

Environment

$$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

s_t, a_t, r_t, s_{t+1}

Relay Buffer

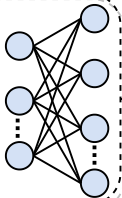
R

$a_t \sim \pi(\cdot|s_t)$

SAC Agent

Critic

Policy evaluation:
 Q function update



$V_{\pi}(s_t)$

Actor

Policy improvement:
Policy gradient update

