Use Of GenAI

This homework is completed with the help of Windsurf VS code extension.`https://windsurf.com/`

What is used:

- Autofill feature to generate syntactically correct latex code (each tab key pressed filled no more than 100 characters, at most 20% of the predicted text is adapted) for the homework with human supervision.
- Use AI to debug the latex code and find unclosed parentheses or other syntax errors.
- Use AI to autofill the parts that follows the same structure as the previous parts (example: case by case proofs).
- Use AI to auto correct misspelled words or latex commands.

What is not used:

- Directly use AI to generate the solutions in latex document.
- Use AI to ask for hint or solution for the problems.
- Select part of the document and ask AI to fill the parts missing.

1.3 Deliveries

1.3.1 Create two graphs:

- In the first graph, compare the learning curves (average return vs. number of environment steps) for the experiments running with batch size of 1000. (The small batch experiments.) (15 pts)

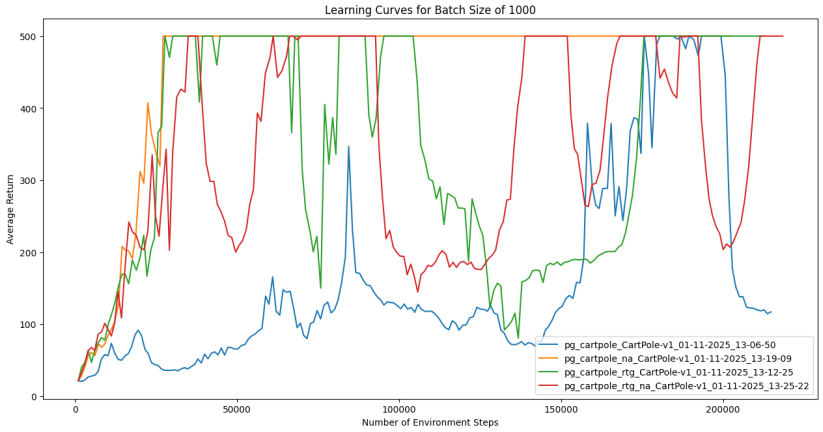

Figure 1: Learning Curves for Batch Size of 1000

- In the second graph, compare the learning curves for the experiments running with batch size of 4000. (The large batch experiments.) (15 pts)

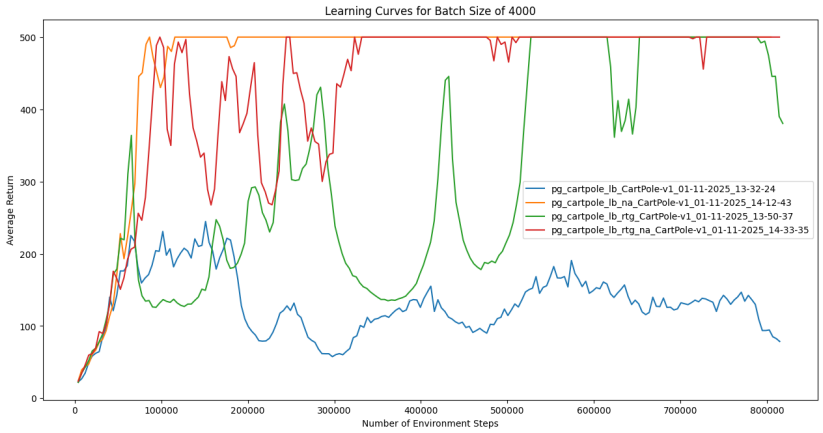

Figure 2: Learning Curves for Batch Size of 4000

Note that the x-axis should be number of environment steps, not number of policy gradient iterations.

1.3.2 Answer the following questions briefly:

Provide the exact command line configurations you used to run your experiments, including any parameters changed from their defaults.

The best configuration in both the small and large batch size cases should converge to a maximum score of 500.

- Which value estimator has better performance without advantage normalization: the trajectory-centric one, or the one using reward-to-go? Why? (10 pts)

  The reward-to-go one has better performance without advantage normalization.

  The reward-to-go has more fine-grained control over the learning process by using the rewards after the current timestep to estimate the Q-value for the current state-action pair.

- Did advantage normalization help? (10 pts)

  Yes, advantage normalization helps.

  The advantage normalization helps the learning process by stabilizing the learning rate and preventing the policy from overfitting to the data.

- Did the batch size make an impact? (10 pts)

  Yes, the batch size makes an impact.

  The larger batch size allows the agent to learn from more data in each update, which can help the agent to converge to a better policy, especially when the normalization and reward-to-go are used.