Advantage Actor-Critic (A2C)

Shusen Wang

Actor and Critic

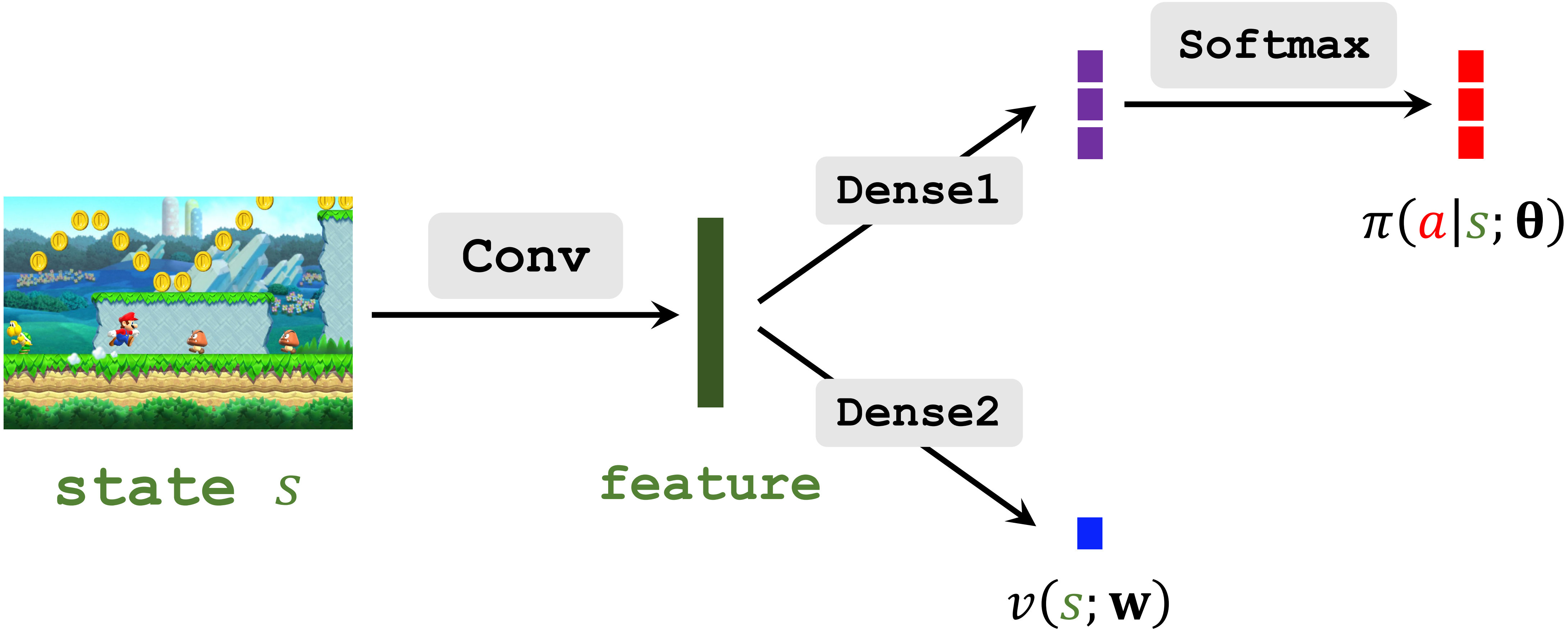
- **Policy network (actor):** $\pi(a|s;\boldsymbol{\theta})$
  - It is an approximation to the policy function, $\pi(a|s)$.
  - It controls the agent.

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

- Observing 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:

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

- 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:

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

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

Property of Action-Value Function

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

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

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

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

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

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

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

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

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

Monte Carlo Approximations

Approximation to Action-Value

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

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

- 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 with baseline:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \Big(Q_\pi(s_t, a_t) - V_\pi(s_t)\Big).$$

Advantage function

Policy Gradient with Baseline

Stochastic policy gradient with baseline:

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

Unknown

MC Approximation to Action-Value

Approximate policy gradient with baseline:

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

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

MC Approximation to Action-Value

Approximate policy gradient with baseline:

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

Function Approximation to State-Value

Approximate policy gradient with baseline:

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

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

Function Approximation to State-Value

Approximate policy gradient with baseline:

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

Function Approximation to State-Value

Approximate policy gradient with baseline:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \Big(\underbrace{r_t + \gamma \cdot v(s_{t+1};\mathbf{w})}_{\text{Denote it by } y_t} - v(s_t;\mathbf{w})\Big).$$

Function Approximation to State-Value

Approximate policy gradient with baseline:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} \approx \frac{\partial \ln \pi(a_t|s_t;\boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \Big(\underbrace{r_t + \gamma \cdot v(s_{t+1};\mathbf{w})}_{\text{Denote it by } y_t} - v(s_t;\mathbf{w})\Big).$$

Policy gradient ascent:

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

Updating Value Network

TD Target

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

TD Target

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

- Thus, $v(s_t; \mathbf{w}) \approx r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.

TD target y_t

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.

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: $\boxed{\frac{\partial\, \delta_t^2/2}{\partial\, \mathbf{w}}}$

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:

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

Summary of Algorithm

- Observing 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:

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

- Update the value network (critic) by:

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

Thank you!