# Advantage Actor-Critic (A2C)

**Shusen Wang**

http://wangshusen.github.io/

# Actor and Critic

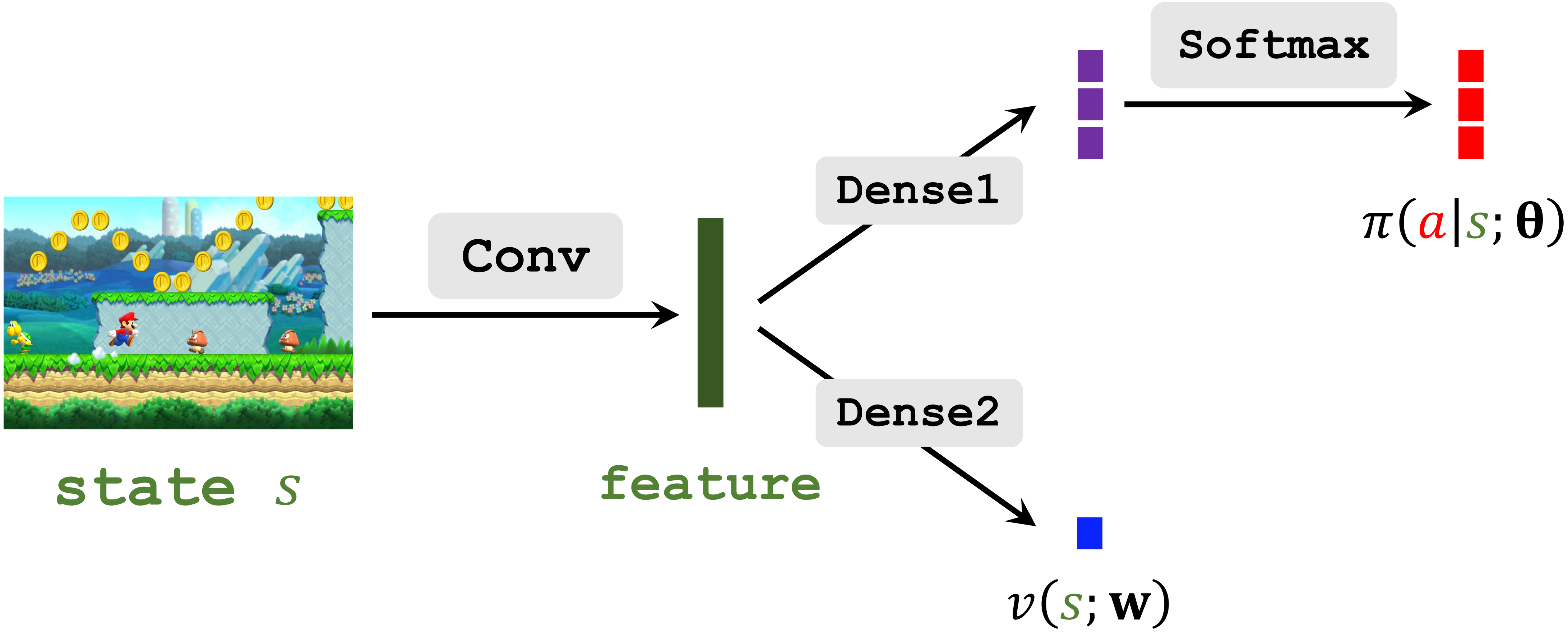
- **Policy network (actor):** $\pi(a|s;\boldsymbol{\theta})$.
  - It is an approximation to the policy function, $\pi(a|s)$.
  - It controls the agent.

# Actor and Critic

- **Policy network (actor):** $\pi(a|s;\boldsymbol{\theta})$.
  - It is an approximation to the policy function, $\pi(a|s)$.
  - It controls the agent.

- **Value network (critic):** $v(s;\mathbf{w})$.
  - It is an approximation to the state-value function, $V_\pi(s)$.
  - It evaluates how good the state $s$ is.

# Actor and Critic



# Training of A2C

- Observe a transition $(s_t, a_t, r_t, s_{t+1})$.
- TD target: $y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.

# Training of A2C

- Observe a transition $(s_t, a_t, r_t, s_{t+1})$.
- TD target: $y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.
- TD error: $\delta_t = v(s_t; \mathbf{w}) - y_t$.

# Training of A2C

- Observe a transition $(s_t, a_t, r_t, s_{t+1})$.
- TD target: $y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.
- TD error: $\delta_t = v(s_t; \mathbf{w}) - y_t$.
- Update the policy network (actor) by:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \cdot \delta_t \cdot \frac{\partial \ln \pi(a_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}}.$$

# Training of A2C

- Observe a transition $(s_t, a_t, r_t, s_{t+1})$.
- TD target: $y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.
- TD error: $\delta_t = v(s_t; \mathbf{w}) - y_t$.
- Update the policy network (actor) by:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \cdot \delta_t \cdot \frac{\partial \ln \pi(a_t \mid s_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

- Update the value network (critic) by:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \delta_t \cdot \frac{\partial\, v(s_t; \mathbf{w})}{\partial\, \mathbf{w}}.$$

# Outline

1. Value functions and Monte Carlo approximations.
2. Updating policy network.
3. Updating value network.

# Properties of Value Functions

# Value Functions

- Discounted return:

$$U_t = R_t + \gamma \cdot R_{t+1} + \gamma^2 \cdot R_{t+2} + \gamma^3 \cdot R_{t+3} + \cdots$$

- Action-value function:

$$Q_\pi(s_t, a_t) = \mathbb{E}[U_t \mid s_t, a_t].$$

- State-value function:

$$V_\pi(s_t) = \mathbb{E}_{A_t}[Q_\pi(s_t, A_t) \mid s_t].$$

# Property of Action-Value Function

**Identity:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}}[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})].$

# Property of Action-Value Function

**Identity:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}}\left[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})\right].$

# Property of Action-Value Function

**Identity:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}}\left[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})\right].$

- Thus, $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}\left[R_t + \gamma \cdot \mathbb{E}_{A_{t+1}}[Q_\pi(S_{t+1}, A_{t+1})]\right]$

# Property of Action-Value Function

**Identity:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}}\left[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})\right].$

- Thus, $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}\left[R_t + \gamma \cdot \mathbb{E}_{A_{t+1}}\left[Q_\pi(S_{t+1}, A_{t+1})\right]\right].$

$$= V_\pi(s_{t+1}).$$

# Property of Action-Value Function

**Identity:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}}[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})].$

- Thus, $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}\left[R_t + \gamma \cdot \mathbb{E}_{A_{t+1}}[Q_\pi(S_{t+1}, A_{t+1})]\right]$

$$= \mathbb{E}_{S_{t+1}}\left[R_t + \gamma \cdot \underline{V_\pi(S_{t+1})}\right].$$

# Property of Action-Value Function

**Identity:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}}[R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1})].$

- Thus, $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}\Big[R_t + \gamma \cdot \mathbb{E}_{A_{t+1}}[Q_\pi(S_{t+1}, A_{t+1})]\Big]$

$$= \mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})].$$

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})].$

# Property of Action-Value Function

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})]$.

# Property of State-Value Function

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})]$.

- By definition, $V_\pi(s_t) = \mathbb{E}_{A_t}[Q_\pi(s_t, A_t)]$

# Property of State-Value Function

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})]$.

- By definition, $V_\pi(s_t) = \mathbb{E}_{A_t}[Q_\pi(s_t, A_t)]$

$$= \mathbb{E}_{A_t}\left[\mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})]\right]$$

# Property of State-Value Function

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})]$.

- By definition, $V_\pi(s_t) = \mathbb{E}_{A_t}[Q_\pi(s_t, A_t)]$

$$= \mathbb{E}_{A_t}\left[\mathbb{E}_{S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})]\right]$$

**Theorem 2:** $V_\pi(s_t) = \mathbb{E}_{A_t, S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})]$.

# Monte Carlo Approximations

# Approximation to Action-Value

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}\left[R_t + \gamma \cdot V_\pi(S_{t+1})\right].$

# Approximation to Action-Value

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}\left[R_t + \gamma \cdot V_\pi(S_{t+1})\right].$

- Suppose we know $(s_t, a_t, r_t, s_{t+1})$.
- Unbiased estimation:

$$Q_\pi(s_t, a_t) \approx r_t + \gamma \cdot V_\pi(s_{t+1}).$$

# Approximation to Action-Value

**Theorem 1:** $Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}\left[R_t + \gamma \cdot V_\pi(S_{t+1})\right].$

- Suppose we know $(s_t, a_t, r_t, s_{t+1})$.
- Unbiased estimation:

$$Q_\pi(s_t, a_t) \approx r_t + \gamma \cdot V_\pi(s_{t+1}).$$

# Approximation to State-Value

**Theorem 2:** $V_\pi(s_t) = \mathbb{E}_{A_t, S_{t+1}}\left[R_t + \gamma \cdot V_\pi(S_{t+1})\right].$

# Approximation to State-Value

**Theorem 2:** $V_\pi(s_t) = \mathbb{E}_{A_t, S_{t+1}}[R_t + \gamma \cdot V_\pi(S_{t+1})].$

- Suppose we know $(s_t, a_t, r_t, s_{t+1})$.
- Unbiased estimation:

$$V_\pi(s_t) \approx r_t + \gamma \cdot V_\pi(s_{t+1}).$$

# Approximations

- Approximation to action-value:

$$Q_\pi(s_t, a_t) \;\approx\; r_t + \gamma \cdot V_\pi(s_{t+1}).$$

- Approximation to state-value:

$$V_\pi(s_t) \;\approx\; r_t + \gamma \cdot V_\pi(s_{t+1}).$$

# Updating Policy Network

# Policy Gradient with Baseline

**Stochastic policy gradient:**

$$\mathbf{g}(a_t) = \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \big( Q_\pi(s_t, a_t) - V_\pi(s_t) \big).$$

**Advantage function**

# Policy Gradient with Baseline

**Stochastic policy gradient:**

$$\mathbf{g}(a_t) = \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \big(Q_\pi(s_t, a_t) - V_\pi(s_t)\big).$$

Unknown

# MC Approximation to Action-Value

**Stochastic policy gradient:**

$$\mathbf{g}(a_t) = \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \big( Q_\pi(s_t, a_t) - V_\pi(s_t) \big).$$

$$\approx r_t + \gamma \cdot V_\pi(s_{t+1}).$$

# MC Approximation to Action-Value

**Approximate stochastic policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \big( r_t + \gamma \cdot V_\pi(s_{t+1}) - V_\pi(s_t) \big).$$

# Function Approximation to State-Value

**Approximate stochastic policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \big(r_t + \gamma \cdot \underline{V_\pi(s_{t+1})} - \underline{V_\pi(s_t)}\big).$$

- Approximate $V_\pi(s)$ by the value network $\underline{v(s; \mathbf{w})}$.

# Function Approximation to State-Value

**Approximate stochastic policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial \,\boldsymbol{\theta}} \cdot \left( r_t + \gamma \cdot \boxed{v(s_{t+1};\mathbf{w})} - \boxed{v(s_t;\mathbf{w})} \right).$$

# Updating Policy Network

**Approximate stochastic policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \Big(\boxed{r_t + \gamma \cdot v(s_{t+1};\mathbf{w})} - v(s_t;\mathbf{w})\Big).$$

Denote it by $y_t$

# Updating Policy Network

**Approximate stochastic policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \Big(\boxed{r_t + \gamma \cdot v(s_{t+1};\mathbf{w})} - v(s_t;\mathbf{w})\Big).$$

Denote it by $y_t$

**Policy gradient ascent:**

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta \cdot \frac{\partial \ln \pi(a_t \mid s_t;\boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \Big(y_t - v(s_t;\mathbf{w})\Big).$$

# Updating Value Network

# Derive TD Target

**MC approximation:** $V_\pi(s_t) \approx r_t + \gamma \cdot V_\pi(s_{t+1})$.

# Derive TD Target

**MC approximation:** $\underline{V_\pi(s_t)} \approx r_t + \gamma \cdot \underline{V_\pi(s_{t+1})}$.

$$V_\pi(s_t) \downarrow v(s_t; \mathbf{w}) \qquad V_\pi(s_{t+1}) \downarrow v(s_{t+1}; \mathbf{w})$$

# Derive TD Target

**Approximation:** $v(s_t; \mathbf{w}) \approx r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.

# Derive TD Target

**Approximation:** $v(s_t; \mathbf{w}) \approx \underline{r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})}$.

TD target $y_t$

**TD learning:** Encourage $\underline{v(s_t; \mathbf{w})}$ to approach $y_t$.

# Updating Value Network

**TD learning:** Encourage $v(s_t; \mathbf{w})$ to approach $y_t$.

# Updating Value Network

**TD learning:** Encourage $v(s_t; \mathbf{w})$ to approach $y_t$.

- TD error: $\delta_t = v(s_t; \mathbf{w}) - y_t$.
- Gradient: $\frac{\partial\, \delta_t^2/2}{\partial\, \mathbf{w}}$

# Updating Value Network

**TD learning:** Encourage $v(s_t; \mathbf{w})$ to approach $y_t$.

- TD error: $\delta_t = v(s_t; \mathbf{w}) - y_t$.
- Gradient: $\frac{\partial\, \delta_t^2/2}{\partial\, \mathbf{w}} = \delta_t \cdot \frac{\partial\, v(s_t;\mathbf{w})}{\partial\, \mathbf{w}}$.

# Updating Value Network

**TD learning:** Encourage $v(s_t; \mathbf{w})$ to approach $y_t$.

- TD error: $\delta_t = v(s_t; \mathbf{w}) - y_t$.
- Gradient: $\frac{\partial\, \delta_t^2/2}{\partial\, \mathbf{w}} = \delta_t \cdot \frac{\partial\, v(s_t;\mathbf{w})}{\partial\, \mathbf{w}}$.
- Update value network by gradient descent:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \delta_t \cdot \frac{\partial\, v(s_t;\mathbf{w})}{\partial\, \mathbf{w}}.$$

# More Explanations

# Approximate Policy Gradient

**Approximate policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \boxed{\big(r_t + \gamma \cdot v(s_{t+1};\mathbf{w}) - v(s_t;\mathbf{w})\big)}$$

evaluation made by the critic

# Approximate Policy Gradient

**Approximate policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(r_t + \gamma \cdot v(s_{t+1}; \mathbf{w}) - \boxed{v(s_t; \mathbf{w})}\right).$$

Approximation to $\mathbb{E}[\underline{U_t} \mid s_t]$

At time $t$, the critic evaluates how good $s_t$ is.

# Approximate Policy Gradient

**Approximate policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \Big(\boxed{r_t + \gamma \cdot v(s_{t+1};\mathbf{w})} - v(s_t;\mathbf{w})\Big).$$

Approximation to $\mathbb{E}[U_t \mid s_t, s_{t+1}]$

At time $t+1$, the critic evaluates how good $s_t$ is.

# Approximate Policy Gradient

**Approximate policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \left( \boxed{r_t + \gamma \cdot v(s_{t+1};\mathbf{w})} - \boxed{v(s_t;\mathbf{w})} \right).$$

Both are approximations to $\mathbb{E}[U_t]$.

Both evaluate how good $s_t$ is.

# Approximate Policy Gradient

**Approximate policy gradient:**

$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \Big( \boxed{r_t + \gamma \cdot v(s_{t+1};\mathbf{w})} - \boxed{v(s_t;\mathbf{w})} \Big).$$

Depends on $a_t$

Independent of $a_t$

If $a_t$ is good, their difference is positive.

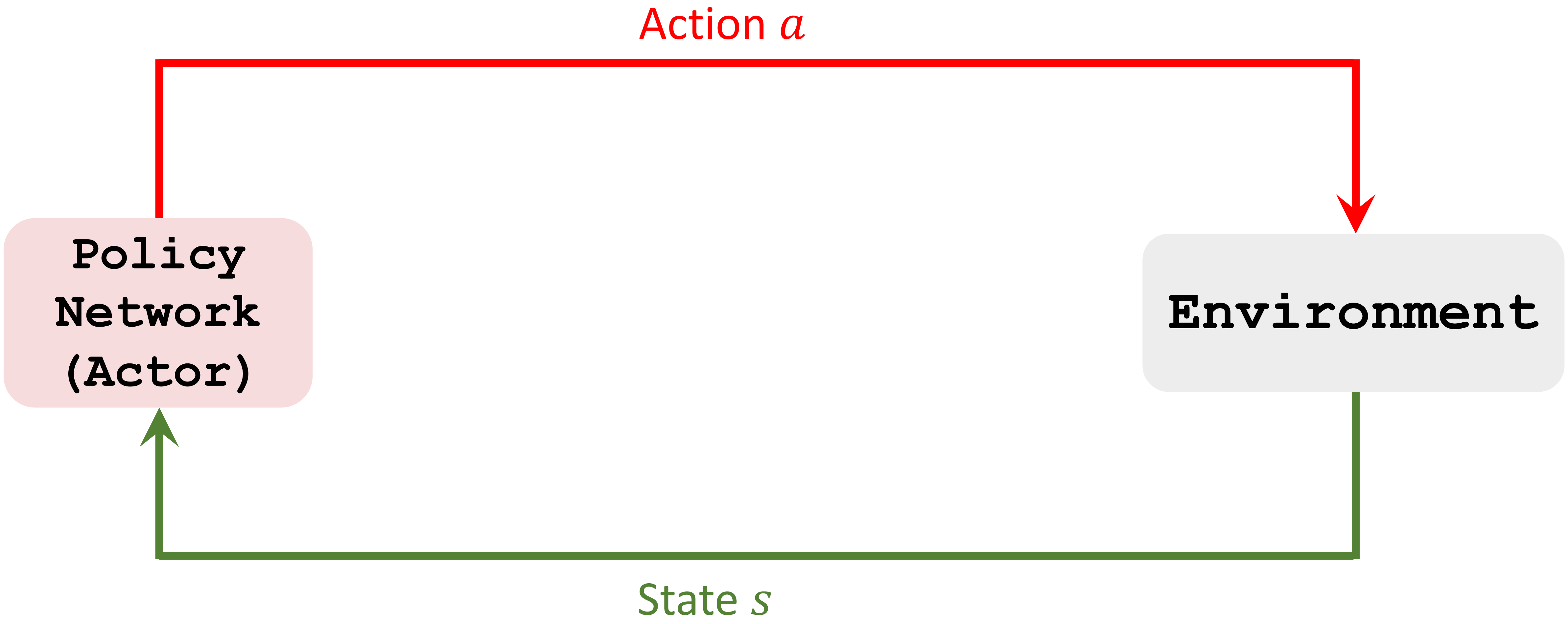
# Approximate Policy Gradient

**Approximate policy gradient:**

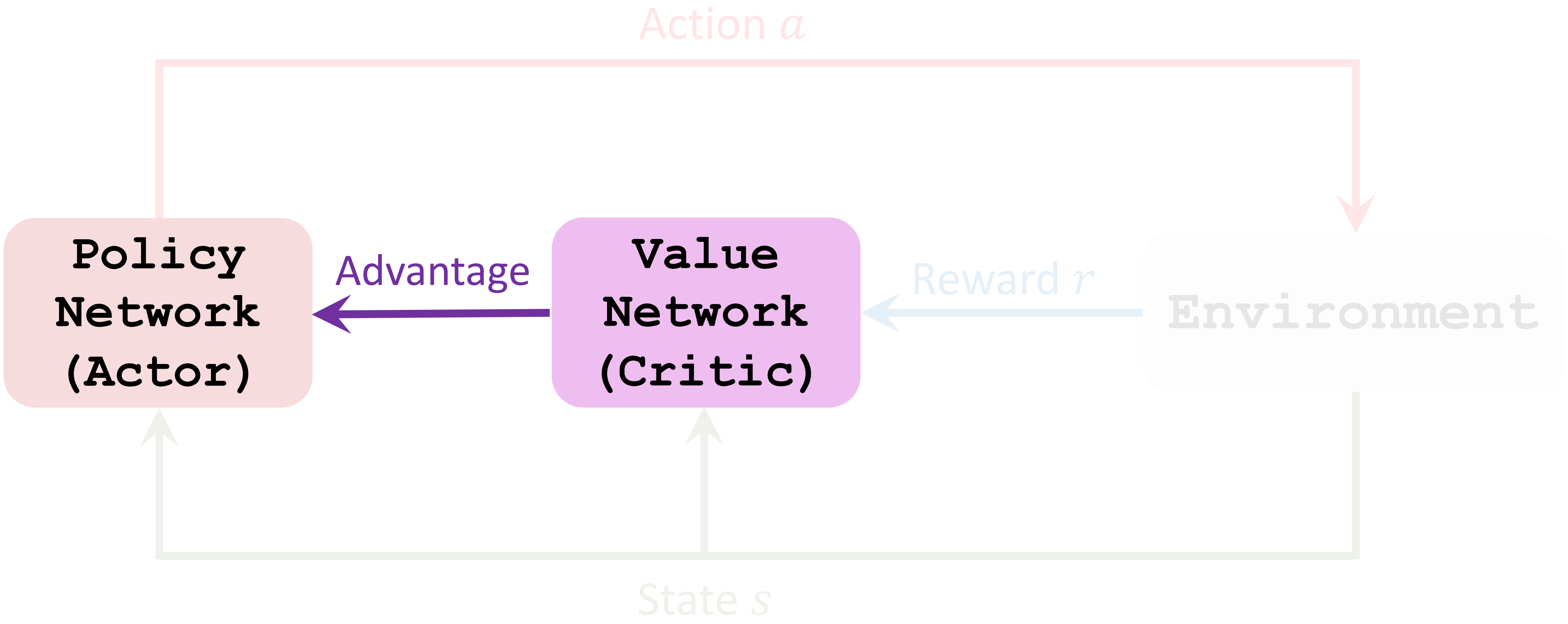
$$\mathbf{g}(a_t) \approx \frac{\partial \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial \, \boldsymbol{\theta}} \cdot \boxed{\left(r_t + \gamma \cdot v(s_{t+1}; \mathbf{w}) - v(s_t; \mathbf{w})\right)}$$

evaluation made by the critic

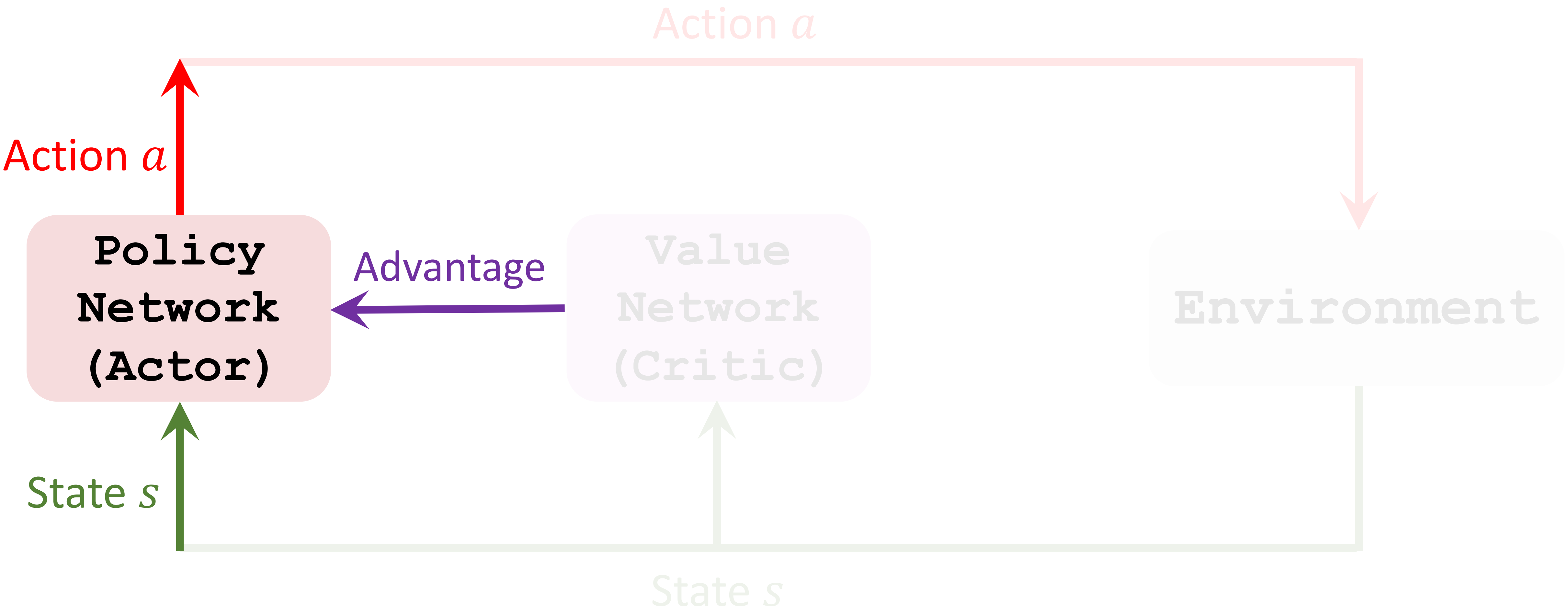
# Advantage Actor-Critic (A2C)



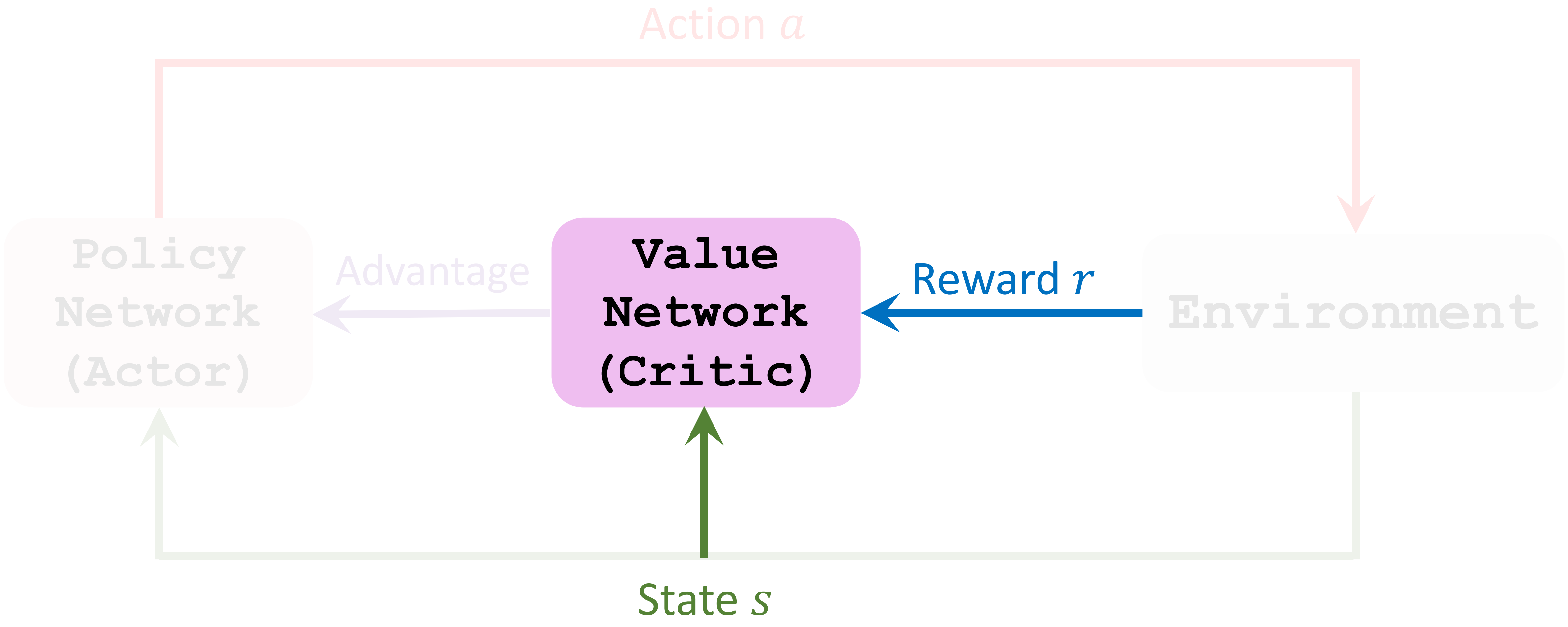
# Advantage Actor-Critic (A2C)



# Advantage Actor-Critic (A2C)



# Advantage Actor-Critic (A2C)



# Thank you!