Dr. Aditya Krishna Menon
Google Research, NYC
adityakmenon@google.com

28th July 2025

Reviewer's opinion
on Ph.D. dissertation authored by
Marek Wydmuch
entitled:
Addressing the long-tail problem in extreme multi-label classification

1. Problem and its impact

The dissertation considers the problem of *multi-label learning* in the presence of *label skew*. Multi-label learning is a fundamental problem in the area of supervised machine learning, and concerns settings where the number of candidate labels is of high cardinality (e.g., order of millions). Label skew further refers to settings where the vast majority of candidate labels appear as "positive" for only a small fraction of examples. Such settings frequently arise in real-world applications (e.g., information retrieval, wherein many documents are only associated with a tiny fraction of queries), and have been the subject of sustained theoretical and empirical study in the machine learning community for more than a decade.

2. Contribution

The dissertation presents novel analyses and techniques for improving both the quality and efficiency of multi-label learning methods under label skew. In particular, the dissertation characterises the theoretical optimal solutions for a broad class of *instance-wise* and *label-wise* metrics for this problem (Chapter 3, 4, 5), and provides efficient algorithms for performing training (Chapter 6, 7) and inference (Chapter 8) with them. These provide a systematic, unified treatment of foundational problems in multi-label learning.

Central to the dissertation's approach is the theoretical observation that the studied metrics admit optimal solutions that are derivable from the *marginal conditional label probabilities*, $\eta_i(x)$. Accurate estimation of these probabilities is shown to provably translate to good performance according to the studied metrics (e.g., Theorem 6.6.1). This significantly simplifies the effort required in multi-label learning, as the naive alternative of estimating the *joint* label vector probability is intractable. Even so, when the number of candidate labels is large, even a linear dependence on the label size can be infeasible. To this end, the dissertation takes care to design practically effective algorithms that can enable *sub-linear* complexity (e.g., Chapter 8). This combination of theoretical grounding and empirical effectiveness is highly commendable.

The dissertation is based on several (5+) published works in prestigious machine learning venues, including NeurIPS, ICML, and ICLR. Several of these works have received a solid number of citations (e.g., >100 for the NeurIPS 2018 paper, "A no-regret generalization of hierarchical softmax to extreme multi-label classification"). These serve as further validation to the novelty and significance of the work, as judged by the broader academic community.

3. Correctness

To the best of my knowledge, the results claimed in the dissertation are correct. I have checked the proofs for the major theoretical results (most of which have been published in peer-reviewed conferences), and did not find any logical or mathematical errors.

A salient feature of the dissertation is that it presents a clean, precise mathematical framework for studying a range of different metrics (e.g., instance precision@k, macro precision@k): each thread of analysis starts with a clear statement of the population-level metric, a quantification of the theoretical optimal solution, and then a means of effectively estimating this solution from finite samples. This provides a simple but remarkably powerful framework for systematically breaking down and studying complex problems.

The code for replicating the experimental results in Chapter 9 has been provided, although I did not independently verify it. Generally, the results here made sense and were in line with the theoretical analyses. Further, the results reported for the baseline methods are in line with what is generally understood in the literature. Consequently, I tend to have high confidence that the experimental results are also correct.

4. Knowledge of the candidate

The dissertation demonstrates a solid general knowledge in the discipline of Information and Communication Technology, more specifically, Machine Learning and Computer Science.

Chapters 1 and 2 cover background material in the area of multi-label classification, and do a commendable job of summarising the formal statement of the problem – including subtleties such as missing data, and expected test utility versus population utility – and surveying relevant prior work. In the course of this analysis, the candidate shows familiarity with core machine learning concepts such as population loss, regret bounds, generalisation bounds, marginal versus joint probabilities, and so on. The subsequent chapters build on and add to this formalism.

The dissertation's references cover the relevant, important works in the area. I did not find any essential citations that were missing. The candidate may *optionally* consider citing works that relate to the long-tail learning setting in multi-*class* classification, e.g.,

Cao et al., Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss, NeurIPS 2019

The candidate may also *optionally* consider citing and discussing works related to "negative mining", which are generally employed to ensure computational efficiency of training methods for optimising recall@k and related metrics. e.g.,

Reddi et al., Stochastic Negative Mining for Learning with Large Output Spaces, AISTATS 2018
Karpukhin et al., Dense Passage Retrieval for Open-Domain Question Answering, EMNLP 2020

Certain forms of negative sampling also have a connection to *multi-class long-tail learning*, as studied in, e.g.,

Rawat et al., Disentangling sampling and labeling bias for learning in large-output spaces. ICML 2021.

5. Other remarks

The candidate may wish to consider the following:

- Page 26, Equation 2.5, it was a little unclear initially that the first part of the equation specifies the rows of the matrix.
- Page 39, there is probably a typo in the definition of ILIR: should the "P" in the denominator be omitted?
- Chapter 4 introduces the "missing labels" setting for instance-level losses. It is slightly asymmetrical, in that there is not an analogue for the label-level losses. Is this intended? Perhaps, it is worth including a comment on this point after Chapter 5.
- Chapter 5 and Chapter 6 could perhaps be merged, as currently Chapter 5 is fairly short compared to the other chapters.
- Page 61, at this stage the distinction between the macro Precision@k and the instance-level Precision@k studied in Chapter 3 could be useful to spell out explicitly.
- Section 6.7, the analysis of Coverage@k is nice. It is not clear though, whether there are a more broader class of discontinuous measures which also admit such an analysis?
- Chapter 10, a brief mention of open questions for future work to consider could be valuable.
- Page 137, Guo et al. citation appears to be repeated.

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the A*ct of 20 July 2018 - The Law on Higher Education and Science* (with amendments)[1], my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

Moreover, taking into account the impressive breadth and depth of its contributions – spanning both theoretical and empirical advances in the fundamental study of multi-label learning – I **recommend to distinguish** the dissertation for its quality.

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276

Aditya K. Menon

Signature