

IEOR 8100 Lab 4: Pseudocode for Policy Gradient

TA: Yunhao Tang

Algorithm 1 Actor Critic PG

- 1: INPUT: number of episodes to rollout N , learning rate for policy α , learning rate for baseline β , discount factor γ , total number of iterations E
- 2: INITIALIZE: Policy network $\pi_\phi(s)$ with parameters ϕ , baseline (value function) network $V_\theta(s)$ with parameters θ
- 3: **for** $e = 1, 2, 3 \dots E$ **do**
- 4: **Rollouts**
- 5: Collect rollouts using the current policy $\pi_\phi(s)$, collect a total of N trajectories, each trajectory is a sequence $\{s_0, a_0, s_1, a_1 \dots a_{T-1}, s_T\}$ with corresponding rewards $r_i, 0 \leq i \leq T$.
- 6: **Train Baseline**
- 7: From the collected trajectories, compute monte carlo estimate of value functions. For state s_i in a given trajectory, its estimated value function is

$$\hat{V}(s_i) = \sum_{t=i}^T r_t \gamma^{t-i}$$

- 8: Compute value function loss
- $$L_v = \frac{1}{N(T+1)} \sum_i (V_\theta(s_i) - \hat{V}(s_i))^2$$
- 9: Update $\theta \leftarrow \theta - \beta \nabla_\theta L_v$
 - 10: **Training policy**
 - 11: (Similar as above) From the collected trajectories, compute monte carlo estimate of action value functions. For state action pair (s_i, a_i) in a given trajectory, its estimated action value function is

$$\hat{Q}(s_i, a_i) = \sum_{t=i}^T r_t \gamma^{t-i}$$

- 12: Compute baseline for each state $V_\theta(s_i)$ and compute advantage $\hat{A}(s_i, a_i) = \hat{Q}(s_i, a_i) - V_\theta(s_i)$
- 13: Compute surrogate loss

$$L_p = -\frac{1}{N(T+1)} \sum_i \hat{A}(s_i, a_i) \log \pi(a_i | s_i)$$

- 14: Update $\phi \leftarrow \phi - \alpha \nabla_\phi L_p$
-

There are multiple ways to improve the efficacy of the algorithm.

- **Number of Samples to collect:** Instead of specifying the number of trajectories to collect, it may make more sense to specify the number of time steps to collect (a more balanced sample set at each iteration).
- **Advanced PG:** More advanced policy gradient methods include: Natural PG, Trust Region Policy Optimization (TRPO), Proximal Policy Optimization (PPO). More stable in training.
- **Advantage Estimate:** Generalized Advantage Estimation. Tradeoff between variance and bias.