Value-based Approach

Learning a Critic

Critic

- A critic does not determine the action.
- Given an actor, it evaluates the how good the actor is

An actor can be found from a critic.

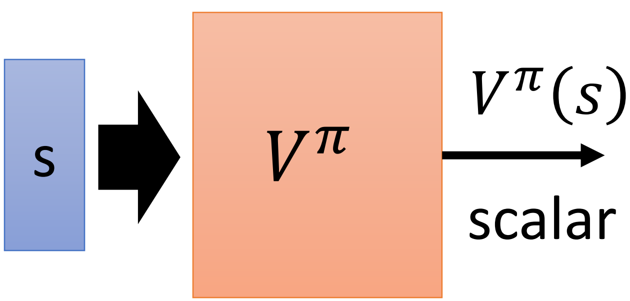
e.g. Q-learning

(not today)

http://combiboilersleeds.com/picaso/critics/critics-4.html

Three kinds of Critics

- A critic is a function depending on the actor π it is evaluated
  - The function is represented by a neural network
- State value function $V^\pi(s)$
  - When using actor π, the *cumulated* reward expects to be obtained after seeing observation (state) s

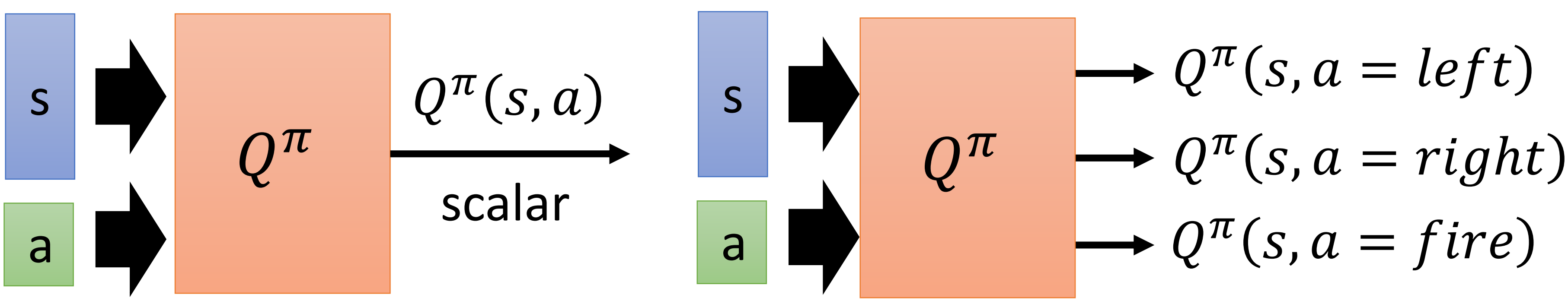


$V^\pi(s)$ is large

$V^\pi(s)$ is smaller

Three kinds of Critics

- State-action value function $Q^\pi(s,a)$
  - When using actor π, the *cumulated* reward expects to be obtained after seeing observation s and taking a



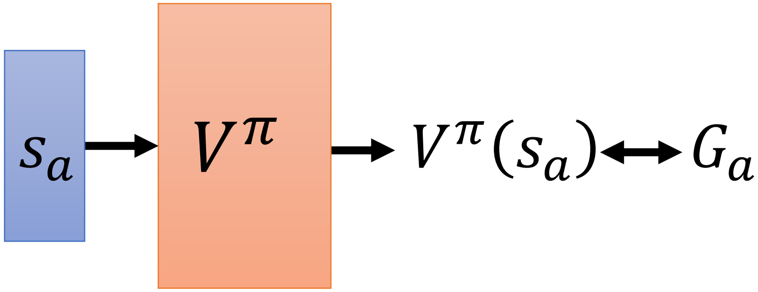
for discrete action only

How to estimate $V^\pi(s)$

- Monte-Carlo based approach
  - The critic watches π playing the game

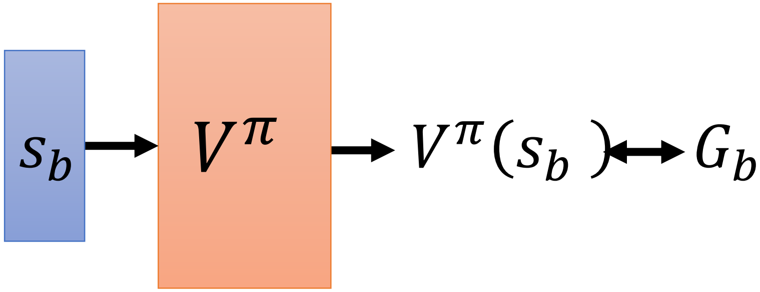
After seeing s_a,

Until the end of the episode, the cumulated reward is G_a



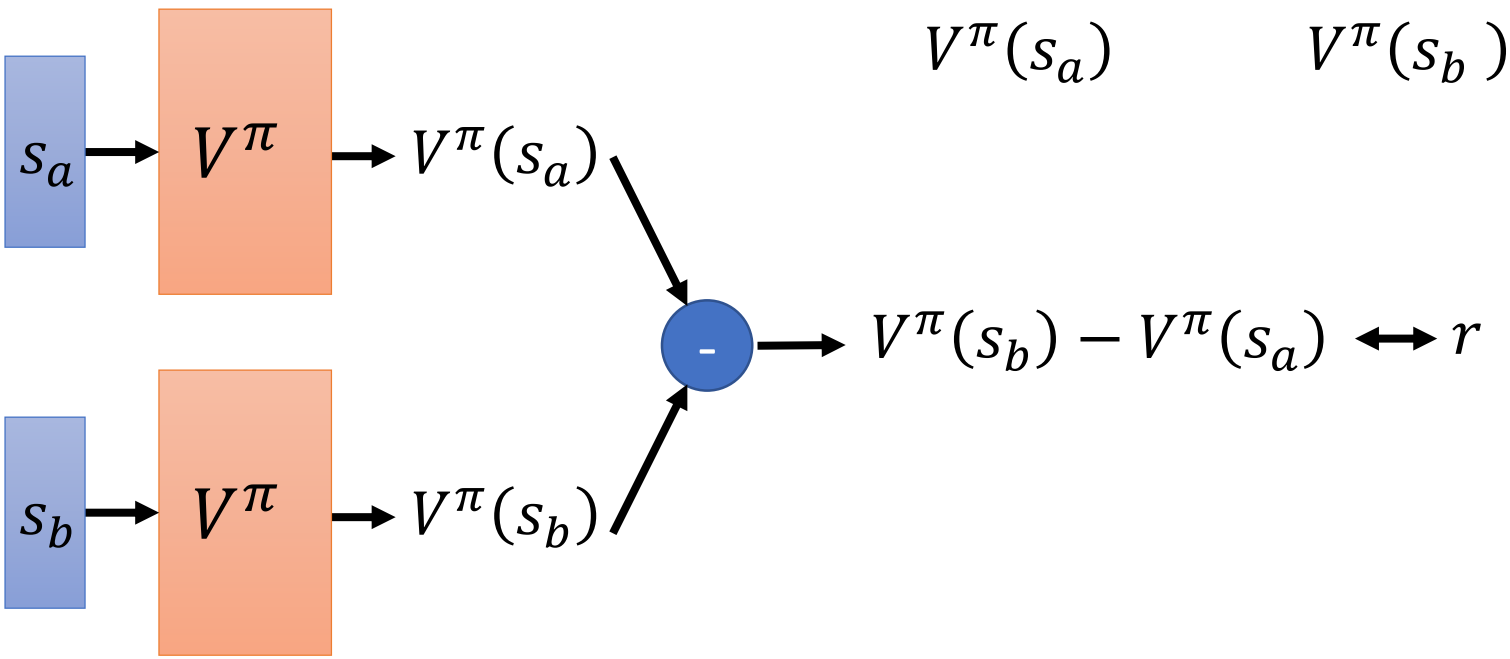
After seeing s_b,

Until the end of the episode, the cumulated reward is G_b



How to estimate $V^{\pi}(s)$

- Temporal-difference approach $\cdots s_a, a, r, s_b \cdots$



Some applications have very long episodes, so that delaying all learning until an episode's end is too slow.

How to estimate $V^{\pi}(s)$

[Sutton, v2, Example 6.4]

- The critic has the following 8 episodes
  - $s_a, r = 0, s_b, r = 0$, END
  - $s_b, r = 1$, END
  - $s_b, r = 1$, END
  - $s_b, r = 1$, END
  - $s_b, r = 1$, END
  - $s_b, r = 1$, END
  - $s_b, r = 1$, END
  - $s_b, r = 0$, END

$$V^{\pi}(s_b) = 3/4$$

$V^{\pi}(s_a) = ?$ 0? 3/4?

Monte-Carlo: $V^{\pi}(s_a) = 0$

Temporal-difference:

$$V^{\pi}(s_a) + r = V^{\pi}(s_b)$$

3/4 0 3/4

(The actions are ignored here.)