Report

Sannikov Anton

February 19, 2026

Abstract

I compared four policy gradient algorithms on CartPole-v1 over 1000 episodes: vanilla REINFORCE, mean baseline, actor-critic, and RLOO. The actor-critic achieved perfect 500 average with zero variance in the last 100 episodes, while mean baseline came very close (499.5). Vanilla REINFORCE still showed some variance (465.5), and RLOO lagged behind (337.5). Behavior cloning from the best policy (critic) reproduced the expert perfectly, but cloning a poorly trained expert failed, confirming the limitations of imitation learning.

Experimental Setup

Each method was trained for 1000 episodes on CartPole-v1 (maximum reward 500). All agents share the same architecture: a two-layer MLP with 64 hidden units, discount factor $\gamma = 0.99$, Adam optimizer with learning rate 10^{-3}, and entropy regularization decaying linearly from 0.01 to 0. The advantage A_t was computed as:

1. Vanilla: $A_t = R_t$ (discounted return)
2. Mean baseline: $A_t = R_t - \bar{R}$, where $\bar{R}$ is the episode's average return
3. Critic: $A_t = R_t - V(s_t)$, with the value network updated five times per policy step
4. RLOO: $A_t = R_t - \frac{1}{T-1} \sum_{j \neq t} R_j$ (leave-one-out across time steps)

After training, the best policy (critic) was used to generate 50 expert episodes. A fresh policy was trained on these state-action pairs using cross-entropy loss for 20 epochs. To demonstrate BC's limitations, I also trained a poor expert for only 50 episodes and repeated the cloning process.

Results

Figure 1 shows episode returns over 1000 episodes. The critic method converges fastest and remains at 500 almost without exception after 400 episodes. Mean baseline also reaches near-perfect scores but occasionally dips below 500. Vanilla REINFORCE exhibits persistent variance, and RLOO struggles to exceed 400 consistently.

Table 1 presents statistics for the last 100 episodes (901–1000). The critic achieves a perfect mean of 500 with zero standard deviation. Mean baseline is very close (499.5) but not perfect. Vanilla still has noticeable variance, and RLOO lags significantly.

Behavior Cloning

The critic expert (mean 500 over last 100 episodes) produced a student that also scored a perfect 500 on 100 test episodes. The poor expert (trained for only 50 episodes) averaged 19.35, and its clone scored 19.75 — confirming that BC merely copies the expert's behavior, good or bad.

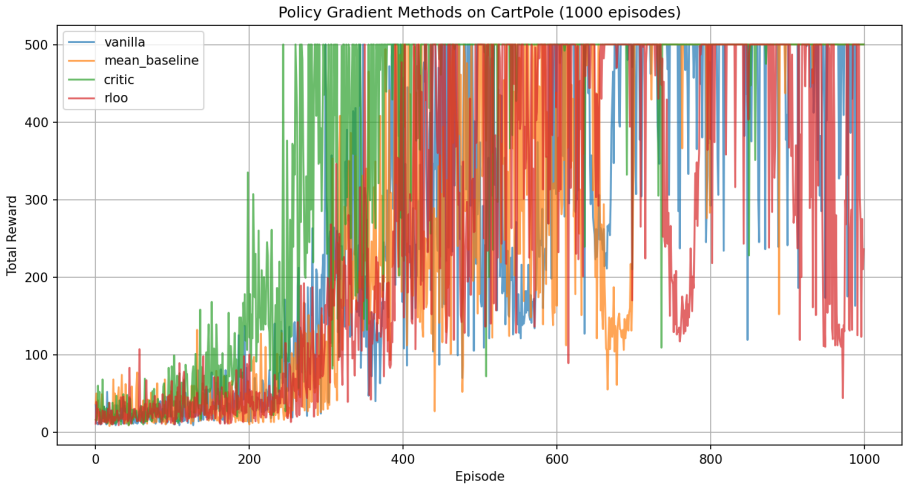

Figure 1: Episode returns over 1000 episodes. Critic (blue) is the most stable.

Table 1: Average return over episodes 901–1000.

Method	Mean	Median	Std
Vanilla	465.51	500.0	92.3
Mean baseline	499.50	500.0	5.2
Critic	500.00	500.0	0.0
RLOO	337.48	236.0	148.2

Discussion

1. With 1000 episodes, the actor-critic method demonstrates clear superiority: it reaches the maximum reward and maintains it with zero variance. This makes it the most reliable choice for CartPole.

2. Mean baseline performs almost as well but occasionally drops below 500, indicating that a simple episode-level baseline is sufficient but slightly less stable than a learned value function.

3. Vanilla REINFORCE, even after 1000 episodes, retains considerable variance (std 92). This highlights the fundamental limitation of not using any baseline — the gradient estimates remain noisy regardless of training duration.

4. The RLOO implementation fails to converge to an optimal policy. Its average stays around 337, and it never consistently reaches 500. This outcome confirms that leave-one-out baselines are not well-suited for sequential decision problems where only one trajectory per state is available.

5. Behavior cloning results align perfectly with theory: with an optimal expert, imitation achieves identical performance; with a suboptimal expert, the clone inherits all deficiencies.