Reinforcement Learning Feedback

Dr Chris G. Willcocks and Dr Robert Lieck

Last Modified: March 20, 2023

Individual feedback and marks

Here are your (cvhm34) marks for deep learning, reinforcement learning, and your final grade:

Percentages	
Convergence:	85.89%
Sophistication:	80.0%
Video intuition:	90.0%

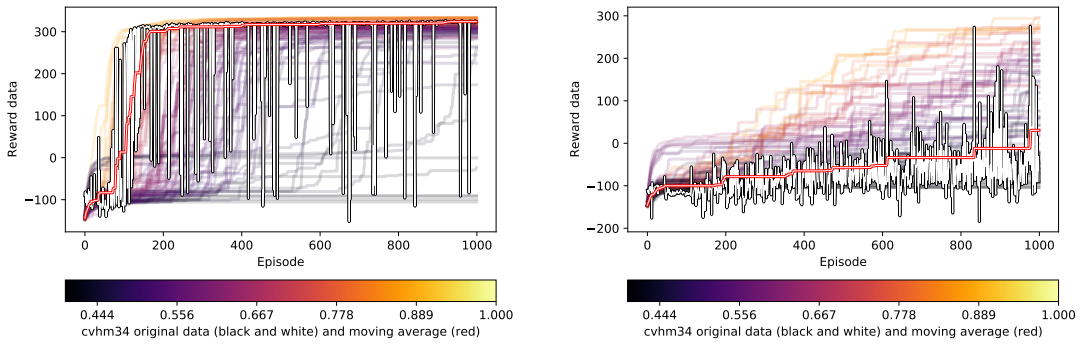
Student code:	cvhm34
Convergence marks:	43/50
Sophistication marks:	24/30
Video marks:	18/20
Final grade:	**85**/100

Comments

Normal environment: The agent has outstanding convergence, ranking within the top 20% of the class. It achieves very high scores extremely quickly within relatively few episodes. *Hardcore environment convergence:* The hardcore agent has very satisfactory convergence, but it could definitely be more sample efficient as it still takes a relatively lot of episodes before it starts to get a good score. *Sophistication and mastery of the domain:* The report gives background on the SAC method and describes issues with stable convergence. No other methods are tested and only little experimentation was performed to further increas performance, but the SAC method show very good convergence in some runs. *Intuition of the video:* The normal agent learns an excellent policy that is moving fast and smoothly. The hardcore agent also learns a reliable policy that passes all obstacles.

Convergence data

These graphs show your individualised convergence data:



The left chart shows your (cvhm34) personalised log data relative to everyone else for the normal environment. The right chart shows the same for the hardcore environment. The colour in both graphs relates to your final convergence mark.

General feedback

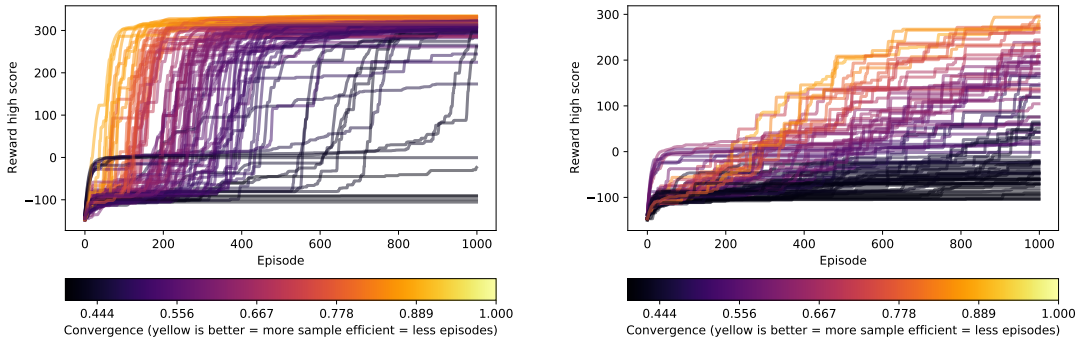
We're happy that the level of successful reinforcement learning solutions was very high this year, where nearly everyone was able to train an agent that was eventually able to learn to walk.

The successful strategy for a 2:1 was to implement the recommended TD3 agent and then carefully tune the hyperparameters inlight of exploration-exploitation and the overestimation bias issues. Several students attempted to alter the agent by incorporating novel ideas from other papers before doing robust parameter sweeps, and were often beaten in the convergence ranks by well-tuned baseline TD3 implementations.

A large group of students implemented FORK, a forward looking extension to TD3. This generally did well in the hardcore environment, where it could anticipate objects, but tended to not perform as well as people who focused on more rigorous optimisation of the original TD3 within the normal environment.

The students with the most efficient agents generally surveyed the recent literature that addresses overstimation bias with a focus on high-quality off-policy learning. In particular, they used distributional RL approaches such as TQCs, REDQs and ACCs with ensembles of distributional critics. They also tended to have rigorous experimental setups that considered AUCs over different intervals to predict a high-quality estimate of the final convergence from early on in training.

Marking of the convergence ranks was automated based on the log data. Can you reverse engineer the score function based on the figure?



These graphs show the results for everyone who submitted valid log data for the normal environment (left) and the hardcore environment (right). The interval on the x-axis is different to represent the task difficulty.

Closing Comment

We hope you found this assignment rewarding and that it has helped set realistic expectations into what can and cannot be achieved with these large models. We also hope that you appreciate the importance in keeping up-to-date with the never-ending stream of new literature in these fast-changing fields.
