

THE SPECTRAL UNDERPINNING OF WORD2VEC

ARIEL JAFFE, YUVAL KLUGER, OFIR LINDENBAUM, JONATHAN PATSENKER,
EREZ PETERFREUND, AND STEFAN STEINERBERGER

ABSTRACT. Word2vec due to Mikolov *et al.* (2013) is a word embedding method that is widely used in natural language processing. Despite its great success and frequent use, theoretical justification is still lacking. The main contribution of our paper is to propose a rigorous analysis of the highly nonlinear functional of word2vec. Our results suggest that word2vec may be primarily driven by an underlying spectral method. This insight may open the door to obtaining provable guarantees for word2vec. We support these findings by numerical simulations. One fascinating open question is whether the nonlinear properties of word2vec that are not captured by the spectral method are beneficial and, if so, by what mechanism.

1. INTRODUCTION

1.1. Introduction. word2vec was introduced by Mikolov *et al.* [17] as an unsupervised scheme for embedding words based on text corpora. We will try to introduce the idea in the simplest possible terms and refer to [6, 7, 17] for the way it is usually presented. Let $\{x_1, x_2, \dots, x_n\}$ be a set of elements for which we aim to compute a numerical representation. These can be, for example, words, documents or nodes in a graph. Our input consists of a $n \times n$ matrix P with non-negative elements P_{ij} , which encode the relationship between x_i and x_j by a numerical value. The larger the value of P_{ij} , the larger the connection between x_i and x_j , for example, this could be the probability that a word appears in the same sentence as another word. Assuming a uniform prior over the n elements, the energy function $L : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ introduced by Mikolov *et. al* [17] can be written as

$$(1) \quad L(w, v) = \langle w, Pv \rangle - \sum_{i=1}^n \log \left(\sum_{j=1}^n \exp(w_i v_j) \right).$$

Word2vec is based on trying to maximize this expression over all $(w, v) \in \mathbb{R}^n \times \mathbb{R}^n$

$$(w^*, v^*) = \underset{(w, v)}{\operatorname{argmax}} L(w, v).$$

There is no reason to assume that the maximum is unique. It has been observed that if x_i and x_j are similar elements in the data set (say, words that frequently occur in the same sentence), then v_i, v_j or w_i, w_j tend to have similar numerical values. Thus, the values $\{w_1, \dots, w_n\}$ are useful for embedding $\{x_1, \dots, x_n\}$. One

Y. K. is supported in part by NIH grants R01HG008383, R01GM131642, and P50CA121974.

E. P. is partially supported by the Federmann Research Center (Hebrew University) and the Israeli Science Foundation research grant no. 1523/16. Part of the work was carried out while E.P. was visiting Yale University.

S.S. is supported by the NSF (DMS-1763179) and the Alfred P. Sloan Foundation.

could also try to maximize the symmetric loss that arises from enforcing $w = v$ and is given by $L : \mathbb{R}^n \rightarrow \mathbb{R}$

$$(2) \quad L(w) = \langle w, Pw \rangle - \sum_{i=1}^n \log \left(\sum_{j=1}^n \exp(w_i w_j) \right).$$

Here, the interpretation of the functional is somewhat more straight-forward: we wish to pick $w \in \mathbb{R}^n$ in a way that makes $\langle w, Pw \rangle$ large. If P were diagonalizable, that would mean that we want w to be a linear combination of the leading eigenvectors of P (i.e. the eigenvectors associated to the largest eigenvalues of P). At the same time the exponential function places a penalty over large entries in w . Our paper initiates a rigorous study of the energy functional $L(w)$, however, we also emphasize that a complete description of the energy landscape $L(w)$ remains an interesting open problem.

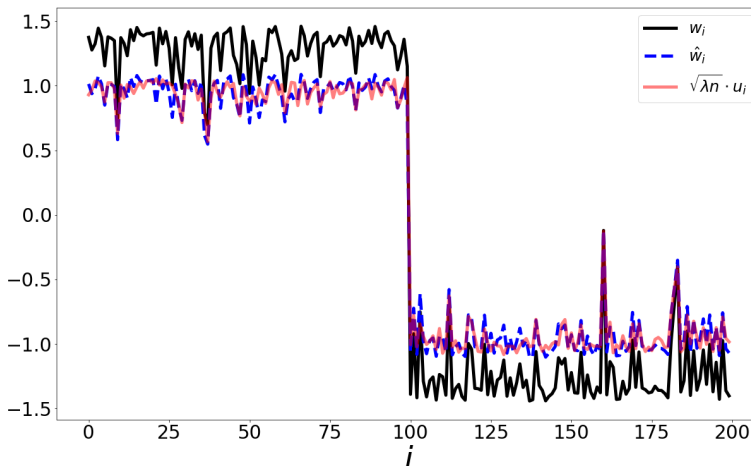


FIGURE 1. Illustration of point set drawn from two distinct Gaussian distributions. The result of maximizing over the word2vec functional (black) is closely tracked (up to scale) by the optimizer of the spectral method (blue) and the eigenvector (red).

We emphasize that our analysis has direct implications for computational aspects as well: for instance, if one was interested in maximizing the nonlinear functional, the maximum of its linear approximation (which is easy to compute) is a natural starting point. A simple example is shown in Figure 1: the underlying dataset are 200 points in \mathbb{R}^{10} where the first 100 points are drawn from a Gaussian distribution and the second 100 points are drawn from a second Gaussian distribution. The matrix P is the row-stochastic matrix induced by a Gaussian kernel $K_{ij} = \exp(-\|x_i - x_j\|^2/\alpha)$ where α is a scaling parameter. We observe that, up to scaling, the maximizer of the energy functional (black) is well approximated by the spectral methods introduced below.

1.2. Related works. Optimizing over energy functions such as (1) to obtain vector embeddings is done for various applications, such as documents [12], words [18] and graphs [19]. Surprisingly, very few works addressed analytic aspects of optimizing over the word2vec functional (1). Hashimoto et. al. [8] derived a relation between word2vec and stochastic neighbor embedding [9]. Cotterell et. al. [5] showed that when P is sampled according to a multinomial distribution, optimizing over (1) is equivalent to exponential family PCA [4]. If the number of elements is large, optimizing over (1) becomes impractical. As an efficient alternative, Mikolov et. al. [18] suggested a variation based on negative sampling. Levy and Goldberg [14] showed that if the embedding dimension is sufficiently high, then optimizing over the negative sampling functional suggested in [18] is equivalent to factorizing the shifted Pointwise Mutual Information matrix. This work was extended in [22], where similar results were derived for additional embedding algorithm such as [7, 21, 23]. Decomposition of the PMI matrix was also justified by Arora et. al. [1], based on a generative random walk model. A different approach was introduced by Landgraf [11], that related the negative sampling loss function to logistic PCA. In this work we approximate the highly nonlinear word2vec functional by Taylor expansion. This approach gives a natural connection between word2vec and classical, spectral dimensionality reduction methods such as [2, 3].

2. RESULTS

We now state our main results. In §2.1. we establish that the energy functional $L(v, w)$ has a nice asymptotic expansion around $(v, w) = (0, 0) \in \mathbb{R}^n \times \mathbb{R}^n$ and corresponds naturally to a spectral method in that regime. Naturally, such an asymptotic expansion is only feasible if one has some control over the size of the entries of the extremizer. We establish in §2.2. that the vectors maximizing the functional are not too large. The results in §2.2. are closely matched by numerical results: in particular, we observe that $\|w\| \sim \sqrt{n}$ in practice, a logarithmic factor smaller than our upper bound. The proofs are given in §3 and explicit numerical examples are shown in §4.

2.1. First order approximation for small data. The main idea is simple: we make an ansatz assuming that the optimal vectors are roughly of size $\|w\|, \|v\| \sim 1$. If we assume that the vectors w, v are fairly ‘typical’ vectors of size ~ 1 , then each individual entry can be expected to be at approximate scale $\sim n^{-1/2}$. Our main observation is that this regime is governed by a regularized spectral method.

Theorem 2.1 (Spectral Expansion). *If $\|v\|_\infty, \|w\|_\infty \lesssim n^{-1/2}$, then*

$$L(w, v) = \langle w, Pv \rangle - \frac{1}{n} \left(\sum_{i=1}^n w_i \right) \left(\sum_{j=1}^n v_j \right) - \frac{1}{n} \sum_{i,j=1}^n \frac{w_i^2 v_j^2}{2} - n \log n + \mathcal{O}(n^{-1}).$$

Naturally, since we are interested in maximizing this quantity, the constant factor $n \log n$ plays no role. The leading terms can be rewritten as

$$\langle w, Pv \rangle - \frac{1}{n} \left(\sum_{i=1}^n w_i \right) \left(\sum_{j=1}^n v_j \right) = \left\langle w, \left(P - \frac{1}{n} \mathbf{1} \right) v \right\rangle,$$

where $\mathbf{1}$ is the matrix all of whose entries are 1. This suggests that the optimal v, w maximizing the quantity should simply be the singular vectors associated to the

matrix $P - \frac{1}{n}\mathbf{1}$. The full expansion has a quadratic term that serves as an additional regularizer. The symmetric case (with ansatz $v = w$) is particularly simple, since we have

$$L(w) = \langle w, Pw \rangle - \frac{1}{n} \left(\sum_{i=1}^n w_i \right)^2 - \frac{\|w\|^4}{2n} - n \log n + \mathcal{O}(n^{-1}).$$

Assuming P is similar to a symmetric matrix, the optimal w should be well described by the leading eigenvector of $(P - \frac{1}{n}\mathbf{1})$ with an additional regularization term ensuring that $\|w\|$ is not too large. We consider this simple insight to be the main contribution of this paper since it explains succinctly why an algorithm like word2vec has a chance to be successful. We also give a large number of numerical examples showing that in many cases the result obtained by word2vec is extremely similar to what we obtain from the associated spectral method. By rephrasing word2vec as a spectral method in the ‘small vector limit’, one gains access to a large number of tools that allow to rigorously establish for which framework word2vec has a chance of coming with provable guarantees. We have not pursued this line of reasoning here since rigorous bounds for spectral methods are, nowadays, classical.

2.2. Optimal vectors are not too large. Another basic question is as follows: how big is the norm of the vector(s) maximizing the energy function? This is of obvious importance in practice, however, as seen in Theorem 2.1, it also has some theoretical relevance: if w has large entries, then clearly one cannot hope to capture the exponential nonlinearity with a polynomial expansion. Assuming $\|P\| \leq 1$, the global maximizer w^* of the second-order approximation

$$(3) \quad L_2(w) = \langle w, Pw \rangle - \frac{1}{n} \left(\sum_{i=1}^n w_i \right)^2 - \frac{\|w\|^4}{2n} - n \log n,$$

satisfies

$$\|w^*\| \leq \sqrt{2n}.$$

This can be seen as follows: if $\|P\| \leq 1$, then $\langle w, Pw \rangle \leq \|w\|^2$. Plugging in $w = 0$ shows that the maximal energy is at least size $-n \log n$. For any vector exceeding $\sqrt{2n}$ in size, we see that the energy is less than that establishing the bound. We obtain similar boundedness properties for the fully nonlinear problem for a fairly general class of matrices.

Theorem 2.2 (Generic Boundedness.). *Let $P \in \mathbb{R}^{n \times n}$ satisfy $\|P\| < 1$. Then*

$$w = \arg \max_w \langle w, Pw \rangle - \sum_{i=1}^n \log \left(\sum_{j=1}^n \exp(w_i w_j) \right),$$

satisfies

$$\|w\|^2 \leq \frac{n \log n}{1 - \|P\|}.$$

We do not claim that this bound is sharp but it does nicely illustrate that solutions of the optimization problem are necessarily bounded. Moreover, if they are bounded, then so are their entries; more precisely, $\|w\|^2 \lesssim n$ implies that, for ‘flat’ vectors, the typical entry is of size $\lesssim 1$ and thus firmly within the approximations that can be reached by a Taylor expansion. It is clear that some condition like

$\|P\| < 1$ is required for boundedness of solutions. This can be seen by considering the row-stochastic matrix

$$P = \begin{pmatrix} 1 - \varepsilon & \varepsilon \\ \varepsilon & 1 - \varepsilon \end{pmatrix}.$$

Writing $w = (w_1, w_2)$, we observe that the arising functional is quite nonlinear even in this simple case. However, it is fairly easy to understand the behavior of the gradient ascent method on the w_1 -axis since

$$\frac{\partial}{\partial w_1} L(w_1, w_2) \Big|_{w_2=0} = 2w_1 \left(1 - \varepsilon - \frac{e^{w_1^2}}{1 + e^{w_1^2}} \right),$$

which is monotonically increasing until $w_1 \sim \pm\sqrt{\log \varepsilon^{-1}}$ and therefore, a priori, unbounded since ε can be arbitrarily close to 0.

In practice, one often uses the method for matrices whose spectral norm is $\|P\| = 1$ and which have the additional property of being row-stochastic. We also observe empirically that the global optimizer w^* has a mean value close to 0 (and the expansion in Theorem 2.1. suggests why this would be the case). We obtain a similar boundedness theorem in which the only relevant operator norm is that of the operator restricted to the subspace of vectors having mean 0.

Theorem 2.3 (Boundedness for row-stochastic matrices). *Let $P \in \mathbb{R}^{n \times n}$ be a row-stochastic matrix and let*

$$P_S : \{w \in \mathbb{R}^n : w_1 + \dots + w_n = 0\} \rightarrow \mathbb{R}^n,$$

denote the restriction of P to that subspace and suppose that $\|P_S\| < 1$. Let

$$w = \arg \max_w \langle w, Pw \rangle - \sum_{i=1}^n \log \left(\sum_{j=1}^n \exp(w_i w_j) \right).$$

If w has a mean value sufficiently close to 0,

$$\left| \left\langle w, \frac{\mathbf{1}}{\sqrt{n}} \right\rangle \right| \leq \frac{1 - \|P_S\|}{3} \|w\|,$$

where $\mathbf{1} = (1, 1, \dots, 1)$, then

$$\|w\|^2 \leq \frac{2n \log n}{1 - \|P_S\|}.$$

The 2×2 matrix given above illustrates that some restrictions are necessary at least to obtain a nicely bounded gradient ascent. There is some freedom in the choice of the constants in Theorem 2.3. Numerical experiments show that the results is not merely theoretical: extremizing vectors tend to have a mean value sufficiently close to 0 for the theorem to be applicable.

2.3. Outlook. Summarizing, our main arguments are as follows:

- (1) the energy landscape of the word2vec functional is well approximated by a spectral method (or regularized spectral method) as long as the entries of the vector are uniformly bounded. In any compact interval around 0, the behavior of the exponential function can be appropriately approximated by a Taylor expansion of sufficiently high degree.

- (2) Moreover, there are a priori estimates suggesting that the entries of extremizing vectors do not grow quickly; our bounds imply that, for ‘flat’ vectors, the individual entries grow at most like $\sqrt{\log n}$ and presumably this is an artifact of the proof.
- (3) Finally, we present examples in §4 showing that in many cases the embedding obtained by maximizing the word2vec functional are indeed accurately predicted by the second order approximation.

This suggests various interesting lines of research: it would be nice to have refined versions of Theorem 2.2. and Theorem 2.3. (an immediate goal being the removal of the logarithmic dependence and perhaps even pointwise bounds on the entries of w). Numerical experiments indicate that Theorem 2.2 and Theorem 2.3. are at most a logarithmic factor away from being optimal. A second natural avenue of research proposed by our paper is to differentiate the behavior of word2vec and that of the associated spectral method: are the results of word2vec (being intrinsically nonlinear) truly different from the behavior of the spectral method (arising as its linearization)? Or, put differently, is the nonlinear aspect of word2vec that is *not* being captured by the spectral method helpful for embedding methods?

3. PROOFS

3.1. Proof of Theorem 2.1.

Proof. We recall our assumption of $\|w\|_\infty \lesssim n^{-1/2}$ and $\|v\|_\infty \lesssim n^{-1/2}$ (where the implicit constant affects all subsequent constants). We remark that the subsequent arguments could also be carried out for any $\|w\|_\infty, \|v\|_\infty \lesssim n^{-\varepsilon}$ at the cost of different error terms; the arguments fail being rigorous as soon as $\|w\|_\infty \sim 1$ since then, a priori, all terms in the Taylor expansion of e^x could be of roughly the same size. We start with the Taylor expansion

$$\begin{aligned} \sum_{j=1}^n e^{w_i v_j} &= \sum_{j=1}^n \left(1 + w_i v_j + \frac{w_i^2 v_j^2}{2} + \mathcal{O}(n^{-3}) \right) \\ &= n + \sum_{j=1}^n \left(w_i v_j + \frac{w_i^2 v_j^2}{2} \right) + \mathcal{O}(n^{-2}). \end{aligned}$$

In particular, we note that

$$\left| \sum_{j=1}^n \left(w_i v_j + \frac{w_i^2 v_j^2}{2} \right) \right| \lesssim 1.$$

We use the series expansion

$$\log(n+x) = \log n + \frac{x}{n} - \frac{x^2}{2n^2} + \mathcal{O}\left(\frac{|x|^3}{n^3}\right)$$

to obtain

$$\begin{aligned} \log \left(\sum_{j=1}^n e^{w_i v_j} \right) &= \log n + \frac{1}{n} \sum_{j=1}^n \left(w_i v_j + \frac{w_i^2 v_j^2}{2} \right) \\ &\quad - \frac{1}{2n^2} \left(\sum_{j=1}^n w_i v_j + \frac{w_i^2 v_j^2}{2} \right)^2 + \mathcal{O}(n^{-3}). \end{aligned}$$

Here, the second sum can be somewhat simplified since

$$\begin{aligned} \frac{1}{2n^2} \left(\sum_{j=1}^n w_i v_j + \frac{w_i^2 v_j^2}{2} \right)^2 &= \frac{1}{2n^2} \left(\sum_{j=1}^n (w_i v_j + \mathcal{O}(n^{-2})) \right)^2 \\ &= \frac{1}{2n^2} \left(\mathcal{O}(n^{-1}) + \sum_{j=1}^n w_i v_j \right)^2 \\ &= \frac{1}{2n^2} \left(\sum_{j=1}^n w_i v_j \right)^2 + \mathcal{O}(n^{-3}) \\ &= \frac{w_i^2}{2n^2} \left(\sum_{j=1}^n v_j \right)^2 + \mathcal{O}(n^{-3}) \end{aligned}$$

Altogether, we obtain that

$$\begin{aligned} \sum_{i=1}^n \log \left(\sum_{j=1}^n e^{w_i v_j} \right) &= \sum_{i=1}^n \left(\log n + \frac{1}{n} \sum_{j=1}^n \left(w_i v_j + \frac{w_i^2 v_j^2}{2} \right) - \frac{w_i^2}{2n^2} \left(\sum_{j=1}^n v_j \right)^2 + \mathcal{O}(n^{-3}) \right) \\ &= n \log n + \frac{1}{n} \sum_{i,j=1}^n w_i v_j + \frac{1}{n} \sum_{i,j=1}^n \frac{w_i^2 v_j^2}{2} \\ &\quad - \frac{1}{n^2} \left(\sum_{i=1}^n \frac{w_i^2}{2} \right) \left(\sum_{j=1}^n v_j \right)^2 + \mathcal{O}(n^{-2}). \end{aligned}$$

Since $\|w\|_\infty, \|v\|_\infty \lesssim n^{-1/2}$, we have

$$\frac{1}{n^2} \left(\sum_{i=1}^n \frac{w_i^2}{2} \right) \left(\sum_{j=1}^n v_j \right)^2 \lesssim n^{-1}$$

and have justified the desired expansion. \square

Proof of Theorem 2.2. Setting $w = 0$ results in the energy

$$L(w) = -n \log n.$$

Let now w be a global maximizer, then we obtain

$$\begin{aligned} -n \log n &\leq \langle w, Pw \rangle - \sum_{i=1}^n \log \left(\sum_{j=1}^n e^{w_i w_j} \right) \\ &\leq \|P\| \|w\|^2 - \sum_{i=1}^n \log \left(e^{w_i^2} \right) \leq (\|P\| - 1) \|w\|^2 \end{aligned}$$

which is the desired result. \square

Proof of Theorem 2.3. We expand the vector w into a multiple of the constant vector of norm 1, the vector

$$\frac{\mathbf{1}}{\sqrt{n}} = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right),$$

and the orthogonal complement via

$$w = \left\langle w, \frac{\mathbf{1}}{\sqrt{n}} \right\rangle \frac{\mathbf{1}}{\sqrt{n}} + \left(w - \left\langle w, \frac{\mathbf{1}}{\sqrt{n}} \right\rangle \frac{\mathbf{1}}{\sqrt{n}} \right),$$

which we abbreviate as $w = \tilde{w} + (w - \tilde{w})$. We expand

$$\langle w, Pw \rangle = \langle \tilde{w}, P\tilde{w} \rangle + \langle \tilde{w}, P(w - \tilde{w}) \rangle + \langle w - \tilde{w}, P\tilde{w} \rangle + \langle w - \tilde{w}, P(w - \tilde{w}) \rangle.$$

Since P is row-stochastic, we have $P\tilde{w} = \tilde{w}$ and thus $\langle \tilde{w}, P\tilde{w} \rangle = \|\tilde{w}\|^2$. Moreover, we have

$$\langle w - \tilde{w}, P\tilde{w} \rangle = \langle w - \tilde{w}, \tilde{w} \rangle = 0$$

since $w - \tilde{w}$ has mean value 0. We also observe, again because $w - \tilde{w}$ has mean value 0, that

$$\langle \tilde{w}, P(w - \tilde{w}) \rangle = \langle \tilde{w}, P_S(w - \tilde{w}) \rangle.$$

Collecting all these estimates, we obtain

$$\frac{\langle w, Pw \rangle}{\|w\|^2} \leq \frac{\|\tilde{w}\|^2}{\|w\|^2} + \frac{\|\tilde{w}\| \|w - \tilde{w}\|}{\|w\| \|w\|} \|P_S\| + \frac{\|w - \tilde{w}\|^2}{\|w\|^2} \|P_S\|.$$

We also recall the Pythagorean theorem

$$\|\tilde{w}\|^2 + \|w - \tilde{w}\|^2 = \|w\|^2.$$

Abbreviating $x = \|\tilde{w}\|/\|w\|$, we can abbreviate our upper bound as

$$\frac{\langle w, Pw \rangle}{\|w\|^2} \leq x^2 + x\sqrt{1-x^2} \|P_S\| + (1-x^2) \|P_S\|.$$

The function

$$x \rightarrow x\sqrt{1-x^2} + (1-x^2)$$

is monotonically increasing on $[0, 1/3]$. Thus, assuming that

$$x = \frac{\|\tilde{w}\|}{\|w\|} \leq \frac{1 - \|P_S\|}{3},$$

we get, after some elementary computation,

$$\begin{aligned} \frac{\langle w, Pw \rangle}{\|w\|^2} &\leq \left(\frac{1 - \|P_S\|}{9} \right)^2 + \frac{1 - \|P_S\|}{9} \sqrt{1 - \left(\frac{1 - \|P_S\|}{9} \right)^2} \|P_S\| \\ &+ \left(1 - \left(\frac{1 - \|P_S\|}{9} \right)^2 \right) \|P_S\| \leq 0.2 + 0.8\|P_S\|. \end{aligned}$$

However, we also recall from the proof of Theorem 2.2. that

$$\sum_{i=1}^n -\log \left(\sum_{j=1}^n e^{w_i w_j} \right) \leq -\|w\|^2.$$

Altogether, since the energy in the maximum has to exceed the energy in the origin, we have

$$-n \log n \leq \langle w, Pw \rangle - \sum_{i=1}^n \log \left(\sum_{j=1}^n e^{w_i w_j} \right) \leq (0.2 + 0.8\|P_S\|) \|w\|^2 - \|w\|^2$$

and therefore

$$\|w\|^2 \leq \frac{2n \log n}{1 - \|P_S\|}.$$

□

4. EXAMPLES

We validate our theoretical findings by comparing, for various datasets, the representation obtained by the following methods: (i) optimizing over the symmetric functional in Eq. (1), (ii) optimizing over the spectral method suggested by Theorem 2.1 and (iii) computing the leading eigenvector of $P - \frac{1}{n}\mathbf{1}$. We denote by w , \hat{w} and u be the three vectors obtained by (i)-(iii) respectively. The comparison is performed for two artificial datasets, two sets of images, a seismic dataset and a text corpus. For the artificial, image and seismic data, the matrix P is obtained by the following steps: we compute a pairwise kernel matrix

$$K(x_i, x_j) = \exp \left(- \frac{\|x_i - x_j\|^2}{\alpha} \right),$$

where α is a scale parameter set as in [16]. We compute P via

$$P_{ij} = K_{ij} / \sum_{l=1}^N K_{il}.$$

The matrix P can be interpreted as a random walk over the data points, see for example [3]. To support our approximation in Theorem 2.1, we compute the correlation coefficient between w and \hat{w} by

$$\rho(w, \hat{w}) = \frac{(w - \mu)^T (\hat{w} - \hat{\mu})}{\|w\| \|\hat{w}\|},$$

where μ and $\hat{\mu}$ are the means of w and \hat{w} respectively. A similar measure is done for w and u . In addition, we compute the norm $\|w\|$ and compare it to the upper bound in Theorem 2.3.

4.1. Noisy Circle. Here, the elements $\{x_1, \dots, x_{200} \in \mathbb{R}^2\}$ are generated by adding Gaussian noise with mean 0 and $\sigma^2 = 0.1$ to a unit circle, see left panel of Figure 2. The right panel shows the extracted representations w , \hat{w} along with the leading eigenvector u scaled by $\sqrt{\lambda n}$ where λ is the corresponding eigenvalue. The correlation coefficients $\rho(w, u)$ and $\rho(\hat{w}, u)$ are equal to 0.98, 0.99 respectively.

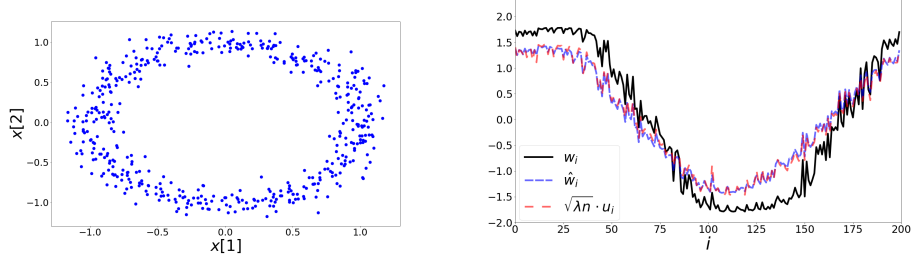


FIGURE 2. Left: 200 elements on the noisy circle data set. Points are generated by adding noise drawn from a two dimensional Gaussian with zero mean and a variance of 0.1. Right: The extracted representations based on the symmetric loss w , second order approximation \hat{w} and leading eigenvector u .

4.2. MNIST. Next, we use a set of 300 images of digits 3 and 4 from the MNIST dataset [13]. Two examples from each category are presented in the left panel of Figure 3. Here, the extracted representations w and \hat{w} match the values of the scaled eigenvector u (see right panel of Figure 3). The correlation coefficients $\rho(w, u)$ and $\rho(\hat{w}, u)$ are both higher than 0.999.

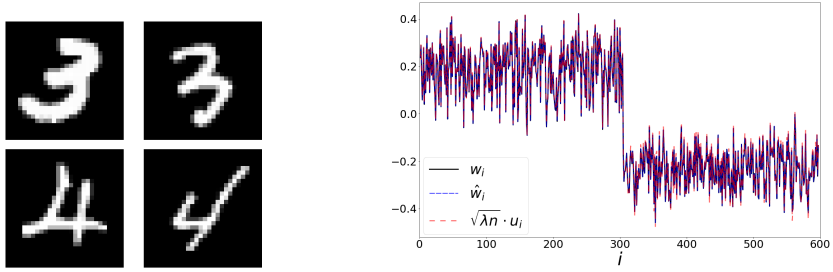


FIGURE 3. Left: handwritten digits from the MNIST dataset. Right: The extracted representations w , \hat{w} and $\sqrt{\lambda n}u$, the leading eigenvector of $P - \frac{1}{n}\mathbf{1}$.

4.3. COIL100. In this example, we use images from Columbia Object Image Library (COIL100) [20]. Our dataset contains 21 images of a cat captured at several pose intervals of 5 degrees, see left panel of Figure 4. We extract the embedding w and \hat{w} and reorder them based on the true angle of the cat at every image. In the right panel, we present the values of the reorders representations w , \hat{w} and u overlayed with the corresponding objects. The values of all representations are strongly correlated with the angle of the object. Moreover, the correlation coefficients $\rho(w, u)$ and $\rho(\hat{w}, u)$, are 0.97 and 0.99 respectively.

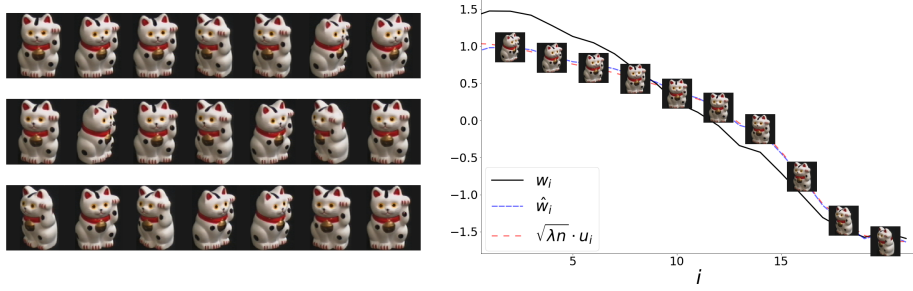


FIGURE 4. Left: 21 samples from COIL100 dataset. The object is captured at several, unorganized angles. Right: The sorted values of the representations w, \hat{w} and u , along with the corresponding object. Here, the representation correlates with the angle of the object.

4.4. Seismic Data. Seismic recordings could be useful for identifying properties of geophysical events.

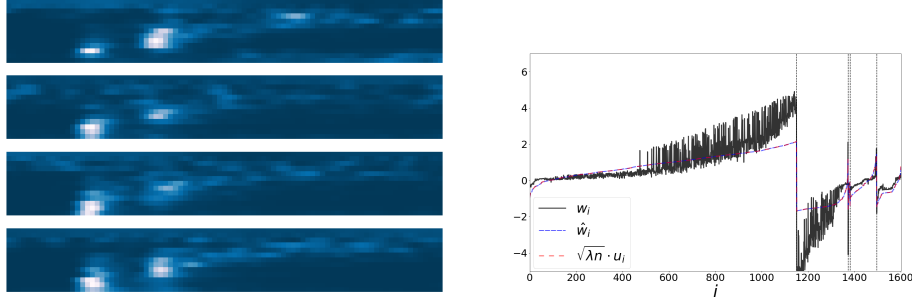


FIGURE 5. Left: 4 samples from the sonogram dataset, of different event types. Right: The values of the representations w, \hat{w} and u . Dashed lines annotate the different categories of the events (based on event type and quarry location). Within each category the representations are ordered based on the value of u .

We use a dataset collected in Israel and Jordan, described in [15]. The data consists of 1632 seismic recordings of earthquakes and explosions from quarries.

Each recording is described by a sonogram with 13 frequency bins, and 89 time bins [10]. See the left panel of Figure 5. Events could be categorized into 5 groups using manual annotations of their origin. We flatten each sonogram into a vector, and extract embeddings w , \hat{w} , and u . In the right panel of this figure, we show the extracted representations of all events. We use dashed lines to annotate the different categories and sort the values within each category based on u . The coefficient $\rho(w, v)$ is equal to 0.89, and $\rho(\hat{w}, v) = 1$.

4.5. Text Data. As a final evaluation we use a corpus of words from the book “Alice in Wonderland”. To define a co-occurrence matrix, we scan the sentences using a window size covering 5 neighbors before and after each word. We subsample the top 1000 words in terms of occurrences in the book. The matrix P is then defined by normalizing the co-occurrence matrix. In Figure 6 we present centered and normalized versions of the representations w , \hat{w} and the leading left singular vector v of $P - \frac{1}{n}\mathbf{1}$. The coefficient $\rho(w, v)$ is equal to 0.77, and $\rho(\hat{w}, v) = 1$.

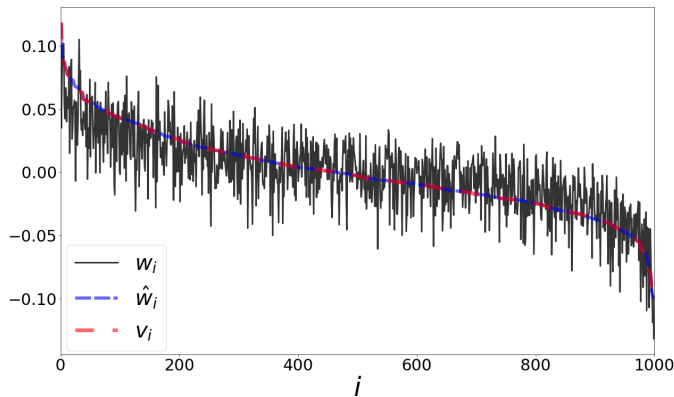


FIGURE 6. Word representation based on “Alice in Wonderland”. The values of the representations w, \hat{w} and v , sorted based on the singular vector v . We normalized all representations to unit norm.

REFERENCES

- [1] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *ArXiv*, abs/1502.03520, 2015.
- [2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
- [3] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.
- [4] Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624, 2002.
- [5] Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. Explaining and generalizing skip-gram through exponential family principal component analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 175–181, 2017.

- [6] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.
- [7] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
- [8] Tatsunori B Hashimoto, David Alvarez-Melis, and Tommi S Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286, 2016.
- [9] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.
- [10] Manfred Joswig. Pattern recognition for earthquake detection. *Bulletin of the Seismological Society of America*, 80(1):170–186, 1990.
- [11] Andrew J Landgraf and Jeremy Bellay. Word2vec skip-gram with negative sampling is a weighted logistic pca. *arXiv preprint arXiv:1705.09755*, 2017.
- [12] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.
- [13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [14] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.
- [15] Ofir Lindenbaum, Yuri Bregman, Neta Rabin, and Amir Averbuch. Multiview kernels for low-dimensional modeling of seismic events. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3300–3310, 2018.
- [16] Ofir Lindenbaum, Moshe Salhov, Arie Yeredor, and Amir Averbuch. Kernel scaling for manifold learning and classification. *arXiv preprint arXiv:1707.01093*, 2017.
- [17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*, 2013.
- [18] 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, 2013.
- [19] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017.
- [20] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). 1996.
- [21] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.
- [22] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 459–467, 2018.
- [23] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015.

PROGRAM IN APPLIED MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, USA
E-mail address: `ariel.jaffe@yale.edu`

DEPARTMENT OF PATHOLOGY, YALE SCHOOL OF MEDICINE, NEW HAVEN, USA
E-mail address: `yuval.kluger@yale.edu`

PROGRAM IN APPLIED MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, USA
E-mail address: `ofir.lindenbaum@yale.edu`

PROGRAM IN APPLIED MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, USA
E-mail address: `jonathan@patsenker.com`

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING, THE HEBREW UNIVERSITY, JERUSALEM,
ISRAEL
E-mail address: `erezpeter@cs.huji.ac.il`

DEPARTMENT OF MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, USA
E-mail address: `stefan.steinerberger@yale.edu`