## PPO Algorithm

## 1: **Input:** initial policy parameters $\theta_0$ , clipping threshold $\epsilon$

2: **for**  $k = 0, 1, 2, \dots$  **do** 

Collect set of partial trajectories 
$$D_k$$
 using policy  $\pi_k = \pi(\theta_k)$   
Estimate advantages  $\hat{A}^{\pi_k}$  using any advantage estimation also

3: Estimate advantages  $\hat{A}_{t}^{\pi_{k}}$  using any advantage estimation algorithm 4:

hate advantages 
$$A_t^{**}$$
 using the policy update:

Compute policy update:

cy update: 
$$\theta_{i}$$

Perform 
$$K$$
 steps of minibatch SGD (via Adam), where:

$$heta_{k+1} = \arg\max_{\theta} L_{\theta_k}^{\text{CLIP}}(\theta)$$

$$\operatorname{ria} \operatorname{Adam}_{\theta}$$

 $L_{\theta_k}^{\text{CLIP}}(\theta) = E_{\tau \sim \pi_k} \left[ \sum_{t=0}^{T} \min \left( r_t(\theta) \hat{A}_t^{\pi_k}, \text{clip}\left( r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t^{\pi_k} \right) \right]$ 

$$L_{\theta_k}$$
 (o). where

6:

7: end for