Skoltech
Skolkovo Institute of Science and Technology

# Jury Member Report – Doctor of Philosophy thesis.

**Name of Candidate:** Alexander Fonarev

**PhD Program:** Computational and Data Science and Engineering

**Title of Thesis:** Matrix Factorization Methods for Training Embeddings in Selected Machine Learning Problems

**Supervisor:** Prof. Ivan Oseledets

**Chair of PhD defense Jury** Prof. Andrzej Cichocki ***Email**: a.cichocki@skoltech.ru*

**Date of Thesis Defense:** 19 September 2018

**Name of the Reviewer:**

| | |
|---|---|
| I confirm the absence of any conflict of interest<br>(Alternatively, Reviewer can formulate a possible conflict) | **Signature:**<br>ACichocki<br>**29-08-2018**<br>**Date: DD-MM-YYYY** |

*The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.*

*If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.*

**Reviewer's Report**

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

# Review of PhD Thesis of Alexander Fonarev

The main objective of the PhD thesis of Alexander Fonarev entitled "Matrix Factorization Methods for Training Embeddings in Selected Machine Learning Problems" was to develop a new framework for training embeddings using low-rank matrix factorizations and to implement related efficient algorithms. In other words, the thesis focuses on the problem of representing complex objects as real-valued low dimensional vectors, so-called embeddings, based on collected statistical data using low-rank matrix factorizations and developing new embedding algorithms that outperform existing approaches in terms of performance for specific applications.

Recently, the embeddings problem became an important area of research in machine learning, whereby most of popular embedding methods are based on neural network approaches, while methods based on low-rank matrix factorizations have not so far be fully explored The author of the thesis focused mostly on this type of methods and demonstrated their superiority over existing in several practical problems. In fact, embeddings have various applications, e.g., Natural language processing, information retrieval, security systems and recommender systems.

There are several sub-problems considered and solved in the thesis which correspond to various applications of embedding:

1. The first sub-problem/task was to learn embeddings of categorical features values in supervised and unsupervised machine learning problems. The goal was to replace categorical features values with their embeddings and use efficient machine learning techniques (especially, Random Forest) without deterioration of performance compared to the state-of-the-art approaches that handle categorical features out-of-the-box (Factorization Machines).

2. The second sub-problem was training word embeddings. The goal was to train such natural language words' embeddings that words with close meanings would have close embeddings in the representation space.

3. The third sub-problem was to obtain embeddings of so called "cold" users and cold items in recommender systems. Since users or items were cold, it was sufficient to ask the environment only for a relatively small piece of information about the cold objects (active learning setup) and their embeddings exploited only this limited information. The main goal was to obtain such embeddings that show relatively good performance in the recommender system problem in terms of popular recommender performance metrics.

The author developed 3 methods and validated and tested his algorithms by measuring their performance:

1. For the first task mentioned above, he used two popular binary classification datasets with categorical features (Section 4.3 for details) and AUC-ROC as a binary classification performance measure.

2. For the second subproblem, he used standard experimental setup from the literature - the benchmark datasets (wordsim-353, simplex-999, men).(Section 5.4.2 ) ;

3. For the third subproblem, he also used the experimental setup that is standard in the literature - two most popular open recommender datasets (MovieLens and Netflix) and standard recommender metrics (Precision@*k*, Recall@*k*, Coverage, Diversity) were used. (Section 6.6).

Furthermore, the author of the thesis developed software packages using the Python programming language and published them on the Web as open source. One of the proposed methods is used in an industrial recommender system at Yandex.

In my opinion there at least 4 main contributions of this PhD thesis

1. Chapter 3 introduces the simple but efficient low-rank factorization framework to train embeddings that generalizes existing approaches to train embeddings using matrix factorizations. Unfortunately, such techniques are usually underestimated by the ML community, despite of the fact that they often show performance that is not achievable the competitors. This framework allowed to develop several new state-of the- art embedding methods described in the thesis;

2. In Chapter 4, the thesis formulates the problem of unsupervised search for embeddings of categorical features' values and introduces a new unsupervisedmethod to train such embeddings;

3. In Chapter 5, the author of the thesis reformulates the Skip-Gram Negative Sampling word embeddings training procedure and, according to this formulation, introduces the new embeddings training method based on Riemannian optimization, which often outperforms existing state-of-the-art approaches. The source code of the method is publicly available on the Web. Moreover, the introduced theoretical findings were recognized as useful and were used and already cited by other researchers .

4. In Chapter 6, the thesis introduces the novel method to obtain embeddings of cold users and cold items in recommender systems based on the rectangular generalization of the Maxvol algorithm. The developed method outperforms existing state-of-the-art approaches in terms of quality and computational complexity.

The source code of the method is published on the Web and was already used by other researchers in the ML community. Furthermore, the method has been successfully applied within an industrial recommender system by Yandex, what additionally proves the practical relevance of the developed approach.

The author of the thesis published partially his results mostly in international conferences (two of them belong top A conferences). Unfortunately the results so far were not published in journals However, I consider that publications in proceedings of conferences are on quite high level and the author of the thesis demonstrated his original and innovative results. I believe that results of the thesis will be further improved and applied for commercial applications and they will be published in extended form in high impact factor journals.

**Provisional Recommendation**

| |
|---|
| x□ *I recommend that the candidate should defend the thesis by means of a formal thesis defense* |
| □ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate’s thesis according to the recommendations of the present report* |
| □ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense* |