PA230 Training Report

Marek Ličko 536662

Team Name: M

Track 1

We used a Double DQN architecture with soft updates to mitigate Q-value overestimation, utilizing a two-layer MLP (128 neurons, ReLU). Training stability was ensured through several mechanisms. We used Polyak averaging ($\tau = 0.005$) for soft target updates, which gave smoother convergence compared to hard updates. Second, by scheduling epsilon decay per episode instead of per step, we ensured consistent exploration horizons. We performed gradient updates every 4 steps to improve computational efficiency and reduce correlation between consecutive updates.

The hyperparameters were picked based on well-performing solutions found online and modified through trial and error.

Track 2

Initial experiments with vanilla REINFORCE were had very high variance. We transitioned to PPO, which incorporates stability mechanisms. Restricted policy updates via a clipped surrogate objective ($\epsilon = 0.2$) were used to prevent destructive gradient steps. To achieve more accurate credit assignment, we balanced bias and variance using GAE ($\lambda = 0.95$). An entropy coefficient of 0.01 helped maintain exploration by preventing deterministic collapse, while advantage normalization stabilized gradients across batches.

The hyperparameters were picked based on well-performing solutions found online and modified through trial and error.

Track 3

We solved the visual control task using PPO with a CNN. A 3-layer CNN (16/32/64 filters) extracted features from 96×96 pixel observations and fed them into separate actor and critic heads. To improve efficiency, we implemented a frame skip ($k = 4$) wrapper, which increased the agent's decision horizon and sped up training 4x. Observations were normalized to $[0, 1]$ and transposed to chanel-first format for PyTorch. Experience collection occurred on CPU, and batch data was transferred to GPU only for training.

Track 4

Same as track 3.