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: Victor Lempitsky, associate professor, Skolkovo Institute of Science and Technology

I confirm the absence of any conflict of interest	**Signature:** **Date: 15-01-2021**

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

Brief evaluation of the thesis quality and overall structure of the dissertation

The thesis describes significant contributions at the intersection of numerical linear algebra and machine learning. The common theme underlying all contributions is the use of fast numerical linear algebra methods for scaling up machine learning and data analysis methods.

The thesis contains a well-written introductory chapter (Chapter 1) describing the basic linear algebra and numeric integration concepts common for all contributions. The following four chapters each describe a significant methodological advance.

Thus, Chapter 2 develops a new approach for the numeric approximation of kernel functions. Kernel methods are a big subfield of machine learning, and their scalability to large-scale datasets remains a challenge. The proposed approach uses the quadrature integration techniques to reduce kernel approaches to linear methods and does so with a notable reduction of approximation errors compared to previously proposed reductions such as random Fourier features. The better quality of approximation translates into higher accuracy of prediction tasks.

Chapter 3 proposes a new way of comparing/describing data manifolds using quadrature integration. Comparison of data manifold (based on the sample sets) is practically important for the assessment of generative modeling methods. The advantages of the newly introduced similarity measure include the fact that it is intrinsic (invariant to isometric transforms) and its ability to compare manifolds across multiple scales.

Chapter 4 proposes a new approach to how numerical integration can be used to estimate two important graph descriptors that are common in graph analysis and can be used to find (dis)similarities between graphs. Significant improvement in approximation quality over previously proposed numerical techniques for the estimation of the same descriptors is demonstrated.

Chapter 5 suggests a new algorithm that estimates node embeddings in large graphs. The algorithm has an attractive computational complexity and allows to estimate node descriptors at *any time* (i.e. after observing a subset of graph). The estimated anytime descriptors are shown to be better suitable for prediction (classification) tasks than descriptors computed with previous node embedding methods.

A short conclusion section and a number of appendices are at the end of the thesis, followed by an extensive list of references containing ~250 works.

All four contributions of the thesis have clear novelty, addresses important problems with clear applications in data analysis. The newly proposed methods are compared to the state-of-the-art through extensive experiments on real data including benchmark data commonly used in the field. The comprehensiveness of empirical evaluation is noteworthy. The comparisons with state-of-the-art are thorough and include comparisons in terms of accuracy and runtime. Strong and weak points of individual methods are highlighted and discussed. In most cases, the comparison in terms of worst-case computational complexity is also conducted. The thesis is very well-written.

The relevance of the topic of dissertation work to its actual content
The relevance of the topic of the dissertation to its actual content is strong, and the contributions described in the thesis are consistent. A similar class of numerical methods (quadrature approaches) are used through chapters 2-4, and a similar application field (graph analysis) is considered in chapters (3-5).

The relevance of the methods used in the dissertation
The thesis brings and adapts state-of-the-art numerical linear algebra techniques to the tasks and applications from data analysis, including machine learning and data analysis of high-dimensional vectorial data, analysis, and machine learning on graphs. To the best of my understanding, state-of-the-art methods, and evaluation protocols are used throughout the thesis.

The scientific significance of the results obtained and their compliance with the international level and current state of the art
The contributions of the thesis advance state-of-the-art as evidenced by the acceptance to the premier publication venues of the data analysis field. Several follow-up works to some of the methods introduced in the thesis (in particular to the approach introduced in Chapter 2) have been presented by other groups. The

The relevance of the obtained results to applications (if applicable)
The methods introduced in the thesis are highly relevant to several application fields, such as recommender systems, web graph analysis, bioinformatics, computational linguistics, etc. The empirical evaluation of the methods considers datasets obtained in all of those application domains.

The quality of publications

The chapters containing contributions (Chapter 2,3,4,5) are based on peer-reviewed papers accepted to the premier and highly-selective conferences in the data analysis field: NeurIPS, ICLR, The Web Conference (previously called WWW), VLDB.

Provisional Recommendation

☒ *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*