Proximal Policy Optimization (PPO) with GAE

Initialization:

A stochastic policy (actor) $\pi(a \mid s; \theta_0)$ with initial parameter θ_0.
A state-value function (critic) $V(s; w_0)$ with initial parameter w_0.
Learning rates $\alpha_\theta, \alpha_w > 0$.
Discount factor γ, GAE parameter λ, clipping parameter ϵ.

Goal:

Learn an optimal policy that maximizes the expected return $J(\theta) = \mathbb{E}_{\pi_\theta}\big[\sum_t \gamma^t r_t\big]$.

TD Error:

$$\delta_t = r_{t+1} + \gamma(1 - d_t)V(s_{t+1}; w_t) - V(s_t; w_t)$$

Generalized Advantage Estimation (GAE):

$$\hat{A}_t = \sum_{k=0}^{\infty} (\gamma\lambda)^k \delta_{t+k}$$

Rollout Collection (On-policy):

At each time step t, collect experience using the current policy:

- Sample action $a_t \sim \pi(\cdot \mid s_t; \theta_{\text{old}})$
- Observe reward r_{t+1}, next state s_{t+1}, and termination flag d_t
- Store transition tuple:
$$(s_t, a_t, r_{t+1}, s_{t+1}, d_t, \log \pi_{\text{old}}(a_t \mid s_t))$$

Rollouts are collected until a fixed buffer size is reached.

Advantage and Target Computation:

After data collection, compute:

$$y_t = \hat{A}_t + V(s_t; w_t)$$

which is used as the TD target for the critic.

Mini-batch PPO Update:

For each PPO epoch, shuffle the rollout buffer and sample mini-batches $\mathcal{B}$:

Actor (Policy Update):

For each mini-batch sample, compute the importance ratio

$$r_t(\theta) = \exp\Big(\log \pi(a_t \mid s_t; \theta) - \log \pi_{\text{old}}(a_t \mid s_t)\Big)$$

The clipped PPO objective is

$$L^{\text{actor}}(\theta) = -\mathbb{E}\Big[\min\big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big] - c_{\text{ent}}\mathcal{H}(\pi)$$

Update the policy parameters:

$$\theta \leftarrow \theta - \alpha_\theta \nabla_\theta L^{\text{actor}}(\theta)$$

Critic (Value Function Update):

Minimize the mean squared error loss:

$$L^{\text{critic}}(w) = \frac{1}{2}\mathbb{E}\big[\big(V(s_t; w) - y_t\big)^2\big]$$

Update the critic parameters:

$$w \leftarrow w - \alpha_w \nabla_w L^{\text{critic}}(w)$$

After all PPO epochs, set $\pi_{\text{old}} \leftarrow \pi$.