Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Marina Munkhoeva

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Fast numerical linear algebra methods for machine learning

Supervisor: Professor Ivan Oseledets

Name of the Reviewer: Andrzej Cichocki

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Signature:** A.Cichocki **Date: 20-02-2021**

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 the Ph.D. thesis of Marina Munkhoeva

"Fast Numerical Linear Algebra Methods for Machine Learning"

Evaluation of the thesis quality and overall structure of the dissertation.
The Ph.D. thesis of Ms. Marina Munkhoeva consists of 5 chapters, which mostly discusses results published in her four excellent papers. Although the topic of each chapter is quite different, they all are connected by powerful linear algebra methods and tools, especially modern and sophisticated methods for high-dimensional approximations and large-scale computational problems. In general the thesis well written and good organized and mathematical formulas are are well explained.

The relevance of the topic of dissertation work to its actual content
In my opinion, this is cutting-edge research since the author of the thesis was able to apply successfully advanced linear algebra tools and to provide new algorithms and extensive computer experiments. It should be emphasized that Marina's research required deep knowledge and skills not only in linear algebra but also in machine learning and deep neural networks. She has clearly demonstrated that Numerical Linear Algebra (NLA) methods have still much to offer to modern Machine Learning (ML), both in terms of problems related to computational complexity and approximation quality in various problems and data settings. She has demonstrated the usefulness of large-scale matrix computations and numerical analysis methods in such machine learning applications and problems as kernel methods, generative model evaluation, and graph summarization.

The relevance of the methods used in the dissertation
The main results and original methods are presented in Chapters 2-4.

Chapter 2, is devoted to the kernel approximation for the most popular class of kernels - shift-invariant kernels. Marina applied quadrature approximation for shift-invariant kernels (e.g., radial basis functions) to derive sparse random features that significantly improve kernel approximation quality in comparison to the state-of-the-art methods, especially random Fourier features, orthogonal random features Gaussian Quadratures, and Quasi-Monte Carlo. Moreover, she has shown that Quadrature-based Features (QF) generalize and improve some existing methods as quadrature rules of varying order. The proposed features rely on butterfly matrices that are products of sparse factors, allowing reduced memory requirements and provide relatively fast matrix-vector multiplications which are useful in resource-constrained settings. Although the author the thesis empirically shown that the kernel approximation approach improves considerably accuracy in downstream tasks, there is still an open question of whether better kernel approximation is explicitly connected to downstream task performance, and related to problems studied in representation learning, e.g., embedding methods.

In Chapter 3, Marina derived a distance measure inspired by the intrinsic geometry of data manifolds. She achieved here new and interesting results due to skillful applying the numerical linear algebra method called Stochastic Lanczos Quadrature (SLQ) for estimating matrix functions. She demonstrated that it is possible to compute the approximate spectral signature of a data manifold by constructing a k-nearest neighbors graph from a data sample, followed by computing a parameterized quantity on its graph Laplacian matrix. The practical metric between two samples can then be found by taking the lower bound on the Gromov-Wasserstein distance between two manifolds by taking an L2-norm distance between respective signatures, so-called Intrinsic Multiscale Distance (IMD).

Marina empirically and extensively tested IMD on generative models evaluation, representation evolution in neural networks, and detecting language kinship through their word embeddings. The results are convincing and quite impressive. In Chapter 3, she did not address a strictly defined mathematical problem, but rather test some new ideas on how we can approximate a certain lower bound on the inter manifold distance with data sampled from these manifolds. In my opinion, this can be considered as an important sub-problem in generative modeling in high-dimensions, especially for implicit likelihood models such as Generative Adversarial Networks (GANs), i.e., how to choose optimal one among trained GANs based on their generated samples? She has shown that current quality metrics (e.g., Inception Score, Frechet Inception Distance) fail in many cases, for example, mode collapse and mode dropping are frequent problems for GANs. The computation of the proposed metric is itself an approximation problem of a trace of a matrix exponential which required careful handling and variance reduction.

For IMD, Marina performed a wide range of experiments, especially testing IMD's scalability and stability, since she wanted to verify how the variance reduction technique works. She also made extensive experiments with unaligned language manifolds as given by word embeddings trained on Wikipedia — IMD was able to reveal language kinship, unlike FID and KID. She applied IMD to see if we can track the change in image representations across neural network layers. IMD showed that initial convolutional blocks do some sort of feature extraction, followed by the convergence of representations in the subsequent layers. In my opinion, the most interesting experiment she did with IMD was comparing the sets of word embeddings derived from Wikipedia data in different languages. IMD was able to reveal language kinship showing closely related language pairs having smaller IMD values between them compared to distant languages.

In Chapter 4, she investigated the SLQ for yet another application - accurate computation of spectral graph descriptors of Web-scale graphs, i.e. huge graphs with billions of nodes and edges. She considered two spectral graph descriptors: Von Neumann Graph Entropy and Network Laplacian Spectral Descriptors. She demonstrated that the approach (called SLaQ) not only provides a few orders of magnitude better approximation than simplistic and heuristic approaches previously used, but it also is scalable allowing comparison of graphs with billions on a single machine in less than an hour. It is important to emphasize that the developed SlaQ approach is a very efficient computation technique for graph descriptors.

She tested whether the developed SlaQ method can provide a better approximation of the graph spectral descriptors against the SOTA methods (FINGER and NetLSD, both having either second-order Taylor approximation or linear interpolation between the spectrum tails). This was tested on rather smaller graphs since she was able to compute exact values only for them (eigenvalue decomposition is practically infeasible for large-scale graphs). She also tested whether the developed method brings benefits of non-local approximation compared to 2nd order Taylor approximation and linear interpolation. She has shown this by using dynamic graphs by computing both descriptors' values for Wikipedia monthly snapshots and detecting anomalous spikes in them that cannot be attributed only to the number of edge deletions/additions (local information). Moreover, she tested the graph classification performance of SLaQ against the same methods as previously. For this, she computed the descriptors and used their features for graphs.

In Chapter 5, she discussed a novel node embedding algorithm based on matrix sketching called FREDE. She showed that this method has some advantages compared to the state-of-the-art graph summarization techniques in terms of implicit optimization objective, and especially for streaming data, distribution, and parallelization. This contribution has a similar limitation to the contribution presented in Chapter 1 (devoted to QF), as although, the proposed embeddings are optimal in terms of the implicit objective (covariance error for SVD-based embedding methods), the downstream performance of any embeddings method is rather not directly connected to the accuracy of its primary objective. She compared FREDE (method for node embeddings) against the best existing representative methods:

- DeepWalk (sampling random walks and applying word2vec algorithm on these sequences/walks of nodes),
- VERSE (a single layer neural network used to obtain personalized PageRank matrix, PPR),
- NetMF (matrix factorization of closed-form DeepWalk matrix),
- NetSMF(sparse matrix factorization),
- LouvainNE (based on hierarchical clustering method).

The scientific significance of the results obtained and their compliance with the international level and current state of the art

Most of the developed methods, algorithms and theoretical analyses were tested, verified, and compared by performing extensive computer experiments for various difficult benchmarks and data sets. All of the methods except for SLaQ (due to open-sourcing restrictions of Google) were provided as open-sourced on Github (QF, IMD, FREDE). Marina mainly used Python and the frameworks, toolboxes, and libraries for scientific computing and machine learning (numpy, scipy, scikit-learn, etc) and deep learning (Tensorflow, PyTorch, and JAX) and she demonstrated excellent programming skills.

I believe that the IMD method is the most important and original result in the thesis as it is one of the first papers in ML addressing data manifold comparison with methods inspired by intrinsic geometry of the manifolds.

In my opinion, the most significant and novel results obtained by Marina Munkhaeva are as follows:

- Propose of novel quadrature-based features method that generalizes the existing SOTA methods for kernel approximation and outperforms these methods in terms of kernel approximation and often downstream tasks.
- Development of a new method for comparing data manifolds that aggregates information about manifold's intrinsic geometry from data sampled from it. It was shown its efficacy in tasks such as generative model comparison, finding similar collections of embedding (language kinship experiment), and tracking the evolution of neural network representations across layers and training epochs.
- Development of a method for accurate and fast approximation of spectral descriptors for very large-scale Web-scale graphs (graphs with millions and even billions of nodes and edges).
- Derivation and implementation of a node embedding algorithm based on matrix sketching that provides desirable properties of anytime optimality guarantees and mergeability which are essential in the streaming and distributed processing of Web-scale graphs.

The quality of publications

The quality of all four papers published by Marina Munkhoeva as a leading coauthor is very high and the results presented there are quite impressive. They were published and presented in top Machine Learning international conferences, especially in NeurIPS and ICLR. They are above high level on international standard.

The relevance of the obtained results to applications

One weak point of the thesis is that most materials are devoted to rather theoretical and fundamental ML problems and it was not made any attempt to apply them to any industrial or practical real-life applications.

It would be helpful to provide a list of important symbols, notations and abbreviations used in the in the thesis directly after list of Tables. It would be great if the text will be revised by native English speaker.

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*