ISSN (Print): 2077-9879
ISSN (Online): 2617-2658

Eurasian Mathematical Journal

2022, Volume 13, Number 4

Founded in 2010 by
the L.N. Gumilyov Eurasian National University
in cooperation with
the M.V. Lomonosov Moscow State University
the Peoples' Friendship University of Russia (RUDN University)
the University of Padua

Starting with 2018 co-funded
by the L.N. Gumilyov Eurasian National University
and
the Peoples' Friendship University of Russia (RUDN University)

Supported by the ISAAC
(International Society for Analysis, its Applications and Computation)
and
by the Kazakhstan Mathematical Society

Published by

the L.N. Gumilyov Eurasian National University
Astana, Kazakhstan

Aims and Scope

The Eurasian Mathematical Journal (EMJ) publishes carefully selected original research papers in all areas of mathematics written by mathematicians, principally from Europe and Asia. However papers by mathematicians from other continents are also welcome.

From time to time the EMJ publishes survey papers.

The EMJ publishes 4 issues in a year.

The language of the paper must be English only.

The contents of the EMJ are indexed in Scopus, Web of Science (ESCI), Mathematical Reviews, MathSciNet, Zentralblatt Math (ZMATH), Referativnyi Zhurnal – Matematika, Math-Net.Ru.

The EMJ is included in the list of journals recommended by the Committee for Control of Education and Science (Ministry of Education and Science of the Republic of Kazakhstan) and in the list of journals recommended by the Higher Attestation Commission (Ministry of Education and Science of the Russian Federation).

Information for the Authors

Submission. Manuscripts should be written in LaTeX and should be submitted electronically in DVI, PostScript or PDF format to the EMJ Editorial Office through the provided web interface (www.enu.kz).

When the paper is accepted, the authors will be asked to send the tex-file of the paper to the Editorial Office.

The author who submitted an article for publication will be considered as a corresponding author. Authors may nominate a member of the Editorial Board whom they consider appropriate for the article. However, assignment to that particular editor is not guaranteed.

Copyright. When the paper is accepted, the copyright is automatically transferred to the EMJ. Manuscripts are accepted for review on the understanding that the same work has not been already published (except in the form of an abstract), that it is not under consideration for publication elsewhere, and that it has been approved by all authors.

Title page. The title page should start with the title of the paper and authors' names (no degrees). It should contain the Keywords (no more than 10), the Subject Classification (AMS Mathematics Subject Classification (2010) with primary (and secondary) subject classification codes), and the Abstract (no more than 150 words with minimal use of mathematical symbols).

Figures. Figures should be prepared in a digital form which is suitable for direct reproduction.

References. Bibliographical references should be listed alphabetically at the end of the article. The authors should consult the Mathematical Reviews for the standard abbreviations of journals' names.

Authors' data. The authors' affiliations, addresses and e-mail addresses should be placed after the References.

Proofs. The authors will receive proofs only once. The late return of proofs may result in the paper being published in a later issue.

Offprints. The authors will receive offprints in electronic form.

Publication Ethics and Publication Malpractice

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the EMJ implies that the work described has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The EMJ follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/NewCode.pdf). To verify originality, your article may be checked by the originality detection service CrossCheck http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the EMJ.

The Editorial Board of the EMJ will monitor and safeguard publishing ethics.

The procedure of reviewing a manuscript, established by the Editorial Board of the Eurasian Mathematical Journal

1. **Reviewing procedure**

1.1. All research papers received by the Eurasian Mathematical Journal (EMJ) are subject to mandatory reviewing.

1.2. The Managing Editor of the journal determines whether a paper fits to the scope of the EMJ and satisfies the rules of writing papers for the EMJ, and directs it for a preliminary review to one of the Editors-in-chief who checks the scientific content of the manuscript and assigns a specialist for reviewing the manuscript.

1.3. Reviewers of manuscripts are selected from highly qualified scientists and specialists of the L.N. Gumilyov Eurasian National University (doctors of sciences, professors), other universities of the Republic of Kazakhstan and foreign countries. An author of a paper cannot be its reviewer.

1.4. Duration of reviewing in each case is determined by the Managing Editor aiming at creating conditions for the most rapid publication of the paper.

1.5. Reviewing is confidential. Information about a reviewer is anonymous to the authors and is available only for the Editorial Board and the Control Committee in the Field of Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan (CCFES). The author has the right to read the text of the review.

1.6. If required, the review is sent to the author by e-mail.

1.7. A positive review is not a sufficient basis for publication of the paper.

1.8. If a reviewer overall approves the paper, but has observations, the review is confidentially sent to the author. A revised version of the paper in which the comments of the reviewer are taken into account is sent to the same reviewer for additional reviewing.

1.9. In the case of a negative review the text of the review is confidentially sent to the author.

1.10. If the author sends a well reasoned response to the comments of the reviewer, the paper should be considered by a commission, consisting of three members of the Editorial Board.

1.11. The final decision on publication of the paper is made by the Editorial Board and is recorded in the minutes of the meeting of the Editorial Board.

1.12. After the paper is accepted for publication by the Editorial Board the Managing Editor informs the author about this and about the date of publication.

1.13. Originals reviews are stored in the Editorial Office for three years from the date of publication and are provided on request of the CCFES.

1.14. No fee for reviewing papers will be charged.

2. **Requirements for the content of a review**

2.1. In the title of a review there should be indicated the author(s) and the title of a paper.

2.2. A review should include a qualified analysis of the material of a paper, objective assessment and reasoned recommendations.

2.3. A review should cover the following topics:

- compliance of the paper with the scope of the EMJ;

- compliance of the title of the paper to its content;

- compliance of the paper to the rules of writing papers for the EMJ (abstract, key words and phrases, bibliography etc.);

- a general description and assessment of the content of the paper (subject, focus, actuality of the topic, importance and actuality of the obtained results, possible applications);

- content of the paper (the originality of the material, survey of previously published studies on the topic of the paper, erroneous statements (if any), controversial issues (if any), and so on);

- exposition of the paper (clarity, conciseness, completeness of proofs, completeness of bibliographic references, typographical quality of the text);

- possibility of reducing the volume of the paper, without harming the content and understanding of the presented scientific results;

- description of positive aspects of the paper, as well as of drawbacks, recommendations for corrections and complements to the text.

2.4. The final part of the review should contain an overall opinion of a reviewer on the paper and a clear recommendation on whether the paper can be published in the Eurasian Mathematical Journal, should be sent back to the author for revision or cannot be published.

Web-page

The web-page of the EMJ is www.emj.enu.kz. One can enter the web-page by typing Eurasian Mathematical Journal in any search engine (Google, Yandex, etc.). The archive of the web-page contains all papers published in the EMJ (free access).

Subscription

Subscription index of the EMJ 76090 via KAZPOST.

E-mail

eurasianmj@yandex.kz

The Eurasian Mathematical Journal (EMJ)
The Astana Editorial Office
The L.N. Gumilyov Eurasian National University
Building no. 3
Room 306a
Tel.: +7-7172-709500 extension 33312
13 Kazhymukan St
010008 Astana, Kazakhstan

The Moscow Editorial Office
The Peoples' Friendship University of Russia
(RUDN University)
Room 473
3 Ordzonikidze St
117198 Moscow, Russia

EURASIAN MATHEMATICAL JOURNAL
ISSN 2077-9879
Volume 13, Number 4 (2022), 70 – 81

CONVERGENCE OF THE PARTITION FUNCTION IN THE STATIC WORD EMBEDDING MODEL

K. Mynbaev, Zh. Assylbekov

Communicated by S. Sagitov

Key words: word embeddings, partition function, neural networks, WORD2VEC, asymptotic distribution.

AMS Mathematics Subject Classification: 68T50.

Abstract. We develop an asymptotic theory for the partition function of the word embeddings model WORD2VEC. The proof involves a study of properties of matrices, their determinants and distributions of random normal vectors when their dimension tends to infinity. The conditions imposed are mild enough to cover practically important situations. The implication is that for any word i from a vocabulary $\mathcal{W}$, the context vector $\mathbf{c}_i$ is a reflection of the word vector $\mathbf{w}_i$ in approximately half of the dimensions. This allows us to halve the number of trainable parameters in static word embedding models.

DOI: https://doi.org/10.32523/2077-9879-2022-13-4-70-81

1 Introduction and main results

Today, it is impossible to imagine natural language processing (NLP) without vector representations of words, which can be obtained by pre-training neural language models on large amounts of text. Early models [15, 16, 13] produced the so-called *static* word embeddings—each word from the vocabulary was mapped into one single vector, regardless of context. This leads to difficulties in the case of polysemous (having multiple meanings) words. For example, given the vector of the word *bank*, it is not clear whether we are talking about a financial institution or about a river bank. Modern models [17, 5] produce *contextualized* embeddings, i.e. they map the word together with its context into a vector. Thus, the same word in different contexts will have different vectors.

Despite the fact that the latter approach is mainstream today, static vectors remain relevant for a number of reasons:

- their models are more interpretable (as opposed to deep contextualizers),
- they can be trained much faster (in a few hours, and not in several days),
- they need much less computing power (1 GPU instead of 8–16 GPUs).

Moreover, static vectors are an integral part of (masked) language models that produce contextualized embeddings. Therefore, we believe it is important to continue studying and better understanding static vectors.

Another important advantage of static embeddings is the abundance of theoretical studies of their properties [13, 3, 10, 8, 19, 6, 2, 1, 4, 20]. For us, the starting point is the work of Zh. Assylbekov and R. Takhanov [4], in which some key assumptions are made about the nature of word vectors and text generation, and then a useful property is stated, which allows us to halve the number of trainable parameters in static word embedding models without sacrificing quality. We will provide an overview of key aspects of their article to better motivate our work. To begin with, we introduce the necessary notation.

1.1 Notation

We let $\mathbb{R}$ denote the set of all real numbers. Bold-faced lowercase letters ($\mathbf{x}$) denote vectors in the Euclidean space $\mathbb{R}^d$, bold-faced uppercase letters ($\mathbf{X}$) denote matrices, plain-faced lowercase letters (x) denote scalars, plain-faced uppercase letters (X) denote scalar random variables, 'i.i.d.' stands for 'independent and identically distributed'. We use the sign $\sim$ to abbreviate the phrase 'distributed as', and the sign $\propto$ to abbreviate 'proportional to'. $X_n \xrightarrow{p} Y$ denotes convergence of X_n to Y in probability. $\mathrm{Tr}(\mathbf{A})$ is used to denote the trace of a matrix $\mathbf{A}$. For any matrix $\mathbf{A}$ with elements a_{ij}, we denote

$$\|\mathbf{A}\|_2 = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}, \qquad \|\mathbf{A}\| = \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2\,, \qquad \|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|.$$

In particular, for a vector x, $\|x\|_2 = \left(\sum_i x_i^2\right)^{1/2}$. $\mathbf{A}_i$ stands for the i-th row of $\mathbf{A}$ and $\mathbf{A}^{(j)}$ for the j-th column of $\mathbf{A}$. For a square matrix $\mathbf{A}$, $\lambda_j(\mathbf{A})$ and $s_j(\mathbf{A})$ denote its eigenvalues and singular values, respectively, counted as many times as their multiplicities. Further,

$$\|\mathbf{A}\|_{\sigma_p} = \left(\sum_j s_j^p(\mathbf{A})\right)^{1/p} \text{ for } 1 \le p < \infty, \qquad \|\mathbf{A}\|_{\sigma_\infty} = \max_j s_j(\mathbf{A}).$$

Recall that for nonnegative symmetric matrices, $\lambda_j(\mathbf{A}) = s_j(\mathbf{A})$ for all j.

1.2 Word modelling and main conjecture

Broadly speaking, word modelling is a mapping of a vocabulary into some structure that can be analyzed using mathematical tools. The challenge is to develop a mapping that adequately describes how words interact contextually. One such model is the subject of this paper.

Assuming that words have already been converted into indices, let $\mathcal{W} := \{1, \ldots, n\}$ be a finite vocabulary of unique words. In what follows we assume that our dataset $\mathcal{D}$ consists of co-occurence pairs (i, j). We say that "the words i and j co-occur" when they co-occur in a fixed-size window of words. For instance, using a window of size 2 we can convert the text *the cat sat on the mat* into the set of pairs $\mathcal{D} = \{$(*the*, *cat*), (*cat*, *the*), (*cat*, *sat*), (*sat*, *cat*), (*sat*, *on*), (*on*, *sat*), (*on*, *the*), (*the*, *on*), (*the*, *mat*), (*mat*, *the*)$\}$.

In WORD2VEC model, the task is to learn to predict which words are most likely to be near each other in some long corpus of text. For each word *center* in the corpus, the model outputs the probability distribution $P(\mathit{context}|\mathit{center})$ of how likely each other word *context* in the vocabulary is to be within a certain number of words away from *center*. Following [15], we assume that there are *two* vectors for each word i:

- $\mathbf{w}_i \in \mathbb{R}^d$ when $i \in \mathcal{W}$ is a center word, such as the word *sat* in a window of 5 words

  the cat *sat* on the

- $\mathbf{c}_i \in \mathbb{R}^d$ when $i \in \mathcal{W}$ is a context word, such as the words *the*, *cat*, *on*, *the* in the same window of 5 words above.

Word vectors $\{\mathbf{w}_i\}$ are also known as *word embeddings*, while context vectors $\{\mathbf{c}_i\}$ are also known as *context embeddings*.

The assumptions of [4] on the nature of word vectors, context vectors, and text generation are as follows:*)

(i) A priori word vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathbb{R}^d$ are i.i.d. draws from an isotropic multivariate Gaussian distribution: $\mathbf{w}_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{1}{d}\mathbf{I}\right)$, where $\mathbf{I}$ is the $d \times d$ identity matrix.

(ii) Context vectors $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are related to word vectors according to $\mathbf{c}_i = \mathbf{Q}\mathbf{w}_i$, $i = 1, \ldots, n$, for some orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d\times d}$.

(iii) Given a word j, the probability of any word i being in its context is given by

$$p(i \mid j) \propto p_i \cdot e^{\mathbf{w}_j^\top \mathbf{c}_i} \tag{1.1}$$

where $p_i = p(i)$ is the unigram probability for the word i.

Conjecture. For any word $i \in \mathcal{W}$, the context vector $\mathbf{c}_i$ is a simple transformation of the word vector $\mathbf{w}_i$: $\mathbf{c}_i$ is obtained from the word vector $\mathbf{w}_i$ by flipping the signs of approximately half of the elements.

This conjecture is stated in [4] and it is suggested that under Assumptions 1–3 above, the conjecture reduces to the fact that the partition function

$$Z_j := \sum_{i=1}^{n} p_i \cdot e^{\mathbf{w}_j^\top \mathbf{c}_i}. \tag{1.2}$$

converges to its mean. The attempt in [4] at proving this statement was unsuccessful.

Our proof of this statement consists of three steps.

(i) We modify Assumptions 1–3, which, as stated, seem to be insufficient.

(ii) We obtain certain bounds for matrices using properties of Schatten–von Neumann classes from [9]. The fact that those properties have been established in the infinite-dimensional case help us make our bounds uniform with respect to the matrix dimension (which later we let go to infinity, unlike [4]).

(iii) The partition function is investigated using precise formulas for means, variances and the moment generating function of the quadratic form of a normal vector from [18]. As in our setup the dimension tends to infinity, we have to develop asymptotic counterparts of those precise formulas.

This last step involves a study of the asymptotic behavior of some determinants whose dimension tends to infinity.

1.3 Main result

First, we modify the assumption 1 from the previous section in the following form:

*)We refer the reader to the original paper [4] for the motivation behind these assumptions.

Assumption 1. Word vectors $\mathbf{w}_1, \dots, \mathbf{w}_n \in \mathbb{R}^d$ are independent and

$$\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{\Sigma}), \; i = 1, \dots, n, \tag{1.3}$$

where $\mathbf{\Sigma}$ is a $d \times d$ positive definite matrix whose elements may change with d and for some $m > 0$,

$$\|\mathbf{\Sigma}\|_{\sigma\infty} \le f(d)m, \qquad \max_i \|\mathbf{\Sigma}_i\|_2^2 \le mdf^2(d). \tag{1.4}$$

where $f(d)$ is a positive function such that

$$f(d) \cdot d^2 \to 0, \; d \to \infty. \tag{1.5}$$

For potential users of our results we note that for pre-trained WORD2VEC embeddings with the covariance matrix $\mathbf{\Sigma}$ the matrix $d^{-2.1}\mathbf{\Sigma}$ resulting from multiplying embeddings by $d^{-1.05}$ is a realistic choice that satisfies our conditions (this does not lead to a deterioration in the quality of vectors as measured by standard similarity and analogy tasks, see [7, 14]).

The second assumption from [4] is kept almost as it is.

Assumption 2. Context vectors $\mathbf{c}_1, \dots, \mathbf{c}_n \in \mathbb{R}^d$ are images of word vectors

$$\mathbf{c}_i = \mathbf{Q}\mathbf{w}_i, \; i = 1, \dots, n,$$

where $\mathbf{Q}$ is an orthogonal matrix. Its elements may change with d.

Our main result is the following

Theorem 1.1. *Under Assumptions 1 and 2,*

$$\begin{aligned} &\mathbb{E}[Z_j] = 1 + o(1) \quad \text{as} \quad d \to \infty, \\ &\mathbb{V}[Z_j] \to 0 \quad \text{as} \quad d \to \infty. \end{aligned}$$

Note that our setup differs from [4] not only by the modification of Assumption 1, but also by the passage to the limit when the dimension of word vectors tends to infinity. Since the dimension of word vectors is at the same time the width of the hidden layer in static embedding models, we can assume that our result complements the work on the theoretical analysis of neural networks in the infinite width mode [12, 11].

The rest of the paper is organized as follows: auxiliary results are stated and proved in Section 2, convergence of $\mathbb{E}[Z_j]$ to 1 is established in Section 3, while convergence of $\mathbb{V}[Z_j]$ to 0 is shown in Section 4.

2 Auxiliary statements

Two matrices will be particularly important in our work:

$$\mathbf{A} = \frac{1}{2}(\mathbf{Q} + \mathbf{Q}^\top), \qquad \mathbf{B} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top. \tag{2.1}$$

Lemma 2.1. *For the matrices in* (2.1) *we have the following bounds:*

(i) $|\mathrm{Tr}(\mathbf{A}\mathbf{\Sigma})| \le mf(d)d$.

(ii) $|\mathrm{Tr}\,[(\mathbf{A}\mathbf{\Sigma})^2]| \le m^2 f^2(d)d$.

(iii) $\|\mathbf{A}\mathbf{\Sigma}\|_\infty^2 \le mf(d)d$.

(iv) $|\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma})| \le m^2 f(d) d$.

(v) $\|\mathbf{B}\boldsymbol{\Sigma}\|_\infty \le m f(d) d^2$.

Proof. Note that by orthogonality

$$|(\mathbf{A}\mathbf{x},\mathbf{x})| = \left|\frac{1}{2}\left[(\mathbf{Q}\mathbf{x},\mathbf{x}) + (\mathbf{Q}^\top\mathbf{x},\mathbf{x})\right]\right| \le \frac{1}{2}\left[\|\mathbf{Q}\mathbf{x}\|_2\|\mathbf{x}\|_2 + \left\|\mathbf{Q}^\top\mathbf{x}\right\|_2\|\mathbf{x}\|_2\right] = \|\mathbf{x}\|_2^2.$$

Hence,

$$\|\mathbf{A}\|_{\sigma\infty} = \max_j s_j(\mathbf{A}) = \max_j |\lambda_j(\mathbf{A})| = \sup_{\|\mathbf{x}\|_2=1}|(\mathbf{A}\mathbf{x},\mathbf{x})| \le 1.$$

(i) By [9, Theorem 8.5] for any matrix $\mathbf{A}$

$$|\mathrm{Tr}(\mathbf{A})| \le \|\mathbf{A}\|_{\sigma_1}. \tag{2.2}$$

Using also the Hölder inequality [9, equation (7.5)]

$$\|\mathbf{A}\mathbf{B}\|_{\sigma_1} \le \|\mathbf{A}\|_{\sigma_p}\|\mathbf{B}\|_{\sigma_q} \ \text{ for } \frac{1}{p}+\frac{1}{q}=1 \tag{2.3}$$

(which is true for any $\mathbf{A}$, $\mathbf{B}$), for $\mathbf{A}$ defined in (2.1) we have

$$|\mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma})| \le \|\mathbf{A}\boldsymbol{\Sigma}\|_{\sigma_1} \le \|\mathbf{A}\|_{\sigma\infty}\|\boldsymbol{\Sigma}\|_{\sigma_1} \le m f(d) d.$$

The last inequality uses (1.4).

(ii) $\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ implies $\|\mathbf{A}\| \le 1$. Moreover, for any matrices $\mathbf{A}$, $\mathbf{B}$ [9, §7, Section 2]

$$\|\mathbf{A}\mathbf{B}\|_{\sigma_p} \le \|\mathbf{A}\| \cdot \|\mathbf{B}\|_{\sigma_p}. \tag{2.4}$$

Hence, by (2.2), (2.3)

$$\left|\mathrm{Tr}\left[(\mathbf{A}\boldsymbol{\Sigma})^2\right]\right| \le \left\|(\mathbf{A}\boldsymbol{\Sigma})^2\right\|_{\sigma_1} \le \|\mathbf{A}\boldsymbol{\Sigma}\|_{\sigma_2}^2 \le \|\mathbf{A}\|^2\|\boldsymbol{\Sigma}\|_{\sigma_2}^2 \le m^2 f^2(d) d. \tag{2.5}$$

(iii) We start with

$$|(\mathbf{A}\boldsymbol{\Sigma})_{ij}| = \left|\mathbf{A}_i\boldsymbol{\Sigma}^{(j)}\right| \le \|\mathbf{A}_i\|_2\left\|\boldsymbol{\Sigma}^{(j)}\right\|_2. \tag{2.6}$$

Here by orthogonality $\|\mathbf{A}_i\|_2 \le \frac{1}{2}\left(\left\|\mathbf{Q}_i^\top\right\|_2 + \|\mathbf{Q}_i\|_2\right) = 1$. By (1.4) $\|\mathbf{A}\boldsymbol{\Sigma}\|_\infty^2 \le m f(d) d$.

(iv) Apply successively (2.2), (2.3) and (2.4):

$$|\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma})| = \left|\mathrm{Tr}(\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^\top\boldsymbol{\Sigma})\right| \le \left\|\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^\top\boldsymbol{\Sigma}\right\|_{\sigma_1} \le \|\mathbf{Q}\boldsymbol{\Sigma}\|_{\sigma_2}\left\|\mathbf{Q}^\top\boldsymbol{\Sigma}\right\|_{\sigma_2} \le \|\mathbf{Q}\|^2\|\boldsymbol{\Sigma}\|_{\sigma_2}^2 \le m^2 f^2(d) d.$$

The last inequality is as in (2.5).

(v) By analogy with (2.6),

$$\left|(\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}^\top\boldsymbol{\Sigma})_{ij}\right| = \left|(\mathbf{Q}\boldsymbol{\Sigma})_i(\mathbf{Q}^\top\boldsymbol{\Sigma})^{(j)}\right| \le \|(\mathbf{Q}\boldsymbol{\Sigma})_i\|_2 \left\|(\mathbf{Q}^\top\boldsymbol{\Sigma})^{(j)}\right\|_2. \tag{2.7}$$

From $|(\mathbf{Q}\boldsymbol{\Sigma})_{ij}| \le \left\|\boldsymbol{\Sigma}^{(j)}\right\|_2$ (confer (2.6)) it follows that

$$\|(\mathbf{Q}\boldsymbol{\Sigma})_i\|_2^2 = \sum_{j=1}^{d} |(\mathbf{Q}\boldsymbol{\Sigma})_{ij}|^2 \le \sum_{j=1}^{d} \left\|\boldsymbol{\Sigma}^{(j)}\right\|_2^2 \le mf^2(d)d^2.$$

Since a similar bound holds for $\left\|(\mathbf{Q}^\top\boldsymbol{\Sigma})^{(j)}\right\|_2^2$, we conclude from (2.7) that $\|\mathbf{B}\boldsymbol{\Sigma}\|_\infty \le mf(d)d^2$.

□

Lemma 2.2. *For any $i \in \mathcal{W}$,*

$$\mathbf{w}_i^\top\mathbf{Q}\mathbf{w}_i - \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) \xrightarrow{p} 0 \quad \text{as} \quad d \to \infty.$$

Proof. By [18, Theorem 5.2a] we have

$$\mathbb{E}[\mathbf{w}_i^\top\mathbf{Q}\mathbf{w}_i] = \mathbb{E}[\mathbf{w}_i^\top\mathbf{A}\mathbf{w}_i] = \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}).$$

From [18, Theorem 5.2c]

$$\mathbb{V}(\mathbf{w}_i^\top\mathbf{Q}\mathbf{w}_i) = 2\,\mathrm{Tr}\left[(\mathbf{A}\boldsymbol{\Sigma})^2\right].$$

By the Chebyshev inequality and Lemma 2.1.2 for any $\varepsilon > 0$

$$P\left(\left|\mathbf{w}_i^\top\mathbf{Q}\mathbf{w}_i - \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma})\right| > \varepsilon\right) \le \frac{1}{\varepsilon^2}\mathbb{V}(\mathbf{w}_i^\top\mathbf{Q}\mathbf{w}_i) \to 0.$$

□

3 Convergence of means to 1 in Theorem 1.1

Recall that the partition function Z_j is defined by (1.2). We need to find the expectations in

$$\mathbb{E}[Z_j] = \sum_{i\ne j} p_i\mathbb{E}[e^{\mathbf{w}_j^\top\mathbf{c}_i}] + p_j\mathbb{E}[e^{\mathbf{w}_j^\top\mathbf{c}_j}] = \sum_{i\ne j} p_i\mathbb{E}[e^{\mathbf{w}_j^\top\mathbf{c}_i}] + p_j\mathbb{E}[e^{\mathbf{w}_j^\top\mathbf{A}\mathbf{w}_j}].$$

Lemma 3.1. *For any symmetric matrix $\mathbf{A}$ such that*

$$d^3\|\mathbf{A}\boldsymbol{\Sigma}\|_\infty^2 \to 0 \quad \text{as} \quad d \to \infty \tag{3.1}$$

one has

$$\mathbb{E}[e^{\mathbf{w}_j^\top\mathbf{A}\mathbf{w}_j}] = 1 + o(1) \quad \text{as} \quad d \to \infty.$$

Proof. This is the heart of the proof. *Step 1.* Let $M_z(t) = \mathbb{E}[e^{tz}]$ be the moment generating function (mgf) of a random variable z. If $\mathbf{y}$ is distributed as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then [18, Theorem 5.2b]

$$M_{\mathbf{y}^\top\mathbf{A}\mathbf{y}}(t) = |\mathbf{I} - 2t\mathbf{A}\boldsymbol{\Sigma}|^{-1/2} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^\top\left[\mathbf{I} - (\mathbf{I} - 2t\mathbf{A}\boldsymbol{\Sigma})^{-1}\right]\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\}$$

where $|\mathbf{A}|$ is the determinant of matrix $\mathbf{A}$. By (1.3) this gives

$$\mathbb{E}[e^{\mathbf{w}_j^\top\mathbf{A}\mathbf{w}_j}] = M_{\mathbf{w}_j^\top\mathbf{A}\mathbf{w}_j}(1) = |\mathbf{I} - 2\mathbf{A}\boldsymbol{\Sigma}|^{-1/2}. \tag{3.2}$$

Step 2. Denote $\mathbf{H} = 2\mathbf{A}\mathbf{\Sigma}$, $x(d) = \max|h_{ij}|$. We want to prove that for $k \le d$ and any $1 \le i_1 < \ldots < i_k \le d$

$$(1-h_{i_1,i_1})\cdots(1-h_{i_k,i_k}) = 1 + o(1) \quad \text{as} \quad d \to \infty, \tag{3.3}$$

where the $o(1)$ is uniform in $0 \le k \le d$ and for $k = 0$ the product on the left is 1 by definition. By (3.1)

$$|h_{ij}| \le x(d), \quad d^{3/2}x(d) \to 0 \quad \text{as} \quad d \to \infty. \tag{3.4}$$

The case $k = 0$ is trivial. We consider the indices $i_1 = 1, \ldots, i_k = k$, the other cases being similar. Denote $g_k(x_1, \ldots, x_k) = (1-x_1)\cdots(1-x_k)$, $1 \le k \le d$. The Taylor series for g_k is a finite sum

$$(1-h_{11})\cdots(1-h_{kk}) = 1 - \sum_{i=1}^{k} h_{ii} + r_2 \tag{3.5}$$

where

$$r_2 = \sum_{l=2}^{k} \sum_{\substack{l_1+\ldots+l_j=l \\ 0\le l_i \le 1}} \frac{\partial^l g_k(0,\ldots,0)}{\partial x_1^{l_1}\ldots\partial x_j^{l_j}} \frac{h_{11}^{l_1}}{l_1!} \cdots \frac{h_{jj}^{l_j}}{l_j!}.$$

Using the equation $\frac{\partial g_k(x_1,\ldots,x_k)}{\partial x_1} = -g_{k-1}(x_2,\ldots,x_k)$ it is easy to see that in this sum all derivatives have values ± 1. In each term, at least two factors of the form $h_{ii}^{l_i}$ are nontrivial. Hence,

$$|r_2| \le \sum_{l=2}^{k} C_l^k x^l(d). \tag{3.6}$$

Note that

$$C_{l+1}^k x^{l+1}(d) = C_l^k \frac{k-l}{l+1} x^{l+1}(d) \le C_l^k [dx(d)] x^l(d).$$

Since by (3.4) $d \cdot x(d) = o(1)$, there exists $d_0 > 0$ such that $d \cdot x(d) \le 1$ for $d \ge d_0$. Then from (3.6) for such d and all $k \le d$

$$|r_2| \le (k-1)C_2^k x^2(d) \le d\frac{k!}{2!(k-2)!}x^2(d) \le \frac{1}{2}d^3x^2(d). \tag{3.7}$$

Now (3.5) and (3.7) imply $|(1-h_{11})\ldots(1-h_{kk}) - 1| \le dx(d) + \frac{1}{2}d^3x^2(d) = o(1)$. We have proved (3.3).

Step 3. Here we prove that $|\mathbf{I} - \mathbf{H}| = 1 + o(1)$ as $d \to \infty$. By the Leibnitz formula

$$|\mathbf{I} - \mathbf{H}| = \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma) \prod_{i=1}^{d} t_{i,\sigma(i)},$$

where $t_{i,j}$ are the elements of $\mathbf{T} = \mathbf{I} - \mathbf{H}$, S_d is the set of permutations of $\{1, \ldots, d\}$, $\operatorname{sgn}(\sigma)$ is the signature of the permutation σ. Separating the diagonal elements, for which $i = \sigma(i)$ are fixed points of σ, we have

$$|\mathbf{I} - \mathbf{H}| = \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma) \prod_{i=\sigma(i)} (1-h_{ii}) \prod_{i \ne \sigma(i)} h_{i,\sigma(i)}. \tag{3.8}$$

Let $k(\sigma)$ denote the number of fixed points of the permutation σ. Then the number of points that do not stay in place is $d - k(\sigma)$ and by (3.4)

$$\left| \prod_{i \ne \sigma(i)} h_{i,\sigma(i)} \right| \le x^{d-k(\sigma)}(d), \ 0 \le k(\sigma) \le d.$$

For $d \geq 0$ and $0 \leq k \leq d$, the *rencontres* number is defined as the number of permutations of $\{1, ..., d\}$ that have k fixed points. We need the equation[*)]

$$D_{d,k} = \frac{d!}{k!} \sum_{l=0}^{d-k} \frac{(-1)^l}{l!}.$$

It implies

$$|D_{d,k}| \leq \frac{d!}{k!} \sum_{l=0}^{\infty} \frac{1}{l!} = e\frac{d!}{k!}. \tag{3.9}$$

In (3.8) only the identity permutation leaves all indices unchanged. The corresponding term is $(1-h_{11})\ldots(1-h_{dd})$ which is $1+o(1)$ by (3.3). Hence,

$$|\mathbf{I}-\mathbf{H}| = 1 + o(1) + \sum_{\sigma\in S_d,\ k(\sigma)\leq d-1} \operatorname{sgn}(\sigma) \prod_{i=\sigma(i)} (1-h_{ii}) \prod_{i\neq\sigma(i)} h_{i,\sigma(i)}. \tag{3.10}$$

According to (3.3) here $\left|\prod_{i=\sigma(i)}(1-h_{ii})\right| \leq 2$, whereas $\prod_{i\neq\sigma(i)} h_{i,\sigma(i)}$ contains $d-k(\sigma)$ terms. Therefore by (3.4) and (3.9)

$$\begin{aligned}
&\left|\sum_{\sigma\in S_d,\ k(\sigma)\leq d-1} \operatorname{sgn}(\sigma) \prod_{i=\sigma(i)} (1-h_{ii}) \prod_{i\neq\sigma(i)} h_{i,\sigma(i)}\right| \\
\leq\ & 2\sum_{k=0}^{d-1} D_{d,k}x^{d-k}(d) \leq 2e\sum_{k=0}^{d-1} \frac{d!}{k!}x^{d-k}(d) = 2e\sum_{k=0}^{d-1} d(d-1)...(k+1)x^{d-k}(d) \\
\leq\ & 2e\sum_{k=0}^{d-1}(dx(d))^{d-k} \leq 2edx(d)\sum_{j=0}^{\infty}(dx(d))^j = \frac{2edx(d)}{1-dx(d)} = o(1).
\end{aligned}$$

This and (3.10) prove that $|\mathbf{I}-2\mathbf{A}\boldsymbol{\Sigma}| = 1+o(1)$.

Now the statement follows from (3.2) and the fact that $(1-x)^{-1/2} = 1+o(1)$, $x\to 0$. □

Lemma 3.2. *For $i \neq j$*

$$\mathbb{E}[e^{\mathbf{w}_j^\top \mathbf{c}_i}] = 1+o(1). \tag{3.11}$$

Proof. Let $M_{\mathbf{y}}(\mathbf{t}) = \mathbb{E}[e^{\mathbf{t}^\top \mathbf{y}}]$ be the multivariate mgf of a random vector $\mathbf{y}$. If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then [18, Theorem 4.3]

$$M_{\mathbf{y}}(\mathbf{t}) = e^{\mathbf{t}^\top\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}^\top\boldsymbol{\Sigma}\mathbf{t}}.$$

In our case this implies

$$\begin{aligned}
\mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{c}_i} \mid \mathbf{w}_j\right] = \mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_i} \mid \mathbf{w}_j\right] &= M_{\mathbf{w}_i}(\mathbf{Q}^\top\mathbf{w}_j) \\
&= e^{(\mathbf{Q}^\top\mathbf{w}_j)^\top f\boldsymbol{\Sigma}\mathbf{Q}^\top\mathbf{w}_j/2} = e^{\frac{f}{2}\mathbf{w}_j^\top\mathbf{B}\mathbf{w}_j}.
\end{aligned} \tag{3.12}$$

Hence, by the law of iterated expectations

$$\mathbb{E}[e^{\mathbf{w}_j^\top \mathbf{c}_i}] = \mathbb{E}\left[\mathbb{E}[e^{\mathbf{w}_j^\top \mathbf{c}_i} \mid \mathbf{w}_j]\right] = \mathbb{E}e^{\frac{f}{2}\mathbf{w}_j^\top\mathbf{B}\mathbf{w}_j}.$$

Because of condition (1.5) and Lemma 2.1.5 we obtain

$$d^3 f^2(d)\|\mathbf{B}\boldsymbol{\Sigma}\|_\infty^2 \leq m^2 d^7 f^4(d) = m^2(d^2 f(d))^3 df(d) \to 0. \tag{3.13}$$

This allows us to use Lemma 2.1 to prove (3.11). □

Corollary 3.1. *From* (2.1)*, Lemmas* 2.1.3, 3.1 *and* 3.2 *it follows that* $\mathbb{E}[Z_j] = 1+o(1)$ *as* $d\to\infty$.

[*)]see `https://en.wikipedia.org/wiki/Rencontres_numbers`

4 Convergence of variances to zero in Theorem 1.1

Obviously,

$$\mathbb{V}[Z_j] = \sum_{s=1}^{n} p_s^2 \mathbb{V}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_s}\right] + \sum_{s\neq t} p_s p_t \operatorname{Cov}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_s}, e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_t}\right].$$

Proof. We use Lemmas 3.1 and 3.2 which hold under our assumptions for the matrices in (2.1).

1) For $s = j$ by Lemma 3.1 with $\mathbf{A} = (\mathbf{Q} + \mathbf{Q}^\top)/2$

$$\mathbb{V}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_j}\right] = \mathbb{E}[e^{2\mathbf{w}_j^\top \mathbf{A}\mathbf{w}_j}] - \left[\mathbb{E}[e^{\mathbf{w}_j^\top \mathbf{A}\mathbf{w}_j}]\right]^2 = o(1). \tag{4.1}$$

2) For $s \neq j$ by Lemma 3.2

$$\mathbb{V}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_s}\right] = \mathbb{E}[e^{2\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_s}] - \left[\mathbb{E}[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_s}]\right]^2 = o(1).$$

3) Let all three numbers s, t, j be different. Denote by $\operatorname{Cov}_Z(X, Y)$ the covariance between X and Y conditional on Z. By the law of total covariance

$$\begin{aligned}&\operatorname{Cov}\left[e^{\mathbf{w}_j^\top \mathbf{c}_s}, e^{\mathbf{w}_j^\top \mathbf{c}_t}\right] \\ &\qquad = \mathbb{E}\left[\operatorname{Cov}_{\mathbf{w}_j}\left(e^{\mathbf{w}_j^\top \mathbf{c}_s}, e^{\mathbf{w}_j^\top \mathbf{c}_t}\right)\right] + \operatorname{Cov}\left(\mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{c}_s} \mid \mathbf{w}_j\right], \mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{c}_t} \mid \mathbf{w}_j\right]\right).\end{aligned}$$

Conditionally on $\mathbf{w}_j$, the variables $\mathbf{w}_s$ and $\mathbf{w}_t$ are independent, so the first term on the right is zero. For the second term we use (3.12):

$$\begin{aligned}&\operatorname{Cov}\left(\mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{c}_s} \mid \mathbf{w}_j\right], \mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{c}_t} \mid \mathbf{w}_j\right]\right) \\ &\qquad = \operatorname{Cov}\left(e^{\frac{f}{2}\mathbf{w}_j^\top \mathbf{B}\mathbf{w}_j}, e^{\frac{f}{2}\mathbf{w}_j^\top \mathbf{B}\mathbf{w}_j}\right) = \mathbb{V}\left[e^{\frac{f}{2}\mathbf{w}_j^\top \mathbf{B}\mathbf{w}_j}\right].\end{aligned}$$

By (3.11) and (3.13) then

$$\operatorname{Cov}\left(e^{\mathbf{w}_j^\top \mathbf{c}_s}, e^{\mathbf{w}_j^\top \mathbf{c}_t}\right) = \mathbb{E}[e^{f\mathbf{w}_j^\top \mathbf{B}\mathbf{w}_j}] - \left(\mathbb{E}e^{\frac{f}{2}\mathbf{w}_j^\top \mathbf{B}\mathbf{w}_j}\right)^2 = o(1).$$

4) Suppose $s \neq t$ and $s = j$. In the equation

$$\begin{aligned}&\operatorname{Cov}\left(e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_j}, e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_t}\right) \\ &\qquad = \mathbb{E}\left[\operatorname{Cov}_{\mathbf{w}_j}\left(e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_j}, e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_t}\right)\right] + \operatorname{Cov}\left(\mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_j} \mid \mathbf{w}_j\right], \mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_t} \mid \mathbf{w}_j\right]\right)\end{aligned}$$

the first term on the right is zero because, conditionally on $\mathbf{w}_j$, $e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_j}$ is constant. By (3.12)

$$\mathbb{E}\left[e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_t} \mid \mathbf{w}_j\right] = e^{\frac{f}{2}\mathbf{w}_j^\top \mathbf{B}\mathbf{w}_j}.$$

Hence, by (3.1), (3.13) and Lemma 3.1

$$\begin{aligned}\operatorname{Cov}\left(e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_j}, e^{\mathbf{w}_j^\top \mathbf{Q}\mathbf{w}_t}\right) &= \operatorname{Cov}\left(e^{\mathbf{w}_j^\top \mathbf{A}\mathbf{w}_j}, e^{\frac{f}{2}\mathbf{w}_j^\top \mathbf{B}\mathbf{w}_j}\right) \\ &= \mathbb{E}[e^{\mathbf{w}_j^\top (\mathbf{A}+f\mathbf{B}/2)\mathbf{w}_j}] - \left(\mathbb{E}[e^{\mathbf{w}_j^\top \mathbf{A}\mathbf{w}_j}]\right)\left(\mathbb{E}[e^{\mathbf{w}_j^\top f\mathbf{B}\mathbf{w}_j/2}]\right) \\ &= o(1).\end{aligned} \tag{4.2}$$

Summarizing, from (4.1)–(4.2) we get

$$\mathbb{V}[Z_j] = \sum_{s=1}^{n} p_s^2 \cdot o(1) + \sum_{s\neq t} p_s \cdot p_t \cdot o(1) \to 0 \quad \text{as} \quad d \to \infty.$$

□

Acknowledgments

The work is supported by the Nazarbayev University Collaborative Research Programme 091019CRP2109, PI Zhenisbek Assylbekov.

References

[1] C. Allen, I. Balazevic, T.M. Hospedales, *What the vec? towards probabilistically grounded embeddings*, in: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019. Vancouver, BC, Canada, 7465–7475.

[2] C. Allen, T.M. Hospedales, *Analogies explained: Towards understanding word embeddings*, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, 97 (2019), 223–231.

[3] S. Arora, Y. Li, Y. Liang, T. Ma, A. Risteski, *A latent variable model approach to pmi-based word embeddings.* Trans. Assoc. Comput. Linguistics, 4 (2016), 385–399. https://doi.org/10.1162/tacl_a_00106

[4] Zh. Assylbekov, R. Takhanov, *Context vectors are reflections of word vectors in half the dimensions.* J. Artif. Intell. Res., 66 (2019), 225–242. https://doi.org/10.1613/jair.1.11368

[5] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, *BERT: pre-training of deep bidirectional transformers for language understanding*, in: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 1 (2019), 4171–4186.

[6] K. Ethayarajh, D. Duvenaud, G. Hirst, *Towards understanding linear word analogies*, in: Korhonen, A., Traum, D.R., Màrquez, L. (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 1 (2019), 3253–3262.

[7] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, *Placing search in context: the concept revisited*, in: Shen, V.Y., Saito, N., Lyu, M.R., Zurko, M.E. (Eds.), Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5 (2001), 406–414.

[8] A. Gittens, D. Achlioptas, M.-W. Mahoney, *Skip-gram – zipf + uniform = vector additivity*, in: Barzilay, R., Kan, M.-Y. (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, 1 (2017), 69–76.

[9] I. Gohberg, M. Kreĭn, *Introduction to the theory of linear nonselfadjoint operators.* Translations of Mathematical Monographs, v. 18, 1968.

[10] T.B. Hashimoto, D. Alvarez-Melis, T.S. Jaakkola, *Word embeddings as metric recovery in semantic spaces.* Trans. Assoc. Comput. Linguistics, 4 (2016), 273–286. https://doi.org/10.1162/tacl_a_00098

[11] A. Jacot, C. Hongler, F. Gabriel, *Neural tangent kernel: Convergence and generalization in neural networks*, in: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 8580–8589.

[12] J. Lee, Y. Bahri, R. Novak, S.S. Schoenholz, J. Pennington, J. Sohl-Dickstein, *Deep neural networks as gaussian processes*, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.

[13] O. Levy, Y. Goldberg, *Neural word embedding as implicit matrix factorization*, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13, 2014, Montréal, Quebec, Canada, 2177–2185.

[14] T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient estimation of word representations in vector space*, in: Bengio, Y., LeCun, Y. (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.

[15] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, *Distributed representations of words and phrases and their compositionality*, in: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems, 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, 3111–3119.

[16] J. Pennington, R. Socher, C.D. Manning, *Glove: Global vectors for word representation*, in: Moschitti, A., Pang, B., Daelemans, W. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, 1532–1543.

[17] M.W. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, *Deep contextualized word representations*, in: Walker, M.A., Ji, H., Stent, A. (Eds.), Proceedings A shortened title of the paper 13 of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 1 (2018), 2227–2237.

[18] A.C. Rencher, G.B. Schaalje, *Linear models in statistics, 2nd ed.* Wiley, New York, 2008.

[19] R. Tian, N. Okazaki, K. Inui, *The mechanism of additive composition.* Mach. Learn., 106 (2017), no. 7, 1083–1130. https://doi.org/10.1007/s10994-017-5634-8

[20] Z. Zobnin, E. Elistratova, *Learning word embeddings without context vectors*, in: Augenstein, I., Gella, S., Ruder, S., Kann, K., Can, B., Welbl, J., Conneau, A., Ren, X., Rei, M. (Eds.), Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019, 244–249.

Kairat Mynbaev
International School of Economics
Kazakh-British Technical University
59 Tolebi St,
050000 Almaty, Kazakhstan
E-mail: k_mynbayev@ise.ac

Zhenisbek Assylbekov
Department of Mathematics
School of Sciences and Humanities
Nazarbayev University
53 Kabanbay Batyr Ave
010000 Astana, Kazakhstan
E-mail: zhassylbekov@nu.edu.kz

Received: 07.07.2022