

Jury Member Report – Doctor of Philosophy thesis.


Name of Candidate: Yermek Kapushev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Gaussian process models for large-scale problems

Supervisor: Associate Professor Evgeny Burnaev

Name of the Reviewer: Prof. Andrzej Cichocki

<p>I confirm the absence of any conflict of interest</p> <p>(Alternatively, Reviewer can formulate a possible conflict)</p>	<p>Signature:</p>  <p>Date: 15.01.2021</p>
---	---

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 "Gaussian Process Models for Large-Scale Problems" by Yermek Kapushev

PhD thesis of Yermek Kapushev is devoted to modern machine learning methods, especially kernel methods. Yermek focused on regression problems for large scale data sets and development novel methods to build exact and approximate Gaussian Process (GP) models. One of the developed by him method, is quite flexible and general since it can be applied to almost any regression problem, while another alternative approach has some restrictions on the data set structures but it is more efficient. Based on the proposed approaches Yermek developed various algorithms for 3 different tasks: Tensor completion, density estimation using score matching, and SLAM (Simultaneous Localization And Mapping).

The main task was to build Gaussian Process Regression model in a computationally efficient way. Yermek considered two scenarios: In the first one a data sets is a grid and objective was to perform an exact inference for the Gaussian Process model. In the second one there was no restrictions on the data set but objective was to find an optimized approximate solution. The considered research topics and problems are quite important and interesting since GP models can provide uncertainty estimation, which is crucial in some applications. Moreover, uncertainty estimation is required in many other problems, e.g., Bayesian Optimization. Moreover, the GP and kernel methods can be used in combination with other machine learning techniques (especially deep neural networks).

The main motivation of the thesis was to pursue such kind of research that arises in practical applications. Particularly, the author of the thesis considered practical application of his methods to robotics, where it is necessary to estimate the robot trajectory and the map simultaneously. He demonstrated that the developed by him approach unifies and extend some state-of-the-art works in these areas. Also, the case of grid with missing points has been so far rarely considered in the literature, so, his thesis to some extent fill up this gap at least for the Gaussian Process.

I would like to emphasize that Yermek investigated his methods and algorithms for dozens of well-known and difficult benchmarks and data sets (synthetic as well as real-world data). The performance and efficiency of the proposed algorithms were compared to a number of state of the arts algorithms. The developed method was also applied to real-life data sets (e.g., KITTI data set that is used in robotics, rotating disc problem and CMOS ring oscillator).

The method for approximate inference using randomized feature maps is provided by the author on github <https://github.com/maremun/quffka>. It was implemented purely in Python. The method for data sets on grid is implemented in pSeven software package (<https://www.datadvance.net/ru/product/pseven/>) developed by DATADVANCE .

In my opinion, the most original and new results are related to exact inference for the data sets on a grid with missing points and also approximation technique of the kernel function. The author showed that some existing methods can be interpreted as special cases of his method with a suitable selection of specific parameters. The most significant results, in my opinion, include development and implementation of algorithms for kernel approximation. These algorithms indicate excellent performance on benchmarks.

Yermek is leading co-author of 5 papers, including paper presented in top A* conference NeuroIPS and paper published in prestigious SIAM Journal on Scientific Computing. Furthermore, one of his paper was submitted to Neurocomputing journal. All papers are high quality.

The PhD thesis have also some minor limitations or weakness. First of all, the developed methods do not fully cover the case of highly multi-dimensional output. The author attempted to build kernel approximation method that minimizes the kernel approximation error. However, for downstream tasks (regression/classification) it turns out that one needs to minimize other objective functions to obtain better accuracy. These are challenging and quite difficult tasks which should be addressed in a future research. I think, that the most promising research direction would be to combine the methods and ideas presented in this thesis together with the other approaches like a data-dependent approach which would allow to improve further performance for the wide class of data sets.

In summary, the main achievements of the author are as follows:

- Development of the kernel approximation method.
- Rigorous derivation error bounds for the approximation of the kernel function, and demonstrate the connection between the proposed approach and some selective state-of-the art machine learning methods.
- Development of computationally efficient approach for exact inference of Gaussian Process Regression model in case of data sets on grid.
- Development of computationally efficient approach for exact inference of Gaussian Process Regression model in case of data sets on grid with missing points.
- Development of new algorithm for tensor completion, density estimate and SLAM based algorithm.

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