

Actor-Critic Algorithm

Initialize policy parameters θ , critic parameters ϕ

For iteration=1, 2 ... **do**

 Sample m trajectories under the current policy

$\Delta\theta \leftarrow 0$

For $i=1, \dots, m$ **do**

For $t=1, \dots, T$ **do**

$$A_t = \sum_{t' \geq t} \gamma^{t'-t} r_{t'}^i - V_{\phi}(s_t^i)$$

$$\Delta\theta \leftarrow \Delta\theta + A_t \nabla_{\theta} \log(a_t^i | s_t^i)$$

$$\Delta\phi \leftarrow \sum_i \sum_t \nabla_{\phi} ||A_t^i||^2$$

$$\theta \leftarrow \theta + \alpha \Delta\theta$$

$$\phi \leftarrow \phi + \beta \Delta\phi$$

End for