ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Rectifying Probabilistic Predictions of Neural Networks**
Author's name:	**Bc. Tuan Anh Ho**
Type of thesis :	Master's thesis
Faculty/Institute:	Faculty of Electrical Engineering
Department:	Dept. of Cybernetics
Thesis reviewer:	Ing. et Ing. Milan Šulc, Ph.D.
Reviewer's department:	external (Rossum.ai)

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — **Extraordinarily challenging**

Evaluation of thesis difficulty of assignment.

The thesis assignment is very challenging, as it requires a deep understanding of deep learning and commonly used training objectives, different cases of distribution shifts, Bayesian decision theory, etc. It required the student to apply and extend theoretical concepts of calibrating predictions to several practical cases, where miscalibration leads to suboptimal decision making.

Satisfaction of assignment — **Fulfilled**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

The thesis fulfills all points of the assignment and proposes two novel loss functions – Direct Loss and Integral Loss - towards the goals of the assignment. While the experimental results do not demonstrate the superiority of the proposed loss functions in most scenarios, the negative result does not lower the quality of the student's work.

Method of conception — **Outstanding**

Assess that student has chosen correct approach or solution methods.

The thesis addresses calibration by focusing on the downstream tasks, where miscalibration (even beyond the top-1 prediction) can negatively impact decision making. The proposed methods and their derivations are well theoretically grounded.
A minor remark/question from my side is whether it's suitable to use the term "calibration", commonly used for generally improving the reliability of probabilistic estimates, for the tuning of the outputs for specific downstream tasks and decision problems.

Technical level — **A - excellent**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

The thesis is very technically sound, it is building on state-of-the-art literature, and the proposed methods are based on strong theoretical foundations.
A minor remark from my side - rather a preference of interpretation: The thesis text considers cross-entropy a surrogate to the 0/1 loss (wishing to train a decision maker). My preferred interpretation (the MLE is also mentioned in the thesis) is: cross-entropy minimization is the maximum likelihood estimation of parameters of an estimator of a-posteriori probabilities, the goal does not have to be to train the final decision maker.

Formal and language level, scope of thesis — **C - good**

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

My only criticism can go to the language and formal level of the thesis: The English level would benefit from careful proofreading and review. First, more focus on stylistics in technical writing might help the overall readability of the thesis. Second, using a spell checker could identify typos (e.g. "teh true risk").
Additionally, a careful review could identify shortcomings in claims – Proposition 2 (page 28) says: "Decision calibration by

Zhao et al. (2021) is not necessary for decision-making." The author probably wanted to propose it is not necessary for *optimal* decision-making (as it is not surprising that no calibration is necessary for non-optimal decision making).

Selection of sources, citation correctness **A - excellent**

Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.

The thesis works with a good amount of related work, including state-of-the-art publications, many from 2021-2022. All sources are referenced.
Minor remark: The author's ICLR submission is cited as "Tuan Anh Ho, Jiri Matas, and Alexander Shekhovtsov. Calibration for decision making via empirical risk minimization. 2022.", but without any reference - e.g. an URL to technical report or OpenReview would be useful.

Additional commentary and evaluation

Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

The proposed methods and the research publication submitted to ICLR 2023 are an impressive result of a master project. I hope the (partly) negative experimental results will not demotivate the student in his future research endeavors.

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

With some remarks to the language level, showing a potential area for improvement to the student, the thesis fulfills all points of a challenging assignment, and I appreciate the novel calibration loss functions, proposed based on strong theoretical foundations while motivated by practical downstream tasks.

I evaluate the handed thesis with classification grade **A - excellent.**

Questions for the defense:

I agree with the statement that probabilistic predictions (e.g. trained by cross-entropy minimization) can become misleading as a result of overfitting the models during training.
A trivial question: Can you illustrate why overfitting to training data typically leads to overconfident predictors?
A non-trivial question: How does the calibration error term in Eq. 2.17 for "training a calibrated neural network" as proposed by (Kumar et al., 2018) help, when it is computed on the same training data that the predictor overfits?

Date: **February 2, 2023** Signature: