$Q(s, a; \theta)$
θ
θ
$Q^*(s, a) \approx$
$Q(s, a :$
$\theta)$
θ
$L_i(\theta_i) = E\Big(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a : \theta_i)\Big)^2$

s'
$\hat{\theta} \in$
$R^{d'}$
$\pi_\theta(a|s)$
θ
$J(\pi_\theta)$
$J(\pi_\theta)$
J
$J(\pi_\theta) =$
$E[R_t]$
J

$\theta_{t+1} = \theta_t + \alpha \nabla \hat{J}(\theta_t)$

$\nabla \hat{J}(\theta_t)$
J
θ_t
$\nabla J(\pi_\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla_\theta \pi_\theta(a|s)$

μ
S
π_θ
?

$\nabla J(\pi_\theta) = E_\pi \left[R_t \frac{\nabla_\theta \pi_\theta(a|S_t)}{\pi_\theta(a|S_t)} \right]$

$\left[R_t \frac{\nabla_\theta \pi_\theta(a|S_t)}{\pi_\theta(a|S_t)} \right]$
$\nabla J(\pi_\theta)$
θ

$\theta_{t+1} = \theta_t + \alpha R_t \frac{\nabla_\theta \pi_\theta(a|S_t)}{\pi_\theta(a|S_t)} = \theta_t + \alpha R_t \nabla_\theta log \pi_\theta(a|S_t)$

$J(\pi\theta)$
?
θ
$\nabla_\theta log \pi(a_t|s_t :$
$\theta) R_t$
$\nabla_\theta E[R_t]$
$b_t(s_t)$
R_t
$\nabla_\theta log \pi(a_t|s_t; \theta)(R_t -$
$b_t(s_t))$
$b_t(s_t) \approx$
$V^\pi(s_t)$
π
b_t
$b_t(s) =$
$V_{\pi_{\theta_t}}(s)$
$R_t -$
$b_t(s)$
a_t
s_t
$A(a_t, s_t) =$
$Q(a_t, s_t) -$
$V(s_t)$
R_t
$Q^\pi(a_t, s_t)$
b_t
$V^\pi(s_t)$
$Advantage Actor-$
$Critic$
$A2C$
$Advantage Actor-$