Part III: Actor–Critic Results on Two Environments

1 Environments

Acrobot-v1

- **State space:** A 4-dimensional continuous vector $[\theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2]$, where θ_i and $\dot{\theta}_i$ are the angle and angular velocity of link i.
- **Action space:** Three discrete actions $\{0, 1, 2\}$ corresponding to applying torque -1, 0, $+1$ to the second joint.
- **Reward function:** A constant -1 reward at each timestep until the termination condition is met.
- **Goal:** Swing the lower link so that the tip of the second link rises above a fixed height threshold (0.5 m above the base).
- **Episode length:** Maximum of 500 steps; terminates early on success.
- **Agent:** Actor–Critic network with two hidden layers (256→128 units) feeding both policy and value heads.

BipedalWalker-v3

- **State space:** A 24-dimensional continuous vector containing hull angle, joint angles & velocities, two leg contact sensors, and 10 lidar rangefinder readings.
- **Action space:** Four continuous torques in $[-1, 1]$ for the hip and knee motors of each leg.
- **Reward function:**
  - Forward progress reward proportional to distance traveled.
  - Quadratic control cost penalty on torques.
  - Alive bonus of $+0.3$ per step.
  - Fall penalty of -100 if the hull hits the ground.
- **Goal:** Learn a stable walking gait to traverse as far as possible without falling.
- **Episode length:** Maximum of 1600 steps; terminates early on fall.
- **Agent:** Same Actor–Critic architecture as above, scaled to continuous actions.

2 Training Results

2.1 Acrobot-v1

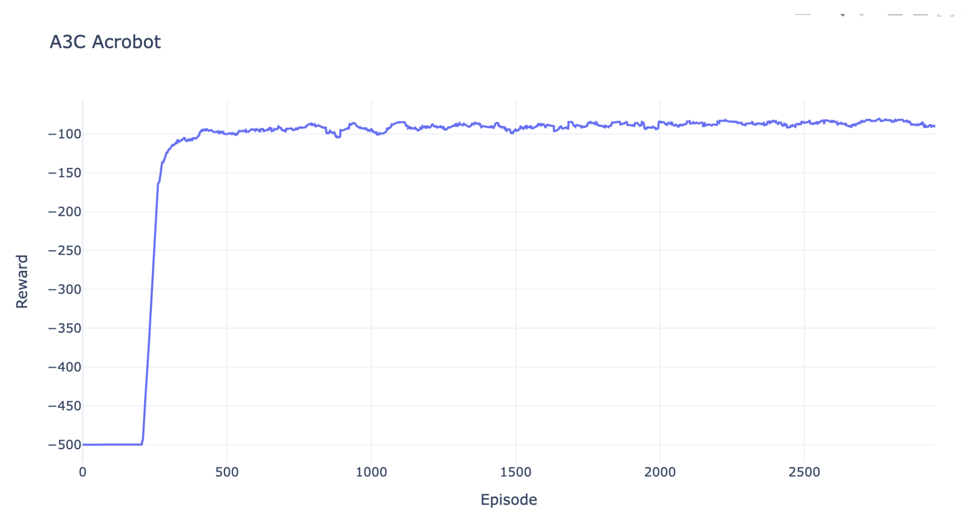

Figure 1: Training: episode return vs. episode number for `Acrobot-v1`.

Discussion: Returns start near -500 (worst case) and improve steadily. By around 400 episodes the average return surpasses -200, and by 3 000 episodes it plateaus near -90. Since each step incurs -1 reward, an average return of -90 means the pendulum is swung up in roughly 90 steps—i.e. the environment is effectively solved.

2.2 BipedalWalker-v3

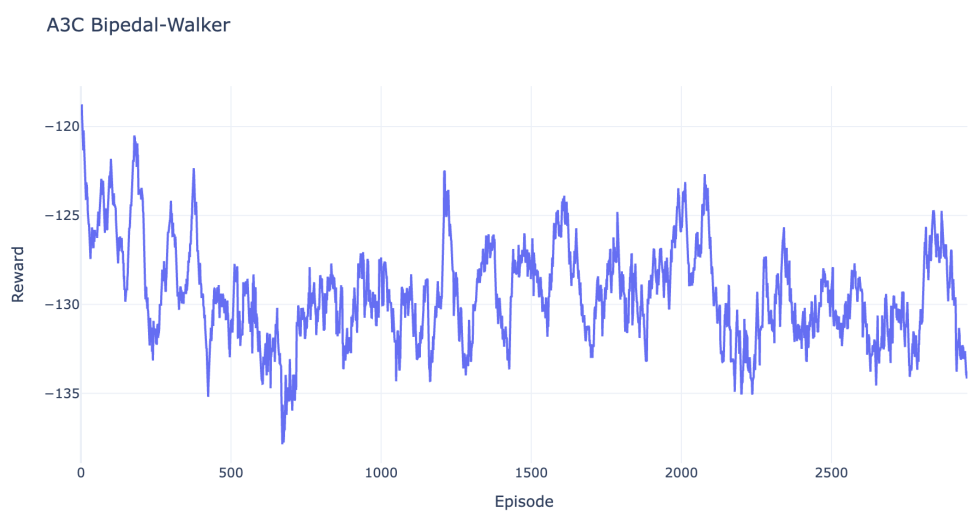

Figure 2: Training: episode return vs. episode number for `BipedalWalker-v3`.

Discussion: Training returns remain around -125 over 3 000 episodes, with only minor variance. The continuous locomotion task is far more challenging and learns very slowly. To attain a positive, stable walking policy typically requires $\geq 10^5$ episodes and more parallel workers; our 4-worker CPU-only setup was insufficient for deeper training.

3 Evaluation Results

3.1 Acrobot-v1

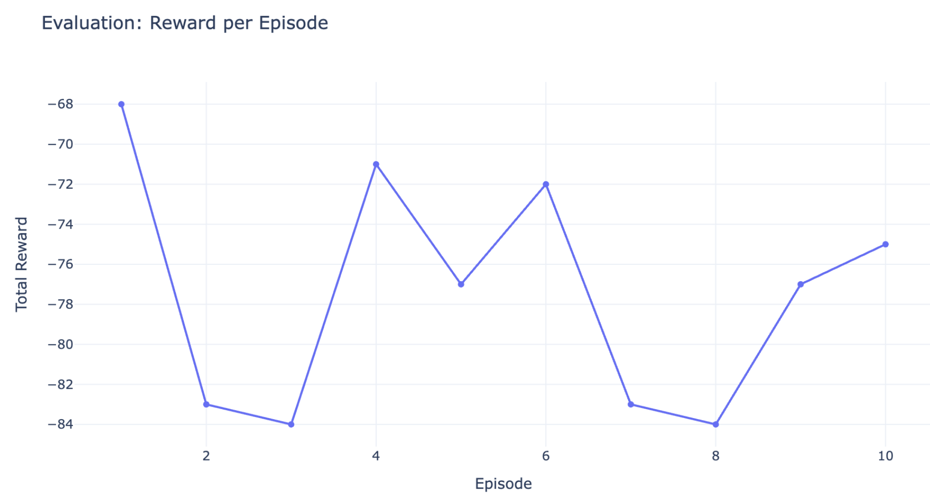

Figure 3: Evaluation: total reward per episode over 10 greedy runs for `Acrobot-v1`.

Discussion: In 10 greedy (exploration-free) episodes, the average return is -79.4, indicating the pendulum consistently reaches the goal in about 79 steps with low variance.

3.2 BipedalWalker-v3

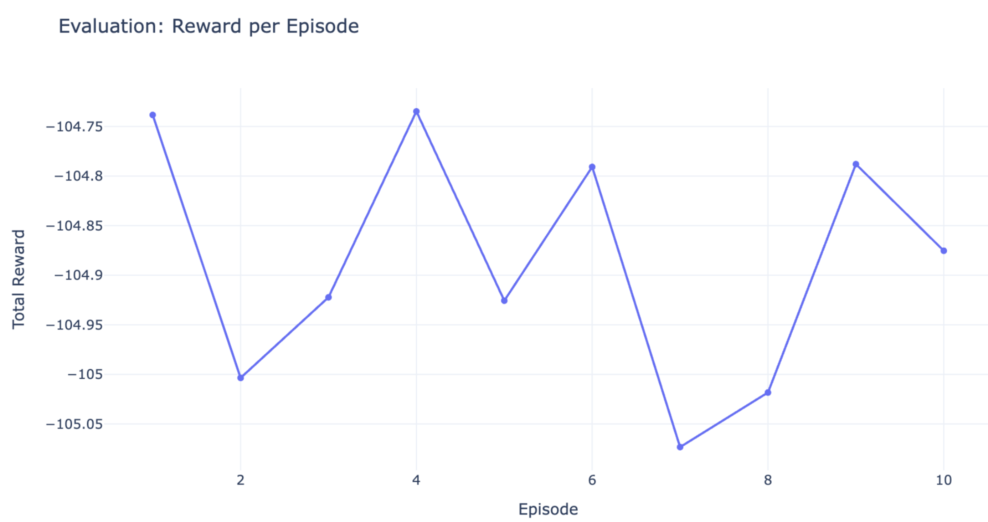

Figure 4: Evaluation: total reward per episode over 10 greedy runs for `BipedalWalker-v3`.

Discussion: Greedy evaluation episodes average around -104, showing no successful walking behavior. A competent goal typically yields rewards > 300, so much more training and compute are required.

4 Author Contributions

Contributor	Contribution (%)
Shreyas Bellary Manjunath	50 %
Ruthvik Vasantha Kumar	50 %

References

[1] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning," in *Proc. ICML*, 2016.

[2] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.

[3] OpenAI Gym Documentation, `https://www.gymlibrary.dev/environments/mujoco/`

[4] N. Heess *et al.*, "Emergence of Locomotion Behaviours in Rich Environments," in *Proc. CoRL*, 2017.