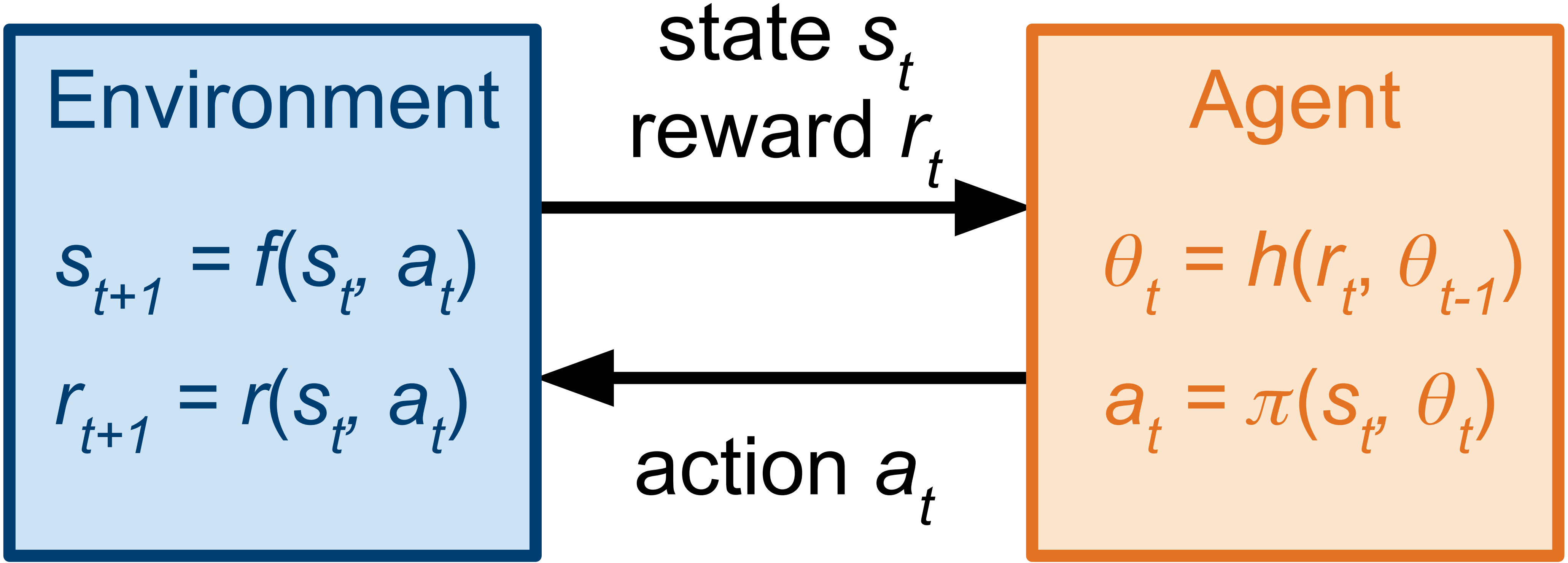
Environment
$s_{t+1} = f(s_t, a_t)$
$r_{t+1} = r(s_t, a_t)$
state s_t
reward r_t
action a_t
Agent
$\theta_t = h(r_t, \theta_{t-1})$
$a_t = \pi(s_t, \theta_t)$