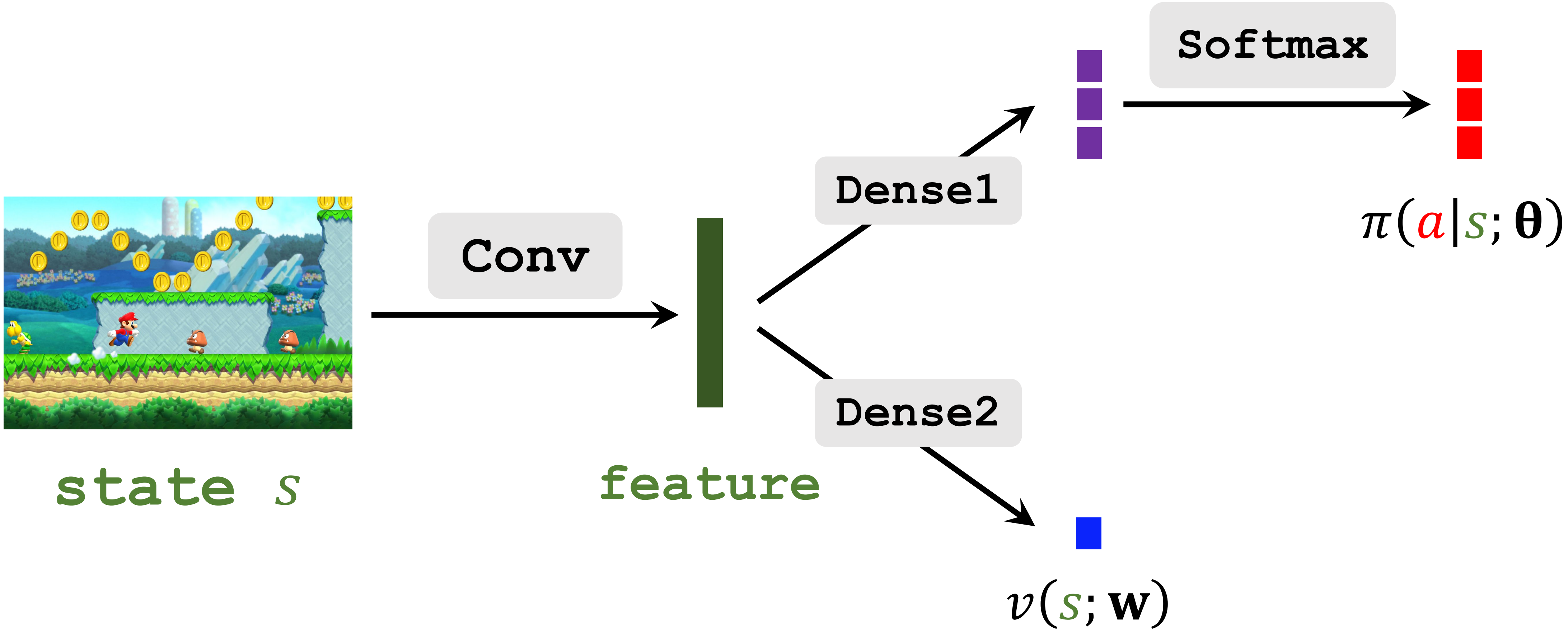
# REINFORCE versus A2C

**Shusen Wang**

http://wangshusen.github.io/

# Policy and Value Networks



state $s$

Conv

feature

Dense1

Dense2

Softmax

$\pi(a|s;\boldsymbol{\theta})$

$v(s;\mathbf{w})$

# A2C with Multi-Step TD Target

# Advantage Actor-Critic (A2C)

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

# Advantage Actor-Critic (A2C)

- 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: $\delta_t = v(s_t; \mathbf{w}) - y_t$.

# Advantage Actor-Critic (A2C)

- 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:

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

- Update the value network (critic) by:

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

# Advantage Actor-Critic (A2C)

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

Use multi-step TD target instead.

- 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:

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

# One-Step VS Multi-Step Target

- Observing a transition $(s_t, a_t, r_t, s_{t+1})$
- **One-step TD target:**

$$y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w}).$$

# One-Step VS Multi-Step Target

- Observing a transition $(s_t, a_t, r_t, s_{t+1})$
- **One-step TD target:**

$$y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w}).$$

- Observing $m$ transitions: $\{(s_{t+i}, a_{t+i}, r_{t+i}, s_{t+i+1})\}_{i=0}^{m-1}$.
- ***m*-step TD target:**

$$y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot v(s_{t+m}; \mathbf{w}).$$

# A2C with Multi-Step TD Target

- Observing a trajectory from time $t$ to $t+m-1$.
- TD target: $\boxed{y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot v(s_{t+m}; \mathbf{w}).}$
- 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:

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

# REINFORCE with Baseline

# REINFORCE with Baseline

- Observing a trajectory from time $t$ to $n$.
- Return: $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$.
- Error: $\delta_t = v(s_t; \mathbf{w}) - u_t$.

# REINFORCE with Baseline

- Observing a trajectory from time $t$ to $n$.
- Return: $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$.
- Error: $\delta_t = v(s_t; \mathbf{w}) - u_t$.
- Update the policy network 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 by:

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

# REINFORCE with Baseline

- Observing a trajectory from time $t$ to $n$.
- Return: $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$.
- Error: $\delta_t = v(s_t; \mathbf{w}) - u_t$.
- 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:

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

# A2C versus REINFORCE

# TD Target versus Return

A2C with $m$-step TD target: $y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot v(s_{t+m}; \mathbf{w})$.

# TD Target versus Return

A2C with one-step TD target: $y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.

↑ Use only one reward ($m = 1$)

A2C with $m$-step TD target: $y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot v(s_{t+m}; \mathbf{w})$.

# TD Target versus Return

A2C with one-step TD target: $y_t = r_t + \gamma \cdot v(s_{t+1}; \mathbf{w})$.

↑ Use only one reward ($m = 1$)

A2C with $m$-step TD target: $y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot v(s_{t+m}; \mathbf{w})$.

↓ Use all the rewards

REINFORCE: $y_t$ becomes $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$.

# A2C versus REINFORCE

- A2C uses m-step TD target (with bootstrapping):

$$y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot v(s_{t+m}; \mathbf{w}).$$

# A2C versus REINFORCE

- A2C uses m-step TD target (with bootstrapping):

$$y_t = \sum_{i=0}^{m-1} \gamma^i \cdot r_{t+i} + \gamma^m \cdot v(s_{t+m}; \mathbf{w}).$$

- REINFORCE uses observed return (without bootstrapping):

$$u_t = \sum_{i=0}^{n-t} \gamma^i \cdot r_{t+i}.$$

# Thank you!