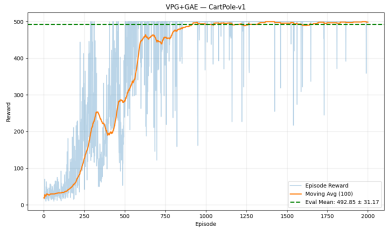
CS778: Foundations of Modern AI

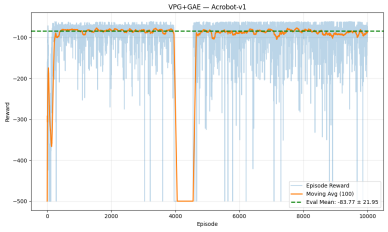
Assignment: Comparative Study of TRPO, PPO, and VPG on Five Benchmarks

Keyansh Vaish
Roll No: 220525

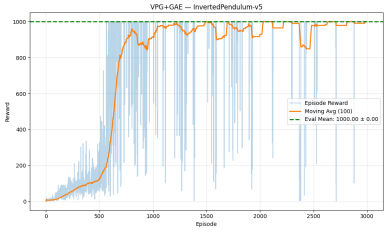
VPG



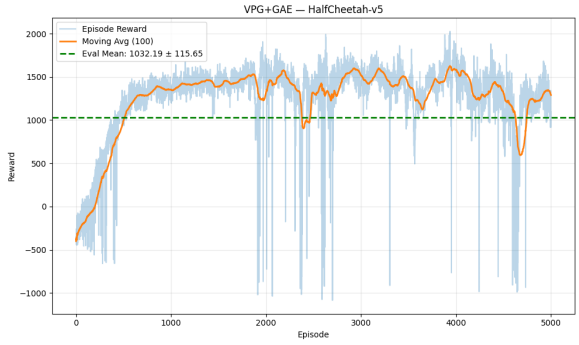
(a) CartPole-v1



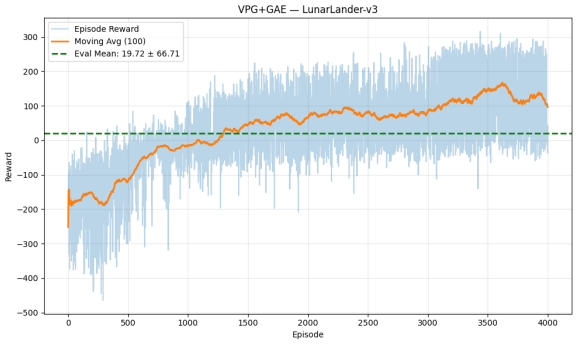
(b) Acrobot-v1



(c) InvertedPendulum-v5



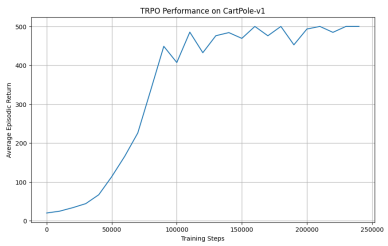
(d) HalfCheetah-v5



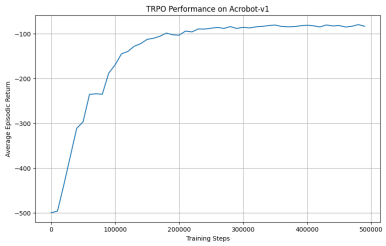
(e) LunarLander-v3

Figure 1: VPG performance across five datasets.

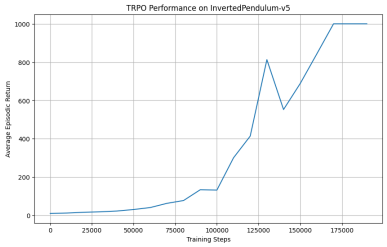
TRPO



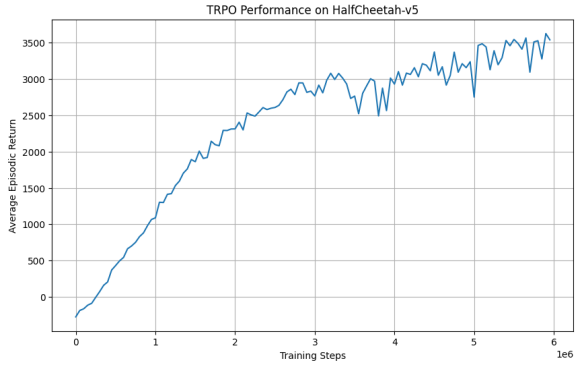
(a) CartPole-v1



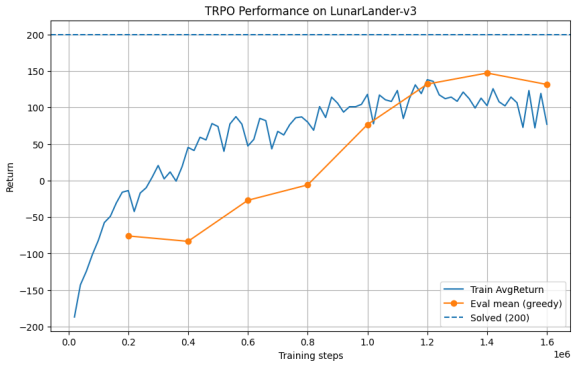
(b) Acrobot-v1



(c) InvertedPendulum-v5



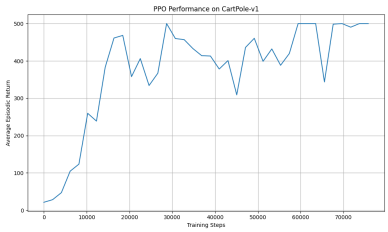
(d) HalfCheetah-v5



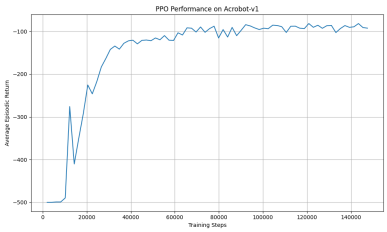
(e) LunarLander-v3

Figure 2: TRPO performance across five datasets.

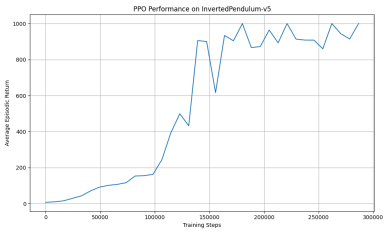
PPO



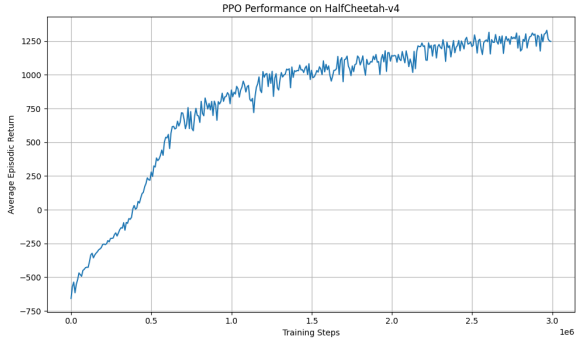
(a) CartPole-v1



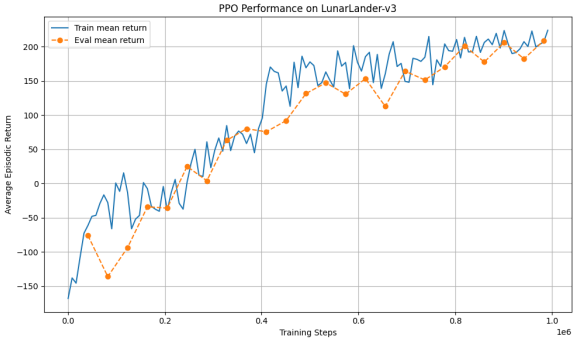
(b) Acrobot-v1



(c) InvertedPendulum-v5



(d) HalfCheetah-v5



(e) LunarLander-v3

Figure 3: PPO performance across five datasets.

Results

Trained VPG, TRPO, and PPO-clip with GAE for each of the five benchmarks using an early-learning budget of 10^4 environment steps per task. Trends are read from the learning curves in the figures above.

- **CartPole-v1**: All improve quickly. PPO and TRPO stabilize earlier. VPG shows higher variance.
- **Acrobot-v1**: Slower gains due to sparse reward. PPO and TRPO progress steadily. VPG plateaus intermittently.
- **InvertedPendulum-v5**: PPO/TRPO reach high returns early. VPG lags in sample efficiency.
- **HalfCheetah-v5**: Limited improvement. PPO rises fastest. TRPO is steadier. VPG often remains near the baseline.
- **LunarLander-v3**: Nonlinear reward shaping yields oscillations. PPO is most consistent. TRPO moderate. VPG noisy.

Across tasks in this budget, **PPO** shows the best stability and sample efficiency, **TRPO** is stable but sometimes slower, and **VPG** is most sensitive to hyperparameters.

Discussion

PPO learned fastest and most stably. TRPO was steady but slower. VPG was the noisiest. This matches an early-learning budget of $\sim 10^4$ steps where variance reduction and conservative updates help.

Why some results looked good

- *CartPole-v1:* Dense reward and simple dynamics. All methods improve quickly; PPO/TRPO stabilize early.
- *InvertedPendulum-v5:* Easy continuous control. PPO/TRPO reach high return fast; VPG lags due to high-variance updates.

Why some results underperformed

- *Acrobot-v1:* Sparse reward. Policies need more exploration; VPG stalls, PPO/TRPO inch up.
- *HalfCheetah-v5:* Progress is limited at 10^4 steps. High-variance, delayed rewards; critic underfit yields oscillations.
- *LunarLander-v3:* Shaped reward plus contact dynamics destabilize policy and value; curves oscillate.

How to improve without changing the setup

- Normalize observations and advantages; consider reward scaling/clipping on continuous tasks.
- Tune GAE and learning rates: $\gamma \approx 0.99$, $\lambda \approx 0.95$; if unstable, reduce LR by 2-10$\times$ and increase batch size.
- Keep entropy high early; decay slowly to avoid premature convergence.
- Strengthen the critic: larger value-loss weight or smaller critic LR; track explained variance.