1: Initialize timestep counter N=0 and network weights θ , θ_v 2: Instantiate set e of n_e environments 3: repeat 4: for t = 1 to t_{max} do Sample a_t from $\pi(a_t|s_t;\theta)$ 5: 6: Calculate v_t from $V(s_t; \theta_v)$ 7: parallel for i=1 to n_e do 8: Perform action $a_{t,i}$ in environment e_i Observe new state $s_{t+1,i}$ and reward $r_{t+1,i}$ 10: end parallel for 11: end for

Algorithm 1 Parallel advantage actor-critic

12:
$$R_{t_{\text{max}}+1} = \begin{cases} 0 & \text{for terminal } s_t \\ V(s_{t_{\text{max}}+1}; \theta) & \text{for non-terminal } s_t \end{cases}$$
13: $\mathbf{for} \ t = t_{\text{max}} \ \text{down to 1 do}$

13: **for**
$$t = t_{\text{max}}$$
 down to 1 **do**
14: $R_t = r_t + \gamma R_{t+1}$

: For
$$t=t_{\max}$$
 down to 1 do
: $R_t=r_t+\gamma R_{t+1}$
: end for

14:
$$R_t = r_t + \gamma R_{t+1}$$
15: end for
16: $d\theta = \frac{1}{r_t} \sum_{n=1}^{n} \sum_{t=n}^{t} \sum_{t=n}^{t} (P_{t+1} - P_{t+1}) \sum_{t=n}^{t} (P_{t+1} - P_{t+1}) \sum_{t=n}^{t} (P_{t+1} - P_{t+1})$

15: **end for**
16:
$$d\theta = \frac{1}{n_e \cdot t_{max}} \sum_{i=1}^{n_e} \sum_{t=1}^{t_{max}} (R_{t,i} - v_{t,i}) \nabla_{\theta} \log \pi(a_{t,i} | s_{t,i}; \theta) + \beta \nabla_{\theta} H(\pi(s_{e,t}; \theta))$$

$$d\theta = \frac{1}{n_e \cdot t_{max}} \sum_{i=1}^{n_e} \sum_{t=1}^{t_{max}} (R_{t,i} - v_{t,i}) \nabla_{\theta} \log \pi(a_{t,i}|s_{t,i};\theta) + \beta \nabla_{\theta} H(\pi(s_{e,t};\theta))$$

$$d\theta_v = \frac{1}{n_e \cdot t_{max}} \sum_{i=1}^{n_e} \sum_{t=1}^{t_{max}} \nabla_{\theta_v} (R_{t,i} - V(s_{t,i};\theta_v))^2$$

17:
$$d\theta_v = \frac{1}{n_e \cdot t_{max}} \sum_{i=1}^{n_e} \sum_{t=1}^{t_{max}} \nabla_{\theta_v} \left(R_{t,i} - V(s_{t,i}; \theta_v) \right)^2$$
18: Update θ using $d\theta$ and θ_v using $d\theta_v$.

18: Update θ using $d\theta$ and θ_v using $d\theta_v$.

19: $N \leftarrow N + n_e \cdot t_{\text{max}}$ 20: until $N \geq N_{max}$