Assignment 2 : Policy gradient

Arij Boubaker, Linghao Zeng, Zhe Huang

28/02/2024

Table of Content

Introduction

Explore and analyze the impact of policy gradients and its variants, neural network baselines, generalized advantage estimation, and hyperparameter tuning on reinforcement learning tasks.

Policy Gradients

Comparing learning curves from experiments using small and large batches in the CartPole environment.

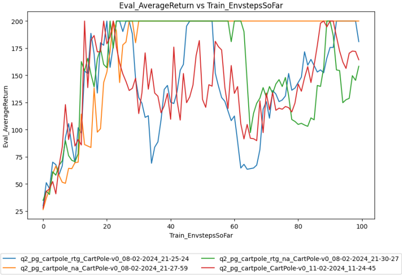

Figure – Learning curves of small batch experiments.

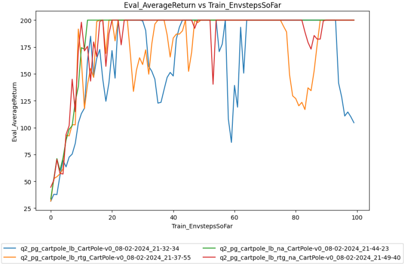

Figure – Learning curves of large batch experiments.

Neural Network Baseline

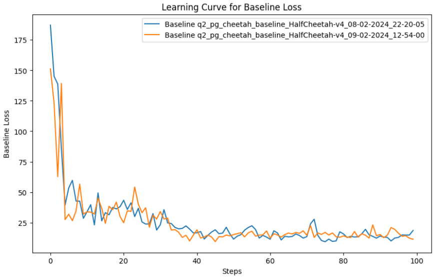

Figure – Learning curves for the baseline loss in the HalfCheetah-v4 environment. The number of baseline gradient steps is 5 and the baseline learning rate is 0.01 be default.

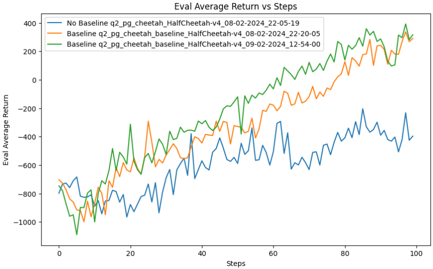

Figure – Comparison of learning curves for the HalfCheetah-v4 environment with and without the use of a baseline.

Neural Network Baseline

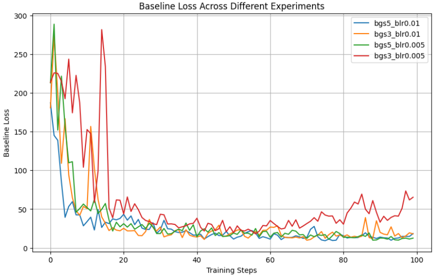

Figure – Comparative baseline loss trends for HalfCheetah-v4 task under different hyperparameter settings.

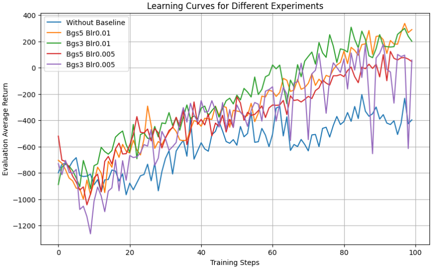

Figure – Evaluation average return for the HalfCheetah-v4 task across different combinations of baseline gradient steps and baseline learning rates.

Neural Network Baseline

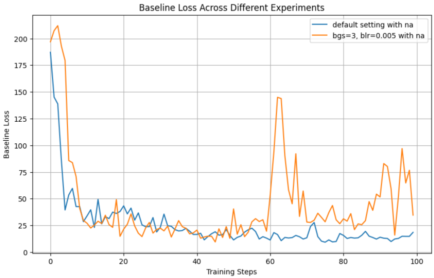

Figure – The baseline loss between experiments conducted with default hyperparameters and those adjusted for normalized advantages.

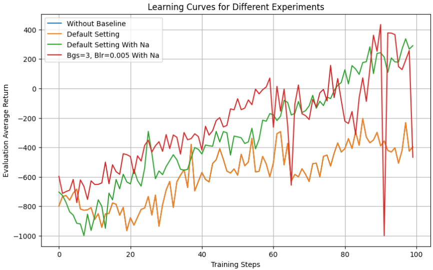

Figure – The learning curves depict the evaluation average return over training steps for the HalfCheetah-v4 environment under various conditions.

Generalized Advantage Estimation

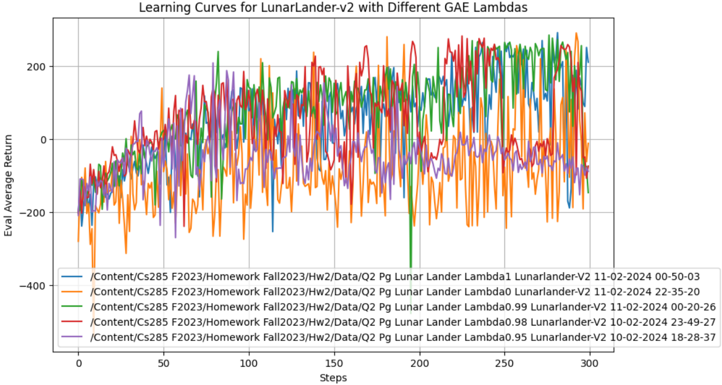

Figure – This figure illustrates the learning curves for the LunarLander-v2 environment using various Generalized Advantage Estimation λ values. Each curve represents the evolution of the evaluation average return over the number of training steps, demonstrating the effect of λ on the learning process.

Generalized Advantage Estimation

- When $\lambda = 0$, the GAE is $A^{\pi}_{\mathsf{GAE}}(s_t, a_t) = \delta_t(s_t, a_t) = r(s_t, a_t) + \gamma V^{\pi}_{\phi}(s_{t+1}) - V^{\pi}_{\phi}(s_t)$, mirroring the characteristics of a single-step advantage estimator that possesses low variance but high bias.
- When $\lambda = 1$, the GAE is $A^{\pi}_{\mathsf{GAE}}(s_t, a_t) = \sum_{t'=t}^{T-1} \gamma^{t'-t} \delta_{t'}$, aligning with the principles of a multi-step actor critic approach, distinguished by its high variance and reduced bias.

Hyperparameter Tuning

Hyperparameter	Default settings	Tuned hyperparameters
Environment	InvertedPendulum-v4	InvertedPendulum-v4
Number of iterations	100	100
Reward-to-Go	Yes	Yes
Use baseline	Yes	Yes
Normalize advantages	Yes	Yes
Batch size	5000	5000
Discount factor	-	0.99
GAE lambda	-	0.99
Network size	-	256
Number of layers	-	3
Learning tate	-	0.001
Baseline learning rate	-	0.002
Baseline gradient steps	-	20

Figure – Hyperparameters used in the default and tuned settings for the InvertedPendulum-v4 experiments.

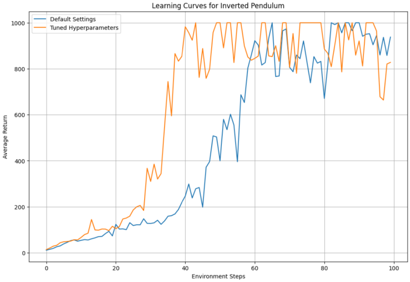

Figure – The plot highlights the trajectories of average returns as a function of environment steps, showcasing the accelerated learning and improved performance achieved through hyperparameter optimization.

Humanoid

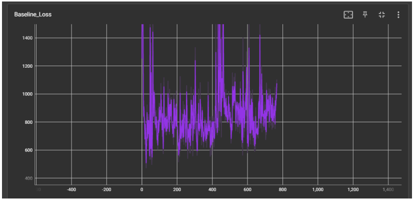

Figure – This plot illustrates the fluctuations in baseline loss during the training of a model. The vertical spikes represent significant changes in loss.

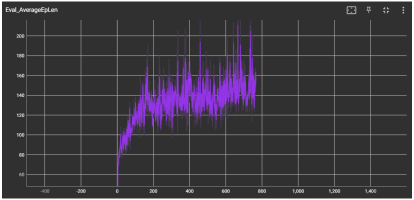

Figure – This figure illustrates the baseline loss throughout the training process.