Generated Model Performance Report

Simon Green

February 7, 2025

1 Results and Model Comparison

This report presents the performance evaluation of four reinforcement learning models: TRPO, PPO, TRPOER, and TRPOR. These models are implemented using **Stable Baselines3** and utilize **mini-batch gradient descent** for optimization, ensuring efficient and stable updates during training. The entropy calculations guiding the models are based on the entropy of each batch, which influences regularization and experience replay mechanisms.

1.1 TRPO (Trust Region Policy Optimization)

Originally proposed by Schulman et al. [1], TRPO is a policy gradient method that constrains updates using a trust region to ensure stability in training.

1.2 PPO (Proximal Policy Optimization)

Introduced by Schulman et al. [2], PPO improves upon TRPO by using a clipped surrogate objective to ensure efficient and stable policy updates.

1.3 TRPOR (TRPO with Entropy Regularization)

This model extends TRPO by introducing entropy regularization only in the policy objective. The entropy coefficient hyperparameter guides the degree of regularization, ensuring a balance between exploration and exploitation. The entropy guiding this model is computed at the batch level, dynamically adjusting policy updates.

1.4 TRPOER (TRPO with Entropy Regularized Experience Replay)

This model extends TRPO by incorporating entropy-based experience replay and an additional policy entropy regularization term. It utilizes a prioritized experience replay buffer sampled according to batch entropy values and a hyperparameter coefficient. The method enables bidirectional adaptive sampling, adjusting both the number and direction of sampled experiences to optimize learning. The adaptive sampling function is formulated as:

$$S = \text{clip}\left((M - m) \times \begin{cases} 1 - \left| \frac{H}{|\lambda + \epsilon|} \right|, & \lambda > 0 \\ \left| \frac{H}{|\lambda + \epsilon|} \right|, & \lambda < 0 \end{cases} + m,\ m,\ M \right) \tag{1}$$

where S is the number of samples, H represents batch entropy, λ is the sampling coefficient, and M, m are the maximum and minimum sample limits.

Model Performance Table

The table below summarizes the models' performance in terms of mean and standard deviation of rewards, along with maximum and minimum rewards recorded during training. A higher mean reward suggests better overall performance, while lower standard deviation indicates increased stability.

Environment Model	Pendulum-v1	InvertedDoublePendulum-v5	Ant-v5	Humanoid-v5
PPO	$-0.02M$ $-207.56\mu \pm 118.01\sigma$ $2892E, 10R$	$9359.93M$ $1090.87\mu \pm 2910.66\sigma$ $6965E, 10R$	$1173.61M$ $515.51\mu \pm 326.48\sigma$ $1895E, 10R$	$1245.44M$ $464.88\mu \pm 298.07\sigma$ $15385E, 9R$
TRPO	$-0.20M$ $-141.97\mu \pm 115.55\sigma$ $2600E, 10R$	$9359.78M$ $7516.76\mu \pm 3879.62\sigma$ $5767E, 10R$	$1960.61M$ $1328.68\mu \pm 284.88\sigma$ $1067E, 10R$	$1247.77M$ $369.87\mu \pm 311.76\sigma$ $9569E, 10R$
TRPOER1	$-0.07M$ $-173.49\mu \pm 156.48\sigma$ $2859E, 10R$	$9351.73M$ $2673.08\mu \pm 2620.77\sigma$ $5070E, 9R$	$1349.85M$ $639.65\mu \pm 465.91\sigma$ $1882E, 8R$	$897.72M$ $371.33\mu \pm 359.99\sigma$ $2799E, 4R$
TRPOR	$-0.22M$ $-197.32\mu \pm 215.34\sigma$ $4508E, 10R$	$9357.32M$ $8045.50\mu \pm 2918.29\sigma$ $9123E, 5R$	$3590.62M$ $1662.87\mu \pm 1171.26\sigma$ $1655E, 7R$	$1416.67M$ $620.72\mu \pm 381.70\sigma$ $9225E, 10R$

2 Performance Analysis Through Plots

The following plots visualize different aspects of model performance.

Learning Stability

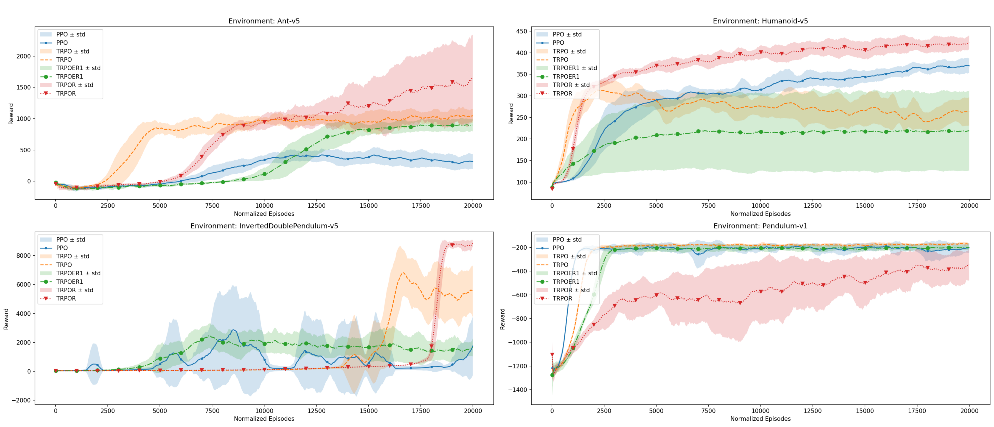

Figure 1: Learning Stability for Different Models

Learning stability is evaluated based on the smoothness of the reward curve. A more stable learning process exhibits a steadily increasing reward trajectory, whereas high variance suggests instability due to sensitivity to hyperparameters.

Learning Stability (Coefficient of Variation)

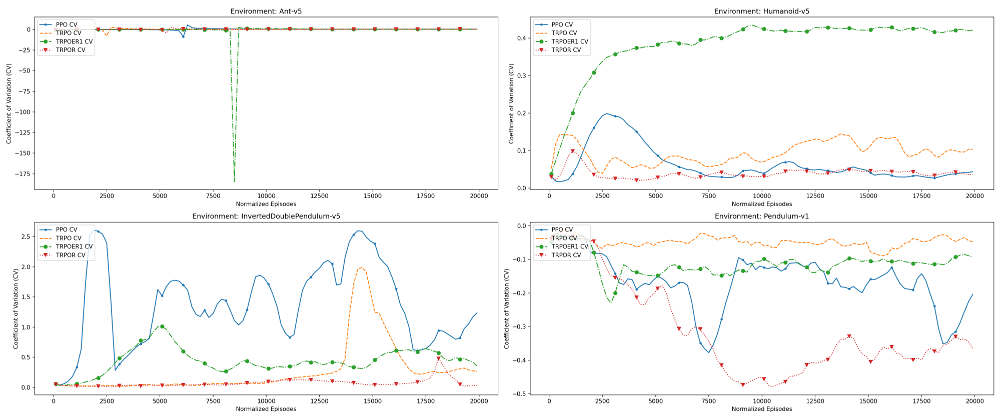

Figure 2: Learning Stability (Coefficient of Variation)

The coefficient of variation (CV) provides a normalized measure of stability. A lower CV signifies less volatile performance, whereas a higher CV indicates inconsistency due to randomness in training.

Sample Efficiency

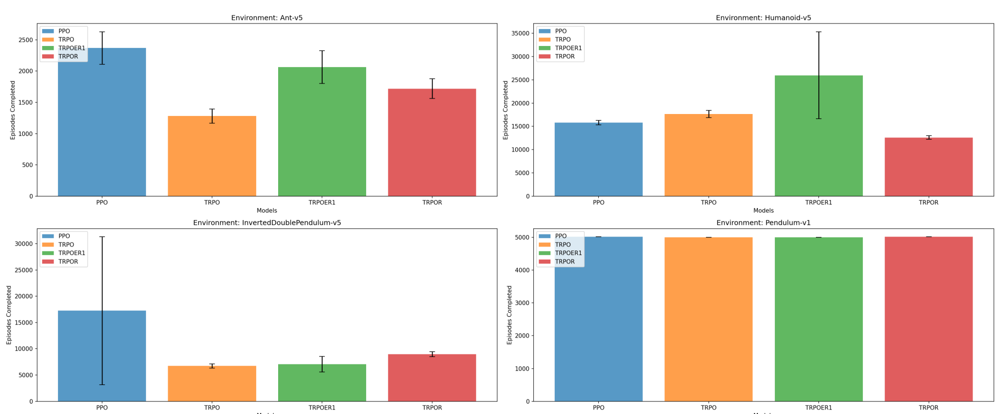

Figure 3: Sample Efficiency Across Models

Sample efficiency measures how quickly a model improves with limited training episodes. Higher sample efficiency is desirable, especially in data-scarce scenarios.

Combined Sample Efficiency Results

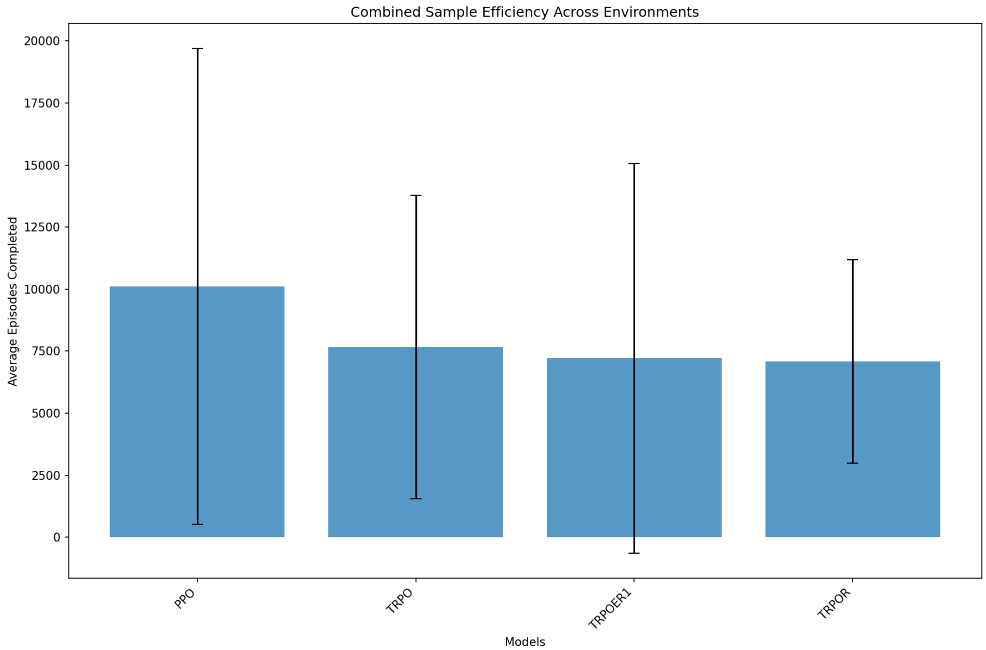

Figure 4: Combined Sample Efficiency Results

The combined sample efficiency plot aggregates results across all environments, showing how different models perform in terms of data efficiency.

Resampled Rewards and Outlier Removal

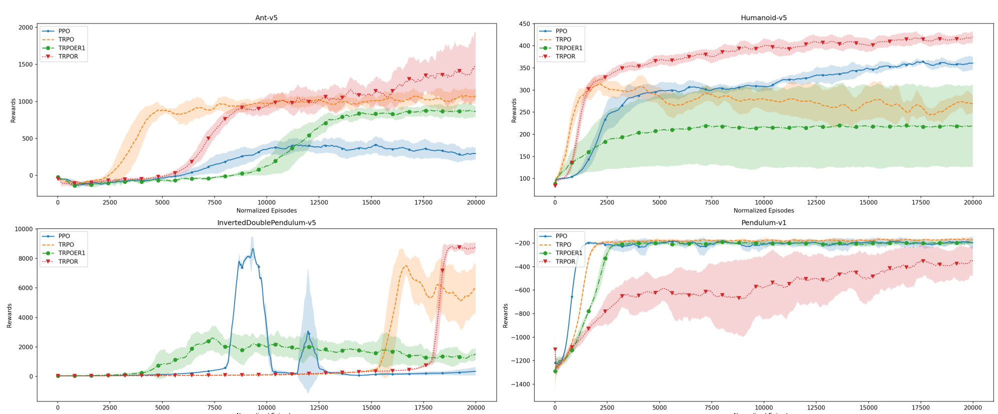

Figure 5: Resampled Rewards with Outlier Removal

This plot presents reward distributions after applying smoothing and outlier removal techniques, filtering out misleading fluctuations.

Raw Data

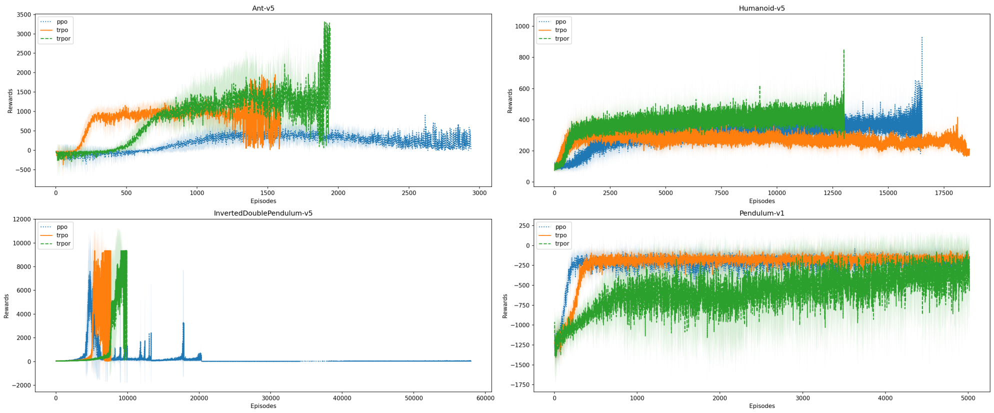

Figure 6: Raw Reward Data for Different Models

The raw data plot displays the recorded reward values without any smoothing. It provides insights into the actual training process and variability in rewards.

References

[1] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust Region Policy Optimization," *International Conference on Machine Learning (ICML)*, 2015.

[2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.