Actor-critic Algorithms

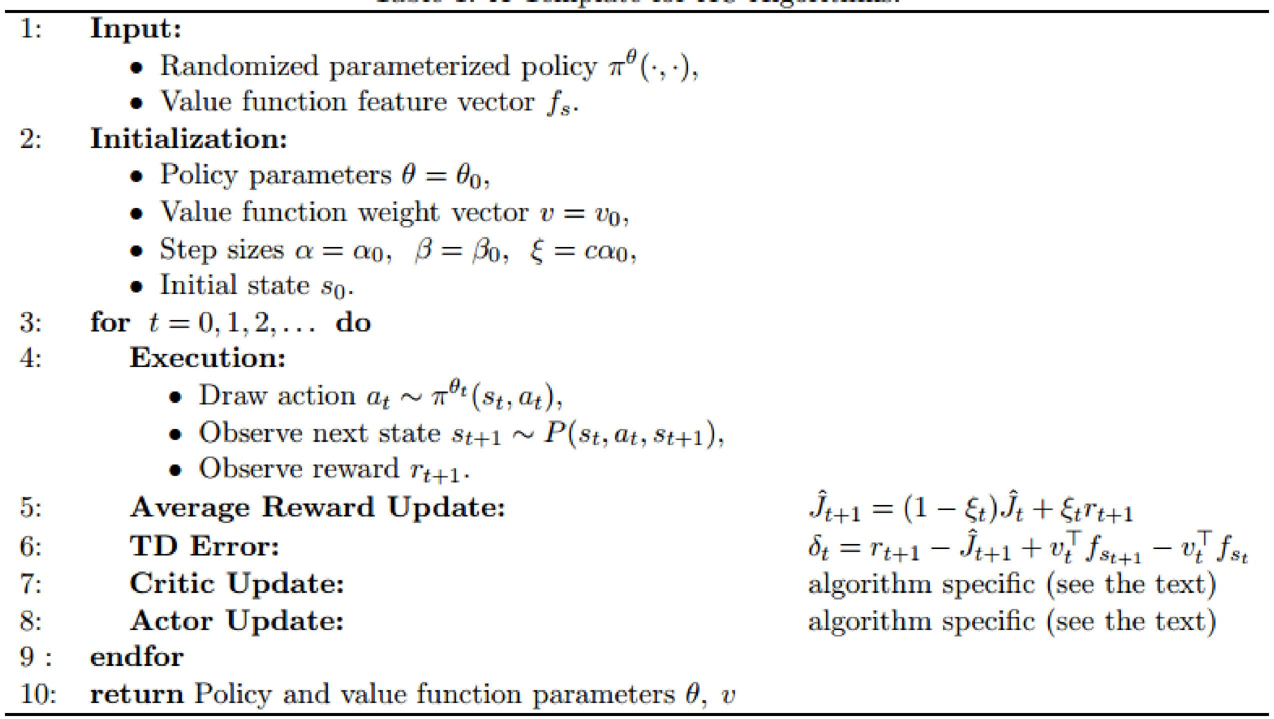
Weiwei Zhang

260684686

Overview of Actor-critic Algorithms

- Actor-critic algorithms use the gradient of the value function to update the policy parameters. The general form of actor-critic is shown as follows.

Table 1: A Template for AC Algorithms.



```
Input:
    • Randomized parameterized policy $\pi^{\theta}(\cdot,\cdot)$,
    • Value function feature vector $f_s$.
Initialization:
    • Policy parameters $\theta = \theta_0$,
    • Value function weight vector $v = v_0$,
    • Step sizes $\alpha = \alpha_0$,  $\beta = \beta_0$,  $\xi = c\alpha_0$,
    • Initial state $s_0$.
for  $t = 0, 1, 2, \ldots$  do
    Execution:
        • Draw action $a_t \sim \pi^{\theta_t}(s_t, a_t)$,
        • Observe next state $s_{t+1} \sim P(s_t, a_t, s_{t+1})$,
        • Observe reward $r_{t+1}$.
    Average Reward Update:        $\hat{J}_{t+1} = (1 - \xi_t)\hat{J}_t + \xi_t r_{t+1}$
    TD Error:                     $\delta_t = r_{t+1} - \hat{J}_{t+1} + v_t^\top f_{s_{t+1}} - v_t^\top f_{s_t}$
    Critic Update:                algorithm specific (see the text)
    Actor Update:                 algorithm specific (see the text)
endfor
return Policy and value function parameters $\theta$, $v$
```

Mountain Car

State Variables

Two dimensional continuous state space.

Velocity = (-0.07, 0.07)

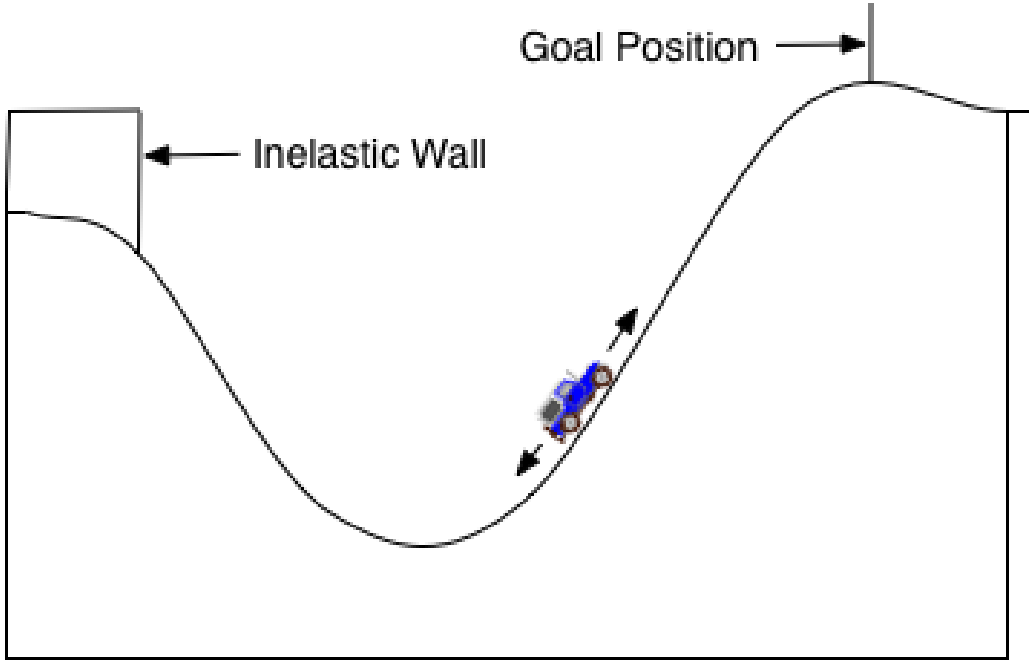
Position = (-1.2, 0.6)

Actions

(reverse, coast, forward)

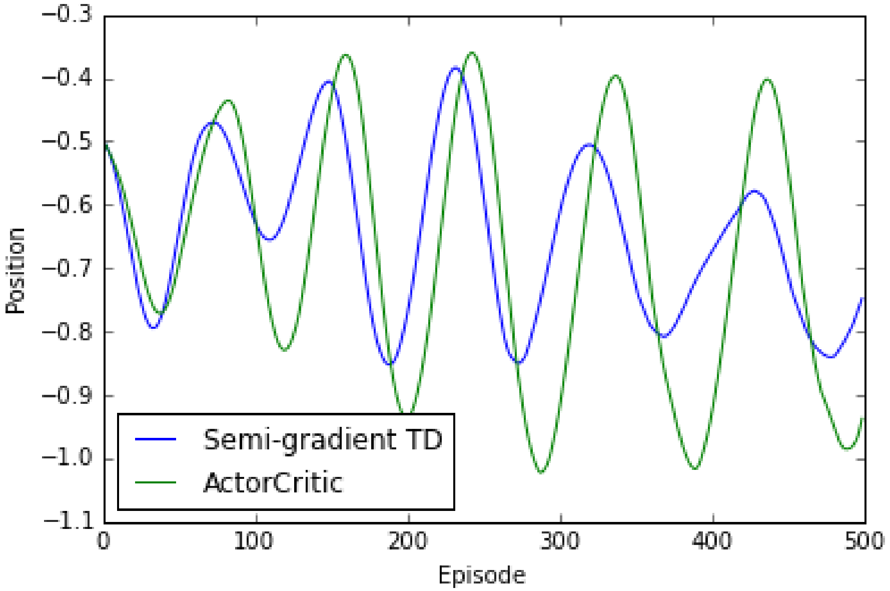
Reward

-1

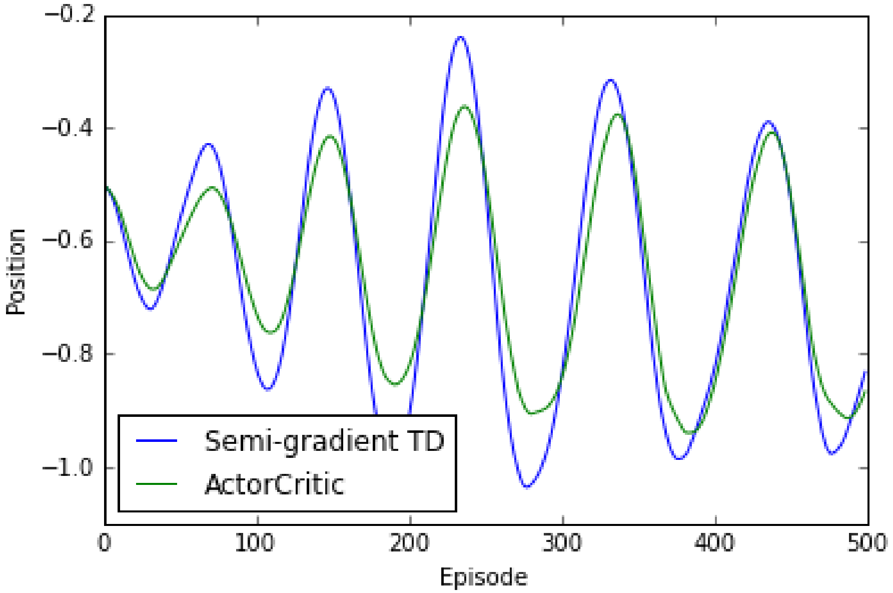


Results

4 tiles



8 tiles


Position
Episode
Semi-gradient TD
ActorCritic