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

Name of the Reviewer: Prof Yarin Gal

I confirm the absence of any conflict of interest Yes (Alternatively, Reviewer can formulate a possible conflict)	**Signature:** **YG** **Date: 03-01-2021**

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

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

The main motivation in the text is the problem of active search - search algorithms that actively interact with the user to choose new objects, and ask the user to evaluate them; they then adapt to the user taking into account user feedback. This leads to the problem of multi-fidelity modelling with categorical outcomes (instead of continuous outcomes, as in current literature). The aim of the work is therefore presented as the development of a multi-fidelity active search framework, posing an interesting problem. The proposed solution involves a new Gaussian process-based Bayesian inference scheme for classification when data comes from several sources with different amounts of noise in the labels, as well as a new active search method.

The quality of the thesis seems good and the overall structure is sound (perhaps the derivations in 2.2.3.1 should be moved to an appendix). The topic (ML, BO) seems relevant to thesis contents, and the methods used in the dissertation seem relevant to the topic chosen (see some comments). The scientific significance of the results is evident from the industrial projects demonstrating the utility of the proposed method in reducing the amount of manual labour for datasets annotation compared to existing alternatives. The results' compliance with international level and current state of the art is satisfactory (although, see notes). The obtained results seem to be of relevance to the discussed application. The quality of publications is high, with numerous high-impact conferences. A summary of issues to be addressed is given below.

Some suggestions:

* If you start from the problem of "let's do classification with multi-fidelity data", then you need to justify why you use GPs to solve the classification problem. GPs are used in the regression case presumably because they offer an analytic solution, which GP classification losses. At the moment the introduction reads like the line of thought was "people use GPs for continuous multi-fidelity data, so let's extend these same GPs to classification" without giving much thought to what the best solution for the downstream would be. Eg, BNNs (or other tools) with HMC inference will also be data efficient and presumably give better trade-offs in certain situations. This needs to be discussed.

* Also, the justification for the Laplace approximation is a reference from 2008. There are many modern techniques that supersede it, eg MC integration of the likelihood using Titisias' VI approximation [see Hensman15 I think]. You need to justify why GP Laplace approximation is appropriate in the context of modern research, and why it preserves the properties of interest (eg data efficiency).

* Some points of ambiguity - is active search in chapter 3 using continuous multi-fidelity GPs? ("so as to learn a continuous relevance score function that captures the user's interests"). The intro motivates the introduction of discrete multi-fidelity GPs exactly because of the problem of active search. There are also numerous language issues.

Provisional Recommendation

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

V *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*