CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS REVIEWER’S REPORT

I. IDENTIFICATION DATA

Thesis title:	Deep Reinforcement Learning on a Modified Car Racing Environment
Author’s name:	**Vojtěch Sýkora**
Type of thesis:	bachelor
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	13133
Thesis reviewer:	RNDr. Martin Suda, PhD
Reviewer’s department:	CIIRC, CTU

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

The thesis deals with the topic of reinforcement learning and the task is 1) to modify the CarRacing environment from Open AI’s Gym to include the aspect of an adjustable wind, 2) to train an agent for the modified environment using the PPO algorithm, 3) to run experiments with the obtained agent(s), observe and analyze the results. This task is potentially challenging due to the theoretical complexity of the PPO algorithm (however, it can also be just applied as a black box) and of the non-trivial environment (which includes physically realistic aspects, such as friction).

Fulfilment of assignment **fulfilled with minor objections**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

The thesis fulfils all the assigned tasks. A minor objection (which could, however, also be regarded as a clever workaround) is that the author decides to simulate wind using a convenient “action wrapper” of the environment, which can modify the steering aspect of each action to go “more to the left” or “more to the right” than intended. This skips the complexity of modifying the environment itself and we end up with a “wind” which is always perpendicular to the car’s direction (rather than a more realistic wind blowing from the “north”/”south”/”east”/”west” as might have been originally imagined. In other words, the wind we get is acting as a force on the steering wheel rather than on the car itself.

Methodology **correct**

Comment on the correctness of the approach and/or the solution methods.

The methodology appears correct and the solution methods are appropriate.

Technical level **B - very good.**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

The technical level is very good. The student is showing good understanding of the topic.

Formal and language level, scope of thesis **A - excellent.**

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The thesis is well-structured into several self-contained chapters and is easy to follow. The language quality is high, with very few typos and minor stylistic transgressions.

Minor points:

- The figure on page 13 could be larger for better readability
- A few inconsistencies in mathematical notation (\$\times\$ vs \$*\$, - vs –)

Selection of sources, citation correctness **B - very good.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the *student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?*

References to related work are adequate. The bibliography could be improved by adding the proceeding/journal names where missing (e.g., "Adam: A Method for Stochastic Optimization. ICLR (2015)") or dates of access for the urls.

Additional commentary and evaluation (optional)

Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.

I enjoyed reading the thesis and learned new things about this exciting and truly modern topic, which deep reinforcement learning is. I appreciate author's bravery in trying to explain PPO in some detail (which is challenging even for the experts), also the hyperparameter summary of chapter four was informative, and the exploration of the behavior of agents' snapshots throughout training led to some interesting observation.

What felt strange to me was the separation of the left and right wind as distinct categories, despite the clear mathematical symmetry that could have also been reflected in the implementation. I think in not pursuing the symmetry also lies the explanation for why the right wind environments were in the experiments easier to both learn and navigate, as explained in detail below. And this in turn probably stems from author's (in my opinion premature) eagerness to normalize the steering action into the interval [0,1] with "go straight" represented by 0.5. ("This removes negative numbers from our equations which is often a well-received change.")

In the end, the left wind is implemented by scaling down the issued value from [0,1] by a factor c from $0<c<1$. This makes sharper right turning impossible (or even going straight if $c < 0.5$) for the agent. On the other hand, right wind is implemented by multiplying the issued value from [0,1] by a factor $d > 1$ and clipping back to [0,1]. So, all possible values of turning are still available to the agent after the transformation (although, e.g., "straight" is now issued by aiming for 0.5/d). I was a bit disappointed that this observation did not show up in the analysis.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

I sum, I think there is a solid piece of work behind the thesis, and I suggest the grade B.

Possible questions:

Q1: Did you consider directly modifying the environment's source code to implement a physically more realistic wind? How much more difficult than using action wrappers it would be?

Q2: You at some point mention that "PPO does not have memory". I would argue that PPO, as a mere training algorithm, is not to blame. Could you think of which other aspect of the environment would have to be modified to simulate memory? (This could, by the way, be quite advantageous for the agent, because with the used setup it has no chance of knowing its current speed, which is not ideal when solving a task such as driving.)

The grade that I award for the thesis is **B - very good.**

Date: **7.6.2023**

Signature: