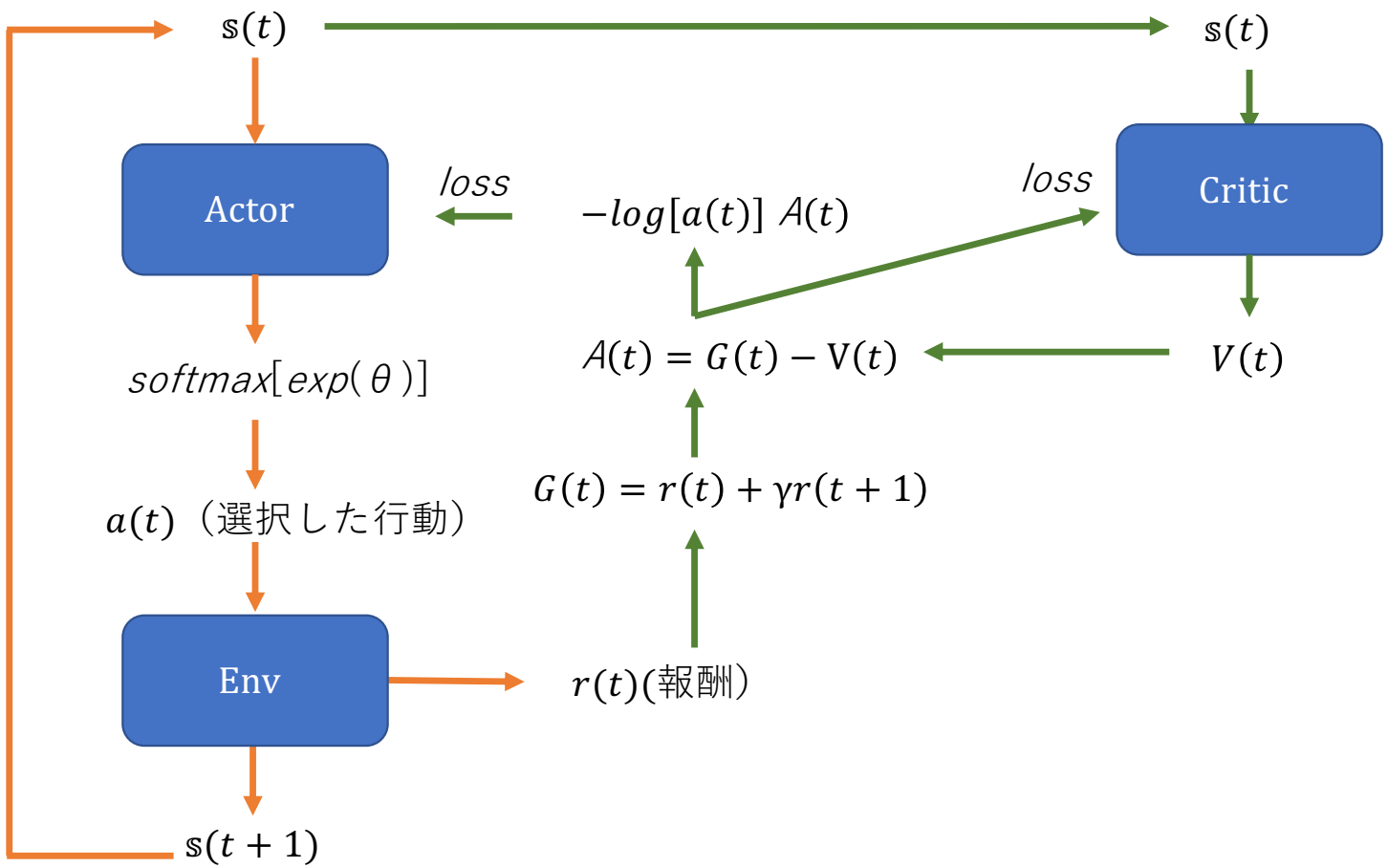
$s(t)$
$s(t)$
Actor
loss
$-log[a(t)]\ A(t)$
loss
Critic
$A(t) = G(t) - V(t)$
$V(t)$
$softmax[exp(\theta)]$
$G(t) = r(t) + \gamma r(t+1)$
$a(t)$（選択した行動）
Env
$r(t)$(報酬)
$s(t+1)$