Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Nikita Klyuchnikov

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Multi-fidelity Classification and Active Search

Supervisor: Associate Professor Evgeny Burnaev, Skoltech

Name of the Reviewer: Maurizio Filippone

I confirm the absence of any conflict of interest	**Signature:** Maurizio Filippone **Date: 24-11-2020**

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

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

The thesis is well written and easy to follow. The structure into 5 chapters makes sense; chapters 1 and 5 contain introduction and conclusions, while chapters 2-4 contain the contributions of the thesis.

The relevance of the topic of dissertation work to its actual content

The thesis investigates a very important problem in Machine Learning and Statistics, that is the one of multi-fidelity. In the multi-fidelity context, where high-fidelity, precise, expensive data or simulations are combined with low-fidelity, fast, cheap data or approximate models. The aim is to combine these in meaningful ways so as to accelerate the learning task compared to the use of individual data/models. In addition to the study of modeling techniques, the thesis considers the problem of active learning, whereby observations are iteratively collected so as to optimize some criteria of interest, such as reducing uncertainty in estimates of model parameters. Furthermore, the thesis considers scenarios where data is heterogeneous, requiring a mix of techniques to handle, e.g., discrete search spaces and structured data.

This problem appears in a variety of relevant applications where quantification of uncertainty matters, or where quantification of uncertainty is necessary to drive the process behind active learning. Therefore the development of solutions to this problem based on advanced modeling in Bayesian Machine Learning is absolutely necessary. The thesis makes significant advancements in this direction. In addition, the thesis considers some compelling applications demonstrating the usefulness of the proposed methods.

The relevance of the methods used in the dissertation

Overall, the thesis builds upon recent advancements in the Machine Learning literature on Gaussian processes and multi-fidelity, and it is therefore of high relevance for this literature. Furthermore, the applications show compelling results, indicating that these approaches might have a good impact on applied disciplines too. Here is a detailed breakdown of the relevance of the contributions for each chapter.

Chapter 2 proposes a novel model for multi-fidelity classification, which is based on co-kriging and Gaussian processes. The proposed multi-fidelity classification model based on co-kriging of latent Gaussian processes is more robust to noise in low-fidelity data source and it is competitive with state-of-the-art alternatives. Chapter 2 develops an approximate Bayesian inference method based on the Laplace approximation, with the premise that this is faster than MCMC-based approaches. Results indicate that the proposed approximations achieves comparable predictions at a lower cost. Finally, the chapter concludes that the proposed algorithm for multi-fidelity active search surpasses single-fidelity methods when the correlation between data sources is high by delivering more relevant results within the same budget on evaluations.

Chapter 3 develops a multi-fidelity classification and active search framework, which integrates user feedback with information from an internal evaluation function via cokriging Gaussian interpolation. This framework allows one to reduce the interaction with the user. The proposed method enjoys some interesting scalability properties, while being limited by the time complexity is cubic only in the size of the set of scored nodes, which in any case is kept small. The experimental campaign on real and simulated user data shows performance above the state-of-the-art, delivering highly relevant information after a few user interactions.

Chapter 4 delves into a number of applications in particle Physics, Geo-statistics,

A notable output of the thesis, along with the interesting publications, is the addition of the proposed methods in software packages for data-driven oil-field development and neural architecture search.

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

The results obtained in the thesis are significant. In particular, they indicate that even though approximate methods are used, they yield results close to the gold-standard offered by Markov chain Monte Carlo methods. The methodological works appear in selective venues in Machine Learning, suggesting that these are of high international standards.

The relevance of the obtained results to applications (if applicable)

The results on Geo-physical applications seem to have more maturity and validation compared to the ones in particle Physics. It would be interesting to see whether the results in section 4.1 will enable some interesting research or discoveries in this domain.

The quality of publications

As already mentioned, the methodological contributions appear in a number of selected venues in Machine Learning, such as Neurocomputing and the KDD conference. This suggests that the advancements proposed in the thesis are of international standards. The results on applications, have been published in applied journals and conferences, which, to the best of my knowledge, are of high quality.

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

The thesis work hinges on the idea of developing a Bayesian modeling framework based on Gaussian processes, which is then approximated using the Laplace approximation. I think this is well motivated and realized, but I'm wondering about the possibility to use other approximations, which have recently found a reasonable success in the literature of Gaussian processes. For example, sparse variational Gaussian processes are widely considered as the state-of-the-art, so it is natural to think that this could be a competitor of the proposed approach. Also, sparse variational Gaussian processes are implemented using automatic differentiation, and I wonder whether this could not be the case for the approach proposed in the thesis. This would considerably simplify the implementation. Perhaps also combining the Laplace approximation with random feature expansions for Gaussian processes could be something of interest?

I have a final comment on alternative ways to approach multi-fidelity, by building on top of the works on multi-task learning (Bonilla and Williams, NIPS 2007) and works that mix Gaussian processes with Physics-based processes (e.g., Kennedy and O'Hagan, JRSS-B 2001). The thesis briefly touches upon these points, but it would be perhaps interesting to elaborate more on these points.

Finally, I'd suggest extending the conclusions by adding some comments on future perspectives for mult-fidelity learning in the context of current trends in Machine Learning.

Minor comments:

- In a few places, I found the English a bit too colloquial, e.g., page 5 uses the word "weird". Also, the readability of some sentences could be improved. Perhaps, a quick proofreading from a native English speaker would help to smooth some of these things out, although the thesis is generally well-written.

- I'd structure sec 4.1 like the others, e.g., adding a subsection about data, conclusions etc…

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