Journal of Machine Learning Research 1 (2021) 1- Submitted 12/20; Revised: 5/21; Published 11/21

Further results on latent discourse models and word embeddings

Sammy Khalife[*] KHALIFE@LIX.POLYTECHNIQUE.FR
LIX, CNRS, Ecole Polytechnique
Institut Polytechnique de Paris
91128, Palaiseau, France

Douglas Gonçalves DOUGLAS@MTM.UFSC.BR
MTM/CFM - Universidade Federal de Santa Catarina
88040-900, Florianópolis, Brazil

Youssef Allouah[†] YOUSSEF.ALLOUAH@POLYTECHNIQUE.EDU
Ecole Polytechnique
Institut Polytechnique de Paris
91128, Palaiseau, France

Leo Liberti LIBERTI@LIX.POLYTECHNIQUE.FR
LIX, CNRS, Ecole Polytechnique
Institut Polytechnique de Paris
91128, Palaiseau, France

Editor: Qiaozhu Mei

Abstract

We discuss some properties of generative models for word embeddings. Namely, (Arora et al., 2016) proposed a latent discourse model implying the concentration of the partition function of the word vectors. This concentration phenomenon led to an asymptotic linear relation between the pointwise mutual information (PMI) of pairs of words and the scalar product of their vectors. Here, we first revisit this concentration phenomenon and prove it under slightly weaker assumptions, for a set of random vectors symmetrically distributed around the origin. Second, we empirically evaluate the relation between PMI and scalar products of word vectors satisfying the concentration property. Our empirical results indicate that, in practice, this relation does *not* hold with arbitrarily small error. This observation is further supported by two theoretical results: (i) the error cannot be exactly zero because the corresponding shifted PMI matrix cannot be positive semidefinite; (ii) under mild assumptions, there exist pairs of words for which the error cannot be close to zero. We deduce that either natural language does not follow the assumptions of the considered generative model, or the current word vector generation methods do not allow the construction of the hypothesized word embeddings.

Keywords: Generative models, latent variable models, asymptotic concentration, natural language processing, matrix factorization

∗. New affiliation: Johns Hopkins University, Department of Applied Mathematics and Statistics
khalife.sammy@jhu.edu

†. New affiliation: Ecole Polytechnique Fédérale de Lausanne (EPFL), IC School
youssef.allouah@epfl.ch

1. Introduction

Context and Motivations

The construction of intermediate representations is essential for language models and their applications. These representations can be cast in two groups. The first consists of vector space models with *static* word embeddings, where a minimal unit of the language, a word, is associated to a fixed and constant representation. This representation encodes the meaning of the word, independently of its context. There exist many examples of static representations, such as word2vec (Mikolov et al., 2013a) or Glove (Pennington et al., 2014) representations. The second group, which we refer to as *contextual* embeddings, maps each word of the vocabulary to a vector which depends on its context. Long short term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), or neural networks with attention mechanisms (e.g. Bidirectional Transformers (Vaswani et al., 2017; Devlin et al., 2019)) are examples of methods to construct contextual representations.

Despite the fact that contextual embeddings are considered to have superseded the use of standard vector space models for applications, most of their properties, in particular the relation with language semantics, remain obscure. On the other hand, the family of static embeddings in (Arora et al., 2016) have been advertised to possess geometric properties related to language semantics, in particular with respect to analogies. In this work, we discuss the foundations of such statements, in particular concerning the properties of a latent model for natural language generation.

Previous Work

The model we will consider has been presented in (Arora et al., 2016): a generative model using prior probability distributions to compute closed form expressions for word statistics. It originally aimed at providing a piece of explanation of the linear structure for analogies (Arora et al., 2016, 2018b). The apparent relation of linear structures of word vectors and semantic analogies has already been studied in (Khalife et al., 2019), going in favor of an incidental phenomenon rather than systematic. For the sake of clarity, we will present in the remaining of this subsection the main assumptions of this generative model.

In the following, $f = O(g)$ (resp. $f = \tilde{O}(g)$) means that f is upper bounded by g (resp. upper bounded ignoring logarithmic factors) in the considered neighborhood. Let d be a strictly positive integer corresponding to the word vectors dimension. The generation of sentences in a given text corpus is made under the following generative assumptions.

- **Assumption 1** *The text generation process is driven by a random walk of a vector $c_t \in \mathbb{R}^d$, called discourse vector, such that if w_t is the word emitted at step t, then*

$$\mathsf{P}(w_t = w | c_t) \propto \exp(\langle c_t, v_w \rangle) \tag{1}$$

  where $v_w \in \mathbb{R}^d$ is the word vector for word w (the vectors c_t and v_w are latent variables of the model). Moreover, the random walk $(c_t \,|\, t \geq 1)$ admits a uniform stationary distribution on the unit sphere.

- **Assumption 2** *The ensemble of word vectors consists of independent and identically distributed (i.i.d.) samples generated by $v = s\,\hat{v}$, where $\hat{v}$ is drawn from the spherical Gaussian distribution in $\mathbb{R}^d$ and s is an integrable random scalar such that $|s| \leq \kappa$, for a positive constant κ.*

- **Assumption 3** $(c_t \,|\, t \geq 1)$ *jumps are small on average. More precisely,* $\exists \epsilon_1 \geq 0$ *such that* $\forall t \geq 1$*:*

$$\mathbb{E}_{c_{t+1}}(\exp(\kappa\sqrt{d}||c_{t+1} - c_t||_2)) \leq 1 + \epsilon_1 \tag{2}$$

Contributions

The contribution of this work is three-fold. First, we present a theoretical result concerning the concentration of a certain partition function Z_c for a generative model following Equation (1). Our statement concerns the behavior of the partition function: it concentrates around its mean, for simpler assumptions than those (1 to 3) aforementioned. Informally, this property means that the variations of Z_c with respect to c (and its randomness) are relatively negligible. This property is very similar to the concentration phenomenon demonstrated in (Arora et al., 2016). Our result suggests that it is not an intrinsic characteristic of word embeddings satisfying the Gaussian prior (Assumption 2), since it holds for other random vectors symmetrically distributed around the origin.

Second, we empirically investigate the relation between the geometry of word vectors and PMI. Although the experiments reported in (Arora et al., 2016, Section 5) support the concentration phenomenon and a linear correlation between squared norms of word vectors and word frequencies (at least, for high frequency words), little is said about the relationship between PMI of word pairs and the scalar product of their word vectors. In this work, we perform a thorough empirical investigation of this relationship. Our extensive experiments strongly support the claim that theoretical relations derived from the considered generative model occur at best in some regimes of the co-occurrence terms.

Finally, we provide evidence that the implicit matrix factorization problem related to the construction of word embeddings is ill-posed for a *symmetric* PMI model, since the shifted PMI matrix (as explored in (Levy and Goldberg, 2014)) is not positive semidefinite. To do so, we establish necessary conditions for the positive definiteness of the shifted symmetric PMI matrix in terms of local pairwise probabilities and show these local conditions can be violated in natural language.

2. Result on the Concentration of the Partition Function

In this section, we discuss a theoretical property presented in (Arora et al., 2016), called the concentration of the partition function. Based on (1), given a discourse vector c, the corresponding partition function value Z_c is defined as:

$$Z_c = \sum_v \exp(\langle v, c \rangle) \tag{3}$$

where v are the word vectors. We remind our reader that the considered generative model treats corpus generation as a dynamic process, where the t-th word is produced at step t. The process is driven by a random walk of a discourse vector c. Its coordinates represent the current topic. In this section, we are interested in an asymptotic property of the partition function Z_c. By analogy with statistical physics, this partition function is the sum of probabilities of the particles state given macroscopic parameters, such as temperature, over all the particles. More precisely, in our context, the particles considered are words and the states are the appearances of a word given a latent discourse vector (which is the analogous with the physical temperature). This latent discourse vector represents a context of fixed length. The aim of this section is to study the variations of Z_c with respect to the

random variable c. This study is motivated by the use of partition concentration as a theoretical basis to demonstrate the relationships between PMI and scalar product of word vectors (Arora et al., 2016).

If the word vectors satisfy Assumptions 1 and 2, and n is the number of words, then the concentration of the partition function is stated as follows (Arora et al., 2016, Lemma 2.1):

$$\mathsf{P}[(1-\epsilon_z)\,Z \;\leq\; Z_c \;\leq\; (1+\epsilon_z)\,Z] \geq 1-\delta \tag{4}$$

for some constant Z (independent of c), $\epsilon_z = \tilde{O}(1/\sqrt{n})$ and $\delta = \exp(-\Omega(\log(n)))$, where $\Omega(g)$ is a function lower bounded by g. We are interested in this property since it is central for the development of all the following theorems and propositions in (Arora et al., 2016), including the relation between PMI of word pairs and the scalar product of their word vectors (Arora et al., 2016, Theorem 2.2). Indeed, based on Equation (1) and applying the law of total expectation, it is possible to derive an estimate for the joint probability of a word pair w, w' (Arora et al., 2016):

$$p(w,w') = \mathbb{E}_{c_t,c_{t+1}}[p(w|c_t)p(w'|c_{t+1})] = \mathbb{E}_{c_t,c_{t+1}}\left[\frac{\exp(\langle v_w, c_t\rangle)}{Z_{c_t}}\frac{\exp(\langle v_{w'}, c_{t+1}\rangle)}{Z_{c_{t+1}}}\right], \tag{5}$$

where $Z_c = \sum_v \exp(\langle v, c\rangle)$, c is a context vector and the sum is over all word vectors v. Furthermore, in the experiments conducted in (Arora et al., 2016, Section 5.1), the property expressed in Equation (4) is empirically evaluated with the histogram of the partition function Z_c (which should concentrate around its mean) for word vectors obtained from common methods, such as GloVe and word2vec. By doing so, this concentration of partition function is implicitly considered as a means to evaluate how well the word vectors follow the generative model. In this section, we will show that the property holds (modulo a small constant) not only for random vectors described by Assumptions 1 and 2, but for a set of random vectors with bounded norm and symmetrically distributed around the origin.

2.1 Preliminaries

Before presenting our main inequality in Section 2.2, we state three lemmata used for its demonstration.

Lemma 1 *Let $\psi : \mathbb{R} \to [0, +\infty[$ be a twice continuously differentiable strictly convex even function, satisfying the following properties:*

1. $\beta \mapsto \psi'(\beta)/\beta$ *is injective on* $\mathbb{R}^{+*}$
2. $\forall \beta \neq 0 \quad \psi''(0) - \psi'(\beta)/\beta > 0$
3. $\forall \beta \neq 0 \quad \psi''(\beta) - \psi'(\beta)/\beta < 0$

Then, the function $\Psi(x) \triangleq \sum_{i=1}^{d} \psi(x_i)$, subject to $\|x\|^2 = R^2$, with $R > 0$, has the following extreme points:

1. $x^* = \pm R e_k, \forall k \in [d]$, *where e_k is the k-th canonical vector of $\mathbb{R}^d$, corresponding to global minimizers;*
2. $x_i^* = \pm \dfrac{R}{\sqrt{d}}$, *for $i = 1, \ldots, d$, corresponding to global maximizers.*

Proof Since we are considering a continuous function over a compact set, it attains a maximum and a minimum in the feasible set. Besides, at any feasible point the linear independence constraint qualification holds (because $x \neq 0$), which ensures that for any minimizer (or maximizer) x, there exists a Lagrange multiplier $\lambda \in \mathbb{R}$ such that

$$\nabla \Psi(x) + \lambda x = 0$$

or equivalently

$$\forall i \in \{1, \dots, d\} \qquad \psi'(x_i) + \lambda x_i = 0 \tag{6}$$

For the remaining of this proof, let $(x, \lambda) \in \mathbb{R}^d \times \mathbb{R}$ be a fixed feasible point and scalar verifying Equation (6). Also notice that since ψ is continuously differentiable and even, we have $\psi'(0) = 0$.

Since x is feasible, there should be components of x verifying $x_i \neq 0$. For the non-zero components of x, Equation (6) must hold for the same λ. First, we remark that $\lambda \neq 0$. Indeed, if $\lambda = 0$, then from Equation (6), $\forall i \ \psi'(x_i) = 0$, but ψ is strictly convex and $\psi'(0) = 0$, which implies that $x_i = 0$ for all i, leading to an infeasible point.

Thus, for the non-zero components of x, from Equation (6), we obtain

$$x_i \neq 0 \implies \lambda = -\frac{\psi'(x_i)}{x_i} \neq 0$$

But, since $\beta \mapsto \psi'(\beta)/\beta$ is injective on $\mathbb{R}^{+*}$, we conclude that the non-zero components of x must be all equal, i.e $\exists \beta^* > 0 \ \text{s.t.} \ \forall i \ x_i \neq 0 \implies x_i = \beta^*$. From the feasibility of x, we conclude that

$$\beta^* = \pm \frac{R}{\sqrt{\|x\|_0}}$$

where $\|x\|_0$ denotes the number of non-zero entries of x.

Let us now analyze the second order conditions for the feasible points verifying Equation (6). Since the objective function is separable, the Hessian of the Lagrangian $\nabla^2_{xx} L(x, \lambda)$ is a diagonal matrix whose diagonal entries verify $\forall i \in \{1, \dots, d\}$:

$$[\nabla^2_{xx} L(x, \lambda)]_{ii} = \psi''(x_i) - \psi'(\beta^*)/\beta^*$$

For the remaining of this proof, for given α and β, let $\delta(\alpha, \beta) \triangleq \psi''(\alpha) - \psi'(\beta)/\beta$. We remind our reader that, by assumption, $\beta \neq 0 \implies \delta(0, \beta) > 0$ and $\delta(\beta, \beta) < 0$.

Therefore, for a given $y \in \mathbb{R}^d$, we have

$$y^T \nabla^2_{xx} L(x, \lambda) y = \delta(0, \beta^*) \sum_{i : x_i = 0} y_i^2 + \delta(\beta^*, \beta^*) \sum_{i : x_i \neq 0} y_i^2$$

If all components of x are non-zero, then we get $\forall y \in \mathbb{R}^d \setminus \{0\}$:

$$y^T \nabla^2_{xx} L(x, \lambda) y = \delta(\beta^*, \beta^*) \sum_{i : x_i \neq 0} y_i^2 < 0$$

This proves that x verifying $\forall i \in \{1, \dots, d\}, \quad x_i = \pm \frac{R}{\sqrt{d}}$ satisfy the second order sufficient conditions for a local maximizer.

Now, let us show that if x has at least one zero component and more than one non-zero components, then x is a saddle-point. Without loss of generality, assume that exactly two entries of x are non-zero, then due to the previous discussion, they must be equal, e.g. $x^T = (0, \dots, 0, \beta, \beta)$. The second order sufficient conditions concern the Hessian of the Lagrangian with respect to primal variables, which should be positive definite when restricted on the linear null space of the Jacobian of the constraints. In this case, this linear space is given by:

$$x^\perp = \{y \in \mathbb{R}^d \ : \ y = (w_1, \dots, w_{d-2}, \alpha, -\alpha), w \in \mathbb{R}^{d-2}, \alpha \in \mathbb{R}\}$$

In particular, choosing

$$y = (w_1, 0, \dots, 0, \alpha, -\alpha) \in x^\perp$$

we obtain

$$y^T \nabla^2_{xx} L(x, \lambda) y = \delta(0, \beta^*) w_1 + 2\delta(\beta^*, \beta^*)\alpha^2$$

Then:

i) $w_1 > 0 \quad \alpha = 0 \implies y^T \nabla^2_{xx} L(x, \lambda) y > 0$

ii) $w_1 = 0 \quad \alpha \neq 0 \implies y^T \nabla^2_{xx} L(x, \lambda) y < 0$

This implies that x is neither a minimizer nor a maximizer.

Finally, if $x = \pm R e_k$, for some canonical vector e_k, we obtain, for every $y \in x^\perp \setminus \{0\}$,

$$y^T \nabla^2_{xx} L(x, \lambda) y = \delta(0, \beta^*) \sum_{i : x_i = 0} y_i^2 + \delta(\beta^*, \beta^*) \times 0 = \delta(0, \beta^*) \sum_{i : x_i = 0} y_i^2 > 0$$

which proves that $x = \pm R e_k$ satisfies the second order sufficient conditions for a local minimizer.

Furthermore, since ψ is even, and the maximizers (and minimizers) described above only differ by the sign of their entries, we can conclude that all of them are global. ■

We present a similar result for the annulus domain:

Lemma 2 *Let η be a strictly positive real, and **1** the vector of ones of appropriate dimension. With the same conditions and notations as in Lemma 1, replacing the sphere of radius R with the annulus Ω_η defined by:*

$$\Omega_\eta = \{x \in \mathbb{R}^d \mid R \leq ||x||_2 \leq R + \eta\} \tag{7}$$

we have that

(i) $\forall k \in [d]$, $x = Re_k$, is a global minimizer of Ψ on Ω_η,

*(ii) $\frac{R+\eta}{\sqrt{d}}$**1** is a global maximizer of Ψ on Ω_η.*

Proof Both (i) and (ii) can be proved in two steps:
(i) Since ψ is even, we limit the study on the set of positive vectors. We show that the maximum of ψ is reached on the sphere of radius $R + \eta$, whereas the minimum is achieved on the sphere of radius R. This can be proved by remarking that:

$$\begin{aligned} x > 0, \quad x \in \mathring{\Omega}_\eta \quad \text{and} \quad R < \lambda ||x|| < R + \eta \\ \implies \lambda x \in \Omega_\eta \quad \text{and} \quad \psi(\lambda x) > \psi(x) \end{aligned}$$

which can be deduced by the fact that ψ is strictly convex and $\psi'(0) = 0$, hence ψ is increasing on $\mathbb{R}^+$. This implies that the minimum of Ψ is reached on the sphere of radius R, and its maximum on the sphere of radius $R + \eta$.
(ii) Then, we use Lemma 1 to conclude. ■

Lemma 3 *Let $L > 0$ and consider the function f: $\mathbb{R}^d \longrightarrow \mathbb{R}$ defined by:*

$$f(c) \quad = \quad \prod_{i=1}^{d} \begin{cases} \frac{\sinh(Lc_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ L & \text{otherwise} \end{cases} \tag{8}$$

Then $\Psi = \log(f)$ verifies the assumptions of Lemma 1.

Proof Left in the Appendix. ■

2.2 Main Inequality

We now present our result concerning the partition function. Proposition 4 shows that the concentration property also holds for very simple distributions of the word and context vectors, independently of Assumptions 1 and 2. More precisely, the concentration property is holding with isotropy and uniformity on word vectors in a centered cube of $\mathbb{R}^d$, and such that latent discourse vectors belong to a sufficiently thin annulus[1]. The analysis of Equation 5 would be difficult if not intractable without the concentration property. We remark this property was empirically verified for certain common word embedding methods (Arora et al., 2016, cf. Section 5.1), and combined with Assumption 3, is sufficient to prove the relations between PMI and the scalar product of word vectors, that we discuss experimentally in Section 3.

Proposition 4 *Let n be the number of words, and let us suppose the word vectors are generated independently and uniformly in a centered cube of $\mathbb{R}^d$. Then, if the discourse vectors belong to the annulus domain Ω_η, for $R \leq 2$, and a sufficiently small η, then there exists $\gamma \ll 1$ such that $\forall \epsilon > 0$, the following inequality holds with probability $1 - \alpha$:*

$$(1 - \epsilon)(1 - \gamma)\,\mathbb{E}[Z_0] \;\leq\; Z_c \;\leq\; (1 + \epsilon)(1 + \gamma)\,\mathbb{E}[Z_0] \tag{9}$$

where $Z_0 = Z(c_0)$, for a constant discourse vector c_0, and $\alpha \leq \exp(-\frac{1}{2}\epsilon^2 n^2)$.

Proof Our proof is decomposed in three steps:
∘ Use Bernstein inequalities to bound $|Z_c - \mathbb{E}[Z_c]|$ with high probability.
∘ Compute a closed form expression of the mapping $c \mapsto \mathbb{E}[Z_c]$.
∘ Study the variation of this function over Ω_η using Lemma 3.

1. A systematic description of the priors for which this concentration occurs is left open for future work.

Let $v, c \in \mathbb{R}^d$ be the word and discourse vectors, respectively, with the following properties:

$$\|v\| \leq \kappa \tag{10}$$

$$\mathbb{E}\left[\langle v, c\rangle\right] = 0 \tag{11}$$

From (10) and Cauchy-Schwarz inequality

$$\langle v, c\rangle \leq |\langle v, c\rangle| \leq \|v\|\|c\| \leq 3\kappa$$

where $||c|| \leq 3$ because $c \in \Omega_\eta$, for $R \leq 2$ and η sufficiently small. It follows that

$$\exp\langle v, c\rangle \leq \exp 3\kappa \tag{12}$$

Since the random vectors v are i.i.d. and by convexity of the exponential, we have from (11)

$$\begin{aligned} \mathbb{E}\left[Z_c\right] = n\mathbb{E}\left[\exp\langle v, c\rangle\right] &\geq n \exp \mathbb{E}\left[\langle v, c\rangle\right] \\ &\geq n \exp(0) = n \end{aligned} \tag{13}$$

Moreover, we are also able to bound the variance of Z_c:

$$\begin{aligned} \text{Var}\left[Z_c\right] = \sum_v \text{Var}\left[\exp\langle v, c\rangle\right] &= n\text{Var}\left[\exp\langle v, c\rangle\right] \\ &\leq n\mathbb{E}\left[\exp 2\langle v, c\rangle\right] \\ &\leq n\mathbb{E}\left[\exp(6\kappa)\right] = \exp(6\kappa)n \end{aligned} \tag{14}$$

Now let Λ be the constant defined as follows:

$$\Lambda = \exp(6\kappa)$$

Let $\epsilon > 0$. Thanks to (12) and (14), we can apply the Bernstein's inequality to the sum of random variables $Z_c = \sum_v \exp\langle v, c\rangle$, to obtain

$$P[|Z_c - \mathbb{E}\left[Z_c\right]| > \varepsilon n] \leq \exp\left(-\frac{\frac{1}{2}\varepsilon^2 n^2}{n\Lambda + \frac{2}{3}\sqrt{\Lambda}\varepsilon n}\right) \tag{15}$$

and from (13)

$$P[|Z_c - \mathbb{E}\left[Z_c\right]| > \varepsilon \mathbb{E}\left[Z_c\right]] \leq \exp\left(-\frac{\frac{1}{2}\varepsilon^2 n^2}{n\Lambda + \frac{2}{3}\sqrt{\Lambda}\varepsilon n}\right) \tag{16}$$

which shows the concentration of Z_c around $\mathbb{E}\left[Z_c\right]$ for *any* fixed unit norm vector c.

Let us show now that $\mathbb{E}\left[Z_c\right]$ does not vary much with c. To this end, we need additional assumptions about the distribution of v apart from (10) and (11). We are interested in $\mathbb{E}[Z_c]$, and in particular the amplitude of its variation with respect to c. If the word vectors admit a density function ξ, then:

$$\mathbb{E}_v[\exp(\langle v, c\rangle)] = \int_\Omega exp(\langle v, c\rangle)\xi(v)dv$$

If the word vectors are independent and identically distributed, it should be noted that:

$$\mathbb{E}_v[Z_c] = n\,\mathbb{E}_v[\exp(\langle v, c\rangle)]$$

where n is the number of words. Firstly, in order to simplify the calculation, we will consider that v is distributed uniformly on Ω which is the cube of $\mathbb{R}^d$ centered in 0, of side length $2L$. Then, integration using Fubini Theorem yields:

$$\mathbb{E}_v[\exp(\langle v, c\rangle)] = \frac{1}{L^d}\prod_{i=1}^{d}\phi(c_i)$$

where

$$\phi(c_i) = \begin{cases} \frac{\sinh(Lc_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ L & \text{otherwise} \end{cases}$$

Consider the function $f\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by $f(c) = \prod_{i=1}^{d}\phi(c_i)$. We will first discuss the variations in the amplitude of f on the sphere $\mathcal{S}_R$ centered at 0 with radius R. The relative amplitude of the variations of f on $\mathcal{S}_R$ is given by:

$$\frac{\max_{c\in\mathcal{S}_R} f(c) - \min_{c\in\mathcal{S}_R} f(c)}{\min_{c\in\mathcal{S}_R} f(c)} \tag{17}$$

Using Lemmas 3 and 1, we can infer the two following properties:

• On the one hand, f reaches its maximum at a point c such that $c_1 = c_2 = \ldots = c_d = \frac{R}{\sqrt{d}}$. And then

$$\max_{c\in\mathcal{S}_R} f(c) = [\frac{\sqrt{d}}{R}\sinh(\frac{LR}{\sqrt{d}})]^d$$

• On the other hand, the minimum of f is reached for a point where every coordinate has been set to 0 except one (such point exists on the sphere), and therefore, f reaches its minimum on a point c such that

$$\phi(c_1) = \ldots = \phi(c_{d-1}) = L$$
$$\text{and} \quad \phi(c_d) = \frac{\sinh(LR)}{R}$$

Hence,

$$\min_{c\in\mathcal{S}_R} f(c) = L^{d-1}\frac{\sinh(LR)}{R}$$

(It is interesting to observe that the minimum of f does not depend on the dimension if $L = 1$.)

It should be noted that the maximum ***relative variation*** of $\mathbb{E}[Z_c] = (n/L^d)\,f(c)$ is the same as that of f, given in Equation (17).

Now, let us observe the behavior of the maximum of f, when the dimension d tends to infinity. The Taylor expansion of order 3 for $\sinh$ at 0 is given by:

$$\sinh(x) = x + \frac{x^3}{6} + o(x^3)$$

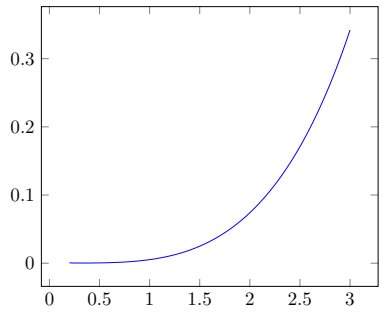


Figure 1: Illustration of the maximum relative variations of $\mathbb{E}[Z_c]$, with the function $\Delta : x \mapsto \frac{e^{\frac{x^2}{6}}}{\frac{\sinh(x)}{x}} - 1$. The x-axis represents the radius considered and th e y- axis the value of the maximum relative variation.

Therefore, using properties of the exponential:

$$\begin{aligned}\max_{c\in\mathcal{S}_R} f(c) &= \left(\frac{\sqrt{d}}{R}\right)^d \left[\frac{LR}{\sqrt{d}} + \frac{1}{6}\left(\frac{LR}{\sqrt{d}}\right)^3 + o\left(\left(\frac{LR}{\sqrt{d}}\right)^3\right)\right]^d \\ &= \left(L + \frac{L^3R^2}{6d} + o(\frac{1}{d})\right)^d = L^d\left(1 + \frac{L^2R^2}{6d} + \frac{1}{L}o(\frac{1}{d})\right)^d \underset{d\to+\infty}{\sim} L^d e^{\frac{L^2R^2}{6}}\end{aligned}$$

Then, if $d \gg 1$ (e.g. $d \geq 50$):

$$\Delta(R) \quad = \frac{\max_{c\in\mathcal{S}_R} f(c) - \min_{c\in\mathcal{S}_R} f(c)}{\min_{c\in\mathcal{S}_R} f(c)} \underset{d\to+\infty}{\sim} L\frac{e^{\frac{L^2R^2}{6}}}{\frac{\sinh(LR)}{R}} - 1$$

This ratio does not depend on the dimension, regardless of the radius of the sphere considered. With slight abuse of notation, let

$$||\Delta||_\infty \triangleq ||\Delta||_{\infty,(0,2]} = \sup_{0<R\leq 2} \Delta(R) \tag{18}$$

The graph of the function $\Delta : R \mapsto \Delta(R)$ for $L = 1$ is drawn in Figure 1. In particular, $||\Delta||_\infty \leq 10^{-1}$. This implies that if $R \leq 2$ (and $L \leq 1$):

$$\frac{\max_{c\in\mathcal{S}_R} \mathbb{E}[Z_c] - \min_{c\in\mathcal{S}_R} \mathbb{E}[Z_c]}{\min_{c\in\mathcal{S}_R} \mathbb{E}[Z_c]} = \Delta(R) \leq 10^{-1} \tag{19}$$

Finally, if Ω_η is replaced by the domain defined by

$$R \leq ||x||_2 \leq R + \eta$$

then the extrema of f on Ω_η can be deduced from Lemma 2 and are given by

$$\min_{c\in\Omega_\eta} f(c) = L^{d-1}\frac{\sinh(LR)}{R}$$

$$\max_{c\in\Omega_\eta} f(c) = [\frac{\sqrt{d}}{R+\eta}\sinh(\frac{L(R+\eta)}{\sqrt{d}})]^d$$

Similarly,

$$\Delta(R) \underset{d\to+\infty}{\sim} L\frac{e^{\frac{L^2(R+\eta)^2}{6}}}{\frac{\sinh(R)}{R}} - 1$$

Let $L = 1$ and denote by

$$\Delta_\eta \triangleq \frac{e^{\frac{(R+\eta)^2}{6}}}{\frac{\sinh(R)}{R}} - 1 \tag{20}$$

such maximum variation for a given η. Plots of Δ_η for several values of η are given in Fig. 2b.

Let Z_0 be a partition function for a constant discourse vector $c_0 \in \mathcal{S}_R$. Notice that

$$|\, Z_c - \mathbb{E}[Z_c]\,| > \epsilon\mathbb{E}[Z_c] \qquad \Longleftrightarrow \qquad \Big|\, \frac{Z_c}{\mathbb{E}[Z_0]} - \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]}\,\Big| > \epsilon\frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \tag{21}$$

and using the previous study (see Eqs. (18) and (19)), we obtain

$$\Big|\frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} - 1\Big| \leq ||\Delta||_\infty$$

which implies that

$$\epsilon\frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \geq \epsilon(1 - ||\Delta||_\infty)$$

From Equation (21):

$$|\, Z_c - \mathbb{E}[Z_c]\,| > \epsilon\mathbb{E}[Z_c] \implies \Big|\, \frac{Z_c}{\mathbb{E}[Z_0]} - \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]}\,\Big| > \epsilon(1 - ||\Delta||_\infty)$$

Let $\mathcal{E}$ be the event corresponding to the right hand side. Then:

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(|Z_c - \mathbb{E}[Z_c]| > \epsilon\mathbb{E}[Z_c]) \leq \alpha \tag{22}$$

where the second inequality is obtained from Equation (16). We recall that ϵ is an arbitrarily small real number, and

$$\alpha = \exp\left(-\frac{\frac{1}{2}\varepsilon^2 n^2}{n\Lambda + \frac{2}{3}\sqrt{\Lambda}\varepsilon n}\right)$$

Hence, with (high) probability $1 - \alpha$:

$$-\epsilon(1 - ||\Delta||_\infty) + \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \;\leq\; \frac{Z_c}{\mathbb{E}[Z_0]} \;\leq\; \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} + \epsilon(1 - ||\Delta||_\infty) \leq \;\frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} + \epsilon(1 + ||\Delta||_\infty)$$

Again, using:

$$1 - ||\Delta||_\infty \leq \frac{\mathbb{E}[Z_c]}{\mathbb{E}[Z_0]} \leq 1 + ||\Delta||_\infty$$

We finally have with probability $1-\alpha$:

$$(1-\epsilon)(1-||\Delta||_\infty)\,\mathbb{E}[Z_0] \;\leq\; Z_c \;\leq\; (1+\epsilon)(1+||\Delta||_\infty)\,\mathbb{E}[Z_0]$$

ϵ is arbitrarily small, and we saw that $||\Delta||_\infty \leq 10^{-1}$, for a domain close to a sphere of radius $R \leq 2$. Setting $\gamma = ||\Delta||_\infty$ concludes the proof.

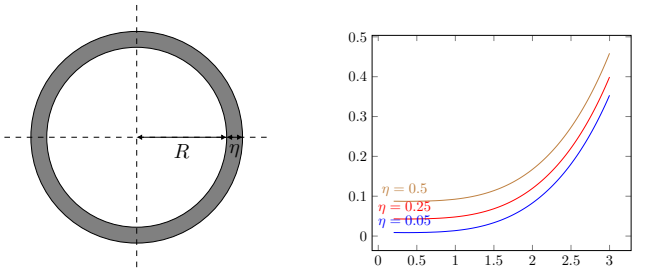
■

Figure 2 illustrates the behavior of the maximum relative variation

$$\Delta_\eta = \frac{\max_{c\in\Omega_\eta} \mathbb{E}\left[Z_c\right] - \min_{c\in\Omega_\eta} \mathbb{E}\left[Z_c\right]}{\min_{c\in\Omega_\eta} \mathbb{E}\left[Z_c\right]}$$

for different values of η as R increases. Such behavior for small enough R and η allows us to apply the Bernstein inequality to arrive at inequality (9) with high probability.

As mentioned earlier, the concentration property combined with Assumption 3 allows to derive the main theoretical results of Arora et al. (2016), a relation between statistical information (pointwise mutual info) and the scalar product of their word vectors. Combined with the assumptions of Property 4, Assumption 1 and 3 also allow to obtain the same relations. This can be verified for instance when the context vectors follow a uniform distribution on the sphere of radius R. In the next section we discuss these relations in more details.



(a) Ω_η (b) Δ_η for $\eta \in \{0.05, 0.25, 0.5\}$

Figure 2: Illustration of the maximum relative variations of $\mathbb{E}[Z_c]$ for $L = 1$ on Ω_η. (a) The annulus domain of width η. In (b), the x-axis represents the radius R considered and the y-axis the value of the maximum relative variation Δ_η (see Equation (20)).

3. Relation between PMI and scalar product

In this section, we provide an empirical evaluation of the main theorem presented in (Arora et al., 2016). First, we state some relations claimed in Arora et al. (2016). Second, we conduct experimental verifications of the aforementioned relations and comment on their results. Third, we discuss the relations in light of the experimental findings.

Let $p(w, w')$ be the probability of words w and w' appearing together in a window of size q in the corpus, $p(w)$ and $p(w')$ be the corresponding marginal probabilities and $v_w, v_{w'} \in \mathbb{R}^d$ the respective word vectors. Theorem 2.2 in (Arora et al., 2016) gives approximations for $\log p(w, w')$

and $\log p(w)$ as linear functions of $\|v_w + v_{w'}\|^2$ and $\|v_w\|^2$ respectively. Such approximations lead to a linear approximation of the Pointwise Mutual Information (PMI) of two words w and w':

$$\mathsf{PMI}(w, w') = \log \frac{p(w, w')}{p(w)p(w')}$$

by the scalar product $\langle v_w, v_{w'} \rangle$ of their word vectors. These results are gathered in the following theorem:

Theorem 5 *(Arora et al., 2016, Theorem 2.2) Suppose the word vectors satisfy the inequality* (4)*, and the window size $q = 2$. Then*

$$\log p(w, w') = \frac{||v_w + v_{w'}||^2}{2d} - 2\log Z \pm \epsilon, \tag{23}$$

$$\log p(w) = \frac{||v_w||^2}{2d} - \log Z \pm \epsilon, \tag{24}$$

for $\epsilon = O(\epsilon_z) + \tilde{O}(1/d) + O(\epsilon_1)$. Jointly, these imply:

$$\mathsf{PMI}(w, w') = \frac{\langle v_w, v_{w'} \rangle}{d} \pm O(\epsilon). \tag{25}$$

In the theorem, $\epsilon_z = \tilde{O}(1/\sqrt{n})$ comes from inequality (4) (Arora et al., 2016, Lemma 2.1) and ϵ_1 is from Assumption 3. See (Arora et al., 2016) for details[2]. For a window size $q > 2$ we have the following:

Corollary 6 *(Arora et al., 2016, Corollary 2.3) Under the assumptions of Theorem 5, and considering $p(w, w')$ and $\mathsf{PMI}(w, w')$ for window size $q > 2$:*

$$\log p(w, w') = \frac{||v_w + v_{w'}||^2}{2d} - 2\log Z + \Gamma \pm \epsilon, \tag{26}$$

$$\mathsf{PMI}(w, w') = \frac{\langle v_w, v_{w'} \rangle}{d} + \Gamma \pm O(\epsilon), \tag{27}$$

where $\Gamma = \log q(q-1)/2$.

In Arora et al. (2016), numerical experiments are presented to empirically verify (24), but little is said about equations (25) and (27). It is just commented that (Arora et al., 2016, Remarks 1): "For PMI however, the noise level $O(\epsilon)$ could be comparable to the leading term, and empirically we also find higher error here". Such remark motivated us to conduct extensive experiments in order to empirically verify equations (25) and (27) (see Section 3.1).

It is also mentioned in (Arora et al., 2016) that relation (27) is consistent with the result of (Levy and Goldberg, 2014), which showed that without dimension constraints, the solution to the

2. We also provide the proof leading to equation (24) in the Appendix (Lemma 9), using Eq. 2.13 given in Arora et al. (2019).

optimization problem in skip-gram with negative sampling (Mikolov et al., 2013b) corresponds to a factorization of a shifted asymmetric PMI matrix:

$$\forall w, w' \quad \mathsf{PMI}(w, w') = \langle \hat{v}_w, \hat{v}_{w'} \rangle - \beta$$

for a suitable constant β. We will discuss this result for the case of a *symmetric* PMI matrix in Section 4. In the following, we will present the results of an experimental verification of Theorem 5 and Corollary 6. Later, a discussion of these results follows.

3.1 Experimental verification

The experimental verification consists in performing a linear regression (we provide the slope, the intercept, and the coefficient of determination R^2, along with the Pearson correlation value) to verify the relations (23)–(27) with $\epsilon = 0$. Thus, when we talk about validity of one of such relations we mean its validity for $\epsilon \approx 0$ (it is intuitive that for large enough ϵ all these relations are valid but they become meaningless).

Word vectors Since Theorem 5 assumes that the word vectors satisfy the concentration property described by Equation (4), we considered GloVe (Pennington et al., 2014) and SN (Squared norm) (Arora et al., 2016) word vectors because they empirically verify such property (Arora et al., 2016, Section 5.1). We recall their respective optimization formulations.

Let $X_{w,w'}$ be the number of times words w and w' co-occur within the same window in the corpus, $f_1(X_{w,w'}) = \min(X_{w,w'}, 100)$ and $f_2(X_{w,w'}) = \min(X_{w,w'}^{3/4}, 100)$.

The SN formulation is given by:

$$\min_{v,C} \sum_{w,w'} f_1(X_{w,w'})(\log X_{w,w'} - \|v_w + v_{w'}\|_2^2 - C)^2 \tag{28}$$

whereas GloVe considers the following optimization problem:

$$\min_{v,s,C} \sum_{w,w'} f_2(X_{w,w'})(\log X_{w,w'} - \langle v_w, v'_w \rangle - s_w - s'_w - C)^2 \tag{29}$$

The SN word embeddings were reproduced using code available at (Arora et al., 2018a) which tries to solve (28) using AdaGrad (Duchi et al., 2011) with initial learning rate 0.05 and 25 training epochs. Pre-trained GloVe word embeddings made available by (Pennington et al., 2014) were used as well.

Datasets The English Wikipedia was used to train the SN word embeddings. The corpus was pre-processed using the standard approach (non-textual elements removed, sentences split, tokenized). Only words appearing more than 1000 times are considered. Three different extracts from the English Wikipedia dump were used. The first corpus (denoted corpus 1) consists of the first 1 million documents of the 2016 Wikipedia dump, deprived of prepositions and pronouns. The second corpus and third corpus (denoted corpus 2 and corpus 3 respectively) consist of the first 1,072,907 and 3,170,407 documents, respectively, of the 2020 Wikipedia dump. A description of the corpora is available in Table 1.

All the results of our experiments are reported in the tables 2, 3, 4. The results of tables 2 and 3 are based solely on corpus 1.

Corpus	Vocabulary size	Number of tokens
1	16,927	266,561,061
2	39,317	1,020,897,871
3	62,051	2,035,545,719

Table 1: Description of the corpora

Table 2: Results for the experimental verification of equations (23) and (26) for SN word embeddings.

Dimension	Window size	Pearson correlation	Slope	Intercept	R^2
50	2	0.73	0.0368	-23.57	0.53
100	2	0.74	0.0243	-24.75	0.55
200	2	0.76	0.0157	-26.11	0.57
300	2	0.76	0.0119	-27.00	0.58
50	10	0.78	0.0517	-27.19	0.61
100	10	0.79	0.0320	-28.40	0.62
200	10	0.79	0.0191	-29.46	0.63
300	10	0.80	0.0139	-30.02	0.64

In these experiments, we consider the following approximations for the underlying probabilities

$$p(w,w') \approx \bar{p}(w,w') \triangleq \frac{X_{w,w'}}{\sum_{(v,v')\in\mathcal{V}\times\mathcal{V}} X_{v,v'}}, \qquad p(w) \approx \bar{p}(w) \triangleq \frac{X_w}{\sum_{v\in\mathcal{V}} X_v}, \tag{30}$$

where $X_w,\ X_{w,w'}$ are respectively the occurrence count of w and the co-occurrence count of (w,w'). Therefore, in the following discussion, the term $\mathsf{PMI}(w,w')$ should be interpreted as its empirical approximation, given by

$$\overline{\mathsf{PMI}}(w,w') \triangleq \log \frac{\bar{p}(w,w')}{\bar{p}(w)\bar{p}(w')} \tag{31}$$

unless stated otherwise.

Table 3: Results for the experimental verification of equation (24) for SN word embeddings. The partial linear regression is based on the 50 points corresponding to words with the largest frequencies.

Regression	Dimension	Window size	Pearson correlation	Slope	Intercept	R^2
full	50	2	0.84	0.115	-16.61	0.70
full	100	2	0.85	0.073	-18.21	0.73
full	200	2	0.89	0.044	-19.69	0.79
full	300	2	0.91	0.03	-20.29	0.83
full	50	10	0.86	0.120	-16.86	0.74
full	100	10	0.85	0.071	-17.96	0.73
full	200	10	0.87	0.040	-18.96	0.75
full	300	10	0.88	0.028	-19.41	0.78
partial	50	10	0.86	0.049	10.77	0.72
partial	100	10	0.85	0.024	10.39	0.73
partial	200	10	0.87	0.012	10.30	0.74
partial	300	10	0.88	0.008	10.38	0.70

Table 4: Results for the experimental verification of equation (25) and (27) for SN and GloVe word embeddings. The partial linear regression is based on points with PMI less than 5.

Corpus	Embedding	Dimension	Regression	Window size	Pearson correlation	Slope	Intercept	R^2
1	SN	50	full	2	0.04	0.0061	0.93	0.002
1	SN	100	full	2	0.09	0.0088	0.89	0.008
1	SN	200	full	2	0.17	0.0109	0.86	0.03
1	SN	300	full	2	0.24	0.0117	0.85	0.06
1	SN	50	full	10	0.09	0.0129	0.32	0.008
1	SN	100	full	10	0.11	0.0094	0.29	0.011
1	SN	200	full	10	0.13	0.0070	0.27	0.02
1	SN	300	full	10	0.15	0.0063	0.27	0.02
2	SN	50	full	2	0.21	0.0360	1.07	0.04
3	SN	50	full	2	0.21	0.0378	1.14	0.04
2	SN	50	partial	2	0.17	0.0269	1.02	0.03
3	SN	50	partial	2	0.16	0.0256	1.03	0.02
2	GloVe	50	full	2	0.05	0.0082	1.17	0.003
2	GloVe	100	full	2	0.03	0.0029	1.20	0.001
3	GloVe	50	full	2	0.10	0.0179	1.21	0.011
3	GloVe	100	full	2	0.08	0.0086	1.23	0.007

Equations (23) **and** (24) From Table 3, it is clear that a high correlation exists between $\log p(w)$ and $||v_w||^2$, as predicted by equation (24), along with a fairly satisfying determination coefficient. However, the experimental slope of this linear relationship from the theoretical $\frac{1}{2d}$ ($= 0.01$ for $d = 50$, for example) for all of the dimensions seen in the experiments. Still, the relationship between the experimental slopes and the theoretical slopes, in an evolution w.r.t the inverse of the dimension, is satisfyingly linear both for (23) and (24). For example in Equation (24), we find the experimental slopes to be ten times the theoretical slopes when dimension grows. This discrepancy cannot be worked around without violating the requirement in Arora et al. (2016) for all word vectors to have a squared $l2$ norm of the order of dimension. On the other hand, this implies that for word vectors that have a large enough $l2$ norm, that is high frequency words, Equation (24) becomes empirically possible together with the norm constraint.

For equation (24), we also performed a partial linear regression based on the 50 points corresponding to words with the largest frequencies. For this regression, the slope approximation from the partial linear regression is much closer to the theoretical one.

For equation (23), Table 2 shows that, although the linear correlation values are satisfyingly large, the experimental slope values for window size 2 do not match with the theoretical $\frac{1}{2d}$.

In order to empirically estimate the theoretical intercept, we approximate[3] $Z \approx 1.67 \times 10^4$, which gives $\log Z \approx 9.72$. The experimental values do not exactly match with the theoretical value of the intercept (approximately -19.44 and -9.72 for equations (23) and (24) respectively with window size 2). The error in the intercept is larger for equation (24). A possible explanation for why the error on the intercept is smaller for equation (23) is that the SN optimization problem (28) tries to fit equation (23)[4].

3. Z was computed as the empirical mean of sampled partition function values Z_c, computed using equation (3), by sampling random context vectors c in the unit sphere.

4. viewing v_w as $\bar{v}_w/\sqrt{2d}$ and considering the approximation $p(w, w') \approx \bar{p}(w, w')$.

Equation (25) As can be observed from Table 4, for the results based on corpus 1, the correlation values are somewhat low and the determination coefficient values are poor. For the latter, it was argued in (Arora et al., 2016, Remark 1) that the high noise $O(\epsilon)$ in Eq. (25) is the principal reason. Figure 3 shows the magnitude of the noise in equation (25).

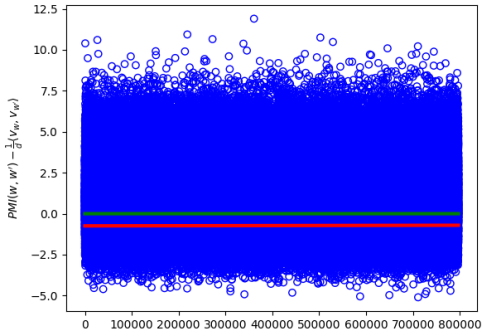


Figure 3: y-axis: $\mathsf{PMI}(w, w') - \frac{1}{d}\langle w, w'\rangle$, x-axis: arbitrary indexation of couples of words (w, w'). Based on corpus 2 and SN word embeddings with dimension 50, window size 2. Green line: constant value 0. Red line: regression line. Only one in every 100 points was plotted.

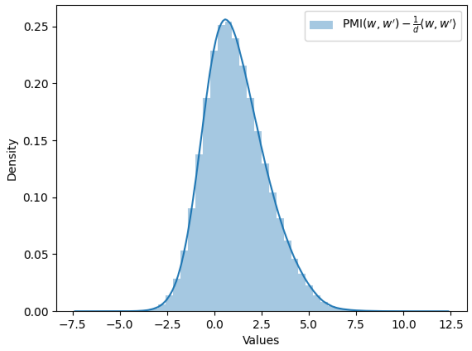
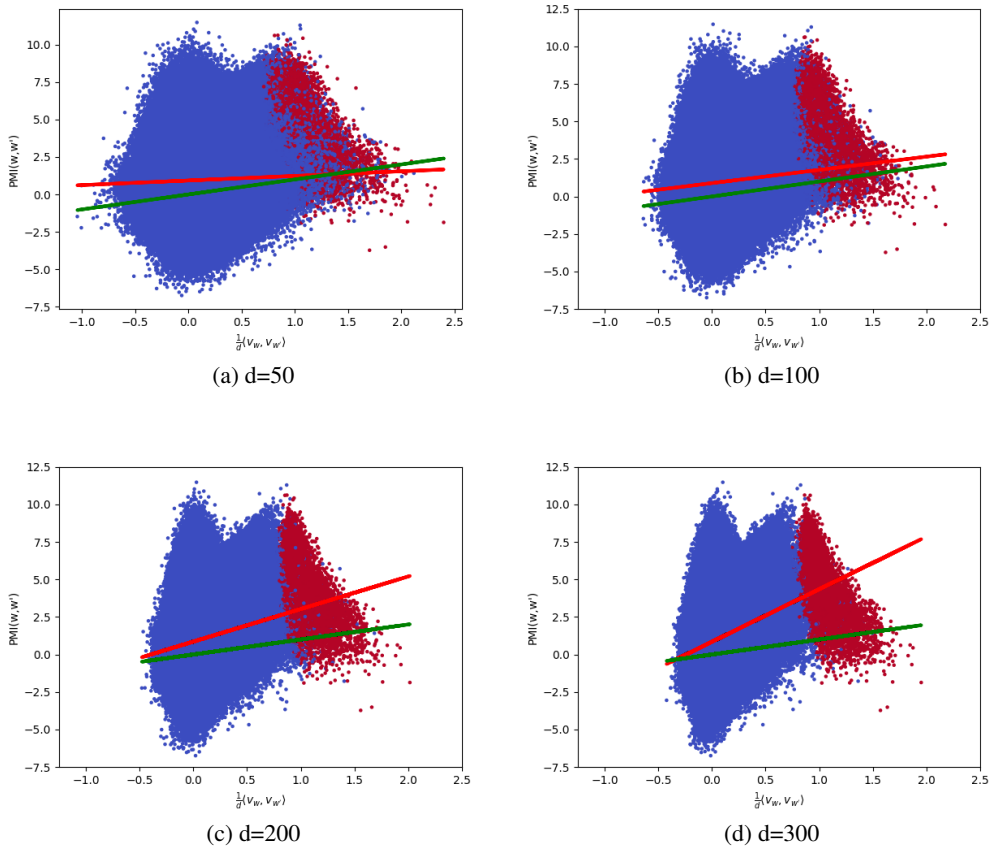


Figure 4: Empirical distribution of $\mathsf{PMI}(w, w') - \frac{1}{d}\langle w, w'\rangle$ with Gaussian density estimation. Based on corpus 2 and SN word embeddings with dimension 50, window size 2.

In fact, for window size 2, the experimental slope is increasing with respect to the dimension (see Figure 5), which is completely contradictory with equation (25). Our experiments show that, as dimension increases, more couples (w, w') tend to concentrate around certain "clusters". One of these is formed by couples of words verifying $\frac{1}{d}\langle v_w, v_{w'}\rangle \approx 1$, while the points in the region with the highest density have $\frac{1}{d}\langle v_w, v_{w'}\rangle \approx 0$. As a matter of fact, the red points group in Figure 5 corresponds to pairs of words occurring with themselves. We later prove that for these couples $\langle v_w, v_{w'}\rangle = ||v_w||^2 \approx d$, when the dimension is large enough (see Lemma 7 in Section 3.2).

We can also prove that $\mathbb{E}[\langle v_w, v'_w\rangle] = 0$ and $\mathbb{V}[\frac{1}{d}\langle v_w, v'_w\rangle] \propto \frac{1}{d}$ as long as v_w, v'_w are independent (Assumption 2). As the couples of words occurring with themselves tend to have medium to large PMI values (as shown by the heat plot of subfigure 6b), they pull the regression line up, hence the experimental slope increases with the dimension. This phenomenon can be observed for window size 10 as well. Although the experimental slopes are decreasing, according to Table 4, if we look at the ratio of the experimental slope to the theoretical one, this ratio increases and goes beyond 1 as dimension grows. Perhaps, when the window size is large enough, the discussed phenomenon is compensated by more points in the blue group. Still, from Table 4, equation (27) seems to hold better than equation (25).

We also remark that for larger corpora the results were slightly better, especially regarding the slope values. For $q = 2$, the intercept values are close to 1, which is not coherent with the theoretical zero value. When the window size is greater than 2, the theoretical intercept is $\gamma = \log q(q-1)/2$ according to equation (27). For window size $q = 10$, the theoretical intercept is $\gamma \approx 3.81$. In all cases, there is a discrepancy between the experimental intercept and the theoretical one.



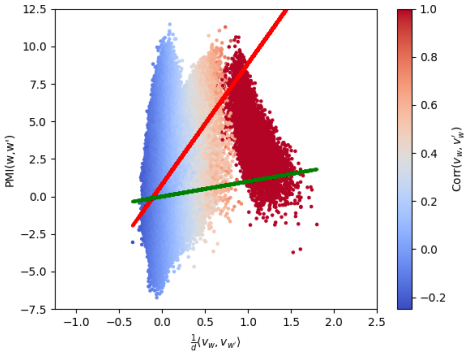
(a) d=50 (b) d=100

(c) d=200 (d) d=300

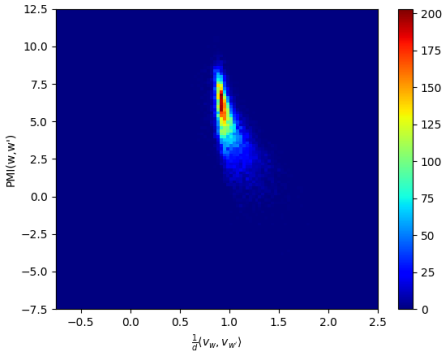
Figure 5: Plot of PMI vs. $\frac{1}{d}\langle v_w, v_{w'}\rangle$ successively for dimensions 50, 100, 200 and 300 based on corpus 1, for window size $q = 2$. In red, couples of words occurring with themselves. This figure shows that experimental slope is increasing with respect to the dimension, for this corpus. We can also observe the separation of two different groups of couples of words.

It remains important to visualize the shape of the plot. Figure 7 provides the plots of the experiments for different corpus sizes. Also, a heat plot is provided in order to consider the density of the plotted points. We observe that the larger the corpus, the larger the upper bound of PMI. And for this part of the plot, that is large PMI values, the linear relationship predicted by equation (25) seems nonexistent. We also observe the high discrepancy of the dot product values when PMI ≈ 0. This point will be discussed further. Finally, the discrepancy observed between the shape of the plots in Figure 7, namely the missing edge in top of the surface of subfigure 7a, is due to the removal of stopwords from corpus 1.

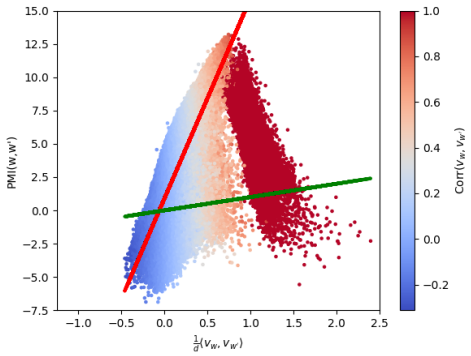
Experiments using GloVe In order to avoid being restricted to SN word embeddings, the relation between PMI and the scalar product was also tested for GloVe word vectors. From Table 4, we can see that the relationship is practically nonexistent for GloVe. In view of the efficiency and popularity of GloVe, it is therefore possible to claim that the relation discussed is not necessary for word vectors to perform well on semantic and syntactic tasks. We recall that the GloVe word



(a) Based on corpus 1, $q = 2$, $d = 1000$.



(b) Heat plot of plot 6a only showing words co-occurring with themselves. It shows that these couples of words typically have a large PMI.



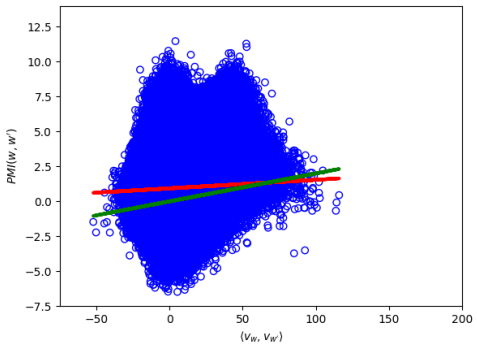
(c) Based on corpus 2, $q = 2$, $d = 1000$.

Figure 6: Plot of PMI vs. $\frac{1}{d}\langle \cdot, \cdot \rangle$ and Pearson's linear correlation denote $\mathsf{Corr}(\cdot, \cdot)$. We can distinguish three groups of points from left to right. The experimental slope seems to be mostly determined by the first and second groups.

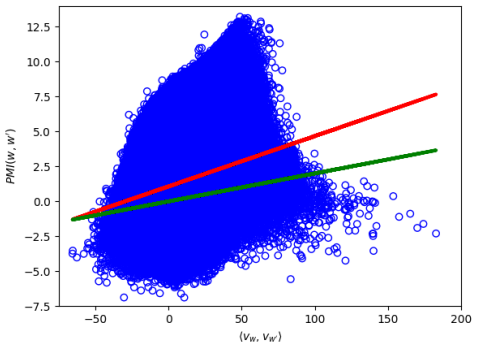
vectors were pretrained on Wikipedia 2014 and the corpus Gigaword 5 for a total of 6 billion tokens. Therefore, as they have not been trained on the same corpus as the SN word vectors, we want to make it clear that we are not comparing the two. Instead, we are checking whether what we know are performant word vectors, like GloVe, verify relation (25).

3.2 Discussion

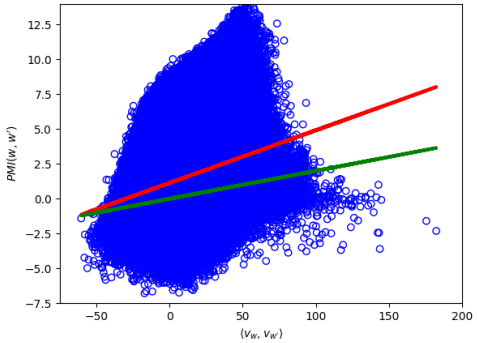
In this subsection, we discuss the relation claimed in Theorem 5 (and Corollary 6) between PMI and the dot product of the model's word embeddings. First, we show that a distribution discrepancy exists in equation (24), restricting the possible domain of validity of this equation. Then, we provide empirical and theoretical arguments to restrict the domain where the claimed theorem can be valid with $\epsilon \approx 0$. Finally, we examine granular examples of some regions of the plots in Figure 7 to argue that equation (25) cannot hold due to the intrinsic difference between PMI and the dot product.



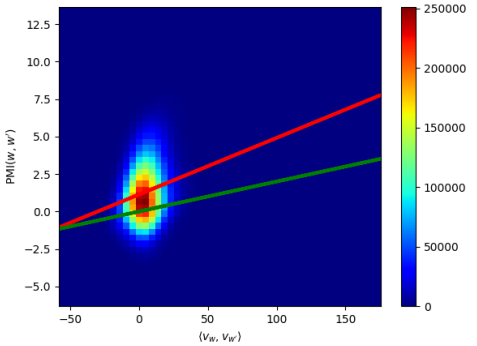
(a) Based on corpus 1



(b) Based on corpus 2



(c) Based on corpus 3. Only a point in two was plotted.



(d) Heat map of the plot based on corpus 3

Figure 7: Experiments on equation (25) for $d = 50$. x-axis: $\langle v_w, v_{w'} \rangle$; y-axis: $\mathsf{PMI}(w, w')$. Green line: theoretical linear relationship predicted by the equation. Red line: result of the linear regression.

DISTRIBUTION DISCREPANCY IN EQUATION (24)

The experiments conducted to verify equation (24) show that it is not empirically verified by infrequent words. Figure 8, similarly to Figure 2 in (Arora et al., 2016), shows that a linear relationship can possibly exist only when $\log p(w) > -9$ or $\frac{1}{d} ||v_w||^2 > 1.5$, although the slope of this linear relationship does not match with the theoretical one.

We also provide a theoretical argument, using Assumption 2 to claim that equation (24) does not hold for infrequent words. To this end, we need an auxiliary lemma.

Lemma 7 *Let $X \in \mathbb{R}^d$ a real-valued random vector drawn from the spherical Gaussian distribution in $\mathbb{R}^d$. Then, for all $z \in \mathbb{R}$,*

$$\mathbb{P}(\frac{1}{d} ||X||^2 \geq z) = 1 - \Phi\left((z-1)\sqrt{\frac{d}{2}}\right) + \mathcal{O}(\frac{1}{\sqrt{d}}) \tag{32}$$

where Φ is the cumulative distribution function of the standard normal distribution.

Proof Left in the Appendix. ■

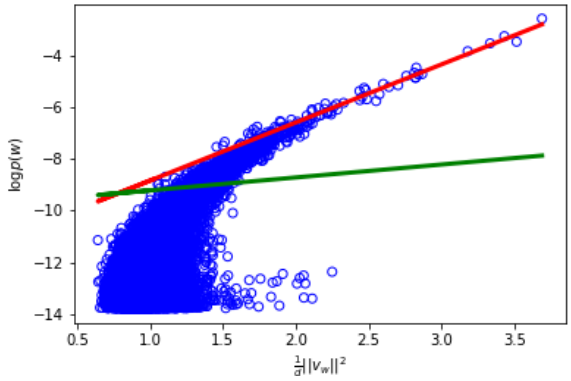


Figure 8: Experiments on equation (24). Green line: theoretical relation predicted. Red line: result of the partial linear regression. Based on corpus 2 and SN word embeddings with $d = 50$, $w = 2$.

Lemma 7 proves that $\frac{1}{d}||v_w||^2$ concentrates around the value 1 for a large enough value of the dimension, for word vectors verifying Assumption 2. On the other hand, from empirical observation, the logarithm of word frequency seems to follow a shifted exponential distribution with a shift very distant from the mean of $\frac{1}{d}||v_w||^2$. Hence, as shown by Figure 9, there is a distribution discrepancy between $2(\log p(w) + \log Z)$ and $\frac{1}{d}||v_w||^2$ which strongly restricts the possible domain of validity of equation (24). In fact, even if the value of $\log Z$ allowed the means of both distributions to be close enough, an important variance discrepancy remains and restricts the range of values where equation (24) can hold. This range is such that $2 \log pZ > 0$, which is equivalent to $\log p > -\log Z \approx -9.72$. This is coherent with Figure 8, which shows that a linear relationship exists when $\log p > -9$.

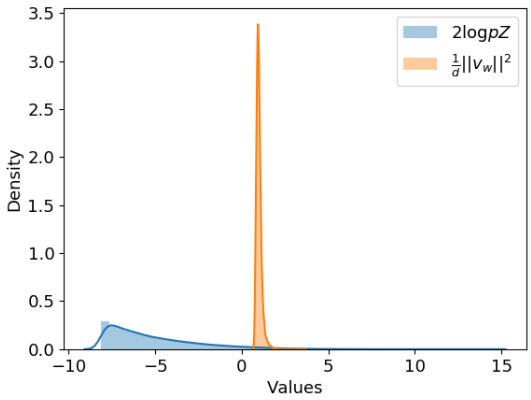


Figure 9: Density estimation for $2(\log p(w) + \log Z)$ and $\frac{1}{d}||v_w||^2$. Based on corpus 2 and SN word embeddings with $d = 50$, $w = 2$.

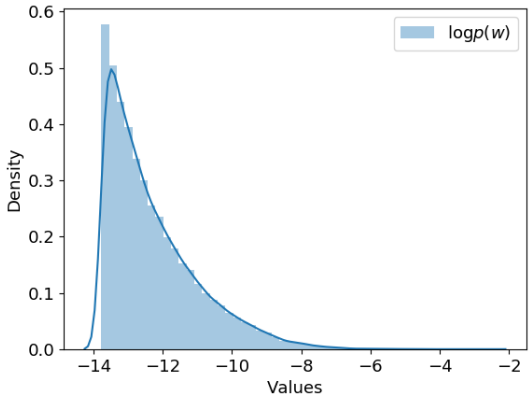


Figure 10: Density estimation for $\log p(w)$. Based on corpus 2 and SN word embeddings with $d = 50$, $w = 2$.

RESTRICTION OF THEOREM 5

From the results displayed in Figure 7c, we can compare the empirical upper bound of PMI and that of $\frac{1}{d}\langle \cdot, \cdot \rangle$. We empirically observe $\max \mathsf{PMI} \approx 15$ and $\max \frac{1}{d}\langle \cdot, \cdot \rangle \approx 4$. This shows that, at least in the region where $\mathsf{PMI} \gg 4$, we cannot have $\mathsf{PMI} \approx \frac{1}{d}\langle \cdot, \cdot \rangle$.

We prove this incompatible range of values to exist for any given corpus, as demonstrated by Proposition 8 below, under mild assumptions about the distribution of occurrences and co-occurrences shown in Assumption 4.

Assumption 4 *(Distributional assumptions for Proposition 8) For every words w, w' in the finite vocabulary $\mathcal{V}$, we consider the real-valued random variables*

$$f(w) \triangleq -\log \bar{p}(w),$$
$$f(w, w') \triangleq -\log \bar{p}(w, w')$$

where $\bar{p}(w), \bar{p}(w, w')$ are from Eq. (30)*.*

We also define

$$f(w|w') \triangleq f(w, w') - f(w'), \qquad \overline{\mathsf{PMI}}(w, w') \triangleq f(w) + f(w') - f(w, w'),$$

and make the following assumptions:

- $-f(w) + s \sim Exp(\lambda)$
- $\forall \beta > 0,\ f(w|w')|(f(w) > \beta) \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$

for some $\lambda, s > 0$, and $\forall \beta > 0,\ \mu_\beta, \sigma_\beta \geq 0$.

Intuitively, $-f(w)$ represents the natural logarithm of frequency for word w and $\overline{\mathsf{PMI}}$ its induced empirical estimation of PMI.

We can experimentally verify that these distributional assumptions hold for the datasets used. For example, for corpus 2, the distributional assumptions hold with a very good approximation for $\beta = 12,\ \sigma = \frac{1}{\lambda} = 1.5,\ s = 13.8$ (cf. Figure 10 for the distributional assumption on $f(w)$).

We introduce the following notations used in Proposition 8:

- $m_1 \triangleq \sum_{w \in \mathcal{V}} X_w$
- $m_2 \triangleq \sum_{(w,w') \in \mathcal{V} \times \mathcal{V}} X_{w,w'}$

Proposition 8 *Let $\mathcal{V}$ be a finite vocabulary. Suppose the word vectors $\{v_w\}_{w \in \mathcal{V}} \subset \mathbb{R}^d$ verify Assumption 2 (with $\kappa = 1$), and Assumption 4. Then, given $\epsilon > 0$, $\exists M \subset \mathcal{V} \times \mathcal{V}$, with $\mathbb{P}((w, w') \in M) > 0$, $\forall (w, w')$, such that if $d = \lceil \frac{1}{\epsilon^{a+2}} \rceil, a > 0$, then with high probability over the word vectors:*

$$\forall (w, w') \in M, \quad \overline{\mathsf{PMI}}(w, w') > \max_{\mathcal{V} \times \mathcal{V}} \frac{1}{d}\langle v_w, v_{w'} \rangle + \epsilon, \tag{33}$$

more precisely

$$\forall (w, w') \in M, \quad \mathbb{P}\left(\overline{\mathsf{PMI}}(w, w') > \max_{\mathcal{V} \times \mathcal{V}} \frac{1}{d}\langle v_w, v_{w'} \rangle + \epsilon \right) \geq 1 - e^{-\frac{1}{4\epsilon^a}} + \mathcal{O}(\epsilon^{1+\frac{a}{2}}) \tag{34}$$

Furthermore, when $m_1, m_2 \to +\infty$, we have:

$$\forall (w, w') \in M, \quad \mathsf{PMI}(w, w') \geq \max_{\mathcal{V}\times\mathcal{V}} \frac{1}{d}\langle v_w, v_{w'}\rangle + \epsilon \tag{35}$$

and

$$\forall (w, w') \in M, \quad \mathbb{P}\left(\mathsf{PMI}(w, w') \geq \max_{\mathcal{V}\times\mathcal{V}} \frac{1}{d}\langle v_w, v_{w'}\rangle + \epsilon\right) \geq 1 - e^{-\frac{1}{4\epsilon^a}} + \mathcal{O}(\epsilon^{1+\frac{a}{2}}) \tag{36}$$

Proof For every $\alpha, \epsilon > 0$, we define

$$M_\alpha \triangleq \{\ (w, w') \in \mathcal{V} \times \mathcal{V} \mid f(w|w') < \alpha,\ f(w) > \alpha + \epsilon + 1\ \}$$

using the notations introduced in Assumption 4.

First, we prove that for a small enough $\epsilon > 0$, there exists $\alpha > 0$ such that:

$$\forall (w, w') \in M_\alpha, \quad \overline{\mathsf{PMI}}(w, w') > 1 + \epsilon \tag{37}$$

and $\mathbb{P}(M_\alpha) \triangleq \mathbb{P}((w, w') \in M_\alpha) > 0$.

Indeed, for every $\epsilon > 0$ and $(w, w') \in M_\alpha$:

$$\overline{\mathsf{PMI}}(w, w') = f(w) + f(w') - f(w, w') = f(w) - f(w|w') > -\alpha + \alpha + 1 + \epsilon = 1 + \epsilon$$

Moreover, from the distributional assumptions (Assumption 4), for $\beta = \alpha + \epsilon + 1$, we have for every $\alpha > 0$

$$\begin{aligned}\mathbb{P}(M_\alpha) &= \mathbb{P}((w, w') \in M_\alpha) = \mathbb{P}(f(w|w') < \alpha | f(w) > \beta) \times \mathbb{P}(f(w) > \beta) \\ &= \Phi_\beta(\alpha)(1 - \exp(\lambda(1 + \epsilon + \alpha - s)))\mathbb{1}_{\alpha \leq s-1-\epsilon}\end{aligned}$$

where Φ_β is the cumulative distribution function of $\mathcal{N}(\mu_\beta, \sigma_\beta^2)$.

Clearly, a small enough ϵ and large enough s such that $\hat{\alpha} = s - 1 - 2\epsilon > 0$ imply $\mathbb{P}(M_{\hat{\alpha}}) > 0$.

Second, we prove that, for any $\epsilon > 0$, we have with high probability over the word vectors

$$\max_{\mathcal{V}\times\mathcal{V}} \frac{1}{d}\langle v_w, v_{w'}\rangle \leq 1 + \epsilon/2 \tag{38}$$

We recall that $\max_{\mathcal{V}\times\mathcal{V}} \frac{1}{d}\langle v_w, v_{w'}\rangle = \max_{\mathcal{V}} \frac{1}{d}||v_w||^2$, due to the Cauchy–Schwarz inequality. Therefore, (38) is equivalent to

$$\max_{\mathcal{V}} \frac{1}{d}||v_w||^2 \leq 1 + \epsilon/2$$

By Lemma 7, we know that

$$\mathbb{P}(\max_{\mathcal{V}} \frac{1}{d}||v_w||^2 \geq 1 + \epsilon/2) = 1 - \Phi\left(\frac{\epsilon}{2}\sqrt{\frac{d}{2}}\right) + \mathcal{O}(\frac{1}{\sqrt{d}}) \tag{39}$$

We then conclude that

$$\lim_{d\to+\infty} \mathbb{P}(\max_{\mathcal{V}} \frac{1}{d}||v_w||^2 \geq 1 + \epsilon/2) = 0$$

showing that (38) holds with high probability over word vectors for a sufficiently large dimension.

Using (37) and (38), we obtain that there exists a set $M \subset \mathcal{V} \times \mathcal{V}$ such that $\mathbb{P}((w, w') \in M) > 0$ and with high probability over word vectors for a sufficiently large dimension:

$$\forall (w, w') \in M, \quad \overline{\mathsf{PMI}}(w, w') > \max_{\mathcal{V}\times\mathcal{V}} \frac{1}{d}\langle v_w, v_{w'}\rangle + \epsilon/2$$

Moreover, we can obtain an asymptotic lower bound of the probability over word vectors for this inequality to hold, using Equation (39) and that $\Phi(x) > 1 - e^{-\frac{x^2}{2}}$, for every $x > 0$. We can then set $d = \lceil \frac{1}{\epsilon^{a+2}} \rceil, a > 0$ to obtain the stated lower bound.

It remains to prove that almost surely

$$\lim_{m_1, m_2 \to +\infty} \overline{\mathsf{PMI}}(w, w') = \mathsf{PMI}(w, w')$$

where $m_1,\ m_2$ denote the total number of occurrences and co-occurrences respectively.

For that, it is enough to prove that for every $w, w' \in \mathcal{V}$, we almost surely have

$$\begin{cases} \lim_{m_1 \to +\infty} f(w) = -\log p(w) \\ \lim_{m_2 \to +\infty} f(w, w') = -\log p(w, w') \end{cases}$$

Let $V_i \triangleq \mathbb{1}_{O_i = w}$ be the variable indicating whether word w appears in token i of the corpus, i.e. equal to 1 if $O_i = w$ and 0 otherwise.

Therefore

$$\exp(-f(w)) = \frac{1}{m_1}(\sum_i \mathbb{1}_{O_i = w})$$

Since $V_1, \ldots, V_{m_1}$ are i.i.d., then due to the strong law of large numbers, we have almost surely:

$$\lim_{m_1 \to +\infty} \exp(-f(w)) = p(w)$$

Thanks to the continuity of $-\log$, we have almost surely:

$$\lim_{m_1 \to +\infty} f(w) = -\log p(w)$$

We similarly prove that

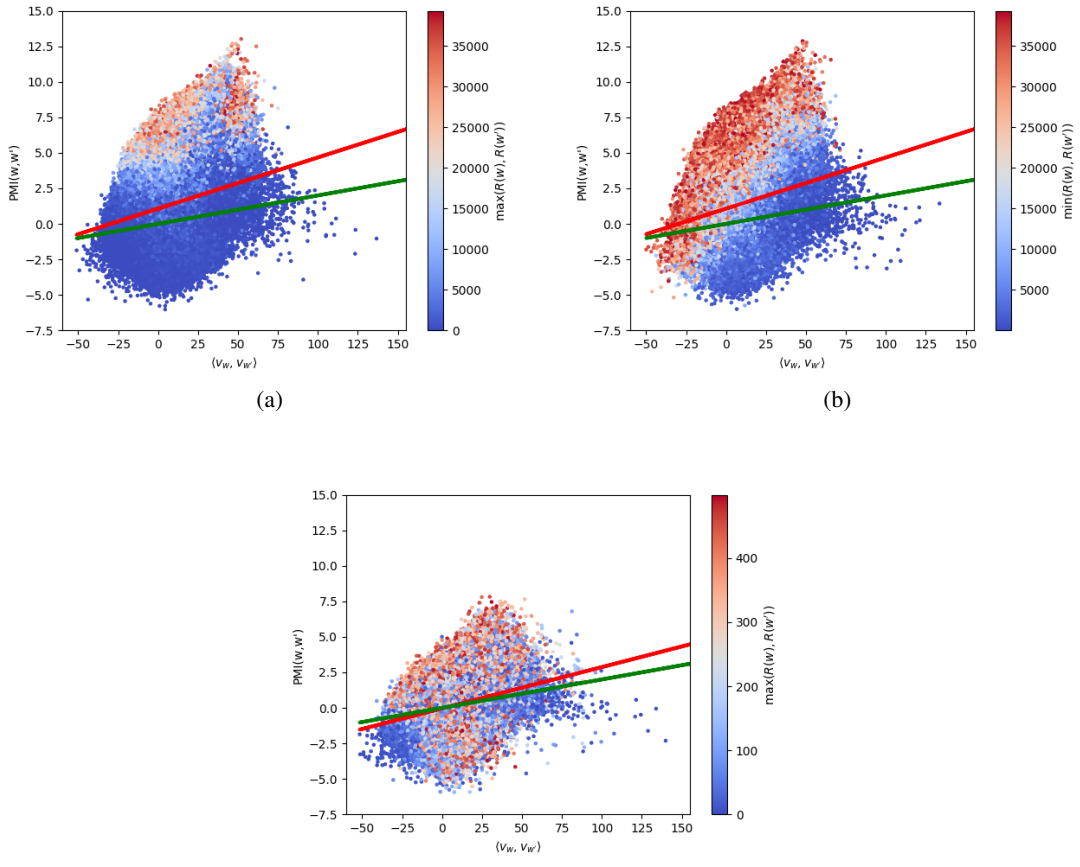
$$\lim_{m_2 \to +\infty} f(w, w') = -\log p(w, w')$$

We then conclude the desired result. ■

Remark on Proposition 8: The statement of Proposition 8 mentions that Inequality (35) is obtained when $m_1, m_2 \to +\infty$. For simplicity, we omitted to mention that this inequality is obtained almost surely. The almost sure convergence stems from the frequency of word counts converging to the probability of word occurrences once $m_1, m_2 \to +\infty$.

Experiments based on corpus 2 show that the set M, found in Proposition 8 with latent parameter $\alpha = 10$ and for $\epsilon = 1$, can be as large as 15% of $\{(w, w') \in \mathcal{V} \times \mathcal{V} |\ p(w, w') > 0\}$. Moreover, the couple of words having the largest error (with respect to Eq. (25)), which has an order of magnitude of 10 for corpus 2 (see Figure 3), can be found in the set M.

In an attempt to find a restricted domain where the claimed relation is valid, we added the frequency rank (denoted $R(w)$ for word w) as a third dimension to the plot of PMI and $\langle \cdot, \cdot \rangle$. The plot displayed in Figure 11a shows that couples (w, w') for which the linear relation of Theorem 5 fails to hold are in general couples of infrequent words. This is not coherent with the fact that, according to (Arora et al., 2016), "very frequent words [...] do not fit our model". However, this is not surprising when we consider that Equations (23) and (24) seem to hold[5] better for very frequent words. In fact, in the objective function for SN word vectors (see Eq. (28)), very frequent pairs of words have the largest weights. Moreover, they are involved in a great number of terms, since they co-occur with a lot of words. Therefore, this can explain why the model fits better for very frequent words.



(a) (b)

(c)

Figure 11: Plot of $\mathsf{PMI}(w, w')$, $\langle v_w, v'_w \rangle$ with $\max(R(w), R(w'))$ in the first and third figures, and with $\min(R(w), R(w'))$ in the second figure, where $R(w)$ is the frequency rank of word w. Based on corpus 2 and SN word embeddings with $d = 50$, $w = 2$. Green line: theoretical relation. Red line: linear regression.

5. from a correlation point of view

When we restrict the third dimension, that is $\max(R(w), R(w'))$, to have a threshold maximum value of 500[6], the relation seems to hold better (see Figure 11c).

Table 5: PMI, scalar product and cosine similarity of words sampled from the corpus. The experiment was made using SN word embeddings on corpus 3.

Word 1	Word 2	PMI	Scalar product	Cosine similarity
notre	dame	10	79	0.85
obama	barack	10	81	0.92

Table 6: Words with low PMI and large scalar product

Word 1	Word 2	PMI	Scalar product	Cosine similarity
many	several	-0.82	134	0.90
march	general	0.77	80	0.55

Table 7: Words with scalar product ≈ 0 and relatively large PMI

Word 1	Word 2	PMI	Scalar product	Cosine similarity
schools	newsweek	3.02	5	0.05
schools	moldovan	2.58	-5	-0.05
schools	ugandan	2.45	8	0.09

ON THE INTRINSIC DIFFERENCE BETWEEN PMI AND THE DOT PRODUCT

We advocate that PMI and the dot product, although they both encode similarity between words, do not encode the same type of similarity. While $\mathsf{PMI}(w, w')$ yields large values for words w and w' which co-occur more often than if they were independent, it merely takes into account "the company [the word] keeps" (Firth, 1957). On the other hand, the scalar product $\langle v_w, v_{w'} \rangle$ of word embeddings trained on the non-zero entries of a global word-word co-occurrence matrix, like SN and GloVe, captures not only co-occurrences of w and w' but also those of other related words as well. Furthermore, PMI usually requires large corpora because of its unreliability with low occurrence words (see (Role and Nadif, 2011) for full details on the difficulties related to handling low occurrence events for PMI).

In order to understand the difference between PMI and the inner product of word vectors, we can distinguish three situations of couples of words: both frequent, both infrequent, one frequent and the other infrequent. Given these types, we can distinguish three regions in the plot of Figure 11a. It can be inferred from the first two plots of Figure 11 that the top region of the surface corresponds in great part to couples of infrequent words. The bottom right region corresponds to couples of frequent words. The bottom left corresponds to couples made of a frequent and an infrequent word.

When both words are infrequent and happen to co-occur, it is very likely that they have large PMI and scalar product values. This explains the shape of the top region of the surface on the plot.

6. That is, a couple of word is left only if at least one of the two words of the couple is in top 500 most frequent words.

This part of the plot is the most interesting as it is where the major outliers of Theorem 5 live. These values always exist in natural language. To illustrate this, Table 5 contains an example of very large values of PMI. Usually, these words would rarely appear without their partner word, thus the large PMI value. The scalar product and cosine similarity are also large which is coherent for such words that rarely appear without the other.

Table 6 contains a sample of words from the region of high discrepancy around the 0 PMI value in Figure 7. This is an example of very similar words, as inferred by the cosine similarity and scalar product, with low PMI values. Especially for the words 'many' and 'several' which are similar but will almost never appear together[7], thus the low PMI value. The scalar product was able to capture the similarity because it had access to all the contexts of 'many' and 'several' and inferred that these were similar.

In Table 7 we can see a sample of words with scalar product ≈ 0 and positive[8] PMI value. We can give the following explanation for 'schools' and 'moldovan' for example: 'moldovan' is a relatively rare word and it happens that it naturally occurs often (relatively to the frequency of 'moldovan') with 'schools', thus the PMI value. But these words are completely different semantically, thus the scalar product value. This is another example of unwanted behavior of PMI for low occurrence words.

Finally, an important difference between the scalar product and PMI can be observed for words occurring with themselves. For this type of co-occurrences, the scalar products are naturally the largest. However, the PMI values can be anywhere from negative to large positive values: words 'the' and 'her' have a dot product of 98 and PMI value of -1.26, while 'as' and 'well' have a dot product of 136 and PMI value of 4.58. In fact, this last example is the exact type of co-occurrences causing the bottom-right edge on the scatter plot (see subfigure (c) of Figure 7). This further justifies how a strict linear relationship between PMI and scalar product can hardly exist unless an unacceptably high error term is tolerated.

The next section demonstrates that if the error term is dropped from Eq. (27), the equality cannot hold because it would imply that the corresponding shifted symmetric PMI matrix should be positive semidefinite which, as we shall show ahead, may be violated by natural language.

4. Relation with implicit matrix factorization

The experimental results of Section 3 point in the direction that Equation (27) is not verified in practice with a small noise level ϵ. In this section we shall prove that, as long as a symmetric PMI matrix is considered, (27) cannot hold if the noise ϵ vanishes. To this end, we will show that the shifted symmetric PMI matrix fails to be positive semidefinite when considering natural language.

In (Levy and Goldberg, 2014), it was shown that the optimization problem solved in skip-gram with negative sampling (Mikolov et al., 2013b) corresponds to an implicit matrix factorization:

$$\forall w, c, \quad \langle v_w, v_c \rangle = \mathsf{PMI}(w, c) - \beta, \tag{40}$$

where $\beta = \log k$, k is the number of "negative samples" and $\mathsf{PMI}(w, c)$ corresponds to an entry of an *asymmetric* (usually rectangular) word-context PMI matrix. Each context is defined by a window of

7. Usually, redundancy is avoided in writings and such interchangeable words would not be used together.
8. These values are relatively large as it is useful to notice that when $\mathsf{PMI}(w, w') \approx 3$, words w and w' are 20 times more likely to co-occur than if they were independent.

size q around each token w_ℓ, i.e $w_{\ell-q}, \ldots, w_{\ell-1}, w_{\ell+1}, \ldots, w_{\ell+q}$ is the context for the word/token w_ℓ. In (40), $v_w, v_c \in \mathbb{R}^d$ for a suitable dimension d.

If we consider a matrix V whose rows are the vectors v_w, and C a matrix whose rows are the vectors v_c, then (40) can be written in matrix form as

$$VC^\top = M - \beta \mathbf{1}_{|V|}\mathbf{1}_{|C|}^\top, \tag{41}$$

where M is a $|V| \times |C|$ matrix with entries $M_{wc} = \mathsf{PMI}(w, c)$ and $\mathbf{1}_m$ denotes the vector of ones in $\mathbb{R}^m$. The singular value decomposition (Golub and Van Loan, 1996) ensures that (41) holds for some $d = \text{rank}(VC^\top) \leq \text{rank}(M) + 1$.

In view of relation (27), one may wonder whether (40) also holds for a *symmetric* PMI matrix: here the vocabularies of words and contexts are the same.

In fact, in (Arora et al., 2016, pg. 389) one finds: "This [Equation (27) in Corollary 6] is also consistent with the shift β for fitting PMI in (Levy and Goldberg, 2014b), which showed that without dimension constraints, the solution of skip-gram with negative sampling satisfies $\mathsf{PMI}(w, w') - \beta = \langle v_w, v_{w'} \rangle$ for a constant β that is related to the negative sampling in the optimization. Our result justifies via a generative model why this should be satisfied even for low dimensional word vectors."

Let us consider a *symmetric* PMI matrix M with entries $M_{ww'} = \mathsf{PMI}(w, w')$ and assume that there exists a scalar β such that

$$\forall w, w', \quad \langle v_w, v_{w'} \rangle = \mathsf{PMI}(w, w') - \beta. \tag{42}$$

Suppose the vocabulary is finite (of size n), and since $\mathsf{PMI}(w, w') = \mathsf{PMI}(w', w)$, we can write (42) in matrix form as

$$VV^\top = M - \beta \mathbf{1}\mathbf{1}^\top,$$

where V is a $n \times d$ matrix whose rows contain the vectors $v_w \in \mathbb{R}^d$, and $\mathbf{1} \in \mathbb{R}^n$ denotes the vector of ones.

Since $VV^\top$ is symmetric positive semidefinite, we obtain that, for every vector $y \in \mathbb{R}^n$

$$0 \leq y^\top (M - \beta \mathbf{1}\mathbf{1}^\top) y = y^\top M y - \beta (\mathbf{1}^\top y)^2.$$

In particular, taking $y \in \{\mathbf{1}\}^\perp$, we have

$$\forall y \in \{\mathbf{1}\}^\perp, \quad y^\top M y \geq 0. \tag{43}$$

Let w and w' be a pair of words for which $p(w, w') > 0$, $p(w, w) > 0$, $p(w', w') > 0$, and choose $y = e_w - e_{w'} \in \{\mathbf{1}\}^\perp$, where $e_w, e_{w'}$ are canonical vectors of $\mathbb{R}^n$. Thus,

$$\begin{aligned} y^\top M y =& (e_w - e_{w'})^\top M (e_w - e_{w'}) = M_{ww} - 2M_{ww'} + M_{w'w'} \\ =& \mathsf{PMI}(w, w) - 2\mathsf{PMI}(w, w') + \mathsf{PMI}(w', w') \\ =& \log \frac{p(w, w)}{p(w)p(w)} - 2 \log \frac{p(w, w')}{p(w)p(w')} + \log \frac{p(w', w')}{p(w')p(w')} \\ =& \log p(w, w) - 2 \log p(w, w') + \log p(w', w'). \end{aligned}$$

From (43) and the last equality we obtain

$$\log \frac{p(w, w)p(w', w')}{p(w, w')^2} \geq 0$$

or, equivalently,

$$p(w,w')^2 \leq p(w,w)p(w',w'). \tag{44}$$

However, this inequality is violated by a pair of words w and w' for which $p(w,w)$ and $p(w',w')$ are quite small when compared to $p(w,w')$, i.e words that appear repeated in very few windows but co-occur considerably more as illustrated in the following examples based on the statistics for the "corpus 2":

- **Example 1:** If we consider $w =$ *professional* and $w' =$ *wrestler*, then $p(w,w') = 4.51 \times 10^{-6}$ and $p(w,w) = 2.09 \times 10^{-7}$ and $p(w',w') = 5.26 \times 10^{-8}$. In this case $p(w,w')^2 > p(w,w)p(w',w')$.
- **Example 2:** If we consider w, w' as the pair of words *well*, *done* (respec.), we have $\log p(w,w) \approx -14.7547$, $\log p(w',w') \approx -17.5806$ and $\log p(w,w') \approx -13.9783$, which shows that $2\log p(w,w') > \log p(w,w) + \log p(w',w')$, i.e inequality (44) does not hold.

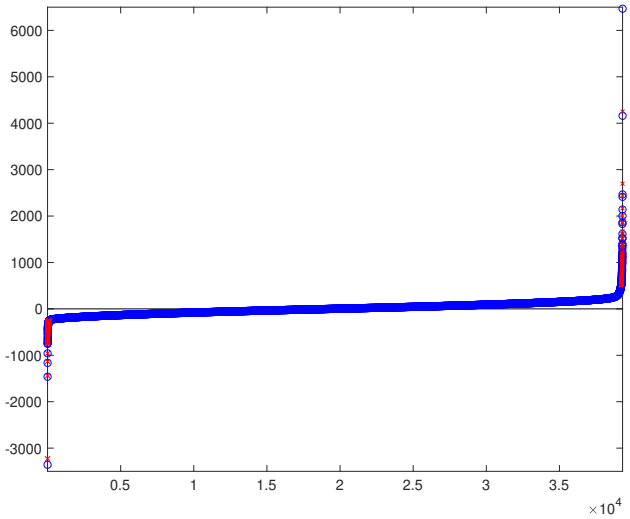


Figure 12: Spectrum of PMI matrix M for "corpus 2" (in blue). In red, the 100 smallest/largest eigenvalues of M restricted to the subspace $\{\mathbf{1}\}^{\perp}$.

Hence, condition (44), which is a necessary condition for (42), can be violated with natural language, thereby invalidating the claim (42), regardless the dimension d and the constant β.

Figure 12 shows the 100 largest/smallest eigenvalues of M restricted to $\{\mathbf{1}\}^{\perp}$ for "corpus 2". The presence of negative eigenvalues shows that (43) is violated.

Since (42) does not hold, independently of d and β, we conclude that Equation (27) cannot hold with an arbitrarily small noise/error $O(\epsilon)$. Interestingly, as n and d increase, the noise ϵ is dominated by $O(\epsilon_1)$, where ϵ_1 is the presumably "small" constant in Assumption 3.

5. Conclusion

The concentration property plays an important role for the theoretical foundations of the relation between PMI and the scalar product of word vectors. It allows to conduct an analysis of the joint distribution of words in contexts, which would be difficult - if not intractable - otherwise. This

relation was further used to justify the analogies conjecture (*"semantic relations=lines"*) of several popular PMI-based word embeddings (Mikolov et al., 2013b; Arora et al., 2016). In this work, we first proved in Section 2 and with Proposition 4, that the concentration property also holds for very simple distributions of the word and context vectors, independently of Assumptions 1 and 2. More precisely, the concentration property is holding with isotropy and uniformity on word vectors in a centered cube of $\mathbb{R}^d$, and such that latent discourse vectors belong to a sufficiently thin annulus. We remark this concentration property was empirically verified for certain common word embedding methods (Arora et al., 2016, cf. Section 5.1). Also, Proposition 4 shows that any generative model such that $\mathsf{P}(w_t = w|c_t) \propto \exp(\langle v_w, c_t\rangle)$, combined with Assumption 3, broadens the relation between PMI and scalar product to a larger class of word and context embeddings.

As the experiments in Arora et al. (2016) did not address the main relation (Equation (25)) claimed by Theorem 5, our work includes extensive experiments on this relation. The empirical verification of the equations listed by Theorem 5 and Corollary 6 strongly suggests that the claimed linear relation between PMI and the inner product of word embeddings does not hold in practice unless a large error term $O(\epsilon)$ is tolerated. Moreover, the statistical discussion in Section 3.2 provides empirical and theoretical evidences of the existence of a range of values where the linear relation cannot hold. These experimental findings concerning the violation of Equation (25) (and Equation (27)) – with error terms dropped – are further corroborated by the theoretical analysis of Section 4 which shows that the desired linear relation $\langle v_w, v_{w'}\rangle = \mathsf{PMI}(w, w') - \beta$ implies in the positiveness of the symmetric PMI matrix in a certain subspace, but such condition can be violated by natural language. Therefore, even when word vectors verify all the assumptions of Arora et al. (2016) (which we have shown not to be necessary for the concentration property to hold) we can observe that they fail to satisfy Equation (27) with small error. We believe that different reasons may be responsible for this failure.

First, the optimization schemes may fail to retrieve the embeddings satisfying the relations between PMI and their scalar product with lowest possible error, which motivates other types of loss functions and optimization schemes.

For example, in Khalife et al. (2021) the loss function is the sum of terms $(\mathsf{PMI}(w, w') - \langle v_w, v_{w'}\rangle)^2$ over all pairs (w, w') for which w or w' is in the top-m most frequent words.

Second, the geometries used to construct the word embeddings may be inadequate. We showed that fitting PMI on Euclidean inner product cannot be done without error, and the current schemes do not allow even to retrieve word embeddings fitting PMI with relative small discrepancy. In this sense, our study suggests to investigate alternative geometries to construct word embeddings where the relation between PMI and Gramian matrices emerges more naturally. An example of alternative geometry we think of is Hyperbolic geometry (cf. for instance Marconi et al. (2020)).

Finally, although the concentration property is empirically verified for common word vectors, suggesting Assumption 2 or the slightly weaker assumptions of Proposition 4 are reasonable in practice, the impossibility of PMI equation to hold for every word pair with arbitrarily small error (cf. Proposition 8 and Section 4), even for n and d sufficiently large, indicates that Assumption 3 is not satisfied with a small ϵ_1. Hence the model assumptions are too strong to have everything matched up exactly. Although this is not enough to fundamentally reject the considered generative model, it certainly exposes its limitations. Nevertheless, the equations in Theorem 5 and Corollary 6 may still be useful and allow to make predictions about high frequency word pairs.

Acknowledgments

This work has been supported by the Agence Nationale de la Recheche project ANR-19-CE45-0019 "MultiBioStruct" and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Process 88887.465828/2019-00.

Appendix A: Proofs of technical lemmata

Lemma 3 *Let $L > 0$ and consider the function f: $\mathbb{R}^d \longrightarrow \mathbb{R}$ defined by:*

$$f(c) \quad = \quad \prod_{i=1}^{d} \begin{cases} \frac{\sinh(Lc_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ L & \text{otherwise} \end{cases} \tag{8}$$

Then $\Psi = \log(f)$ verifies the assumptions of Lemma 1.

Proof In order to simplify the expressions, we will consider that $L = 1$ but the general case can be treated similarly. First, let us consider the function

$$\phi : x \mapsto \begin{cases} \frac{\sinh(c_i)}{c_i} & \text{if} \quad c_i \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

and in the following, let $\psi = \log \phi$.

i) First, ψ is twice continuously differentiable. Indeed, ψ is continuous on $\mathbb{R}$ and $\lim_{x\to 0} \psi'(x) = 0$, so with the derivation extension theorem, ψ is differentiable at 0 and $\psi'(0) = 0$. We use the same reasoning with ψ' and show that ψ is twice differentiable on $\mathbb{R}$, with $\psi''(0) = \frac{1}{3}$.

ii) ψ is strictly convex because ϕ is strictly logarithmically convex. Indeed,

$$\forall x \neq 0 \quad \phi(x)\phi''(x) - \phi'(x)^2 = \frac{-1 - 2x^2 + \cosh(2x)}{2x^4} > 0$$

where the strict inequality can be deduced from the Taylor series of $\cosh$.

iii) ψ is even since ϕ is. Besides, as proved in i), we have $\psi'(0) = 0$ and $\psi''(0) = \frac{1}{3}$. Furthermore, for $x \neq 0$:

$$\psi''(0) - \frac{\psi'(x)}{x} = \frac{1}{3} - \frac{\psi'(x)}{x} = \frac{1}{3} - \frac{1}{\sinh(x)} \left(\frac{\cosh(x)}{x} - \frac{\sinh(x)}{x^2} \right)$$

and

$$\begin{aligned} \psi''(x) - \frac{\psi'(x)}{x} = & - x \frac{\coth(x)}{\sinh(x)} \left(\frac{\cosh(x)}{x} - \frac{\sinh(x)}{x^2} \right) \\ & + \frac{x}{\sinh(x)} \left(-2\frac{\cosh(x)}{x^2} + 2\frac{\sinh(x)}{x^3} + \frac{\sinh(x)}{x} \right) \end{aligned}$$

Now, let us prove:

iv) $\forall x \neq 0, \quad \psi''(0) - \frac{\psi'(x)}{x} > 0$

v) $\forall x \neq 0, \quad \psi''(x) - \frac{\psi'(x)}{x} < 0$

vi) $x \mapsto \frac{\psi'(x)}{x}$ is injective on $\mathbb{R}^{+*}$.

After some algebraic manipulations, we remind that:

$$\phi'(x) = \frac{x\cosh x - \sinh x}{x^2}, \qquad \phi''(x) = \frac{(x^3+2x)\sinh x - 2x^2\cosh x}{x^4}$$

In particular:

$$\begin{aligned}\psi'(x) = \frac{\phi'(x)}{\phi(x)} &= \frac{x\cosh x - \sinh x}{x\sinh x} = \frac{\cosh x}{\sinh x} - \frac{1}{x} = \coth x - \frac{1}{x} \\ &= \left(\frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \cdots\right) - \frac{1}{x} = \frac{x}{3} - \frac{x^3}{45} + \cdots \end{aligned} \tag{45}$$

Now, let us consider the function q be defined as:

$$q(x) = \begin{cases} \dfrac{\psi'(x)}{x} & x \neq 0 \\ \frac{1}{3} & \text{otherwise} \end{cases}$$

After some algebraic manipulation and Taylor series expansion of $\coth$, we obtain

$$\forall x \neq 0 \qquad q(x) = \frac{-1 + x\coth x}{x^2} = \frac{-1 + x\left(\frac{1}{x} + \frac{x}{3} - \cdots\right)}{x^2} = \frac{1}{3} - \frac{x^2}{45} + 2\frac{x^4}{945} - \cdots \tag{46}$$

and

$$q'(x) = \frac{2 - x(\coth x + x\,\mathrm{csch}^2 x)}{x^3}$$

with $\lim_{x\to 0} q'(x) = 0$. Since $x(\coth x + x\,\mathrm{csch}^2 x) > 2$, for $x \neq 0$, we obtain: $\frac{\psi'(x)}{x} = q'(x) < 0$, for all $x > 0$. This proves that the function q is injective on $\mathbb{R}^{+*}$. Property vi) is proved.

Similarly, we obtain $q'(x) > 0$, for all $x < 0$, implying that $q(0) = \frac{1}{3}$ is the global maximum: $\forall x \in \mathbb{R}$, $q(x) \leq 1/3$.

Moreover,

$$\begin{aligned}\psi''(x) &= \frac{\phi''(x)\phi(x) - \phi'(x)^2}{\phi(x)^2} = \frac{1}{x^2} - \mathrm{csch}^2 x = \frac{1}{x^2} - \left(\frac{1}{x^2} - \frac{1}{3} + \frac{x^2}{15} - \cdots\right) \\ &= \frac{1}{3} - \frac{x^2}{15} + \cdots\end{aligned}$$

implying

$$\forall x \neq 0 \quad \psi''(0) - \frac{\psi'(x)}{x} = \frac{1}{3} - q(x) > 0$$

showing Property iv).

Finally, for $x \neq 0$:

$$\psi''(x) - \frac{\psi'(x)}{x} = \frac{2 - x(\coth x + x \operatorname{csch}^2 x)}{x^2} < 0$$

proving v). ■

Lemma 7 *Let $X \in \mathbb{R}^d$ a real-valued random vector drawn from the spherical Gaussian distribution in $\mathbb{R}^d$. Then, for all $z \in \mathbb{R}$,*

$$\mathbb{P}(\frac{1}{d}\,||X||^2 \geq z) = 1 - \Phi\left((z-1)\sqrt{\frac{d}{2}}\right) + \mathcal{O}(\frac{1}{\sqrt{d}}) \tag{32}$$

where Φ is the cumulative distribution function of the standard normal distribution.

Proof Let X_k be the k-th component of vector $X \in \mathbb{R}^d$ for $k \in \{1, \ldots, d\}$. Since X is drawn from the spherical Gaussian distribution, then for all $k \in \{1, \ldots, d\}$, X_k^2 is chi-squared distributed and $\mathbb{E}[X_k^2] = 1, \mathbb{V}[X_k^2] = 2$. Therefore, by using the Central Limit Theorem applied to the squared components of X, we have for all $\hat{z} \in \mathbb{R}$

$$\mathbb{P}(\sqrt{\frac{d}{2}}(\frac{1}{d}\,||X||^2 - 1) \leq \hat{z}) \xrightarrow{d\to\infty} \Phi(\hat{z})$$

Moreover, since $\rho \triangleq \mathbb{E}[|X_k^2 - 1|^3] < \infty$, then thanks to the Berry–Esseen theorem, there exists $C > 0$ such that

$$|F_d(z) - \Phi(z)| \leq \frac{C\rho}{\sigma^3\sqrt{d}}$$

where $F_d(\hat{z}) \triangleq \mathbb{P}(\frac{1}{d}\,||X||^2 \leq 1 + \hat{z}\sqrt{\frac{2}{d}})$ and $\sigma^2 \triangleq \mathbb{E}[(X_k^2 - 1)^2]$.

Since $\mathbb{P}(\frac{1}{d}\|X\|^2 \leq z) = \mathbb{P}(\sqrt{\frac{d}{2}}(\frac{1}{d}\|X\|^2 - 1) \leq (z-1)\sqrt{\frac{d}{2}})$, by using $\hat{z} = (z-1)\sqrt{\frac{d}{2}}$, we then conclude the desired result. ■

Lemma 9 *With the same notations as in Section 3, given the relation (Arora et al., 2016, arxiv version 8., Eq. 2.13):*

$$\mathbb{E}[\langle v_w + v_{w'}, c\rangle] = (1+\epsilon)\exp(\frac{v_w + v_{w'}}{2d}) \tag{47}$$

Then:

$$\log p(w) = \frac{||v_w||^2}{2d} - \log Z \pm \epsilon'$$

for $\epsilon' = O(\epsilon_z) + \tilde{O}(1/d) + O(\epsilon_1)$.

Proof We recall that $p(w, w')$ denotes the probability of words w and w' appearing together in a window of size q in the corpus. Moreover, $p(w)$ and $p(w')$ denote the corresponding marginal probabilities and $v_w, v_{w'} \in \mathbb{R}^d$ the respective word vectors.

From Equation (2.6) in Arora et al. (2016), we have

$$p(w, w') = \mathbb{E}_{c,c'}\left[\frac{\exp\langle v_w, c\rangle}{Z_c} \cdot \frac{\exp\langle v_{w'}, c'\rangle}{Z_{c'}}\right]$$

And since for all $c,\ Z_c = \sum_w \exp\langle v_w, c\rangle$, we then have

$$\begin{aligned}
p(w) &= \sum_{w'} p(w, w') \\
&= \sum_{w'} \mathbb{E}_{c,c'}\left[\frac{\exp\langle v_w, c\rangle}{Z_c} \cdot \frac{\exp\langle v_{w'}, c'\rangle}{Z_{c'}}\right] \\
&= \mathbb{E}_{c,c'}\left[\frac{\exp\langle v_w, c\rangle}{Z_c}\left(\sum_{w'} \frac{\exp\langle v_{w'}, c'\rangle}{Z_{c'}}\right)\right] \\
&= \mathbb{E}_c\left[\frac{\exp\langle v_w, c\rangle}{Z_c}\right]
\end{aligned}$$

The concentration property of the partition function allows $Z_c = (1 \pm \mathcal{O}(\epsilon_z))Z$
Therefore

$$p(w) = \frac{(1 \pm \mathcal{O}(\epsilon_z))}{Z}\mathbb{E}_c\left[\exp\langle v_w, c\rangle\right]$$

by neglecting the event where $Z_c \notin [(1-\epsilon_z)Z, (1+\epsilon_z)Z]$.
We can then write

$$p(w) = \frac{(1 \pm \mathcal{O}(\epsilon_z))}{Z}(1 \pm \epsilon)(1+\epsilon_1)\exp\frac{||v_w||^2}{2d}$$

with $\epsilon = \tilde{\mathcal{O}}(\frac{1}{d})$, thanks to Eq. 47 with $v_{w'} = 0$.
We then conclude the desired equation by introducing the $\log$ function on this last result. ■

References

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*, 4:385–399, 2016.

Sanjeev Arora, Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A compressed sensing view of unsupervised text embeddings, bag-of-n-grams, and lstms. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018a.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018b.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings, 2019. URL https://arxiv.org/abs/1502.03520.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

John Rupert Firth. A synopsis of linguistic theory. 1957.

G. A. Golub and C.F. Van Loan. *Matrix Computations, 3rd edition*. The John Hopkins University Press, London, 1996.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sammy Khalife, Leo Liberti, and Michalis Vazirgiannis. Geometry and analogies: a study and propagation method for word representations. In *International Conference on Statistical Language and Speech Processing*, pages 100–111. Springer, 2019.

Sammy Khalife, Douglas Gonçalves, and Leo Liberti. Distance geometry for word embeddings and applications. 2021.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf.

Gian Marconi, Carlo Ciliberto, and Lorenzo Rosasco. Hyperbolic manifold regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2570–2580. PMLR, 2020.

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

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*, pages 3111–3119, 2013b.

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, 2014.

Francois Role and Mohamed Nadif. Handling the impact of low frequency events on co-occurrence based measures of word similarity. 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.