On The Dimensionality of Word Embedding

Central University of Finance and Economics

Fan Jia

Mar 10th , 2019

Outline:

□ Introduction

□ Preliminaries and Background Knowledge

□ PIP Loss: a Novel Unitary-invariant Loss Function for Embeddings

□ How Does Dimensionality Affect the Quality of Embedding?

□ Two New Discoveries

□ Conclusion

Word Embedding != Word2Vec

Embedding :

- **Mapping:** $\mathcal{F}: \boldsymbol{X} \rightarrow \boldsymbol{Y}$
- **injective:** each Y has a unique X, and vice versa
- **structure-preserving:** if $X_1 < X_2$, then $Y_1 < Y_2$

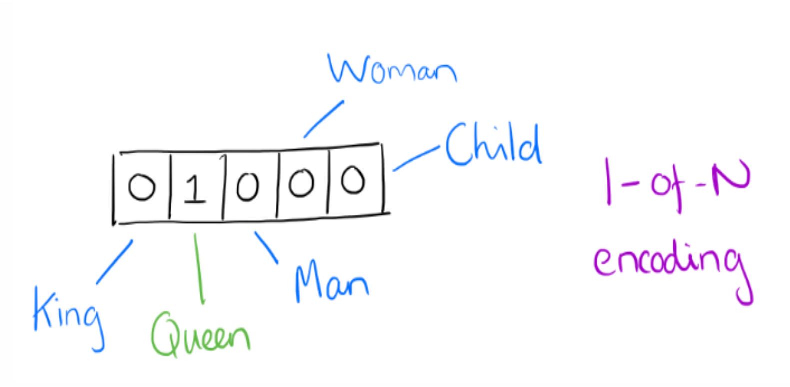
The Meaning Given By Wikipedia

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are **mapped** to vectors of real numbers.

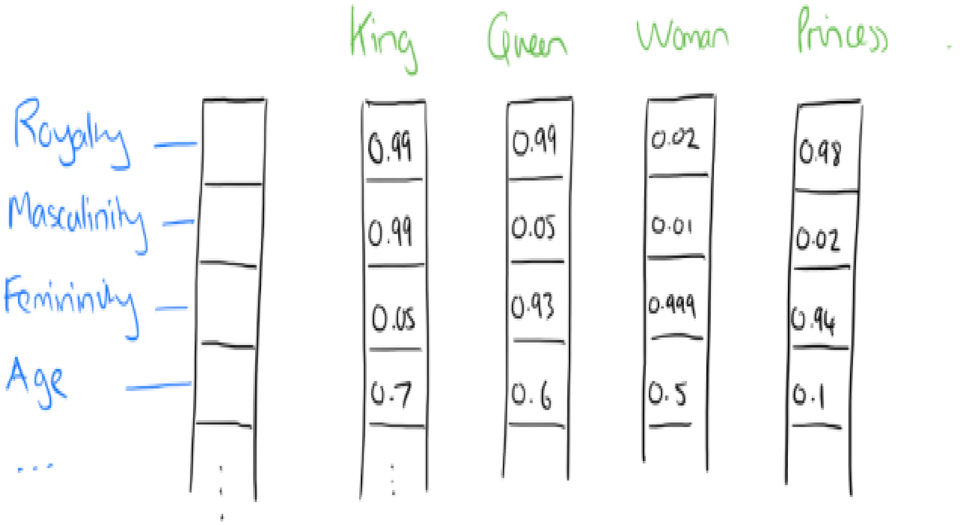
Conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with a **much lower dimension**.

Word2Vec

One-hot representation
(*1-of- N*)



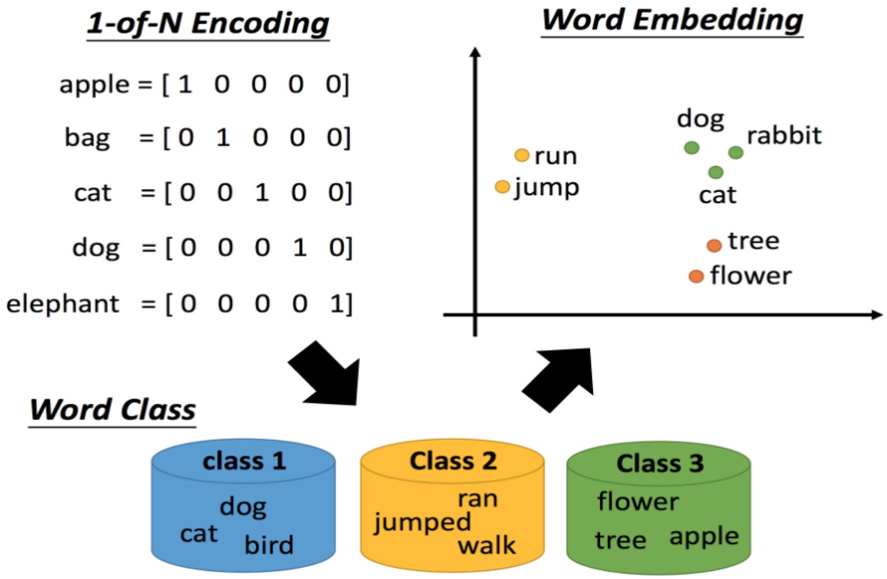
Distributed representation
（Word Embedding）



	King	Queen	Woman	Princess
Royalty	0.99	0.99	0.02	0.98
Masculinity	0.99	0.05	0.01	0.02
Femininity	0.05	0.93	0.999	0.94
Age	0.7	0.6	0.5	0.1
...	⋮			

Word2Vec

1-of-N Encoding can't reflect the relationship between the words (such as can not reflect the cat, the relationship between the dog). While by word embedding, we can learn the **relationship** between the vector easily.



1-of-N Encoding

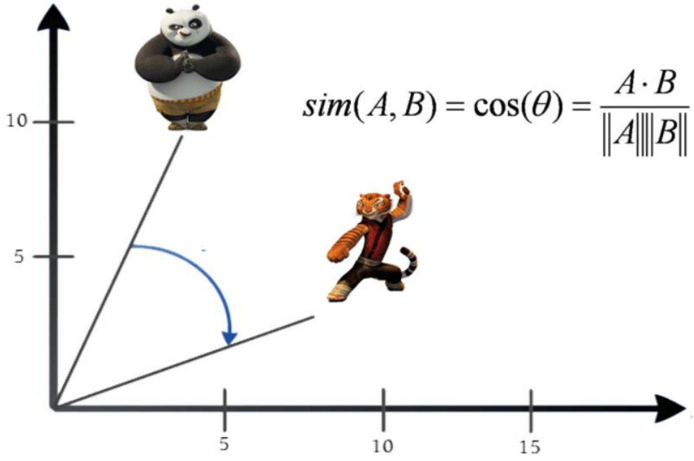
apple = [1 0 0 0 0]
bag = [0 1 0 0 0]
cat = [0 0 1 0 0]
dog = [0 0 0 1 0]
elephant = [0 0 0 0 1]

Word Embedding

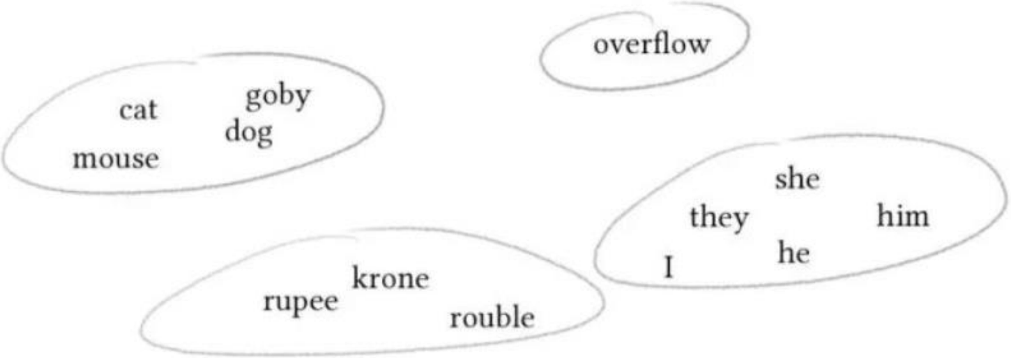
Word Class

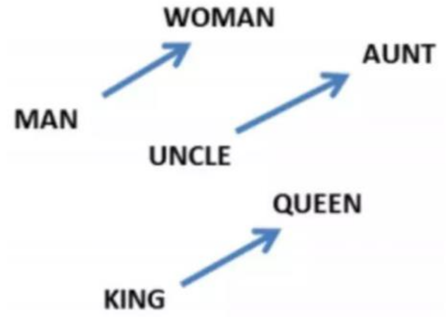
Geometric properties and geometric invariance:

- word similarity
- word analogy
- word clustering



$$sim(A,B)=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|}$$





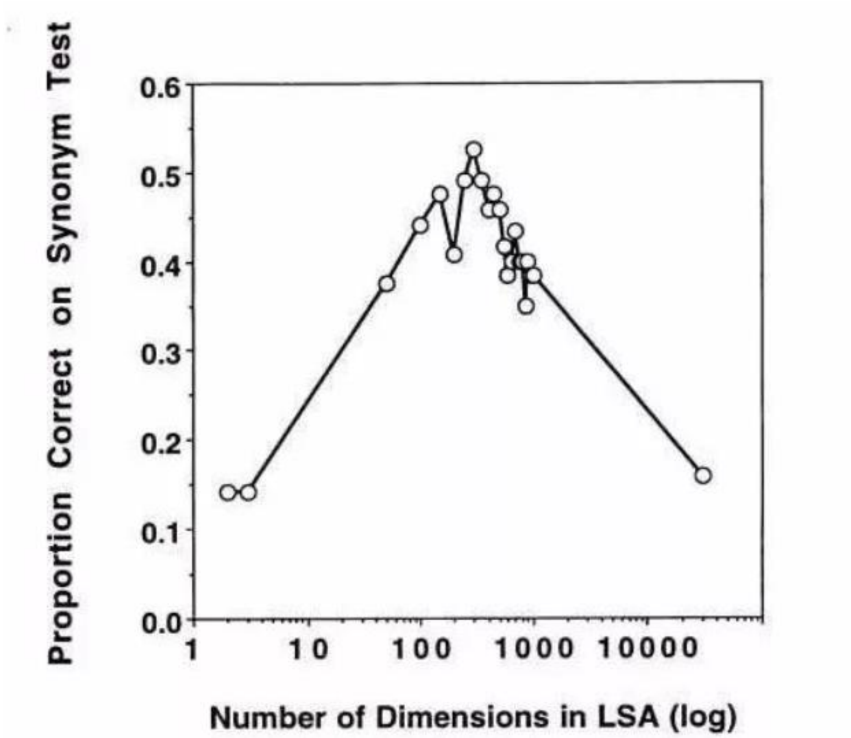
On The Dimensionality of Word Embedding

However, the impact of dimensionality on word embedding has not yet been fully understood.

In most NLP research, dimensionality is either selected **ad hoc** or by **grid search**, either of which can lead to sub-optimal model performances.

For example, 300 is perhaps the most commonly used dimensionality in various studies. This is possibly due to the influence of the groundbreaking paper, which introduced the skip-gram Word2Vec model and chose a dimensionality of 300.

As a critical hyper-parameter, the choice of dimensionality for word vectors has huge influence on the **performance of a word embedding.**



A word embedding with a small dimensionality is typically **not expressive enough** to capture all possible word relations, whereas one with a very large dimensionality suffers from **over-fitting**.

Model Selection Based on Parameter Estimation

- Define a class of estimators:

$$\Theta$$

- Define a loss function:

$$||\theta - \theta^*||$$

- Select the estimator with the least expected loss function:

$$argmin_{\hat{\theta} \in \Theta} E[||\theta - \theta^*||]$$

Embedding Matrix ***E***:

- Defined a **vocabulary** as $V = \{1,2,\ldots,n\}$ with size n
- **A vector representation** $v_i \in R^d$ for each token i
- **The embedding matrix** $E \in R^{n\times d}$ consists of the stacked vectors

v_i, where $E_{i\cdot} = v_i$.

<I> Unitary Invariance of Word Embeddings

- Two embeddings are essentially identical if one can be obtained from the other by performing **a unitary operation**:

$$v' = vU\ ,$$

where $U^T U = U U^T = Id$

- A **unitary transformation** preserves the relative geometry of the vectors, and hence defines an equivalence class of embeddings.

$$E_2 = E_1 U, E_2 E_2^T = (E_1 U)(U^T E_1^T) = E_1 E_1^T$$

PIP Loss

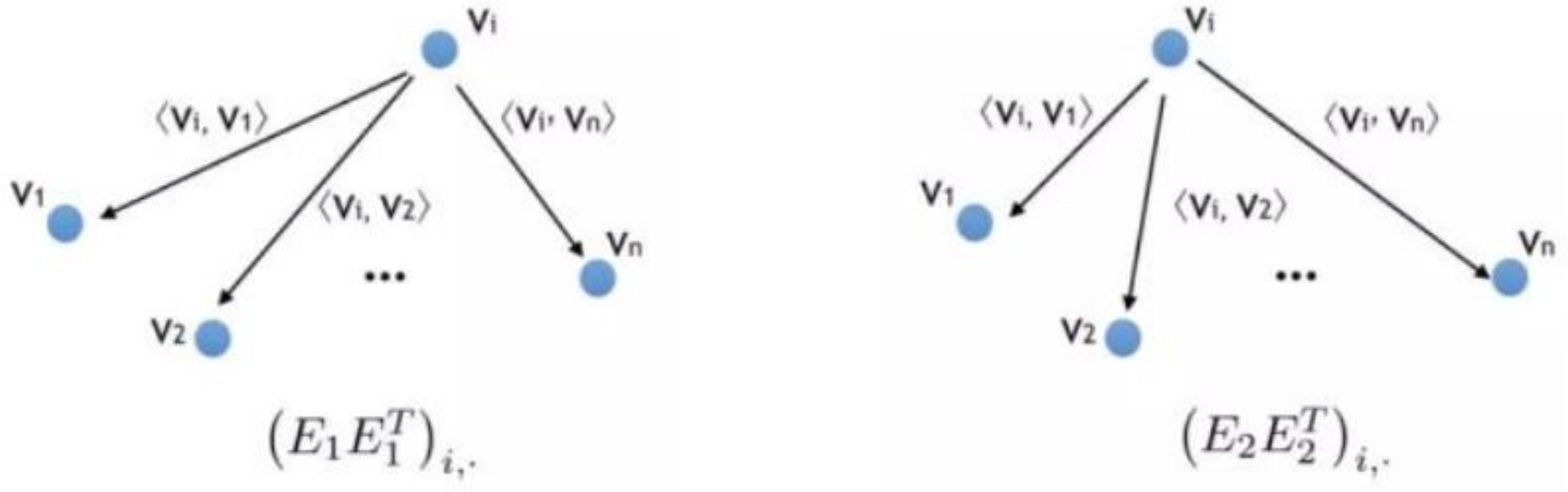
Definition 1 (PIP matrix)**.** Given an embedding matrix $E \in \mathbb{R}^{n \times d}$, define its associated Pairwise Inner Product (PIP) matrix to be

$$\text{PIP}(E) = EE^T$$

It can be seen that the (i, j)-th entry of the PIP matrix corresponds to the inner product between the embeddings for word i and word j, i.e. $\text{PIP}_{i,j} = \langle v_i, v_j \rangle$. To compare E_1 and E_2, two embedding matrices on a common vocabulary, we propose the **PIP loss**:

Definition 2 (PIP loss)**.** The PIP loss between E_1 and E_2 is defined as the norm of the difference between their PIP matrices

$$\|\text{PIP}(E_1) - \text{PIP}(E_2)\| = \|E_1E_1^T - E_2E_2^T\| = \sqrt{\sum_{i,j}(\langle v_i^{(1)}, v_j^{(1)} \rangle - \langle v_i^{(2)}, v_j^{(2)} \rangle)^2}$$



<2> Word Embeddings from Matrix Factorization

- **Explicit Matrix Factorization :**

LSA(Latent Semantic Analysis):

obtained by truncated SVD of a signal matrix M.

- **Implicit Matrix Factorization :**

Although skip-gram Word2Vec and Glo Ve learn word embeddings by optimizing over some objective functions using stochastic gradient methods, they have both been shown to be implicitly performing matrix factorizations.

- **Define M as a signal matrix**

- **The oracle embedding E:**

$$E = f_{\alpha,d}(M)$$

where $f_{\alpha,d}(M) \triangleq U_{\cdot,1:d} D^{\alpha}_{1:d,1:d}$, and $M = UDV^T$ is the SVD.

- **The trained embedding $\widehat{E}$:**

$$\hat{E} = f_{\alpha,d}(\widetilde{M})$$

where $\widetilde{M} = M + Z$ is perturbed by the estimation noise Z.

To ensure $\hat{\mathbf{E}}$ is close to E, we want the PIP loss $||\mathbf{E}\mathbf{E}^{\mathbf{T}} - \hat{\mathbf{E}}\hat{\mathbf{E}}^{\mathbf{T}}||$ to be small.

Dimensionality's Affect the Quality of Embedding

Theorem 3 (Main theorem)**.** Suppose $\tilde{M} = M + Z$, where M is the signal matrix, symmetric with spectrum $\{\lambda_i\}_{i=1}^{d}$. Z is the estimation noise, symmetric with iid, zero mean, variance σ^2 entries. For any $0 \leq \alpha \leq 1$ and $k \leq d$, let the oracle and trained embeddings be

$$E = U_{\cdot,1:d} D^{\alpha}_{1:d,1:d}, \ \hat{E} = \tilde{U}_{\cdot,1:k} \tilde{D}^{\alpha}_{1:k,1:k}$$

where $M = UDV^T$, $\tilde{M} = \tilde{U}\tilde{D}\tilde{V}^T$ are the SVDs of the clean and estimated signal matrices. Then

1. When $\alpha = 0$,

$$\mathbb{E}[\|EE^T - \hat{E}\hat{E}^T\|] \leq \sqrt{d - k + 2\sigma^2 \sum_{r \leq k, s > d} (\lambda_r - \lambda_s)^{-2}}$$

2. When $0 < \alpha \leq 1$,

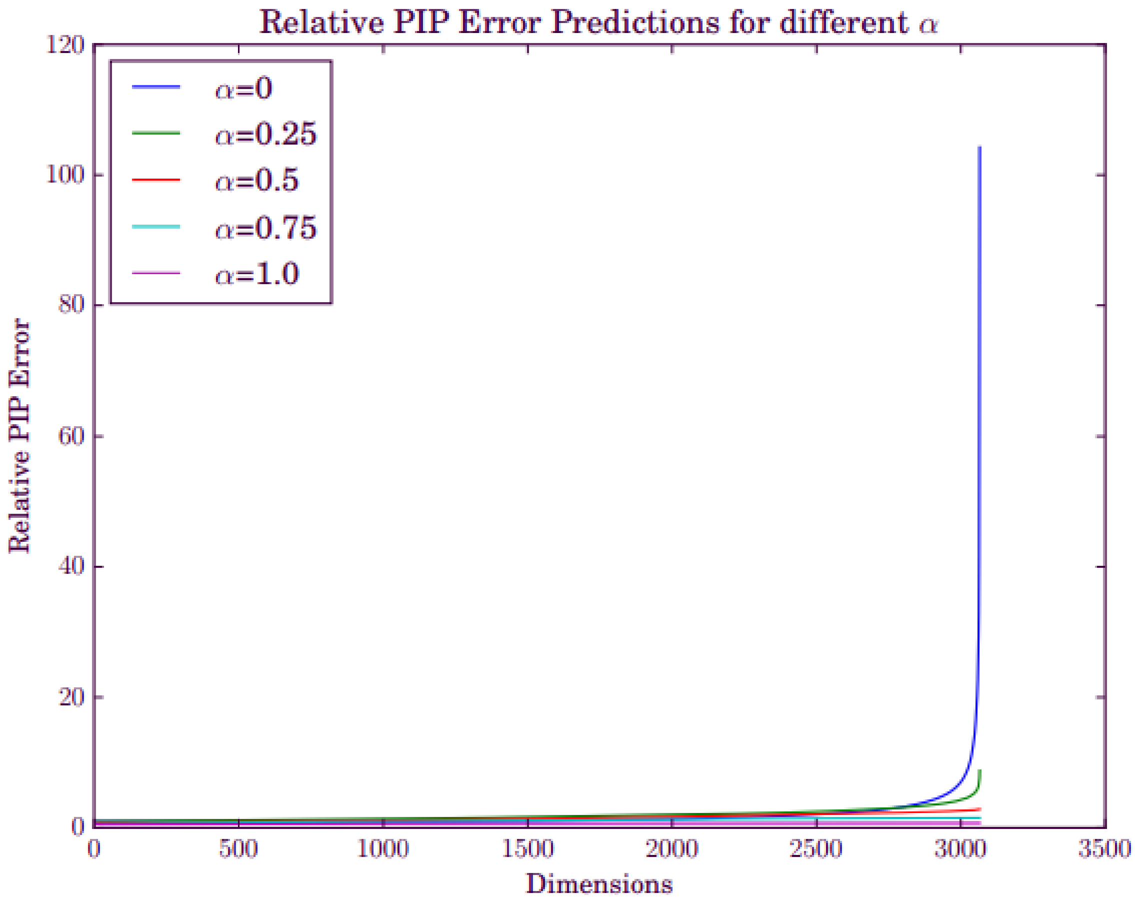
$$\mathbb{E}[\|EE^T - \hat{E}\hat{E}^T\|] \leq \underset{\substack{\downarrow \\ \textit{bias}}}{\sqrt{\sum_{i=k+1}^{d} \lambda_i^{4\alpha}}} + \underset{\substack{\downarrow \\ \textit{variance on} \\ \textit{signal magnitudes}}}{2\sqrt{2n}\alpha\sigma\sqrt{\sum_{i=1}^{k} \lambda_i^{4\alpha-2}}} + \underset{\substack{\downarrow \\ \textit{variance on} \\ \textit{signal directions}}}{\sqrt{2}\sum_{i=1}^{k}(\lambda_i^{2\alpha} - \lambda_{i+1}^{2\alpha})\sigma\sqrt{\sum_{r \leq i < s} (\lambda_r - \lambda_s)^{-2}}}$$

Dimensionality's Affect the Quality of Embedding

When $0 < \alpha \leq 1$,

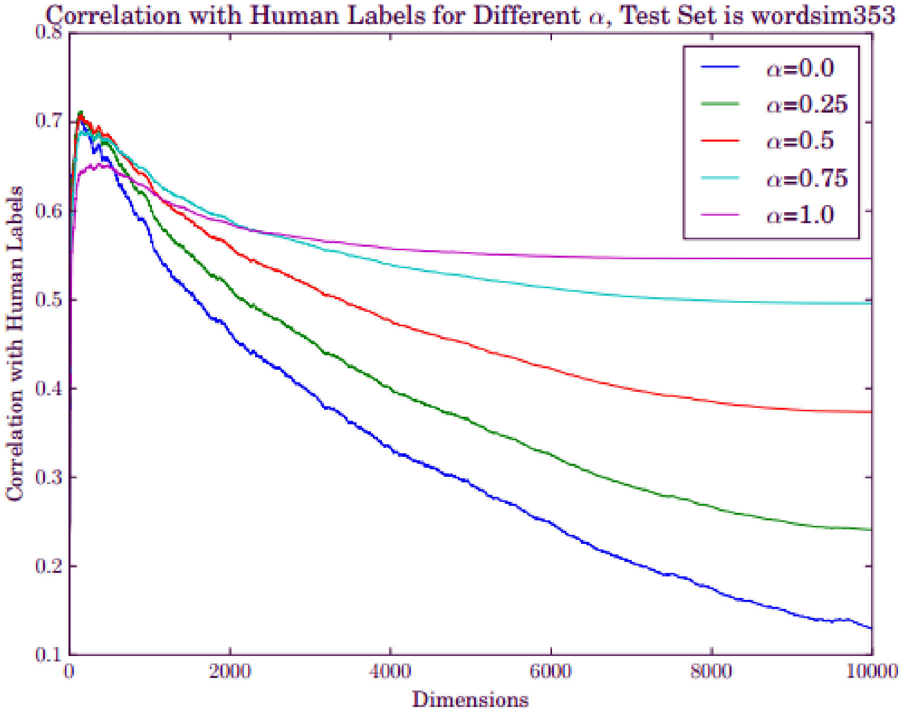
$$\mathbb{E}[\|EE^T - \hat{E}\hat{E}^T\|] \leq \sqrt{\sum_{i=k+1}^{d} \lambda_i^{4\alpha}} + \underbrace{2\sqrt{2n}\alpha\sigma\sqrt{\sum_{i=1}^{k} \lambda_i^{4\alpha-2}}}_{} + \sqrt{2}\sum_{i=1}^{k}(\lambda_i^{2\alpha} - \lambda_{i+1}^{2\alpha})\sigma\sqrt{\sum_{r \leq i < s} (\lambda_r - \lambda_s)^{-2}}$$

For small λ_k, (i.e. $\lambda_k < 1$), the rate $\lambda_k^{2\alpha-1}$ increases as α decreases: when $\alpha < 0.5$, this rate can be very large; When $0.5 \leq \alpha < 1$, the rate is bounded and sub-linear, in which case the PIP loss will be robust to over-parametrization.

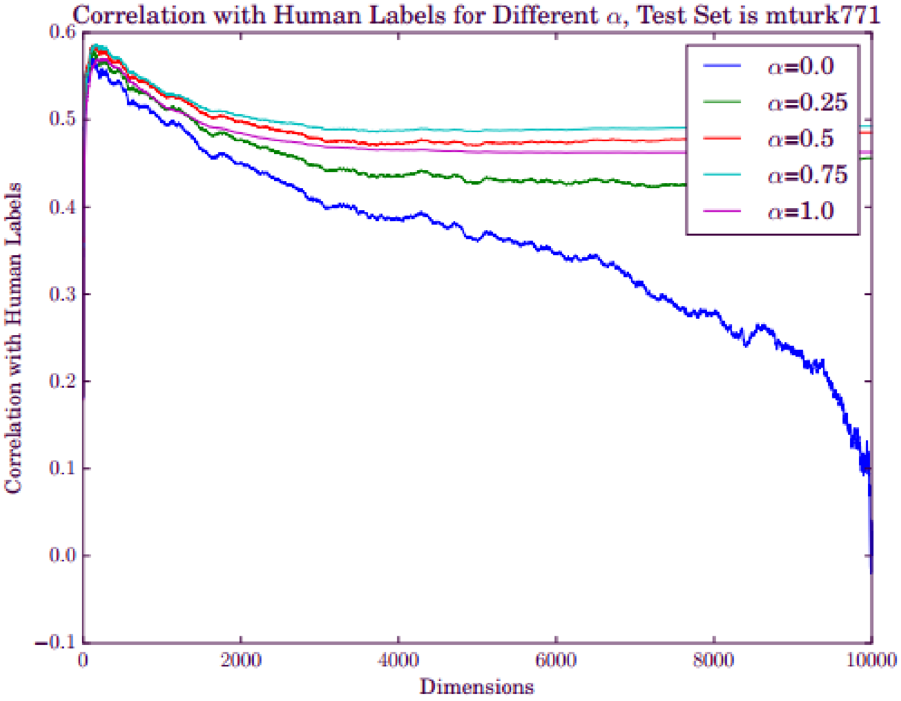


In other words, as α becomes larger, the embedding algorithm becomes **less sensitive** to over-fitting caused by the selection of an excessively large dimensionality k.

Two New Discoveries I



（a） WordSim 353 Test

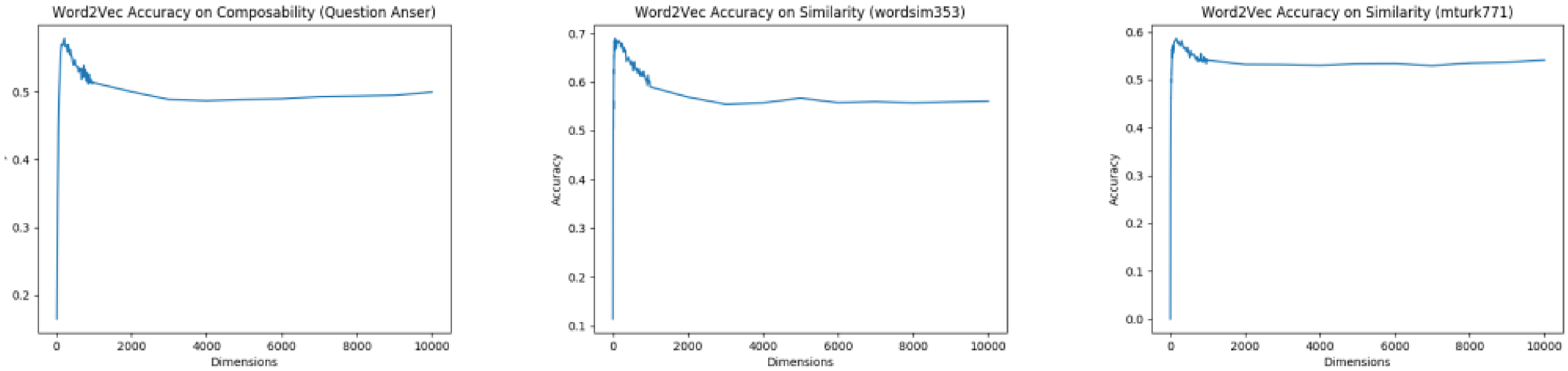


（b） Mturk771 Test

（a） and （b） display the performances (measured by the correlation between vector cosine similarity and human labels) of word embeddings of various dimensionalities from the PPMI LSA algorithm, evaluated on two word correlation tests

Two New Discoveries I

For the popular **skip-gram** [Mikolov et al., 2013b] and GloVe [Pennington et al., 2014], α equals 0.5 as they are implicitly doing a symmetric factorization, and that they are **robust to over-parametrization.**



(a) Google Analogy Test

(b) WordSim353 Test

(c) MTurk771 Test

Even with extreme over-parametrization (up to k = 10000), **skip-gram** still performs within 80% to 90% of optimal performance, for both analogy test [Mikolov et al., 2013a] and relatedness tests (WordSim353 [Finkelstein et al., 2001] and MTurk771 [Halawi et al., 2012]).

Noise Estimation :

1. Randomly **split the data** into two equally large subsets,

2. Get matrices $\widetilde{M}_1 = M_1 + Z_1, \widetilde{M}_2 = M_2 + Z_2,$ in $R_{m\times n}$, where Z_1, Z_2 are two independent copies of noise with variance $2\sigma^2$.

3. $\widetilde{M}_1 = M_1 + Z_1, \widetilde{M}_2 = M_2 + Z_2$ is a random matrix with zero mean and variance $4\sigma^2$. **The sample standard deviation** estimator is :

$$\hat{\sigma} = \frac{1}{2\sqrt{mn}} \left\| \widetilde{M}_1 - \widetilde{M}_2 \right\|$$

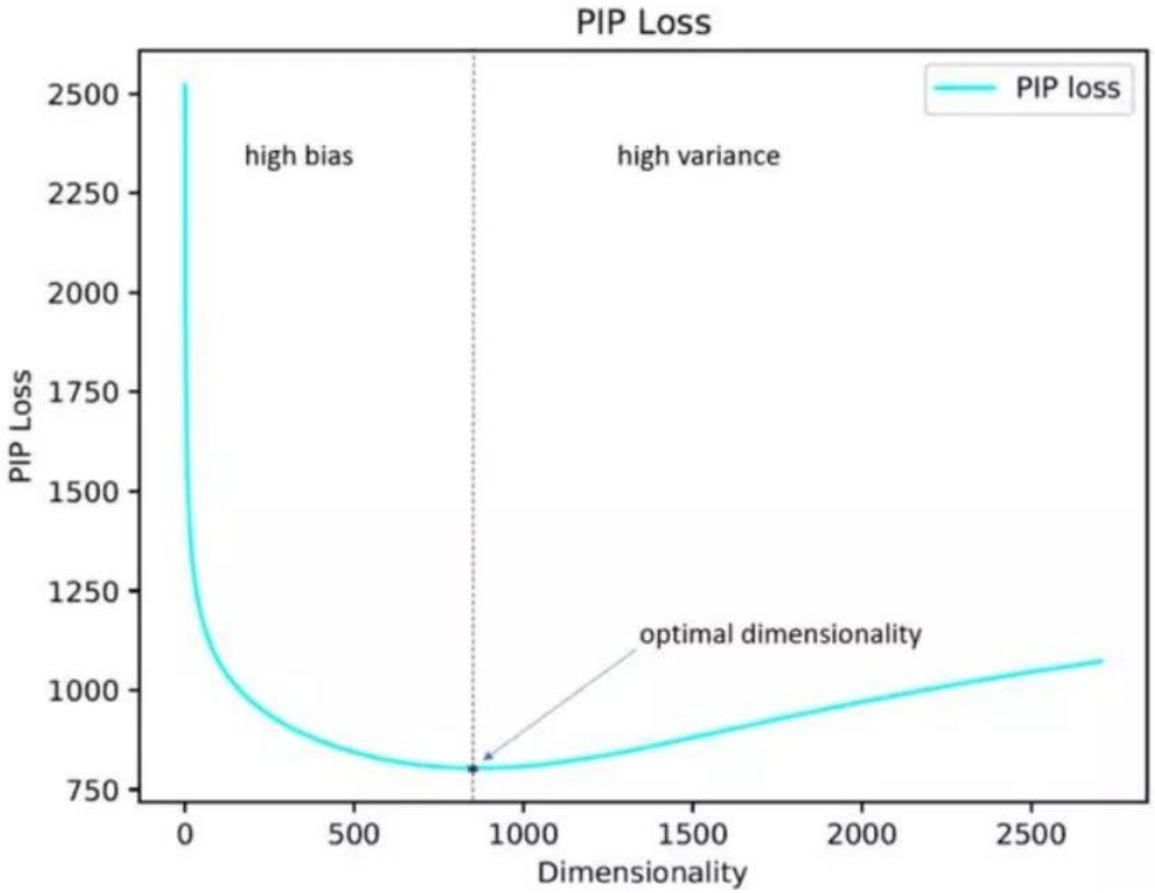
Spectral Estimation

For our experiments, we use the well- established universal singular value thresholding (USVT) proposed by Chatterjee [2015].

$$\widehat{\lambda}_i = (\widehat{\lambda}_i - 2\sigma\sqrt{n})_+$$

By the **deviation and variance's trade-off** , we can directly solve the theoretically optimal dimension.

It is obvious that deviation and variance exist. And note that the deviation and variance are around 700 dimensions



The figure shows the PIP loss function graph of Glo Ve algorithm on Text8 data set.

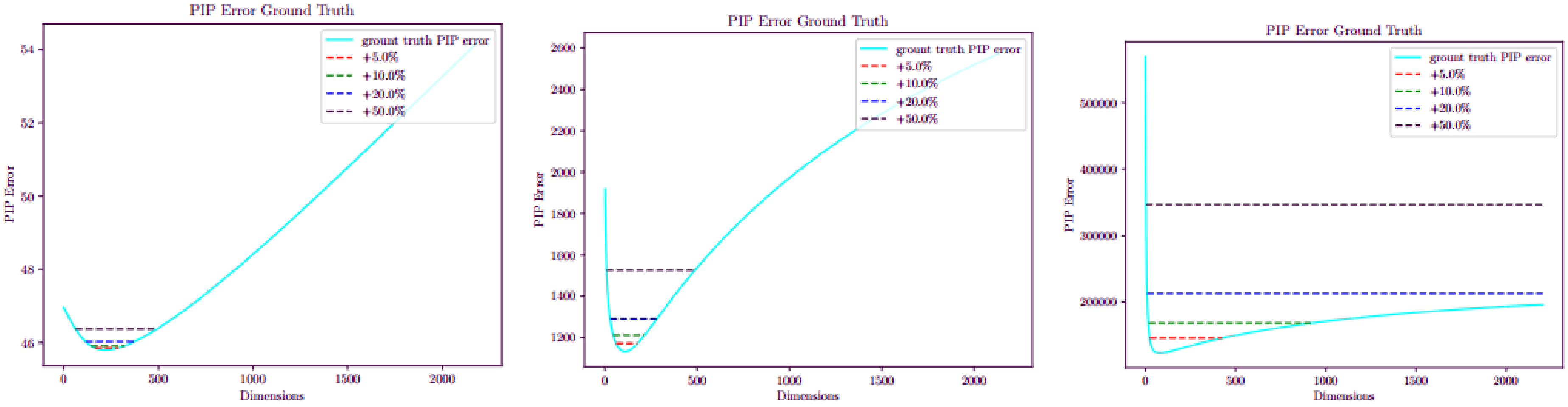
Define the sub-optimality of a particular dimensionality k as the additional PIP loss compared with $k*$:

$$\left\|E_k {E_k}^T - EE^T\right\| - \left\|E_{k*} {E_{k*}}^T - EE^T\right\|$$

And the p% sub-optimal interval as the interval of dimensionalities whose sub-optimality are no more than p% of that of a 1-D embedding.

In other words, if k is within the p% interval, then the PIP loss of a k -dimensional embedding is at most p% worse than the optimal embedding.

Two New Discoveries II



(a) $\alpha = 0$

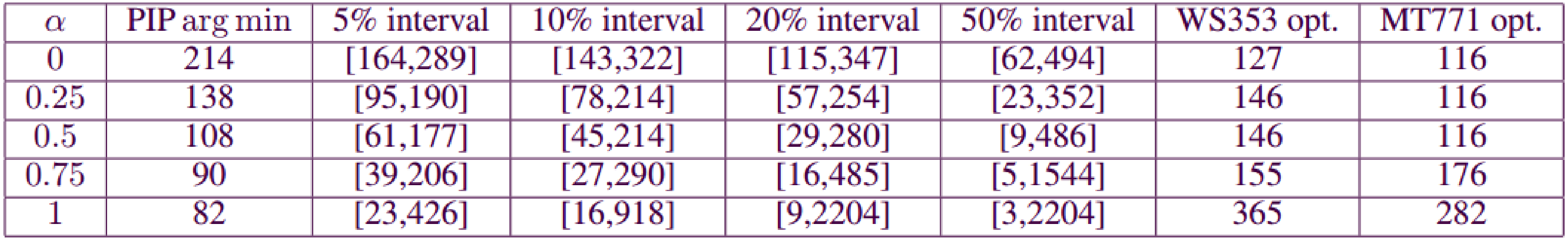
(b) $\alpha = 0.5$

(c) $\alpha = 1$

Figure 4: PIP loss and its bias-variance trade-off allow for explicit dimensionality selection for LSA

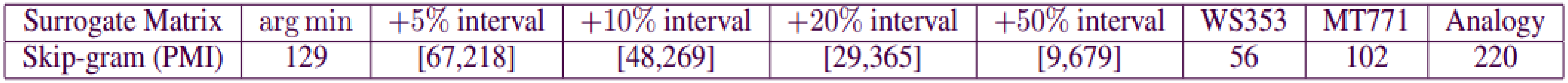
Two New Discoveries II

Table 1: Optimal dimensionalities for word relatedness tests are close to PIP loss minimizing ones



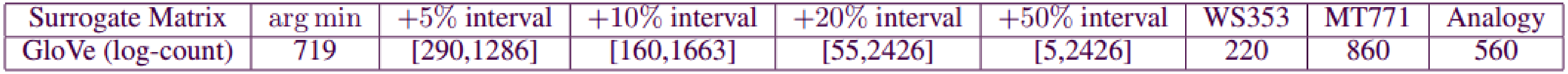
α	PIP arg min	5% interval	10% interval	20% interval	50% interval	WS353 opt.	MT771 opt.
0	214	[164,289]	[143,322]	[115,347]	[62,494]	127	116
0.25	138	[95,190]	[78,214]	[57,254]	[23,352]	146	116
0.5	108	[61,177]	[45,214]	[29,280]	[9,486]	146	116
0.75	90	[39,206]	[27,290]	[16,485]	[5,1544]	155	176
1	82	[23,426]	[16,918]	[9,2204]	[3,2204]	365	282

Table 2: PIP loss minimizing dimensionalities and intervals for Skip-gram on Text8 corpus



Surrogate Matrix	arg min	+5% interval	+10% interval	+20% interval	+50% interval	WS353	MT771	Analogy
Skip-gram (PMI)	129	[67,218]	[48,269]	[29,365]	[9,679]	56	102	220

Table 3: PIP loss minimizing dimensionalities and intervals for GloVe on Text8 corpus



Surrogate Matrix	arg min	+5% interval	+10% interval	+20% interval	+50% interval	WS353	MT771	Analogy
GloVe (log-count)	719	[290,1286]	[160,1663]	[55,2426]	[5,2426]	220	860	560

https://github.com/ziyin-dl/word-embedding-dimensionality-selection

Review:

□ PIP Loss: a Novel Unitary-invariant Loss Function for Embeddings

□ How Does Dimensionality Affect the Quality of Embedding?

□ Two New Discoveries :

- Word Embeddings' Robustness to Over-Fitting Increases with Respect to α
- Optimal Dimensionality Selection: Minimizing the PIP Loss