RAND-WALK: A latent variable model approach to word embeddings

Sanjeev Arora Yuanzhi Li Yingyu Liang Tengyu Ma Andrej Risteski *

Abstract

Semantic word embeddings represent the meaning of a word via a vector, and are created by diverse methods including Vector Space Methods (VSMs) such as Latent Semantic Analysis (LSA), generative text models such as topic models, matrix factorization, neural nets, and energy-based models. Many of these use nonlinear operations on co-occurrence statistics, such as computing Pairwise Mutual Information (PMI). Some use hand-tuned hyperparameters and *term reweighting*.

Often a *generative model* can help provide theoretical insight into such modeling choices, but there appears to be no such model to "explain" the above nonlinear models. For example, we know of no generative model for which the correct solution is the usual (dimension-restricted) PMI model.

This paper gives a new generative model, a dynamic version of the loglinear topic model of Mnih and Hinton (2007), as well as a pair of training objectives called RAND-WALK to compute word embeddings. The methodological novelty is to use the prior to compute *closed form* expressions for word statistics. These provide an explanation for the PMI model and other recent models, as well as hyperparameter choices.

Experimental support is provided for the generative model assumptions, the most important of which is that latent word vectors are *spatially isotropic*.

The model also helps explain why linear algebraic structure arises in low-dimensional semantic embeddings. Such structure has been used to solve analogy tasks by Mikolov et al. (2013a) and many subsequent papers. This theoretical explanation is to give an improved analogy solving method that improves success rates on analogy solving by a few percent.

1 Introduction

Vector representations of words (word embeddings) try to capture relationships between words as distance or angle, and have many applications in computational linguistics and machine learning. They are constructed by various models, all built around the unifying philosophy that the meaning of the word is defined by "the company it keeps" (Firth, 1957)—namely, co-occurrence statistics. The simplest methods use word vectors that explicitly represent co-occurrence statistics. Reweighting heuristics are known to improve these methods, as is dimension reduction (Deerwester et al., 1990). Some reweightings are nonlinear; e.g., taking the *square root* of co-occurrence counts (Rohde et al., 2006), or the *logarithm*, or the related *pairwise mutual information* (PMI) (Church and Hanks, 1990). These are called *Vector space models* (VSMs); a survey appears in (Turney et al., 2010).

Neural network language models are another approach (Hinton, 1986; Rumelhart et al., 1988; Bengio et al., 2006; Collobert and Weston, 2008); the word vector is simply the neural network's internal representation for the word. This method was sharpened and clarified via word2vec, a family of *energy based models* in (Mikolov et al., 2013b;c). The first of these papers also made the surprising discovery that despite being produced via nonlinear methods, these word vectors exhibit *linear* structure, which allows easy solutions to analogy

*Princeton University, Computer Science Department. {arora,yuanzhil,yingyul,tengyu,risteski}@cs.princeton.edu. This work was supported in part by NSF grants CCF-0832797, CCF-1117309, CCF-1302518, DMS-1317308, Simons Investigator Award, and Simons Collaboration Grant. Tengyu Ma was also supported by Simons Award for Graduate Students in Theoretical Computer Science.

questions of the form "*man*:*woman*::*king*:??." Specifically, *queen* happens to be the word whose vector v_{queen} is most similar to the vector $v_{king} - v_{man} + v_{woman}$. (Note that the two vectors may only make an angle of, say, 45 degrees, but that is still a significant overlap in 300-dimensional space.)

This surprising result caused a flurry of follow-up work, including a matrix factorization approach (Pennington et al., 2014), and experimental evidence in Levy and Goldberg (2014b) that these newer methods are related to the older PMI based models, but featuring new hyperparameters and/or term reweightings.

Another approach to word embeddings uses *latent variable probabilistic models* of language, such as *Latent Dirichlet Allocation* (LDA) and its more complicated variants (see the survey (Blei, 2012)), and some neurally inspired nonlinear models (Mnih and Hinton, 2007; Maas et al., 2011). It is worth noting that LDA evolved out of efforts in the 1990s to provide a generative model that "explains" the success of linear methods like Latent Semantic Analysis (Papadimitriou et al., 1998; Hofmann, 1999).

But there has been no corresponding latent variable generative model "explanation" for the PMI family of models: in other words, a latent variable model whose maximum likelihood (MLE) solutions approximate those seen in the PMI models.

Let's clarify this question. The simplest PMI method considers a symmetric matrix whose each row/column is indexed by a word. The entry for (w, w') is $\text{PMI}(w, w') = \log \frac{p(w,w')}{p(w)p(w')}$, where $p(w, w')$ is the empirical probability of words w, w' appearing within a window of size q in the corpus, say $q = 10$, and $p(w)$ is the marginal probability of w. (More complicated models could involve asymmetric PMI matrices with context words, and also do term reweighting.) Word vectors are obtained by low-rank SVD on this matrix or its reweightings. In particular, the PMI matrix is found to be closely approximated by a low rank matrix: there exist word vectors in say 300 dimensions— which is much smaller than the number of words in the dictionary— such that

$$\langle v_w, v_{w'} \rangle \approx \text{PMI}(w, w'). \tag{1.1}$$

(Here $\approx$ should be interpreted loosely.) This paper considers the question: Can we give a generative model "explanation" of this empirical finding? Levy and Goldberg (2014b) give an argument that if there were no dimension constraint on the solutions to the skip-gram with negative sampling model in the word2vec family, then they would satisfy (1.1), provided the right hand side were replaced by $\text{PMI}(w, w') - \beta$ for some scalar β. However, skip-gram is a discriminative model (due to use of negative sampling), not generative. Furthermore, their argument does not imply anything about low-dimensional vectors (constraining the dimension in the algorithm is important for analogy solving).

The current paper gives a probabilistic model of text generation that augments the *loglinear topic model* of Mnih and Hinton (2007) with *dynamics*, in the form of a random walk over a latent *discourse* space. Our new methodogical contribution is to derive—using the model priors—a closed-form expression that directly explains (1.1) (see Theorem 1 and experimental results in Section 4).

Section 2.1 shows the relationship of this generative model to earlier works such as word2vec and GloVe, and gives explanations for some hyperparameters. Section 4 shows good empirical fit to this model's predictions. Our model is somewhat simpler than earlier models —essentially no "knob to turn"–yet the fit to data is good.

The reason low dimension plays a key role in our theory is our assumption that the set of all word vectors (which are latent variables of the generative model) are *spatially isotropic*, which means that they have no preferred direction in space. Having n vectors be isotropic in d dimensions requires $d \ll n$. This isotropy is needed in the calculations (i.e., multidimensional integral) that yield (1.1). It also holds empirically for our word vectors, as shown in Section 4. Conceptually it seems related to the old *semantic field theory* in linguistics (Kittay, 1990).

In fact we need small d for another interesting empirical fact. For most analogies there is only a small difference between the *best* solution, and the *second-best* (incorrect) solution to the analogy, whereas the approximation error in relationship (1.1) —both empirically and according to our theory—is much larger. Why does this approximation error not kill the analogy solving? In Section 3 we explain this by mathematically showing — improving upon earlier intuitions of Levy and Goldberg (2014a) and Pennington et al. (2014)— that the isotropy of word vectors has a "purification" effect that *mitigates* the effect of this approximation error. This can also be seen as a theoretical explanation of the well-known observation (pointed out as early

as (Deerwester et al., 1990)) that dimension reduction improves the quality of word embeddings for various tasks. Section 3 also points out that the intuitive explanation —smaller models generalize better—doesn't apply.

2 Generative model and its properties

We think of corpus generation as a dynamic process, where the t-th word is produced at time t. The process is driven by the random walk of a *discourse* vector which is $c_t \in \Re^d$. Its coordinates represent *what is being talked about*[1]. Each word has a (time-invariant) latent vector $v_w \in \Re^d$ that captures its correlations with the discourse vector. To give an example, a coordinate could correspond to *gender*, with the sign indicating *male/female*. This coordinate could be positive in v_{king} and negative in v_{queen}. When the discourse vector has a positive value of the gender coordinate, it should be more likely to produce words like "*king, man*," and when it has negative value, favor words like "*queen, woman*." We model this bias with a loglinear word production model (equivalently, product-of-experts):

$$\Pr[w \text{ emitted at time } t \mid c_t] \propto \exp(\langle c_t, v_w \rangle). \tag{2.1}$$

The random walk of the discourse vector will be slow, so that nearby words are generated under similar discourses. We are interested in co-occurrence of words near each other, so occasional big jumps in the random walk are allowed because they have negligible effect on these probabilities.

A similar loglinear model appears in Mnih and Hinton (2007) but without the random walk. The linear chain CRF of Lafferty et al. (2001) is more general. The *dynamic topic model* of Blei and Lafferty (2006) utilizes topic dynamics, but with a linear word production model. Belanger and Kakade (2015) have proposed a dynamic model for text using *Kalman Filters*. The novelty here over such past works is a theoretical analysis in the *method of moments* tradition. Assuming a prior on the random walk we *analytically* integrate out the hidden random variables and compute a simple closed form expression that approximately connects the model parameters to the observable joint probabilities (see Theorem 1); this is reminiscent of analysis of similar random walk models in finance (Black and Scholes, 1973).

Model details. Let n denote the number of words, d denote the ambient dimension of the discourse space, where $d = \Omega(\log^2 n)$ and $d = O(\sqrt{n})$. Inspecting (2.1) suggests word vectors need to have varying lengths, to fit the empirical finding that word probabilities satisfy a power law. We assume that the ensemble of word vectors consists of i.i.d draws generated by $v = s \cdot \hat{v}$, where $\hat{v}$ is from the spherical Gaussian distribution[2] and s is a random scalar with expectation and standard deviation less than $\sqrt{d}/\kappa$ and absolute bound of $\kappa\sqrt{d}$ for constant κ. (Dynamic range of word probabilities will roughly equal $\exp(\kappa^2)$, so think of κ as constant like 5.)

We assume that each coordinate of the hidden discourse vectors $\{c_t\}$ lies in $[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]$. The random walk can be in any form so long as the stationary distribution $\mathcal{C}$ of the random walk is uniform on $[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]^d$, and at each step the movement of the discourse vector is at most $O(1/\log^2 n)$ in ℓ_1 norm[3]. This is still fast enough to let the walk mix quickly in the space.

Our main theorem gives simple closed form approximations for $p(w)$, the probability of word w in the corpus, and $p(w, w')$, the probability that two words w, w' occur next to each other (the same analysis works for pairs that appear in a small window, say of size 10). Recall that $\mathrm{PMI}(w, w') = \log(p(w, w')/p(w)p(w'))$

[1] This is a different interpretation of the term "discourse" than in some other settings in computational linguistics.

[2] This generative assumption about word vectors is purely for ease of exposition. It can be replaced with "deterministic" properties that are verified in the experimental section: (i) for most c the sum $\sum_w \exp(\langle v_w, c \rangle)$ is close to some constant Z (Lemma 1) (ii) facts about singular values etc. stated before Theorem 2.

The deterministic versions have the advantage of being compatible with known structure among the word vectors —e.g., clusters, linear relationships etc.—that would be highly unlikely if the vectors were truly drawn from the Gaussian prior.

[3] More precisely, the proof extends to any symmetric product distribution over the coordinates satisfying $\mathbb{E}_c\left[|c|^2\right] = \frac{1}{d}, |c|_\infty \leq \frac{2}{\sqrt{d}}$ a.s., and the steps are such that for all c_t, $\mathbb{E}_{p(c_{t+1}|c_t)}[\exp(4\kappa|c_{t+1} - c_t|_1 \log n)] \leq 1 + \epsilon_2$ for some small ϵ_2.

and we assume the window size $q = 2$ in the theorem below, and remark the extension to general q in the remarks that follow.

Theorem 1. *There is a constant $Z > 0$, and some $\epsilon = \epsilon(n, d)$ that goes to 0 as $d \to \infty$ such that with high probability over the choice of word vectors, for any two different words w and w',*

$$\log p(w, w') = \frac{\|v_w + v_{w'}\|_2^2}{2d} - 2\log Z \pm \epsilon, \tag{2.2}$$

$$\log p(w) = \frac{\|v_w\|_2^2}{2d} - \log Z \pm \epsilon. \tag{2.3}$$

Jointly these imply:

$$PMI\ (w, w') = \frac{\langle v_w, v_{w'} \rangle}{d} \pm O(\epsilon). \tag{2.4}$$

Remarks. (1) Since the word vectors have ℓ_2 norm of the order of $\sqrt{d}$, for two typical word vectors $v_w, v_{w'}$, $\|v_w + v_{w'}\|_2^2$ is of the order of $\Theta(d)$. Therefore the noise level ϵ is very small compared to the leading term $\frac{1}{2d}\|v_w + v_{w'}\|_2^2$. For PMI however, the noise level $O(\epsilon)$ could be comparable to the leading term, and empirically we also find higher error here. (2) When window size $q > 2$, both equation (2.2) and (2.4) need to be corrected with adding constant $\log\left(\frac{q(q-1)}{2}\right)$ on the right hand sides. This is also consistent with the shift β for fitting PMI in Levy and Goldberg (2014b) as remarked below. (3) Levy and Goldberg (2014b) showed that without dimension constraints, the solution to skip-gram with negative sampling satisfies PMI $(w, w') = \langle v_w, v_{w'} \rangle - \beta$ for a constant β. Our result justifies via a generative model why this should be satisfied even for low dimensional word vectors. (4) Variants of (2.2) were hypothesized and empirically supported also in (Globerson et al., 2007) and (Maron et al., 2010).

Proof sketch of Theorem 1 Here we describe a proof sketch, while the complete proof is provided in Appendix A.

Let w and w' be two arbitrary words. We start with integrating out the hidden variables c:

$$p(w, w') = \int_{c,c'} p(c, c') \left[p(w|c)p(w'|c')\right] dcdc' = \int_{c,c'} p(c, c') \frac{\exp(\langle v_w, c \rangle)}{Z_c} \frac{\exp(\langle v_{w'}, c' \rangle)}{Z_{c'}} dcdc' \tag{2.5}$$

where c and c' are the hidden discourse variables that control the emission of two words w, w', and $Z_c = \sum_w \exp(\langle v_w, c \rangle)$ is the *partition function* that is the implied normalization in equation (2.1). Integrals like (2.5) would normally be difficult, because of the partition functions. However, in our case we can prove that the values of the partition functions Z_c's typically don't vary much.

Lemma 1. *There exists Z such that with high probability $(1 - 4\exp(-d^{0.2}))$ over the choice of v_w's and c,*

$$(1 - o(1))Z \leq Z_c \leq (1 + o(1))Z \tag{2.6}$$

Using this lemma, we get that the right-hand side of (2.5) equals

$$\frac{1 \pm o(1)}{Z^2} \int_{c,c'} p(c, c') \exp(\langle v_w, c \rangle) \exp(\langle v_{w'}, c' \rangle) dcdc' \tag{2.7}$$

Our model assumptions state that c and c' cannot be too different. To leverage that, we rewrite (2.7) a little, and get that it equals

$$\frac{1 \pm o(1)}{Z^2} \left(\int_c \exp(\langle v_w, c \rangle) p(c) dc \int_{c'} \exp(\langle v_{w'}, c' \rangle) p(c'|c) dc' \right) = \frac{1 \pm o(1)}{Z^2} \left(\int_c \exp(\langle v_w, c \rangle) p(c) A(c) dc \right)$$

where $A(c) := \int_{c'} \exp(\langle v_{w'}, c' \rangle) p(c'|c) dc'$. We claim that $A(c) = (1 \pm o(1)) \exp(\langle v_{w'}, c \rangle)$. Doing some algebraic manipulations,

$$A(c) = \int_{c'} \exp(\langle v_{w'}, c' \rangle) p(c'|c) dc' = \exp(\langle v_{w'}, c \rangle) \int_{c'} \exp(\langle v_{w'}, c' - c \rangle) p(c'|c) dc'.$$

By the fact that the directions of the vectors v_w are Gaussian distributed, one can show that the maximum absolute value of any coordinate of v_w is $\|v_w\|_\infty = O(\kappa \log n)$. Furthermore, by our model assumptions, $\|c - c'\|_1 = O(1/\log^2 n)$. So

$$\langle v_w, c - c' \rangle \leq \|v_w\|_\infty \|c - c'\|_1 = o(1)$$

and thus $A(c) = (1 \pm o(1)) \exp(\langle v_{w'}, c \rangle)$. Doing this calculation carefully, we have

$$p(w, w') = \frac{1 \pm o(1)}{Z^2} \int_c p(c) \exp(\langle v_w + v_{w'}, c \rangle) dc.$$

Since c has a product distribution, by Taylor expanding $\exp(\langle v, c \rangle)$ (considering the coordinates of v as variables) and bounding the terms with order higher than 2, we can show that

$$p(w, w') = \frac{(1 \pm o(1))}{Z^2} \exp(\|v_w + v_{w'}\|_2^2 / 2d)$$

leading to the desired bound on $\log p(w, w')$ for the case when the window size $q = 2$. The bound on $\log p(w)$ can be shown similarly.

What remains is to prove Lemma 1. Note that for fixed c, when word vectors have Gaussian priors assumed as in the our model, $Z_c = \sum_w \exp(\langle v_w, c \rangle)$ is a sum of independent random variables. Using proper concentration of measure tools[4], it can be shown that the variance of Z_c are relatively small compared to its mean $\mathbb{E}_{v_w}[Z_c]$, and thus Z_c concentrates around its mean. So it suffices to show that $\mathbb{E}_{v_w}[Z_c]$ for different c are close to each other.

Using the fact that the word vector directions have a Gaussian distribution, $\mathbb{E}_{v_w}[Z_c]$ turns out to only depend on the norm of c (which are fairly concentrated around 1). More precisely,

$$\mathop{\mathbb{E}}_{v_w}[Z_c] = f(\|c\|_2^2) \tag{2.8}$$

where f is defined as $f(\alpha) = n \, \mathbb{E}_s[\exp(s^2 \alpha / 2)]$ and s has the same distribution as the norms of the word vectors. We sketch the proof of this. In our model, $v_w = s_w \cdot \hat{v}_w$, where $\hat{v}_w$ is a unit Gaussian vector, and s_w is the norm of v_w. Then

$$\mathop{\mathbb{E}}_{v_w}[Z_c] = n \mathop{\mathbb{E}}_{v_w}[\exp(\langle v_w, c \rangle)] = n \mathop{\mathbb{E}}_{s_w} \left[\mathop{\mathbb{E}}_{v_w | s_w} [\exp(\langle v_w, c \rangle) \mid s_w] \right]$$

where the second line is just an application of the law of total expectation, if we pick the norm of the (random) vector v_w first, followed by its direction. Conditioned on s_w, $\langle v_w, c \rangle$ is a Gaussian random variable with variance $\|c\|_2^2 s_w^2 / d$. Hence, $\mathbb{E}_{v_w}[Z_c] = n \, \mathbb{E}_s[\exp(s^2 \|c\|_2^2 / 2)]$, as we needed.

Using more concentration inequalities, we have that the typical c has $\|c\|_2 = 1 \pm o(1)$ and therefore it can be shown that $f(\|c'\|_2^2) \approx f(\|c\|_2^2)$. In summary, we connected Z_c with $Z_{c'}$ by

$$Z_c \approx \mathop{\mathbb{E}}_{v_w}[Z_c] = f(\|c\|_2^2) \approx f(\|c'\|_2^2) = \mathop{\mathbb{E}}_{v_w}[Z_{c'}] \approx Z_{c'}$$

Bounding the amount of approximation in each step carefully will lead to the desired result as in Lemma 1. Finally we note that the concentration bounds crucially use the fact that d is sufficiently small ($O(\sqrt{n})$), so the low-dimensionality is necessary for our main result.

[4]Note this is quite non-trivial: the random variable $\exp(\langle v_w, c \rangle)$ is not subgaussian nor bounded, since the scaling of w and c is such that $\langle v_w, c \rangle$ is $\Theta(1)$, and therefore $\exp(\langle v_w, c \rangle)$ is in the non-linear regime. In fact, the same concentration phenomenon does not happen for w. The occurrence probability of word w is not necessarily concentrated because the ℓ_2 norm of v_w can vary a lot in our model, which allows the words to have a large dynamic range.

2.1 RAND-WALK training objective and relationship to other models

To get a training objective out of Theorem 1, we reason as follows. Let L be the corpus size, and $X_{w,w'}$ the number of times words w, w' co-occur within a context of size 10 in the corpus. If the random walk mixes fairly quickly, then the set of $X_{w,w'}$'s over all word pairs is distributed (up to a very close approximation) as a multinomial distribution $\mathsf{Mul}(L, \{p_{w,w'}\})$ where $p_{w,w'} \propto \|v_w + v_{w'}\|_2^2$.

Then by a simple calculation (see Appendix B), the maximum likelihood values for the word vectors correspond to

$$\min_{\{v_w\},C} \sum_{w,w'} X_{w,w'} \left(\log(X_{w,w'}) - \|v_w + v_{w'}\|_2^2 - C\right)^2 \qquad \text{(Objective } \mathbf{SN}\text{)} \tag{2.9}$$

As usual, empirical performance is improved by weighting down very frequent word pairs, which is done by replacing the weighting $X_{w,w'}$ by its truncation $\min\{X_{w,w'}, X_{\max}\}$ where $X_{\max}$ is a constant such as 100. This is possibly because the very frequent words like "*the*" do not fit our model. We call this objective with the truncated weights the **SN** objective (**S**quared **N**orm). A similar objective **PMI** can be obtained from (2.4), using the same weighting of terms as above. (This is an approximate MLE, using that the error between the emprical and true value of $PMI(w, w')$ is driven by this error in $\Pr(w, w')$, not $\Pr[w], \Pr[w']$.)

Both objectives involve something like *Weighted SVD* which is NP-hard, but empirically seems solvable in our setting via AdaGrad.

Connection to GloVe Compare (2.9) with the objective used by GloVe (Pennington et al., 2014):

$$\sum_{w,w'} f(X_{w,w'})(\log(X_{w,w'}) - \langle v_w, v_{w'}\rangle - s_w - s_{w'} - C)^2 \tag{2.10}$$

with $f(X_{w,w'}) = \min\{X_{w,w'}^{3/4}, 100\}$. Their weightings and the need for *bias* terms $s_w, s_{w'}, C$ were experimentally derived; here they are all predicted and given meanings due to Theorem 1. In particular, our objective has essentially no "knobs to turn."

Connection to word2vec(CBOW) The CBOW model in word2vec posits that the probability of a word w_{k+1} as a function of the previous k words $w_1, w_2, \ldots, w_k$ is

$$\Pr\left[w_{k+1} \middle| \{w_i\}_{i=1}^k\right] \propto \exp(\langle v_{w_{k+1}}, \frac{1}{k}\sum_{i=1}^k v_{w_i}\rangle).$$

Assume a simplified version of our model, where a small window of k words is generated as follows: sample $c \sim \mathcal{C}$, where $\mathcal{C}$ is a uniformly random unit vector, then sample $(w_1, w_2, \ldots, w_k) \sim \exp(\langle \sum_{i=1}^k v_{w_i}, c\rangle)/Z_c$. Furthermore, assume $Z_c = Z$ for any c.

Lemma 2. *In the simplified version of our model, the Maximum-a-Posteriori (MAP) estimate of c given $(w_1, w_2, \ldots, w_k)$ is $\frac{\sum_{i=1}^k v_{w_i}}{\|\sum_{i=1}^k v_{w_i}\|_2^2}$.*

Proof. The c maximizing $\Pr[c|(w_1, w_2, \ldots, w_k)]$ is the maximizer of $\Pr[c]\Pr[(w_1, w_2, \ldots, w_k)|c]$. Since $\Pr[c] = \Pr[c']$ for any c, c', and we have $\Pr[(w_1, w_2, \ldots, w_k)|c] = \exp(\langle \sum_i v_{w_i}, c\rangle)/Z$, the maximizer is clearly $c = \frac{\sum_{i=1}^k v_{w_i}}{\|\sum_{i=1}^k v_{w_i}\|_2^2}$. □

Thus using the MAP estimate of c_t gives essentially the same expression as CBOW apart from the rescaling. (Empirical works often omit rescalings due to computational efficiency.)

3 Linear algebraic structure of concept classes

As mentioned, word analogies like "*a*:*b*::*c*:??"can be solved via a linear algebraic expression:

$$\operatorname*{argmin}_d \|v_a - v_b - v_c + v_d\|_2^2, \tag{3.1}$$

where vectors have been normalized so that $\|v_d\|_2 = 1$. (A less linear variant of (3.1) called 3COSMUL can slightly improve success rates (Levy and Goldberg, 2014a).) This suggests that the semantic relationships in question are characterized by a straight line: for each such relationship R there is a direction μ_R in space such that $v_a - v_b$ lies along roughly along μ_R. Below, we mathematically prove this fact. In Section 4 this explanation is empirically verified, as well as used to improve success rates at analogy solving by a few percent,

Previous papers have also tried to prove the existence of such a relationship among word vectors from first principles, and we now sketch what was missing in those attempts. The basic issue is approximation error: the difference between the best solution and the 2nd best solution to (3.1) is typically small, whereas the approximation error in the objective in the low-dimensional solutions is larger. Thus in principle the approximation error could kill the method (and the emergence of linear relationship) but it doesn't. (Note that expression (3.1) features 6 inner products, and all may suffer from this approximation error.)

Prior explanations Pennington et al. (2014) try to invent a model where such linear relationships should occur *by design*. They posit that *queen* is a solution to the analogy "*man*:*woman*::*king*:??" because

$$\frac{p(\chi \mid king)}{p(\chi \mid queen)} \approx \frac{p(\chi \mid man)}{p(\chi \mid woman)}, \tag{3.2}$$

where $p(\chi \mid king)$ denotes the conditional probability of seeing word χ in a small window of text around *king*. (Relationship (3.2) is intuitive since both sides will be ≈ 1 for gender-neutral χ, e.g., "*walks*" or "*food*", will be > 1 when χ is, e.g., "*he, Henry*" and will be < 1 when χ is, e.g., "*dress, she, Elizabeth.*" This was also observed by Levy and Goldberg (2014a).) Then they posit that the correct model describing word embeddings in terms of word occurrences must be a *homomorphism* from $(\Re^d, +)$ to $(\Re^+, \times)$, so vector differences map to ratios of probabilities. This leads to the expression

$$p_{w,w'} = \langle v_w, v_{w'} \rangle + b_w + b_{w'},$$

and their method is a (weighted) least squares fit for this expression. One shortcoming of this argument is that the homomorphism assumption *assumes* the linear relationships instead of explaining them from a more basic principle. More importantly, the empirical fit to the homomorphism has nontrivial approximation error, high enough that it does not imply the desired strong linear relationships.

Levy and Goldberg (2014b) show that empirically, skip-gram vectors satisfy

$$\langle v_w, v_{w'} \rangle \approx \mathrm{PMI}(w, w') \tag{3.3}$$

up to some shift. They also give an argument suggesting this relationship must be present if the solution is allowed to be very high-dimensional. Unfortunately that argument doesn't extend at all to low-dimensional embeddings. But again, empirically the approximation error is high.

Our explanation. The current paper has given a generative model to theoretically explain the emergence of relationship (3.3), but, as noted after Theorem 1, the issue of high approximation error does not go away either in theory or in the empirical fit. We now show that the isotropy of word vectors (assumed in the theoretical model and verified empirically) implies that even a weak version of (3.3) is enough to imply the emergence of the observed linear relationships in low-dimensional embeddings.

A side product of this argument will be a *mathematical* explanation of the superiority –empirically well-established— of low-dimensional word embeddings over high-dimensional ones in this setting. (Many people

seem to assume this superiority follows theoretically from generalization bounds: smaller models generalize better. But that theory doesn't apply here, since analogy solving plays no role in the training objective. Generalization theory does not apply easily to such unsupervised settings, and there is no *a priori* guarantee that a more succinct solution will do better at analogy solving —just as there is no guarantee it will do well in some other arbitrary task.)

This argument will assume the analogy in question involves a relation that obeys Pennington et al.'s suggestion in (3.2). Namely, for such a relation R there exists function $\nu_R(\cdot)$ depending only upon R such that for any a, b satisfying R there is a *noise function* $\xi_{a,b,R}(\cdot)$ for which:

$$\frac{p(\chi \mid a)}{p(\chi \mid b)} = \nu_R(\chi) \cdot \xi_{a,b,R}(\chi) \tag{3.4}$$

For different words χ there is huge variation in (3.4), so the multiplicative noise may be large.

Our goal is to show that the low-dimensional word embeddings have the property that there is a vector μ_R such that for every pair of words a, b in that relation, $v_a - v_b = \mu_R +$ noise vector, where the noise vector is small. (*This is the mathematical result missing from the earlier papers.*)

Taking logarithms of (3.4) gives:

$$\log\left(\frac{p(\chi \mid a)}{p(\chi \mid b)}\right) = \log(\nu_R(\chi)) + \zeta_{a,b,R}(\chi) \tag{3.5}$$

Theorem 1 implies that the left side simplifies to $\log\left(\frac{p(\chi|a)}{p(\chi|b)}\right) = \frac{1}{d}\langle v_\chi, v_a - v_b\rangle + \epsilon_{a,b}(\chi)$ where ϵ captures the small approximation errors induced by the inexactness of Theorem 1. This adds yet more noise! Denoting by V the $n \times d$ matrix whose rows are the v_χ vectors, we rewrite (3.5) as:

$$V(v_a - v_b) = d\log(\nu_R) + \zeta'_{a,b,R} \tag{3.6}$$

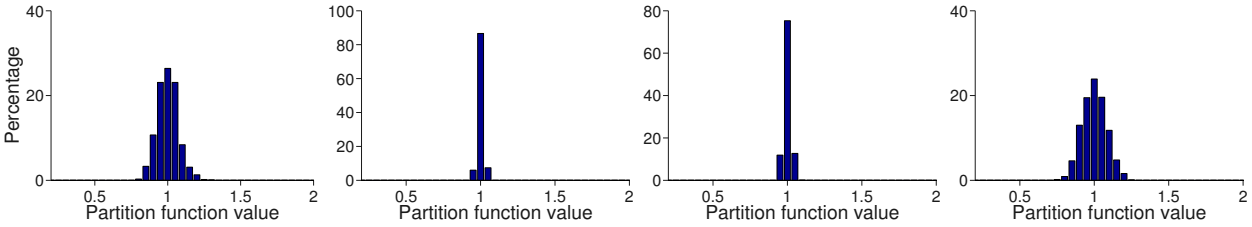
where $\log(\nu_R)$ in the entry-wise log of vector ν_R and $\zeta'_{a,b,R} = d(\zeta_{a,b,R} - \epsilon_{a,b,R})$ is the noise.

In essence, (3.6) is a linear regression with $n \gg d$ and we hope to recover $v_a - v_b$. Here V, the *design matrix* in the regression, is the matrix of all word vectors, which in our model (as well as empirically) satisfies an isotropy condition. This makes it random-like, and thus solving the regression by left-multiplying by $V^\dagger$, the pseudo-inverse of V, ought to "denoise" effectively. We now show that it does.

Our model assumed the set of all word vectors is drawn from a scaled Gaussian, but the next proof will only need the following weaker properties. (1) The smallest non-zero singular values of V is larger than some constant c_1 times the average of the squared singular values, namely, $\|V\|_F/\sqrt{d}$. (Empirically we find $c_1 \approx 1/3$ holds; see Section 4.) (2) The left singular vectors behave like random vectors with respect to $\zeta'_{a,b,R}$ —i.e., have inner product at most $c_2\|\zeta'_{a,b,R}\|/\sqrt{n}$ for some constant c_2. (3) The max norm of a row in V is $O(\sqrt{d})$.

Theorem 2 (Noise reduction)**.** *Under the conditions of the previous paragraph, the noise in the dimension-reduced semantic vector space satisfies* $\|\bar{\zeta}_{a,b,R}\|_2 \lesssim \|\zeta'_{a,b,R}\|_2 \frac{\sqrt{d}}{n}$*. As a corollary, the relative error in the dimension-reduced space is* $\sqrt{d/n}$ *factor smaller.*

Proof. The proof uses the standard analysis of linear regression. Let $V = P\Sigma Q^T$ be the SVD of V and let $\sigma_1, \ldots, \sigma_d$ be the left singular values of V (the diagonal entries of Σ). For notational ease we omit the subscript in $\bar{\zeta}$ and ζ since they are not relevant for this proof. We have $V^\dagger = Q\Sigma^{-1}P^T$ and therefore $\bar{\zeta} = V^\dagger\zeta' = Q\Sigma^{-1}P^T\zeta'$. By the second assumption we have $\|P^T\zeta'\|_\infty \le \frac{c_2}{\sqrt{n}}\|\zeta'\|_2$ and therefore $\|P^T\zeta'\|_2^2 \le \frac{c_2^2 d}{n} \cdot \|\zeta'\|_2^2$. Furthermore, $\|\bar{\zeta}\|_2 \le \sigma_d^{-1}\|P^T\zeta\|_2$. But, we claim $\sigma_d^{-1} \le \sqrt{\frac{1}{c_1 n}}$: indeed, $\sum_{i=1}^d \sigma_i^2 = O(nd)$, since the average squared norm of a word vector is d. The claim then follows from the first assumption. Continuing, we get $\sigma_d^{-1}\|P^T\zeta\|_2 \le \sqrt{\frac{1}{c_1 n}}\sqrt{\frac{c_2^2 d}{n}\|\zeta'\|_2^2} = \frac{c_2\sqrt{d}}{\sqrt{c_1}n}\|\zeta'\|_2$ as desired. The last statement follows because the norm of the "signal" (which is $d\log(\nu_R)$ originally, and is $V^\dagger d\log(\nu_R) = v_a - v_b$ after dimension reduction) also gets reduced by a factor of $\sqrt{n}$. □



(a) **SN** (b) GloVe (c) CBOW (d) skip-gram

Figure 1: The partition function Z_c. The figure shows the histogram of Z_c for 1000 random vectors c of appropriate norm, as defined in the text. The x-axis is normalized by the mean of the values. The values Z_c for different c concentrate around the mean, mostly in $[0.9, 1.1]$. This concentration phenomenon is predicted by our analysis.

4 Experimental verification

In this section, we provide experiments empirically supporting our generative model.

Corpus All word embedding vectors are trained on the English Wikipedia (March 2015 dump). It is pre-processed by standard approach (removing non-textual elements, sentence splitting, and tokenization), leaving about 3 billion tokens. Words that appeared less than 1000 times in the corpus are ignored, resulting in a vocabulary of $68,430$. The co-occurrence is then computed using windows of 10 tokens to each side of the focus word.

Training method Our embedding vectors are trained by optimizing the **SN** objective (2.9) using AdaGrad (Duchi et al., 2011) with initial learning rate of 0.05 and 100 iterations. The **PMI** objective derived from (2.4) was also used. **SN** has average (weighted) term-wise error of 5%, and **PMI** has 17%. We observed that **SN** vectors typically fit the model better and have better performance, which can be explained by larger errors in PMI, as implied by Theorem 1. So, we only report the results for **SN**.

For comparison, GloVe and two variants of word2vec (skip-gram and CBOW) vectors are trained. GloVe's vectors are trained on the same co-occurrence as **SN** with the default parameter values[5]. word2vec vectors are trained using a window size of 10, with other parameters set to default values[6].

4.1 Model verification

Experiments were run to test our modeling assumptions. First, we tested two counter-intuitive properties: the isotropy property of the word vectors (see Theorem 2) and the concentration of the partition function Z_c for different discourses c (see Theorem 1). For comparison we also tested these properties for word2vec and GloVe vectors, though they are trained by different objectives. Finally, we tested the linear relation between the squared norms of our word vectors and the logarithm of the word frequencies, as implied by Theorem 1.

Isotropy For the isotropy condition, the quadratic mean of the singular values is 34.3, while the minimum non-zero singular value of our word vectors is 11. Therefore, the ratio between them is a small constant, consistent with our model. The ratios for GloVe, CBOW, and skip-gram are 1.4, 10.1, and 3.1, respectively, which are all small constants.

[5] http://nlp.stanford.edu/projects/glove/
[6] https://code.google.com/p/word2vec/

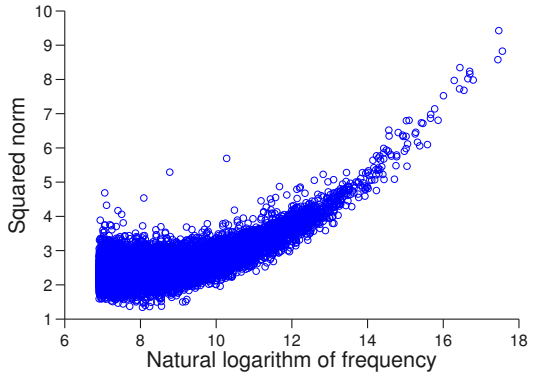


Figure 2: The linear relationship between the squared norms of our word vectors and the logarithms of the word frequencies. Each dot in the plot corresponds to a word, where x-axis is the natural logarithm of the word frequency, and y-axis is the squared norm of the word vector. The Pearson correlation coefficient between the two is 0.75, indicating a significant linear relationship, which strongly supports our mathematical prediction, that is, equation (2.3) of Theorem 1.

Partition function Our proofs predict the concentration of the partition function $Z_c = \sum_{w'} \exp(c^\top w')$ for a random discourse vector c (see Lemma 1). This is verified empirically by picking a uniformly random direction, of norm $\|c\| = 4/\mu_w$, where μ_w is the average norm of the word vectors[7]. Figure 1(a) shows the histogram of Z_c for 1000 such randomly chosen c's for our vectors. The values are concentrated, mostly in the range $[0.9, 1.1]$ times the mean. Concentration is also observed for other types of vectors, especially for GloVe and CBOW.

Squared norms v.s. word frequencies Figure 2 shows a scatter plot for the squared norms of our vectors and the logarithms of the word frequencies. A linear relationship is observed (Pearson correlation 0.75), thus supporting Theorem 1. The correlation is stronger for high frequency words, possibly because the corresponding terms have higher weights in the training objective.

4.2 Performance on analogy tasks

We compare the performance of our word vectors on analogy tasks, specifically the two testbeds GOOGLE and MSR (Mikolov et al., 2013a;c). The former contains 7874 semantic questions such as "*man*:*woman*::*king*:??", and 10167 syntactic ones such as "*run*:*runs*::*walk*:??." The latter has 8000 syntactic questions for adjectives, nouns, and verbs.

To solve these tasks, we use linear algebraic query[8]. That is, first normalize the vectors to unit norm and then solve "a:b::c:??" by

$$\operatorname*{argmin}_d \|v_a - v_b - v_c + v_d\|_2^2 . \tag{4.1}$$

The algorithm succeeds if the best d happens to be correct.

The performance of different methods is presented in Table 1. Our vectors achieve performance comparable to the others. On semantic tasks, our vectors achieve similar accuracy as GloVe, while word2vec has lower performance. On syntactic tasks, they achieve accuracy 0.04 lower than GloVe and skip-gram, while

[7]Note that our model uses the inner products between the discourse vectors and word vectors, so it is invariant if the discourse vectors are scaled by s while the word vectors are scaled by $1/s$ for any $s > 0$. Therefore, one needs to choose the norm of c properly. We assume $\|c\|\mu_w = \sqrt{d}/\kappa \approx 4$ for a constant $\kappa = 5$ so that it gives a reasonable fit to the predicted dynamic range of word frequencies according to our theory; see model details in Section 2.

[8]One can instead use the 3COSMUL in (Levy and Goldberg, 2014a), which increases the accuracy by about 3%. But it is not linear while our focus here is the linear algebraic structure.

	Relations	**SN**	GloVe	CBOW	skip-gram
G	semantic	0.84	0.85	0.79	0.73
	syntactic	0.61	0.65	0.71	0.68
	total	0.71	0.73	0.74	0.70
M	adjective	0.50	0.56	0.58	0.58
	noun	0.69	0.70	0.56	0.58
	verb	0.48	0.53	0.64	0.56
	total	0.53	0.57	0.62	0.57

Table 1: The accuracy on two word analogy task testbeds: G (the GOOGLE testbed); M (the MSR testbed). Performance is close to state of the art despite using a generative model with provable properties.

relation	1	2	3	4	5	6	7
1st	0.65 ± 0.07	0.61 ± 0.09	0.52 ± 0.08	0.54 ± 0.18	0.60 ± 0.21	0.35 ± 0.17	0.42 ± 0.16
2nd	0.02 ± 0.28	0.00 ± 0.23	0.05 ± 0.30	0.06 ± 0.27	0.01 ± 0.24	0.07 ± 0.24	0.01 ± 0.25
relation	8	9	10	11	12	13	14
1st	0.56 ± 0.09	0.53 ± 0.08	0.37 ± 0.11	0.72 ± 0.10	0.37 ± 0.14	0.40 ± 0.19	0.43 ± 0.14
2nd	0.00 ± 0.22	0.01 ± 0.26	0.02 ± 0.20	0.01 ± 0.24	0.07 ± 0.26	0.07 ± 0.23	0.09 ± 0.23

Table 2: The verification of relation directions on the GOOGLE testbed. For each relation, take $v_{ab} = v_a - v_b$ for pairs (a, b) in the relation, and then calculate the top singular vectors of the matrix formed by these v_{ab}'s. The row with label "1st"/"2nd" shows the cosine similarities of individual v_{ab} to the 1st/2nd singular vector (the mean and standard deviation).

CBOW typically outperforms the others. This is not suprising since our model is tailored for modeling semantics, and lacks specific features for syntactic relations. For example, a word "*she*" can affect the context by a lot and can determine if the next word is "*thinks*" rather than "*think*." Incorporating such features in the model is left for future work.

4.3 Directions corresponding to relations

The theory in Section 3 predicts the existence of a direction for a relation. To verify this, we took $v_{ab} = v_a - v_b$ for word pairs (a, b) in a relation, calculated the top singular vectors of the matrix formed by these v_{ab}'s, and computed the cosine similarity of individual v_{ab} to the singular vectors. Table 2 shows the mean similarities and standard deviations on the first and second singular vectors. This shows that most $(v_a - v_b)$'s are close to the first singular vector, which is approximately the vector for the relation. Their similarities to the second singular vectors are centered around 0 with larger deviations, indicating that these components look like random noise, in line with our model.

Next, we indicate how to use this linear structure to improve analogy solving.

Cheating solver for analogy testbeds As a proof of concept we first design cheating way to improve performance on the analogy testbed. This uses the fact that the same semantic relationship (e.g., masculine-feminine, singular-plural) is tested many times in the testbed. If a relation R is represented by a direction μ_R then the cheating algorithm can learn this direction (via rank 1 SVD) after seeing a few examples of the relationship. Then use the following method of solving "a:b::c:??": look for a word d such that $v_c - v_d$ has the largest projection on μ_R, the relation direction for (a, b). This can boost success rates by about 10%.

The testbed can try to combat such cheating by giving analogy questions in a random order. But the cheating algorithm can just *cluster* the presented analogies to learn which of them rest the same relation. Thus the final algorithm, named analogy solver with relation direction by clustering (**RD-c**), is: take all vectors $v_a - v_b$ for all the word pairs (a, b) presented among the analogy questions and do k-means clustering

	SN	GloVe	CBOW	skip-gram
w/o **RD**	0.71	0.73	0.74	0.70
RD-c ($k = 20$)	0.74	0.77	0.79	0.75
RD-c ($k = 30$)	0.79	0.80	0.82	0.80
RD-c ($k = 40$)	0.76	0.80	0.80	0.77

Table 3: The accuracy of **RD-c** algorithm (i.e., the cheater method) on the GOOGLE testbed. The algorithm is described in the text. For comparison, the row "w/o **RD**" shows the accuracy of the old method without using relation direction.

	SN	GloVe	CBOW	skip-gram
w/o **RD**	0.71	0.73	0.74	0.70
RD-nn ($k = 10$)	0.71	0.74	0.77	0.73
RD-nn ($k = 20$)	0.72	0.75	0.77	0.74
RD-nn ($k = 30$)	0.73	0.76	0.78	0.74

Table 4: The accuracy of **RD-nn** algorithm on the GOOGLE testbed. The algorithm is described in the text. For comparison, the row "w/o **RD**" shows the accuracy of the old method without using relation direction.

on them; for each (a, b), estimate the relation direction by taking the first singular vector of its cluster, and substitute that for $v_a - v_b$ in (4.1) when solving the analogy. Table 3 shows the performance on GOOGLE with different values of k; e.g. using our **SN** vectors and $k = 30$ leads to 0.79 accuracy. Thus future designers of analogy testbeds should remember not to test the same relationship too many times! This still leaves other ways to cheat, such as learning the directions for interesting semantic relations from other collections of analogies.

Non-cheating solver for analogy testbeds Now we show that even if a relationship is tested only once in the testbed, there is a way to use the above structure. Given "a:b::c:??," the solver first finds the top 300 nearest neighbors of a and those of b, and then finds among these neighbors the top k pairs (a', b') so that the cosine similarities between $v_{a'} - v_{b'}$ and $v_a - v_b$ are largest. Finally, the solver uses these pairs to estimate the relation direction (via rank 1 SVD), and substitute this (corrected) estimate for $v_a - v_b$ in (4.1) when solving the analogy. This algorithm is named *analogy solver with relation direction by nearest neighbors* (**RD-nn**).

Table 4 shows its performance, which consistently improves over the old method by about 3%.

5 Conclusions

A simple generative model has been given to "explain" the classical PMI based word embedding models, as well as recent variants involving energy-based models and matrix factorization. Though our RAND-WALK training objective has almost no "knobs to turn", the model fits surprisingly well with the word pair co-occurrence data and solves analogies almost as well as prior discriminative models.

The spatial isotropy of word vectors is both an assumption in our model, and also a new empirical finding of our paper. We feel it may help with further development of language models. It is important for explaining the success of solving analogies via low dimensional vectors. It also implies that semantic relationships among words manifest themselves as special directions among word embeddings (Section 3), which leads to a cheater algorithm for solving analogy testbeds.

Our model is tailored to capturing semantic similarity, more akin to a loglinear dynamic topic model. In particular, local word order is unimportant. Designing similar generative models (with provable and interpretable properties) with linguistic features is left for future work.

Acknowledgments

We would like to thank Yann LeCun, Christopher D. Manning, and Sham Kakade for numerous helpful discussions throughout various stages of this work.

References

David Belanger and Sham M. Kakade. A linear dynamical system model for text. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 833–842. JMLR.org, 2015.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. 2006.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, pages 637–654, 1973.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei and John D Lafferty. Dynamic topic models. In *International Conference on Machine Learning*, 2006.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning*, 2008.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391–407, 1990.

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

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

Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8:2265–2295, 2007.

Geoffrey E Hinton. Distributed representations. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1986.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

Eva Feder Kittay. *Metaphor: Its cognitive force and linguistic structure*. Oxford University Press, 1990.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 2014a.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014b.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics*, pages 142–150, 2011.

Yariv Maron, Michael Lamar, and Elie Bienenstock. Sphere embedding: An application to part-of-speech induction. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations*, 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*, 2013b.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013c.

Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine Learning*, pages 641–648. ACM, 2007.

Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 159–168, New York, NY, USA, 1998. ACM. ISBN 0-89791-996-3. doi: 10.1145/275487.275505. URL http://doi.acm.org/10.1145/275487.275505.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing*, 2014.

Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut. An improved model of semantic similarity based on lexical co-occurence. *Communications of the ACM*, 2006.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 1988.

Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

A Proofs of Theorem 1

In this section we prove Theorem 1 and 2 (restated below) .

Theorem 1. *Assume that the hidden contexts are at stationary distribution, with high probability over the choice of v_w's, we have that for any two different words w and w'*

$$\log p(w, w') = \frac{1}{2d}\|v_w + v_{w'}\|^2 - 2\log Z \pm o(1) \tag{A.1}$$

for some fixed constant Z. Moreover,

$$\log(p(w)) = \frac{1}{2d}\|v_w\|_2^2 - \log Z \pm o(1). \tag{A.2}$$

Lemma 1. *There exists Z such that for any context c with $|\|c\| - 1| \le d^{-0.4}$, with high probability $(1 - 2e^{-2n^{0.4}})$) over the choice of v_w's,*

$$(1 - o(1))Z \le Z_c \le (1 + o(1))Z.$$

We first prove Theorem 1 using Lemma 1, and Lemma 1 will be proved in Section A.1. For the intuition of the proof, please see Section 2 of the main paper.

Proof of Theorem 1. Let c be the hidden context that determines the probability of word w, and c' be the next one that determines w'. We use $p(c'|c)$ to denote the Markov kernel (transition matrix) of the Markov chain. Let $\mathcal{C}$ be the stationary distribution of context vector c. We marginalize over the contexts c, c' and then use the independence of w, w' conditioned on c, c',

$$\begin{aligned} p(w, w') &= \int_{c,c'} p(w|c)p(w'|c')p(c, c')dcdc' \\ &= \int_{c,c'} \frac{\exp(\langle v_w, c\rangle)}{Z_c}\frac{\exp(\langle v_{w'}, c'\rangle)}{Z_{c'}}p(c)p(c'|c)dcdc' \end{aligned} \tag{A.3}$$

We first get rid of the partition function Z_c using Lemma 1, which says that there exists Z such that, with probability $1 - 4\exp(-d^{0.2})$,

$$(1 - \epsilon_z)Z \le Z_c \le (1 + \epsilon_z)Z. \tag{A.4}$$

where $\epsilon_z = o(1)$.

Let $\mathcal{F}$ be the event that both c and c' satisfy (A.4)and $\overline{\mathcal{F}}$ be its negation, and let $\mathbf{1}_{\mathcal{F}}$ be the indicator function for the event $\mathcal{F}$. Therefore we have $\Pr[\mathcal{F}] \ge 1 - 4\exp(-d^{0.8})$.

We first decompose the integral (A.3) into the two parts according to whether event $\mathcal{F}$ happens,

$$\begin{aligned} p(w, w') = &\int_{c,c'} \frac{1}{Z_c Z_{c'}}\exp(\langle v_w, c\rangle)\exp(\langle v_{w'}, c'\rangle)p(c)p(c'|c)\mathbf{1}_{\mathcal{F}}dcdc' \\ &+ \int_{c,c'} \frac{1}{Z_c Z_{c'}}\exp(\langle v_w, c\rangle)\exp(\langle v_{w'}, c'\rangle)p(c)p(c'|c)\mathbf{1}_{\overline{\mathcal{F}}}dcdc' \end{aligned} \tag{A.5}$$

We bound the first quantity on RHS by using (A.4) and the definition of $\mathcal{F}$.

$$\begin{aligned} &\int_{c,c'} \frac{1}{Z_c Z_{c'}}\exp(\langle v_w, c\rangle)\exp(\langle v_{w'}, c'\rangle)p(c)p(c'|c)\mathbf{1}_{\mathcal{F}}dcdc' \\ &\le (1 + \epsilon_z)^2\frac{1}{Z^2}\int_{c,c'} \exp(\langle v_w, c\rangle)\exp(\langle v_{w'}, c'\rangle)p(c)p(c'|c)\mathbf{1}_{\mathcal{F}}dcdc' \end{aligned} \tag{A.6}$$

and for the second one we use the fact that $Z_c \geq n$ and $\exp(\langle v_w, c\rangle) \leq \exp(2\kappa\sqrt{d})$ (by assumption $\|v_w\| \leq \kappa\sqrt{d}$ and $\|c\| \leq 2$), and conclude

$$\begin{aligned}&\int_{c,c'} \exp(\langle v_w, c\rangle) \exp(\langle v_{w'}, c'\rangle) p(c)p(c'|c)\mathbf{1}_{\overline{\mathcal{F}}} dcdc' \\ &\leq \Pr[\overline{\mathcal{F}}] \cdot \exp(4\kappa\sqrt{d}) \leq \exp(-d^{0.7}) \end{aligned} \tag{A.7}$$

For the last inequality we use $\Pr[\overline{\mathcal{F}}] \leq 4\exp(-d^{0.2})$. Combining (A.5), (A.6) and (A.7), we obtain

$$\begin{aligned}p(w, w') &\leq (1+\epsilon_z)^2 \frac{1}{Z^2} \int_{c,c'} \exp(\langle v_w, c\rangle) \exp(\langle v_{w'}, c'\rangle) p(c)p(c'|c)\mathbf{1}_{\mathcal{F}} dcdc' + \exp(-d^{0.2}) \\ &\leq (1+\epsilon_z)^2 \frac{1}{Z^2} \left(\int_{c,c'} \exp(\langle v_w, c\rangle) \exp(\langle v_{w'}, c'\rangle) p(c)p(c'|c) dcdc' + \delta_0 \right)\end{aligned}$$

where $\delta_0 = \exp(-d^{0.2})Z^2 \leq \exp(-d^{0.1})$. This is because $Z \leq \exp(2\kappa)n$ and $d = \omega(\log^2 n)$, and κ is a constant.

On the other hand, we can lowerbound similarly

$$\begin{aligned}p(w, w') &\geq (1-\epsilon_z)^2 \frac{1}{Z^2} \int_{c,c'} \exp(\langle v_w, c\rangle) \exp(\langle v_{w'}, c'\rangle) p(c)p(c'|c)\mathbf{1}_{\mathcal{F}} dcdc' \\ &\geq (1-\epsilon_z)^2 \frac{1}{Z^2} \left(\int_{c,c'} \exp(\langle v_w, c\rangle) \exp(\langle v_{w'}, c'\rangle) p(c)p(c'|c) dcdc' - \exp(-d^{0.7}) \right) \\ &\geq (1-\epsilon_z)^2 \frac{1}{Z^2} \left(\int_{c,c'} \exp(\langle v_w, c\rangle) \exp(\langle v_{w'}, c'\rangle) p(c)p(c'|c) dcdc' - \delta_0 \right)\end{aligned}$$

Taking logarithm, the multiplicative error translates to a additive error

$$\log p(w, w') = \log \left(\int_{c,c'} \exp(\langle v_w, c\rangle) \exp(\langle v_{w'}, c'\rangle) p(c)p(c'|c) dcdc' \pm \delta_0 \right) - 2\log Z + 2\log(1 \pm \epsilon_z)$$

For the purpose of exploiting the fact that c, c' should be close to each other, we further rewrite $\log p(w, w')$ by re-organizing the integrals,

$$\begin{aligned}\log p(w, w') &= \log \left(\int_c \exp(\langle v_w, c\rangle) p(c) dc \int_{c'} \exp(\langle v_{w'}, c'\rangle) p(c'|c) dc' \pm \delta_0 \right) - 2\log Z + 2\log(1 \pm \epsilon_z) \\ &= \log \left(\int_c \exp(\langle v_w, c\rangle) p(c) A(c, c') dc \pm \delta_0 \right) - 2\log Z + 2\log(1 \pm \epsilon_z)\end{aligned} \tag{A.8}$$

where the inner integral which is denoted by $A(c, c')$,

$$A(c, c') := \int_{c'} \exp(\langle v_{w'}, c'\rangle) p(c'|c) dc'$$

Note that by Lemma 3, we have that for any $w \in W$, $\|v_w\|_\infty \leq 4\kappa \log n$. Therefore we have that $\langle v_w, c - c'\rangle \leq \|v_w\|_\infty \|c - c'\|_1 \leq 4\kappa \log n \|c - c'\|_1$.

Then we can bound $A(c, c')$ by

$$
\begin{aligned}
A(c,c') &= \int_{c'} \exp(\langle v_{w'}, c' \rangle) p(c'|c) dc' \\
&= \exp(\langle v_{w'}, c \rangle) \int_{c'} \exp(\langle v_{w'}, c' - c \rangle) p(c'|c) dc' \\
&\leq \exp(\langle v_{w'}, c \rangle) \int_{c'} \exp(4\kappa |c' - c|_1 \log n) p(c'|c) \\
&= \exp(\langle v_{w'}, c \rangle) \mathop{\mathbb{E}}_{p(c'|c)} [\exp(4\kappa |c' - c|_1 \log n)] \\
&\leq (1 + \epsilon_2) \exp(\langle v_{w'}, c \rangle)
\end{aligned}
$$

For the lower bound of $A(c, c')$, we first observe that

$$
\mathop{\mathbb{E}}_{p(c'|c)} [\exp(4\kappa |c' - c|_1 \log n)] + \mathop{\mathbb{E}}_{p(c'|c)} [\exp(-4\kappa |c' - c|_1 \log n)] \geq 2
$$

Therefore it follows model assumption that

$$
\mathop{\mathbb{E}}_{p(c'|c)} [\exp(-4\kappa |c' - c|_1 \log n)] \geq 1 - \epsilon_2
$$

Therefore,

$$
\begin{aligned}
A(c,c') &= \exp(\langle v_{w'}, c \rangle) \int_{c'} \exp(\langle v_{w'}, c' - c \rangle) p(c'|c) dc' \\
&\geq \exp(\langle v_{w'}, c \rangle) \int_{c'} \exp(-4\kappa \|c' - c\| \log n) p(c'|c) dc' \\
&= \exp(\langle v_{w'}, c \rangle) \mathop{\mathbb{E}}_{p(c'|c)} [\exp(-4\kappa \|c' - c\| \log n)] \\
&\geq (1 - \epsilon_2) \exp(\langle v_{w'}, c \rangle)
\end{aligned}
$$

Therefore, we obtain that $A(c, c') = (1 \pm \epsilon_2) \exp(\langle v_{w'}, c \rangle)$. Plugging the estimation of $A(c, c')$ into the equation A.8, we obtain that

$$
\begin{aligned}
\log p(w, w') &= \log \left(\int_c \exp(\langle v_w, c \rangle) p(c) A(c, c') dc \pm \delta_0 \right) - 2 \log Z + 2 \log(1 \pm \epsilon_z) \\
&= \log \left(\int_c \exp(\langle v_w, c \rangle) p(c) (1 \pm \epsilon_2) \exp(\langle v_{w'}, c \rangle) dc \pm \delta_0 \right) - 2 \log Z + 2 \log(1 \pm \epsilon_z) \\
&= \log \left(\int_c \exp(\langle v_w, c \rangle) p(c) \exp(\langle v_{w'}, c \rangle) dc \pm \delta_0 \right) - 2 \log Z + 2 \log(1 \pm \epsilon_z) + \log(1 \pm \epsilon_2) \\
&= \log \left(\int_c \exp(\langle v_w + v_{w'}, c \rangle) p(c) dc \pm \delta_0 \right) - 2 \log Z + 2 \log(1 \pm \epsilon_z) + \log(1 \pm \epsilon_2) \\
&= \log \left(\mathop{\mathbb{E}}_{c \sim \mathcal{C}} [\exp(\langle v_w + v_{w'}, c \rangle)] \pm \delta_0 \right) - 2 \log Z + 2 \log(1 \pm \epsilon_z) + \log(1 \pm \epsilon_2)
\end{aligned}
$$

Now it suffices to compute $\mathbb{E}_{c \sim \mathcal{C}} [\exp(\langle v_w + v_{w'}, c \rangle)]$. Let $t = v_w + v_{w'}$. By our assumption, $\mathcal{C}$ is a product distribution across the coordinates. Therefore we can write

$$
\mathop{\mathbb{E}}_{c \sim \mathcal{C}} [\exp(\langle v_w + v_{w'}, c \rangle)] = \prod_{i=1}^{d} \mathop{\mathbb{E}}_{c_i} \exp(t_i c_i)
$$

Using lemma 4 for $t_i c_i \leq 1$(we used the fact that $t_i \leq 8\kappa \log n$ for all $t = v_w + v_{w'}$, $c_i \leq \frac{2}{\sqrt{d}}$ (see Lemma 3); In our setting κ is a constant, $d = \omega((\log n)^2)$), we can estimate $\mathbb{E}_{c_i} \exp(t_i c_i)$ by

$$\mathop{\mathbb{E}}_{c_i} \exp(t_i c_i) = 1 + \frac{t_i^2}{2d} + O(t_i^4/d^2)$$

Using the fact that $x - \frac{x^2}{2} \leq \ln(1+x) \leq x$.

$$\begin{aligned}
\log \mathop{\mathbb{E}}_{c}[\exp(\langle v_w + v_{w'}, c\rangle)] &= \sum_{i=1}^{d} \log \mathop{\mathbb{E}}_{c_i} \exp(t_i c_i) = \sum_{i=1}^{d} \log\left(1 + \frac{t_i^2}{2d} + O(t_i^4/d^2)\right) \\
&= \sum_{i=1}^{d} \frac{t_i^2}{2d} + O(t_i^4/d^2) \\
&= \|t\|^2/(2d) + O(\|t\|_4^4/d^2)
\end{aligned}$$

Putting altogether, we have that

$$\begin{aligned}
\log p(w, w') &= \log\left(\mathop{\mathbb{E}}_{c\sim\mathcal{C}}[\exp(\langle v_w + v_{w'}, c\rangle)] \pm \delta_0\right) - 2\log Z + 2\log(1 \pm \epsilon_z) + \log(1 \pm \epsilon_2) \\
&= \|v_w + v_{w'}\|^2/(2d) + O(\|v_w + v_{w'}\|_4^4/d^2) + O(\delta_0') - 2\log Z \pm 2\epsilon_z \pm \epsilon_2 \\
&= (1+\delta)\|v_w + v_{w'}\|^2/(2d) - 2\log Z \pm 2\epsilon_z \pm \epsilon_2
\end{aligned}$$

where $\delta_0' = \delta_0 \cdot (\mathbb{E}_{c\sim\mathcal{C}}[\exp(\langle v_w + v_{w'}, c\rangle)])^{-1} = o(1)$ and $\delta = d^{-0.4}$, where in the last step we used Lemma 3 that $\|v_w + v_{w'}\|_4^4/d^2 < \|v_w + v_{w'}\|^2/d^{1.6}$.

Note that ϵ_z, ϵ_2 are on the order of $o(1)$, and $\delta(\|v_w + v_{w'}\|^2/(2d)) = o(1)$ for all w and w' by Lemma 3, we obtain the desired bound,

$$\log p(w, w') = \frac{1}{2d}\|v_w + v_{w'}\|^2 - 2\log Z \pm o(1).$$

□

Lemma 3. *With high probability over the choice of v_w's, we have that for any $w \in W$ and any i, $(v_w)_i \leq 4\kappa \log n$, and for any pair of words w, w',*

$$\|v_w + v_{w'}\|_4^4/d^2 < \|v_w + v_{w'}\|^2/d^{1.6}$$

Proof. Recall that we assume v_w are generated independently as $v_w = s_w \cdot \hat{v}_w$ where $s_w \leq \kappa\sqrt{d}$ for some constant κ and $\hat{v}_w$ is from a spherical Gaussian distribution (each coordinate is i.i.d $N(0, 1/d)$).

Let's do each of the claims separately.

For a standard Gaussian distribution, we know that

$$\Pr\left[|(\hat{v}_w)_i| \geq 4d^{-0.5}\log n\right] \leq e^{-\frac{16}{2}\log^2 n}$$

Since $s_w \leq \kappa\sqrt{d}$, we know that

$$\Pr\left[|(s_w \cdot \hat{v}_w)_i| \geq 4\kappa \log n\right] \leq e^{-\frac{16}{2}\log^2 n}$$

Union bounding, we have that with probability $1 - dne^{-\frac{16}{2}\log^2 n} = 1 - o(1)$ (recall that $d < n^{0.5}$), for all words w, for every coordinate i, $(v_w)_i \leq 4\kappa \log n$.

The second claim is not much more difficult. For standard Gaussian distribution, we know that

$$\Pr\left[|(\hat{v}_w)_i| \geq d^{-0.3}/\kappa\right] \leq e^{-\frac{d^{0.4}}{2\kappa^2}}$$

Since $s_w \leq \kappa\sqrt{d}$, we know that

$$\Pr\left[|(s_w \cdot \hat{v}_w)_i| \geq d^{0.2}\right] \leq e^{-\frac{d^{0.4}}{2\kappa^2}}$$

Taking a union bound, we have: with probability $1 - dne^{-\frac{d^{0.4}}{2\kappa^2}}$ (note in our setting $dne^{-\frac{d^{0.4}}{2\kappa^2}} = o(1)$ for large enough d), for all words w and their coordinate i, $|(v_w)_i| \leq d^{0.2}$. In this case we have:

$$(v_w + v_{w'})_i^4/d^2 \leq (v_w + v_{w'})_i^2/d^{1.6}$$

which easily implies the claim we want.

□

Lemma 4. *If a real random variable X is symmetric and $\mathbb{E}[X^2] = 1/d$ and $|X| \leq 2/\sqrt{d}$ a.s. Then for $t < \sqrt{d}/10$, we have*

$$1 + \frac{t^2}{2d} \leq \mathop{\mathbb{E}}_X \exp(tX) \leq 1 + \frac{t^2}{2d} + \frac{4}{3}\left(\frac{t}{\sqrt{d}}\right)^4$$

Proof. By the moment generating function of X, we have

$$\mathop{\mathbb{E}}_X \exp(tX) = \sum_{j=0}^{\infty} \frac{t^j}{(j)!}\,\mathbb{E}[X^j]$$

Therefore by the assumption that X is symmetric and $\mathbb{E}[X^2] = \frac{1}{d}$, we have that

$$\mathop{\mathbb{E}}_X \exp(tX) \geq 1 + \frac{t^2}{2d}$$

On the other hand, using the fact that $|X| < \frac{2}{\sqrt{d}}$ a.s.

$$\mathop{\mathbb{E}}_X \exp(tX) = \sum_{j=0}^{\infty} \frac{t^{2j}}{(2j)!}\,\mathbb{E}[X^{2j}] = 1 + \frac{t^2}{2d} + \sum_{j=2}^{\infty} \frac{1}{(2j)!}\left(\frac{2t}{\sqrt{d}}\right)^{2j} \leq 1 + \frac{t^2}{2d} + \frac{4}{3}\left(\frac{t}{\sqrt{d}}\right)^4$$

where the last inequality is because we choose $t < \sqrt{d}/10$: $\frac{t}{\sqrt{d}} < 1/10$; hence

$$\sum_{j=2}^{\infty} \frac{1}{(2j)!}\left(\frac{2t}{\sqrt{d}}\right)^{2j} \leq \left(\frac{t}{\sqrt{d}}\right)^4 \sum_{j=0}^{\infty}\left(\frac{1}{5}\right)^{2j} \leq \frac{4}{3}\left(\frac{t}{\sqrt{d}}\right)^4$$

□

A.1 Analyzing partition function Z_c

In this section, we prove Lemma 1. We basically first prove that for the means of Z_c are all $(1 + o(1))$-close to each other, and then prove that Z_c is concentrated around its mean. It turns out the concentration part is non trivial because the random variable of concern, $\exp(\langle v_w, c\rangle)$ is not well-behaved in terms of the tail. Note that $\exp(\langle v_w, c\rangle)$ is NOT sub-gaussian for any variance proxy. This essentially disallows us to use an existing concentration inequality directly. We get around this issue by considering the truncated version of $\exp(\langle v_w, c\rangle)$, which is bounded, and have similar tail properties as the original one, in the regime that we are concerning.

Proof of Lemma 1. Recall that by definition

$$Z_c = \sum_w \exp(\langle v_w, c\rangle).$$

We fix context c and view v_w as random variables throughout this proof. For convenience, we denote the norm of c by $\ell = \|c\|$. Recall that v_w is composed of $v_w = s_w \cdot \hat{v}_w$, where s_w is the scaling and $\hat{v}_w$ is from spherical Gaussian with covariance $\frac{1}{d} I_{d\times d}$ and thus almost a unit vector.

Just as a warm-up, we lowerbound the mean of Z_c as follows:

$$\mathbb{E}[Z_c] = n\,\mathbb{E}\left[\exp(\langle v_w, c\rangle)\right] \geq n\,\mathbb{E}\left[1 + \langle v_w, c\rangle\right] = n$$

On the other hand, to upperbound the mean of Z_c, we condition on the scaling s_w,

$$\begin{aligned}\mathbb{E}[Z_c] &= n\,\mathbb{E}[\exp(\langle v_w, \tilde{v}_c\rangle)] \\ &= n\,\mathbb{E}\left[\mathbb{E}\left[\exp(\langle v_w, \tilde{v}_c\rangle) \mid s_w\right]\right]\end{aligned}$$

Note that conditioned on s_w, we have that $\langle v_w, \tilde{v}_c\rangle$ is a Gaussian random variable with variance $\sigma^2 = \|c\|^2 s_w^2/d$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\exp(\langle v_w, \tilde{v}_c\rangle) \mid s_w\right] &= \int_x \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2}) \exp(x) dx \\ &= \int_x \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\sigma^2)^2}{2\sigma^2} + \sigma^2/2) dx \\ &= \exp(\sigma^2/2)\end{aligned}$$

It follows that

$$\mathbb{E}[Z_c] = \mathbb{E}[\exp(\sigma^2/2)] = \mathbb{E}[\exp(s_w^2\|c\|^2/2)] = \mathbb{E}[\exp(s^2\|c\|^2/2)]$$

Let $Z := \mathbb{E}[\exp(|s|^2/2d)]$. By Proposition 5, we have that $1 - o(d^{-0.4}) \leq \|c\| \leq 1 + o(d^{-0.4})$. Therefore, for any c,

$$\mathbb{E}[Z_c] = \mathbb{E}[\exp(s^2\|c\|^2/2d)] \leq \mathbb{E}[\exp(s^2/2) \cdot \exp(o(d^{-0.4})s^2/2d)] \leq (1 + o(d^{-0.4})\kappa^2/2)Z = (1 + o(1))Z$$

Similarly we can prove that $\mathbb{E}[Z_c] \geq (1 - o(1))Z$ for any c.

We calculate the variance of Z_c as follows:

$$\begin{aligned}\mathbb{V}[(Z_c - \mathbb{E}Z_c)^2] &= \sum_w \mathbb{V}\left[\exp(\langle v_w, c\rangle)^2\right] \leq n\,\mathbb{E}[\exp(2\langle v_w, c\rangle)] \\ &= n\,\mathbb{E}\left[\mathbb{E}\left[\exp(2\langle v_w, c\rangle) \mid s_w\right]\right]\end{aligned}$$

By a very similar calculation as above, using the fact that $\langle v_w, c\rangle$ is a Gaussian random variable with variance $\sigma^2 = \ell^2\|s_w\|^2/d$,

$$\begin{aligned}\mathbb{E}\left[\exp(\langle v_w, c\rangle^2) \mid s_w\right] &= \int_x \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2}) \exp(2x) dx \\ &= \int_x \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-2\sigma^2)^2}{2\sigma^2} + 2\sigma^2) dx \\ &= \exp(2\sigma^2)\end{aligned}$$

Therefore, we have that

$$\begin{aligned}\mathbb{E}[(Z_c - \mathbb{E}Z_c)^2] &\le n\,\mathbb{E}\left[\mathbb{E}\left[\exp(2\langle v_w, \tilde{v}_c\rangle) \mid s_w\right]\right] \\ &= n\,\mathbb{E}\left[\exp(2\sigma^2)\right] = n\,\mathbb{E}\left[\exp(2\ell^2\|s_w\|^2/d)\right] \le \Lambda n\end{aligned}$$

For $\Lambda = \exp(8\kappa^2)$ being a constant. Therefore, the standard deviation of Z_c is $\sqrt{\Lambda n}$ is much less than n. Also note that $\mathbb{E}[Z_c] \ge n$, therefore we should expect with good probability over the choice of v_w's, we have that Z_c is within $\mathbb{E}[Z_c] \pm \sqrt{\Lambda n} = \mathbb{E}[Z_c](1 + o(1))$.

However, observe that $\exp(\langle v_w, c\rangle)$ is not sub-Gaussian or bounded. This disallows us to apply the usual concentration inequalities. The rest of the proof deals with this issue in a slightly more specialized manner.

Let's define $\mathcal{F}_w$ be the event that $\exp(\langle v_w, c\rangle) < d^{0.2}$. Observe that $\mathcal{F}$ is a very high probability event with $\Pr[\mathcal{F}_w] \ge 1 - \exp(-d^{0.2}/\kappa^2)$. Let random variable X_w have the same distribution as $\exp(\langle v_w, c\rangle)|_{\mathcal{F}_w}$.

We prove concentration inequality for $Z'_c = \sum_w X_w$. Observe that mean of Z'_c is lowerbounded

$$\mathbb{E}[Z'_c] = n\,\mathbb{E}\left[\exp(\langle v_w, c\rangle)|_{\mathcal{F}_w}\right] \ge n\exp(\mathbb{E}\left[\langle v_w, c\rangle|_{\mathcal{F}_w}\right]) = n$$

and the variance is upperbounded by

$$\begin{aligned}\mathbb{V}[Z'_c] &\le n\,\mathbb{E}\left[\exp(\langle v_w, c\rangle)^2|_{\mathcal{F}_w}\right] \\ &\le \frac{1}{\Pr[\mathcal{F}_w]}\,\mathbb{E}\left[\exp(\langle v_w, c\rangle)^2\right] \\ &\le \frac{1}{\Pr[\mathcal{F}_w]}\Lambda n \le 1.1\Lambda n\end{aligned}$$

where the second line uses the fact that

$$\begin{aligned}\mathbb{E}\left[\exp(\langle v_w, c\rangle)^2\right] \quad &= \Pr[\mathcal{F}_w]\,\mathbb{E}\left[\exp(\langle v_w, c\rangle)^2|\mathcal{F}_w\right] + \Pr[\overline{\mathcal{F}}_w]\,\mathbb{E}\left[\exp(\langle v_w, c\rangle)^2|\overline{\mathcal{F}}_w\right] \\ &\ge Pr[\mathcal{F}_w]\,\mathbb{E}\left[\exp(\langle v_w, c\rangle)^2|\mathcal{F}_w\right].\end{aligned}$$

Moreover, by definition, for any w, $|X_w| \le d^{0.2}$. Therefore by Bernstein's inequality, we have that

$$\Pr\left[|Z'_c - \mathbb{E}[Z'_c]| > 4\sqrt{\Lambda n} + 12n^{0.7}\right] \le e^{-2n^{0.4}}$$

By the fact that $\mathbb{E}[Z'_c] \ge n$, we have that for $\epsilon = n^{-0.3} \le d^{-0.6}$ (we use the fact that $d < n^{0.5}$)

$$\Pr\left[|Z'_c - \mathbb{E}[Z'_c]| > \epsilon\,\mathbb{E}[Z'_c]\right] \le 2e^{-2n^{0.4}}$$

Let $\mathcal{F} = \cap_w \mathcal{F}_w$ be the union of all $\mathcal{F}_w$. We have that by definition, Z'_c have the same distribution as $Z_c|_{\mathcal{F}}$. Therefore, we have that

$$\Pr[|Z_c - \mathbb{E}[Z'_c]| > \epsilon\,\mathbb{E}[Z'_c] \mid \mathcal{F}] \le 2e^{-2n^{0.4}}$$

and therefore

$$\Pr[|Z_c - \mathbb{E}[Z'_c]| > \epsilon\,\mathbb{E}[Z'_c]] \le \frac{1}{\Pr[\mathcal{F}]} \cdot \Pr[|Z_c - \mathbb{E}[Z'_c]| > \epsilon\,\mathbb{E}[Z'_c] \mid \mathcal{F}] \le 2e^{-2n^{0.4}}$$

Finally we show that $\mathbb{E}[Z'_c]$ are close to each other as well. We take c that satisfies that $\|c\| = 1 \pm d^{-0.4}$ and consider $\mathbb{E}[Z_c] = \mathbb{E}[\exp(\langle v, c\rangle)|\mathcal{F}]$, where v is from the same distribution where v_w is generated, and $\mathcal{F}$ is the event that $\exp(\langle v, c\rangle) \le d^{0.2}$. Note that random variable $\exp(\langle v, c\rangle)|\mathcal{F}$ is really rational invariant with

respect to c. Therefore we have that $\mathbb{E}[Z'_c] = \mathbb{E}[\exp(\langle v, \|c\| z\rangle)|\mathcal{F}]$, where z is any unit vector in the space, and $\mathcal{F}'$ is the event that $\exp(\langle v, z\rangle) \leq d^{0.2}$.

$$\begin{aligned}
\mathbb{E}[Z'_{c_1}] &\leq \mathbb{E}[\exp(\langle v, \|c\| z\rangle)|\mathcal{F}'] \leq \mathbb{E}[\exp(\langle v, z\rangle)|\mathcal{F}'] \sup\{\exp(\langle v, (\|c\| - 1)z\rangle)|\mathcal{F}'\} && \text{(A.9)}\\
&= \mathbb{E}[\exp(\langle v, \|c\| z\rangle)|\mathcal{F}'] \exp(d^{-0.2}) && \text{(A.10)}\\
&= (1 + o(1))\, \mathbb{E}[\exp(\langle v, \|c\| z\rangle)|\mathcal{F}'] && \text{(A.11)}
\end{aligned}$$

Similarly we can prove that

$$\mathbb{E}[Z'_{c_1}] \geq (1 - o(1))\, \mathbb{E}[\exp(\langle v, \|c\| z\rangle)|\mathcal{F}'] \quad \text{(A.12)}$$

Therefore, let $Z = \mathbb{E}[\exp(\langle v, \|c\| z\rangle)|\mathcal{F}']$, we have the desired result.

□

Proposition 5. *When $c \sim \mathcal{C}$ is at stationary distribution of the random walk, we have that*

$$\Pr_{c\sim\mathcal{C}}\left[|\|c\| - 1| > 2d^{-0.4}\right] \leq 2\exp(-d^{0.2})$$

Proof. By assumption, each coordinate of c is independent with $\mathbb{E}[c_i^2] = \frac{1}{d}$ and $|c_i|^2 \leq \frac{4}{d}$, so the Proposition 5 follows from standard Chernoff bound. □

B Maximum likelihood estimator for co-occurrence

In this section, we present a simple calculation to provide justification for the weighting in our training objective. Let L to be the corpus size, and W be the set of all words. We assume that the co-occurrence counts $X_{w,w'}$ for word pairs are generated according to a multinomial distribution $\mathsf{Mul}(m, p(w, w')_{w,w'\in W})$. Denoting $\{X_{w,w'}\}$ the set of random variables corresponding to co-occurrence counts for all of the word pairs (w, w') and $\{p_{w,w'}\}$ the set of corresponding probabilities, we show that $\log \Pr[\{X_{w,w'}\}|\{p(w, w')\}]$ is of the form:

$$\log \Pr[\{X_{w,w'}\}|\{p(w, w')\}] \approx C - \frac{1}{2L} \sum_{w,w'} X_{w,w'} \left(x_{w,w'} - \log(X_{w,w'})\right)^2$$

where C is a constant that depends on the data but not $\{p_{w,w'}\}$, $x_{w,w'} = \log(Lp(w, w'))$, and the approximation ignores the lower order terms obtained from the Taylor expansion.

More, precisely:

Theorem 3. *Suppose the random variables $\{X_{w,w'}\}$ are generated from $\mathsf{Mul}(m, p(w, w')_{w,w'\in W})$. Then:*

$$\log \Pr[\{X_{w,w'}\} \mid \{p(w, w')\}] = C - \frac{1}{2L}\left(\sum_{w,w'} X_{w,w'}(x_{w,w'} - \log X_{w,w'})^2\right)$$

$$+O\left(\frac{1}{L} X_{a,b}(x_{a,b} - \log X_{a,b})^2 + \frac{1}{L}\sum_{w,w'} X_{w,w'}(x_{w,w'} - \log X_{w,w'})^3 + \frac{1}{L^2}\left(\sum_{w,w'} X_{w,w'}(x_{w,w'} - \log X_{w,w'})\right)^2\right)$$

where $x_{w,w'} = \log(Lp(w, w'))$, and C only depends on X but not p, and (a, b) is an arbitrary pair such that $X_{a,b} \neq 0$. Typically, the terms inside the big-oh are much smaller than $\frac{1}{2L}\left(\sum_{w,w'} X_{w,w'}(x_{w,w'} - \log X_{w,w'})^2\right)$.

Proof. For any pair (a, b) such that $X_{a,b} \neq 0$, it holds that:

$$\log \Pr\left[\{X_{w,w'}\} \mid \{p(w, w')\}\right] = \log\left(C_1 \left(\prod_{(w,w')\neq(a,b)} p(w, w')^{X_{w,w'}}\right)\left(1 - \sum_{(w,w')\neq(a,b)} p(w, w')\right)^{X_{a,b}}\right)$$

where C_1 is the number of documents of size L such that co-occurence count of each w, w' is $X_{w,w'}$. Substituting $x_{w,w'} = \log[Lp(w,w')]$, we have:

$$\begin{aligned}\log \Pr\left[\{X_{w,w'}\} \mid \{p(w,w')\}\right] &= \log C_1 - L\log L + \frac{1}{L}\left(\sum_{(w,w')\neq(a,b)} X_{w,w'} x_{w,w'}\right) \\ &\quad + X_{a,b}\log\left(1 - \sum_{(w,w')\neq(a,b)} p_{w,w'}\right) \end{aligned} \tag{B.1}$$

The last term can be expanded as follows. Let ξ denote $O([x_{w,w'} - \log X_{w,w'}]^3)$.

$$\begin{aligned}
& \log\left(1 - \sum_{(w,w')\neq(a,b)} p_{w,w'}\right) \\
= \;& \log\left(1 - \frac{1}{L}\sum_{(w,w')\neq(a,b)} e^{x_{w,w'}}\right) \\
= \;& \log\left(1 - \frac{1}{L}\sum_{(w,w')\neq(a,b)} X_{w,w'} e^{x_{w,w'} - \log X_{w,w'}}\right) \\
= \;& \log\left(1 - \frac{1}{L}\sum_{(w,w')\neq(a,b)} X_{w,w'}\left(1 + [x_{w,w'} - \log X_{w,w'}] + [x_{w,w'} - \log X_{w,w'}]^2/2 + \xi\right)\right) \\
= \;& \log\left(X_{a,b} - \frac{1}{L}\sum_{(w,w')\neq(a,b)} X_{w,w'}\left([x_{w,w'} - \log X_{w,w'}] + [x_{w,w'} - \log X_{w,w'}]^2/2 + \xi\right)\right) \\
= \;& \log X_{a,b} + \log\left(1 - \frac{1}{LX_{a,b}}\sum_{(w,w')\neq(a,b)} X_{w,w'}\left([x_{w,w'} - \log X_{w,w'}] + [x_{w,w'} - \log X_{w,w'}]^2/2 + \xi\right)\right) \\
= \;& \log X_{a,b} - \frac{1}{LX_{a,b}}\sum_{(w,w')\neq(a,b)} X_{w,w'}\left([x_{w,w'} - \log X_{w,w'}] + [x_{w,w'} - \log X_{w,w'}]^2/2 + \xi\right) \\
& + O\left(\frac{1}{L^2 X_{a,b}^2}\left(\sum_{(w,w')\neq(a,b)} X_{w,w'}(x_{w,w'} - \log X_{w,w'})\right)^2\right).
\end{aligned}$$

Putting this into (B.1), we get the final bound. □