Advantage Actor-Critic

서울시립대학교
UNIVERSITY OF SEOUL

CIDA Lab.

Presentation: Baekryun Seong

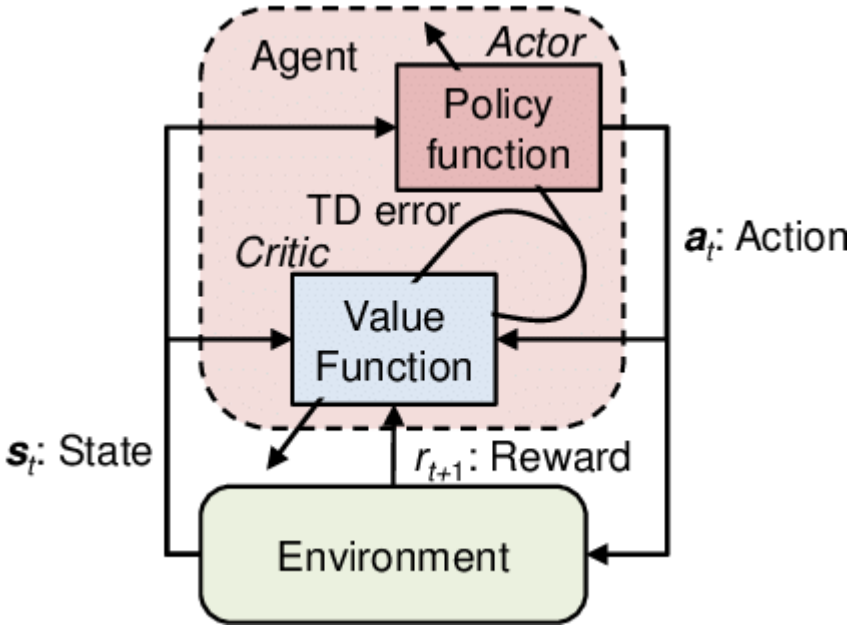
2024.07.10

Table of Contents

In A2C, we will use the policy gradient and a learned value function.

In REINFORCE, A policy is reinforced by reward, directly.

But in Actor-Critic, a policy is reinforced with a learned reinforcing signal.



Ch 6.1
The Actor

In REINFORCE, we used the policy gradient of:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_t[R_t(\tau)\nabla_\theta \log \pi_\theta(a_t|s_t)].$$

Now in actor-critic, we will use:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_t[A_t^\pi \nabla_\theta \log \pi_\theta(a_t|s_t)]$$

where A_t^π is an advantage function.

Ch 6.2
The Critic

Policy gradient is very useful, but it has large variance. So, we introduce an advantage function.

$$A^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) - V^{\pi}(s_t)$$

where

$$\mathbb{E}_{a \in A}[A^{\pi}(s_t, a)] = 0.$$

Advantage function reinforces the actions with rewards over expectation.

	$Q^{\pi}(s,a)$	$V^{\pi}(s)$	$A^{\pi}(s,a)$
Case 1	110	100	10
Case 2	-90	-100	10

Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

To get advantage function, we need to estimate Q^{π}, V^{π}.

We generally estimate V^{π} first and infer Q^{π} from it and reward.

There is two main reason:

1. Q^{π} is more complicated function.

2. It is hard to get V^{π} from Q^{π} because it requires summation over all states.

With approximated value function V^{π}, we get:

$$A^{\pi}_{NSTEP}A(s_t, a_t) = Q^{\pi}(s_t, a_t) - V^{\pi}(s_t)$$

$$\approx r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^n r_{t+n} + \gamma^{n+1}\hat{V}^{\pi}(s_{t+n+1}) - \hat{V}^{\pi}(s_t)$$

Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

Ch 6.2.1.2 Generalized Advantage Estimation

Like SARSA or TD(λ), we can approximate advantage utilizing different steps.

$$A^{\pi}_{GAE}(s_t, a_t) = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$$

$$where\ \delta_t = r_t + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t).$$

δ_t is usually called *TD error*.

λ regulates bias-variance tradeoff. For further explanation, see this.

Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

Now we train advantage function using V^{π} target, n-step or GAE.

V^{π} target can be set differently. In one step,

$$V_{tar}^{\pi}(s) = r + \gamma \hat{V}^{\pi}(s;\theta).$$

In n-step, MC and GAE, targets are:

$$V_{tar}^{\pi}(s_t) = r_t + \gamma r_{t+1} + \cdots + \gamma^n r_{t+n} + \gamma^{n+1}\hat{V}^{\pi}(s_{t+n+1}),$$

$$V_{tar}^{\pi}(s_t) = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'},$$

$$V_{tar}^{\pi}(s_t) = A_{GAE}^{\pi}(s_t, a_t) + \hat{V}^{\pi}(s_t).$$

Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

Ch 6.3
A2C Algorithm

Ch 6.3 A2C Algorithm

Algorithm 6.1 A2C algorithm

Set $\beta \geq 0$ # entropy regularization weight
Set $\alpha_A \geq 0$ # actor learning rate
Set $\alpha_C \geq 0$ # critic learning rate
Randomly initialize the actor and critic parameters θ_A, θ_C[4]
for *episode* $= 0 \ldots MAX\_EPISODE$ **do**
    Gather and store data (s_t, a_t, r_t, s'_t) by acting in the environment using the current policy
    for $t = 0 \ldots T$ **do**
        Calculate predicted V-value $\hat{V}^\pi(s_t)$ using the critic network θ_C
        Calculate the advantage $\hat{A}^\pi(s_t, a_t)$ using the critic network θ_C
        Calculate $V^\pi_{\text{tar}}(s_t)$ using the critic network θ_C and/or trajectory data
        Optionally, calculate entropy H_t of the policy distribution, using the actor network θ_A. Otherwise, set $\beta = 0$
    end for
    Calculate value loss, for example using MSE:
    $L_{\text{val}}(\theta_C) = \frac{1}{T} \sum_{t=0}^{T} (\hat{V}^\pi(s_t) - V^\pi_{\text{tar}}(s_t))^2$
    Calculate policy loss:
    $L_{\text{pol}}(\theta_A) = \frac{1}{T} \sum_{t=0}^{T} (-\hat{A}^\pi(s_t, a_t) \log \pi_{\theta_A}(a_t \mid s_t) - \beta H_t)$
    Update critic parameters, for example using SGD:[5]
    $\theta_C = \theta_C + \alpha_C \nabla_{\theta_C} L_{\text{val}}(\theta_C)$
    Update actor parameters, for example using SGD:
    $\theta_A = \theta_A + \alpha_A \nabla_{\theta_A} L_{\text{pol}}(\theta_A)$
end for

Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

Ch 6.5
Network Architecture

In A2C, we have to train two different networks: actor and critic.

But we can assume that two networks can share some information.

Sharing parameters makes the actor network to utilize state representation of critic net.

But it makes training very unstable, and scale problems may appear.

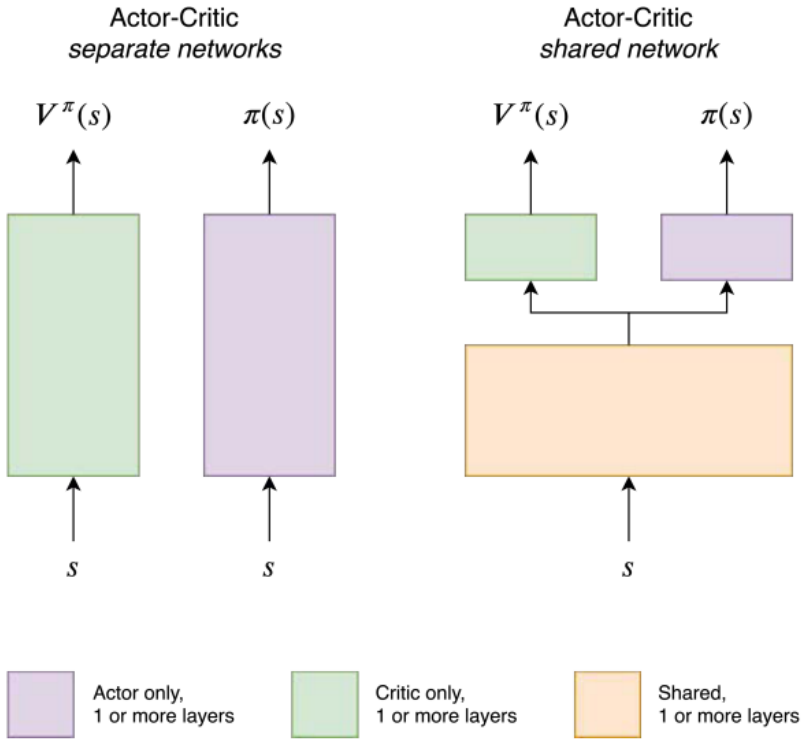


Figure 6.1 Actor-Critic network architectures: shared vs. separate networks

Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

Appendix
Actor-Critic in Neuroscience

The learning rule of real neurons is generally called Hebbian learning.

The most famous learning rule of Hebb is Spike-Timing Dependent Plasticity. (STDP)

Researchers found that TD error exists in brain: *TD Reward Prediction Error.* (TD RPE)

This is called *RPE hypothesis of dopamine neuron activity*.

Dopamine level in brain is very similar to TD error: $R_t + \gamma V(S_{t+1}) - V(S_t)$.

It is very interesting: reinforcement learning have started from dynamic programming, but it also appears in neuroscience.

Researchers also suggests the structure of actor-critic in real brain.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Neural Actor-Critic

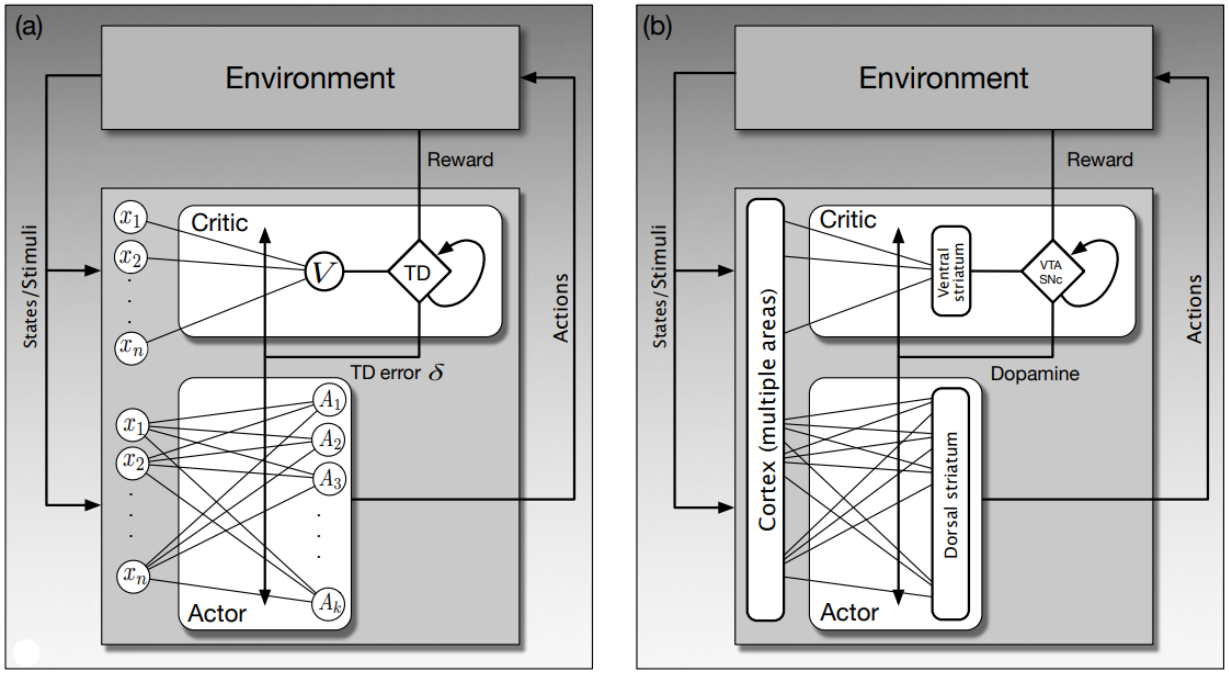


Figure 15.5: Actor–critic ANN and a hypothetical neural implementation. a) Actor–critic algorithm as an ANN. The actor adjusts a policy based on the TD error δ it receives from the critic; the critic adjusts state-value parameters using the same δ. The critic produces a TD error from the reward signal, R, and the current change in its estimate of state values. The actor does not have direct access to the reward signal, and the critic does not have direct access to the action. b) Hypothetical neural implementation of an actor–critic algorithm. The actor and the value-learning part of the critic are respectively placed in the dorsal and ventral subdivisions of the striatum. The TD error is transmitted by dopamine neurons located in the VTA and SNpc to modulate changes in synaptic efficacies of input from cortical areas to the ventral and dorsal striatum. Adapted from Frontiers in Neuroscience, vol. 2(1), 2008, Y. Takahashi, G. Schoenbaum, and Y. Niv, Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/Critic model.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.