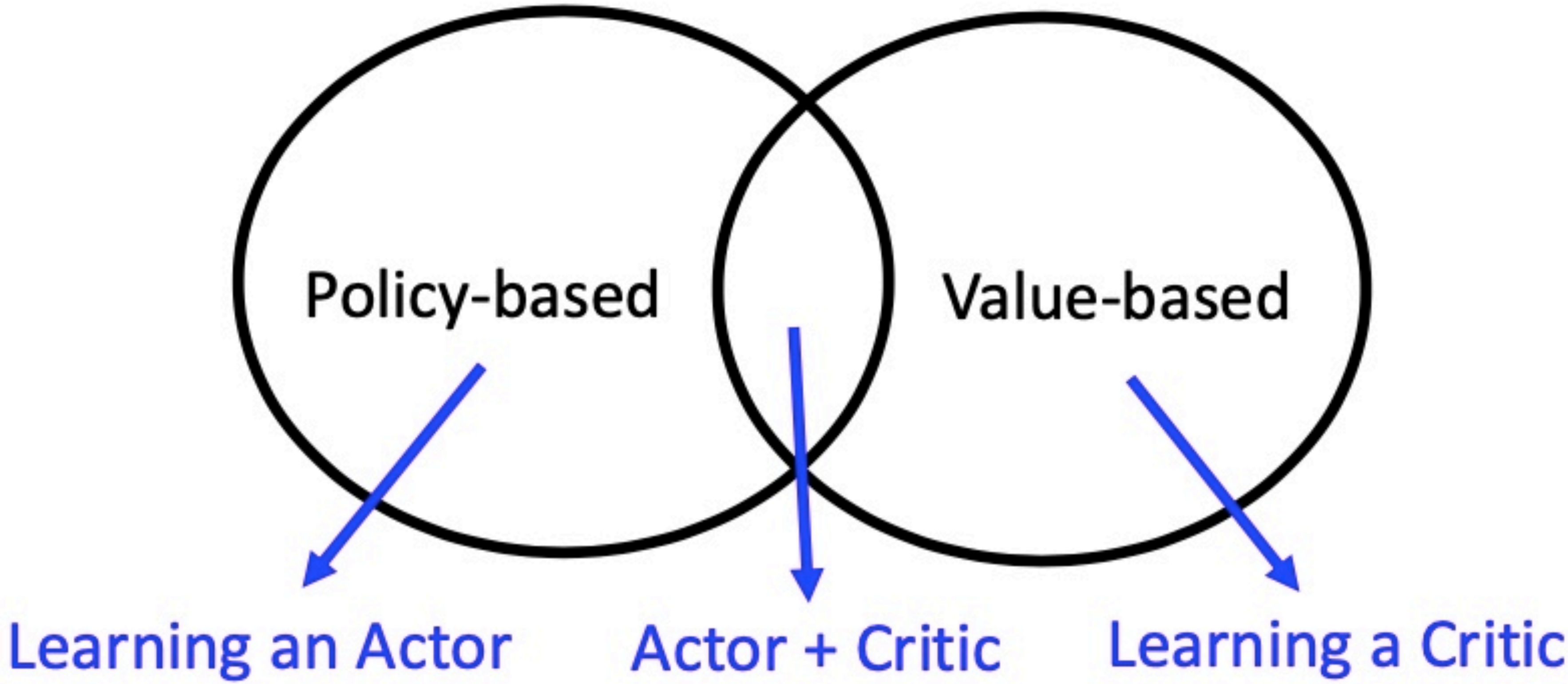
Actor-Critic

Hung-yi Lee


Policy-based
Value-based
Learning an Actor
Actor + Critic
Learning a Critic

Asynchronous Advantage Actor-Critic (A3C)

Advantage Actor-Critic (A2C)

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, Koray Kavukcuoglu, "Asynchronous Methods for Deep Reinforcement Learning", ICML, 2016

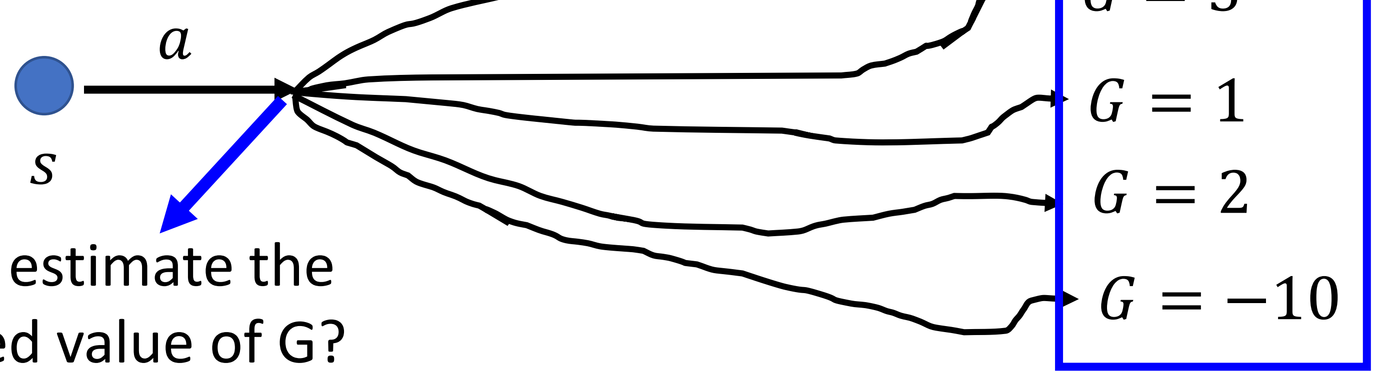
導入 Actor：在 Policy Gradient 中，主要在 Maximize Rbar

Review – Policy Gradient

$$\nabla \bar{R}_\theta \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\left(\boxed{\sum_{t'=t}^{T_n}\gamma^{t'-t}r_{t'}^{n}} - \overset{\text{baseline}}{\underline{b}}\right)\nabla log \boxed{p_\theta(a_t^n|s_t^n)}$$

G_t^n : obtained via interaction
Very unstable

With sufficient samples,
approximate the expectation of G.



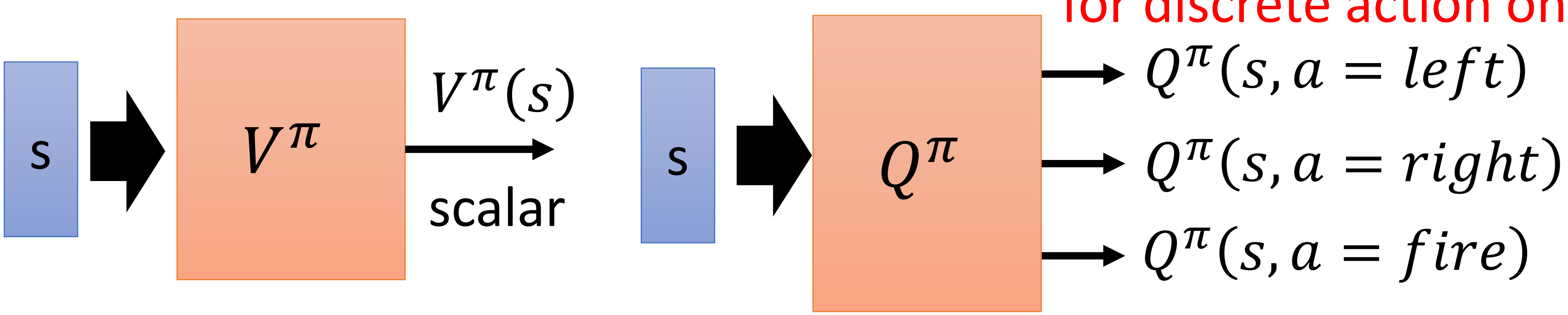
Can we estimate the
expected value of G?

導入 Critic：主要有 State Value Function V(s) 與 State-Action Value Function Q(s, a)

Review – Q-Learning

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

for discrete action only



Estimated by TD or MC

將 Actor 與 Critic 結合！

Actor-Critic

3.

$$Q^{\pi_\theta}(s_t^n, a_t^n) - V^{\pi_\theta}(s_t^n)$$

2.

$$\boxed{V^{\pi_\theta}(s_t^n)}$$

baseline

$$\nabla \bar{R}_\theta \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\boxed{\left(\underbrace{\sum_{t'=t}^{T_n}\gamma^{t'-t} r_{t'}^n}_{G_t^n} - \underline{b}\right)}\nabla \log p_\theta(a_t^n|s_t^n)$$

G_t^n : obtained via interaction

1.

$$\boxed{E[G_t^n] = Q^{\pi_\theta}(s_t^n, a_t^n)}$$

Advantage Actor-Critic

這樣需要用到兩個 Network => 增加誤差

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

Estimate two networks? We can only estimate one.

$$r_t^n + V^{\pi}(s_{t+1}^n) - V^{\pi}(s_t^n)$$

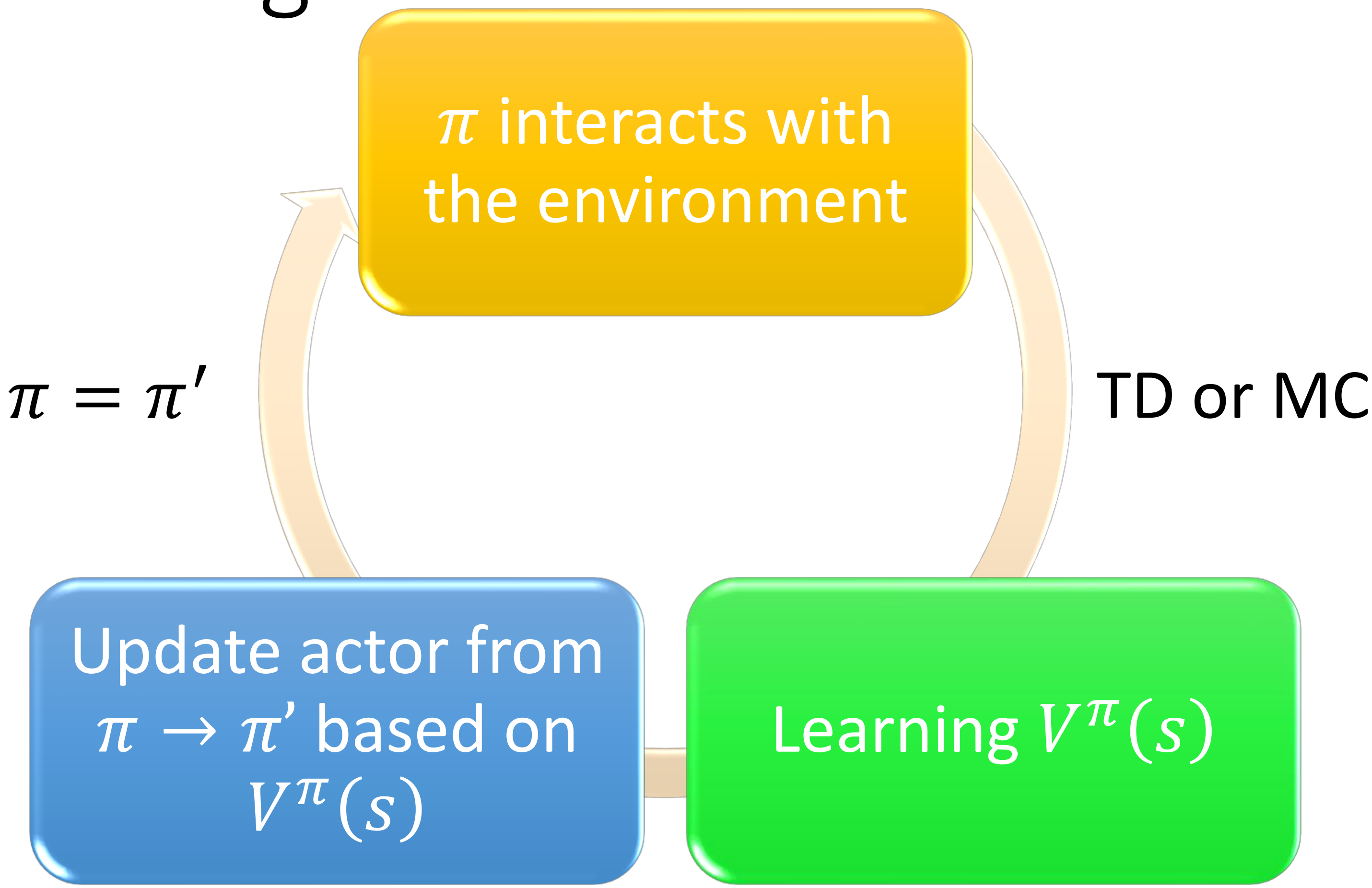
Only estimate state value

A little bit variance

$$Q^{\pi}(s_t^n, a_t^n) = E[r_t^n + V^{\pi}(s_{t+1}^n)]$$

$$Q^{\pi}(s_t^n, a_t^n) = r_t^n + V^{\pi}(s_{t+1}^n)$$

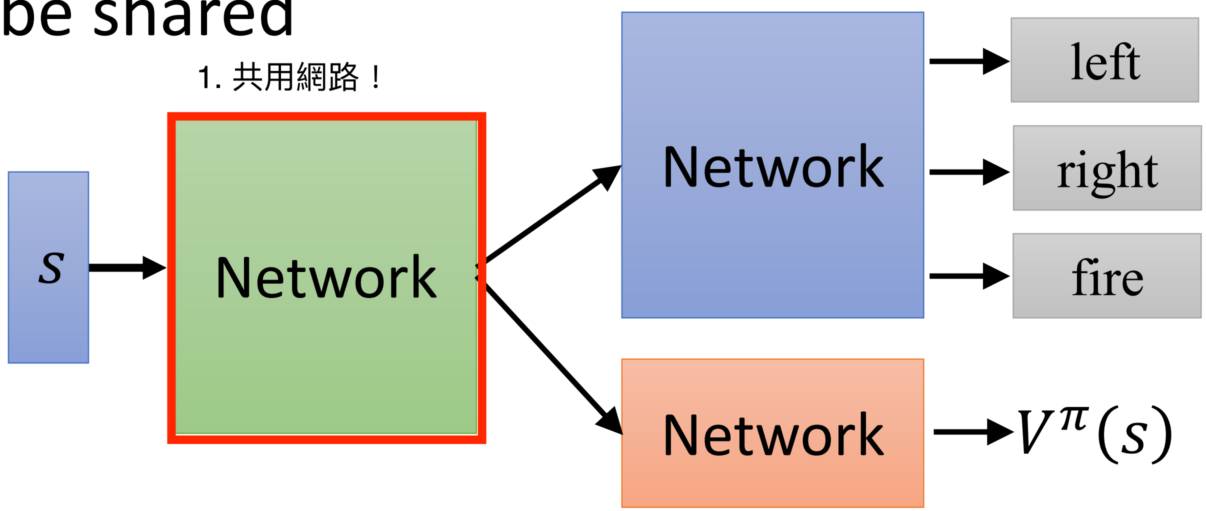
Advantage Actor-Critic



$$\nabla \bar{R}_{\theta} \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\overbrace{\boxed{r_t^n + V^{\pi}(s_{t+1}^n) - V^{\pi}(s_t^n)}}^{\text{Advantage Function}} \right) \nabla log p_{\theta}(a_t^n | s_t^n)$$

Advantage Actor-Critic

- Tips
  - The parameters of actor $\pi(s)$ and critic $V^{\pi}(s)$ can be shared



- Use output entropy as regularization for $\pi(s)$
  - Larger entropy is preferred → exploration

2. Exploration

Asynchronous Advantage Actor-Critic (A3C)

The idea is from 李思叡

Asynchronous

Source of image: https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2#.68x6na7o9

1. Copy global parameters
2. Sampling some data
3. Compute gradients
4. Update global models