their dot product continues to satisfy (12). Their lengths increase from the *minimum length* until their projection in U^i falls in S^i (again shown by s) whereby all constraints are satisfied..

C.1 Geometric and probabilistic interpretations of *minimum length*

In vector terms, word i can have zero *internal angle* and its vectors take their *minimum length* (defined directly by $\text{PMI}_{i,i}$) only if that length is sufficient for the dot product with every other vector to satisfy (9); or if that length is, in some sense, *long enough*. In probabilistic terms, this equates to whether a word's own probability ratio is sufficiently high relative to that word's probability ratios with all other words, i.e. the extent to which it becomes more likely *in its own presence*, compared to the extent to which other words make it more probable. (This can be related back to the term $\mathcal{L}_{i,j}$ in Lemma 1).

Appendix D Contours of probability over the vector space

Having a relationship between properties of vectors and PMI values, allows us to predict how certain probabilities vary across the vector space, i.e. to predict their *contours*. Such predictions can be compared to empirical observations. Evaluating our predictions may be useful for several reasons: (i) by providing empirical support for the theory (or otherwise); (ii) not only providing evidence as to whether the theory may be right or wrong, but that it can be *learned*, e.g. we have not considered local minima or convexity issues at this point; and (iii) it may show how well the vectors directly reflect various probabilistic properties as an indication of whether this might be utilised downstream.

To test predictions empirically, we plot the respective positions of all target words w_i relative to a particular context word w'_j. We consider the direction of vector v'_i as our reference and thus plot w'_j at a distance $\|v'_j\|$ from the origin along the x-axis. We compute the norm of all target word vectors, $\|v_i\|$ and their angle to v'_j by:

$$\theta_{i,j} = \tfrac{{v_i}^\top v'_j}{\|v_i\|\|v'_j\|} \quad .$$

We then plot all target words relative to our context word at a point v_i from the origin with (positive) angle $\theta_{i,j}$ to the x-axis. This projects the entire cloud of word vectors onto 2 dimensions and into the upper half-plane, capturing all the information we need as lengths and angle to v'_j are preserved.

We compute empirical estimates of each probability for all words and 'discretise' words into subsets according to their value. Each subset comprises all words for which the log of the chosen probability sits within a range, the width of which is constant across subsets. Plotting a variety of subsets in different colours, should allow probability contours to be observed. We now predict the shape of contours for probabilities $p(w'_j|w_i)$ and $p(w_i|w'_j)$. The contours of each of these are of interest as they allow us geometrically to answer question such as 'which contexts are most likely having seen X?' or 'which words do we expect next having seen Y?'

D.1 Predictions

Considering $p(w'_j|w_i)$, (9) gives ${v_i}^\top v'_j = \text{PMI}(w_i, w'_j) = \log p(w'_j|w_i) - \log p(w'_j)$. Hence for fixed w'_j, words of equal $p(w'_j|w_i)$ must project onto v'_j at the same point, so lie in the same plane orthogonal to v'_j. Also, $p(w'_j|w_i)$ should increase in the direction of v'_j (positively along the x-axis).

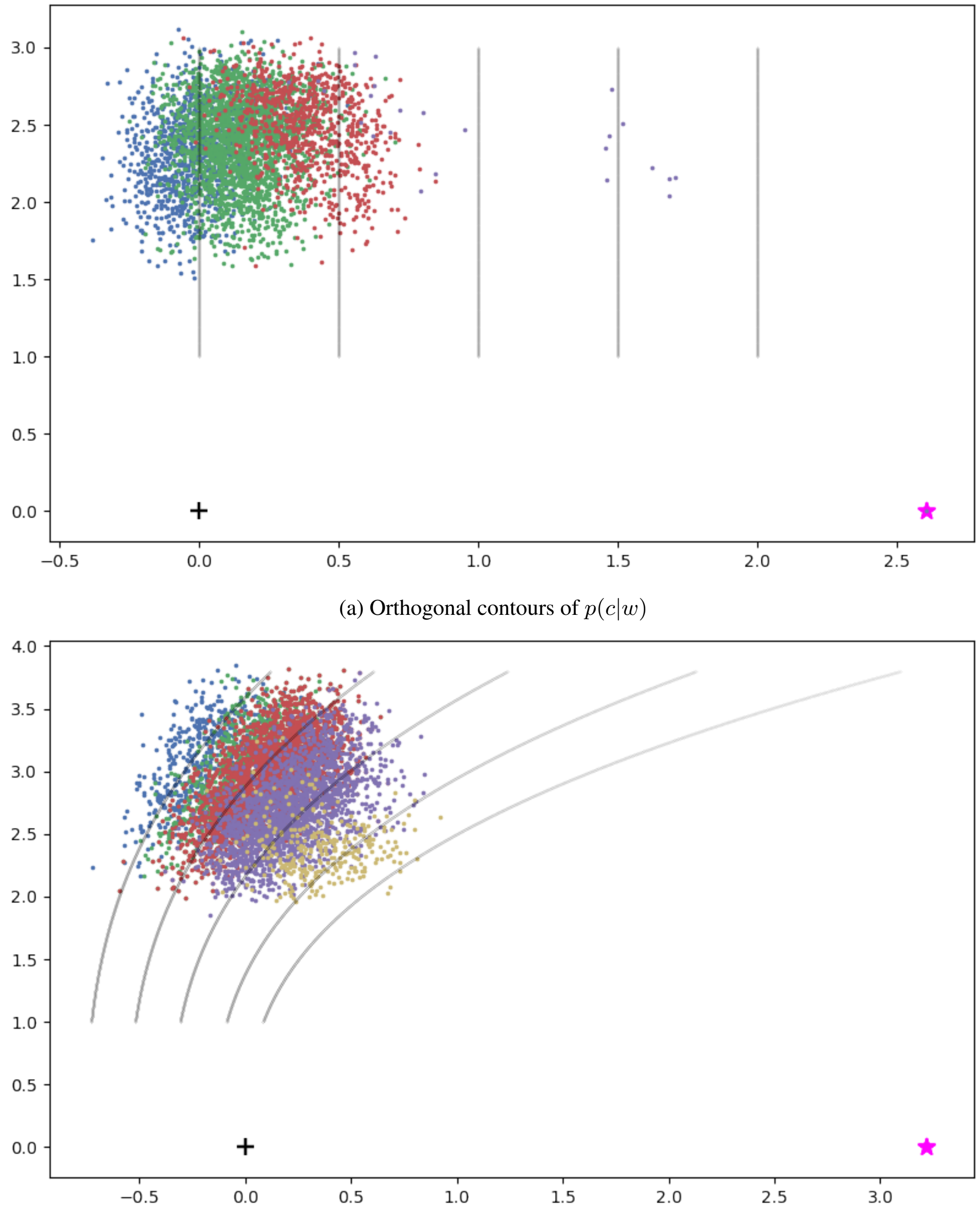
Considering $p(w_i|w'_j)$, we have ${v_i}^\top v'_j = \log p(w_i|w'_j) - \log p(w_i)$, which is less straightforward as for fixed $\log p(w_i|w'_j)$, $\log p(w_i)$ can still vary. Here we can use our understanding that $p(w'_j|w_i)$ increases along the axis to deduce that for fixed $\log p(w_i, w'_j)$ (by constant $\log p(w_i|w'_j)$ and w'_j), $\log p(w_i)$ must decrease along the x-axis. So we have a relationship ${v_i}^\top v'_j \approx \alpha x + \beta$ for constants α, β, and thus $\alpha x - \gamma\|v_i^{(x)}\| \approx \beta$, for constant $\gamma = \|v_j'^{(x)}\|$ and $\|v_i^{(x)}\|$, the projection of v_i onto the x-axis. It can be shown that this relationship induces a curve similar to a horizontal parabola (consider a typical plot of $y = x^2$ rotated 90° clockwise).

D.2 Observations

We construct plots as described for the illustrative context word 'four'. We add contours based on our prediction for comparison (note these are for illustration only and not accurately computed).

Figure (3a) shows the plot for subsets of words with similar $\log p(w'_j|w_i)$, bounded within ± 0.4 of -6.8 (blue), -5.7 (green), -4.5 (red) to -3.2 (purple). This shows the orthogonal contours predicted do exist within the vector space and we have the correct trend of $p(w'_j|w_i)$ increasing as we move along the x-axis.

Figure (3b) shows the relative locations of subsets of all words with $\log p(w_i|w'_j)$ within ± 0.7 of $-\log p(w_i|w'_j)$: -17.8 (blue), 16.0 (green), 15.6 (red), 14.5 (purple), 12.7 (yellow).



(a) Orthogonal contours of $p(c|w)$

(b) Parabolic-like contours of $p(w|c)$

Figure 3: Probability contours of the vector space: plots show word locations, considering vector length and angle, relative to the context vector for the word ‘four’ (pink star) for groups of target words (indicated by colour) with similar empirical estimates of the respective probability. ‘+’ marks the origin.