Course Project

Code of Honor. All external resources used in the project, including research papers, open-source repositories, datasets, and any content or code generated using AI tools, e.g., ChatGPT, GitHub Copilot, Claude, Gemini, must be *clearly cited* in the final submission. The final report must also include *a clear breakdown of individual group member contributions.* Any lack of transparency in the use of external resources or in reporting group contributions will be considered academic dishonesty and will significantly impact the final evaluation.

Topic	*Trust Region* versus *Proximal* Policy Optimization
Category	Deep RL from Scratch

Objective Design and implement two advanced actor–critic policy optimization algorithms: *Trust Region Policy Optimization (TRPO)* and *Proximal Policy Optimization (PPO)* [2, 3]. Compare their stability, sample efficiency, and robustness in classic continuous control tasks.

Motivation TRPO and PPO are among the most influential Deep RL algorithms, widely used in both research and applications such as robotics [1]. TRPO introduced the concept of trust region optimization, while PPO simplified it with clipped objectives, making it practical and scalable. Implementing these algorithms from scratch will give hands-on experience with constrained policy optimization, surrogate loss functions, and stability techniques in Deep RL. You will also observe how theoretical ideas translate into practical design choices.

Requirements The final submission should address the following requirements while the details can be freely decided by the group members.

1. Implementation: in this respect, you should
   - Implement TRPO with its KL-divergence constraint and conjugate gradient solver for policy updates.
   - Implement PPO with clipped surrogate objectives.
   - Use an actor–critic structure with shared or separate neural network backbones for policy and value functions.
2. Environment modification: you can use a pre-implemented Gym environment that is *required to be controlled by Deep RL*. Note that basic environments that are handled easily by tabular RL are **not accepted**. To give the implementation some level of novelty, you **must** modify standard Gym environments with **at least one** of the following modifications:
   - Adding *noise* to observations.
   - *Perturbing* actions before execution.
   - Injecting irrelevant *distractor* features into the state vector.
   - Adding *stochasticity or delays* to rewards.

3. Evaluation: the final project should report key evaluation of the implemented algorithms in the modified environment. In this respect, the results should
    - compare TRPO and PPO learning curves,
    - quantify stability, convergence speed, and sample efficiency, and
    - report how environment modifications affect algorithm performance.
4. The results should be elaborated through
    - performing ablation studies, e.g., removing clipping in PPO or relaxing trust region in TRPO, and
    - providing discussion on trade-offs between complexity and performance.

Milestones The following milestones are to be accomplished through semester.

1. Literature Review and Setup
    - Review TRPO and PPO.
    - Choose environments and finalize environment modifications.
2. Implementation
    - Implement TRPO and validate on a simple environment.
    - Implement PPO and validate similarly.
    - Train both on modified environments.
3. Evaluation and Analysis
    - Collect and plot learning curves.
    - Compare TRPO vs PPO under normal and noisy conditions.
    - Perform ablation experiments.
4. Final Report and Presentation

Submission Guidelines The main body of work is submitted through Git. In addition, each group submits a final paper and gives a presentation. In this respect, please follow these steps.

- Each group must maintain a Git repository, e.g., GitHub or GitLab, for the project. By the time of final submission, the repository should have
    - Well-documented codebase
    - Clear `README.md` with setup and usage instructions
    - A `requirements.txt` file listing all required packages or an `environment.yaml` file with a reproducible environment setup
    - Demo script or notebook showing sample input-output
    - *If applicable,* a `/doc` folder with extended documentation
- A final report (maximum *5 pages*) must be submitted in a PDF format. The report should be written in the provided formal style, including an abstract, introduction, method, experiments, results, and conclusion.
  Important: Submissions that do not use template are considered *incomplete.*
- A 5-minute presentation (maximum *5 slides including the title slide*) is given on the internal seminar on Week 14, i.e., *Dec 1 to Dec 5,* by the group. For presentation, any template can be used.

Final Notes While planning for the milestones please consider the following points.

1. You are encouraged to explore innovative approaches as long as the core objectives are met.
2. While computational resources are limited, the chosen environment can keep training feasible. Trade-offs (e.g., fewer episodes, smaller networks) are expected and should be justified.
3. Teams are expected to manage their computing needs and are advised to perform early tests to estimate runtime and training feasibility. As graduate students, team members can use facilities provided by the university, e.g., ECE Facility. Teams are expected to inform themselves about the limitations of the available computing resources and design accordingly.

References

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[2] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, pages 1889–1897. PMLR, 2015.

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.