REINFORCE with Baseline

Shusen Wang

http://wangshusen.github.io/

Value Functions

- Discounted return:

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

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:

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

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:

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

Approximations to Policy Gradient

Policy Gradient

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\frac{\partial\, \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - V_\pi(s_t)\right)\right].$$

Policy Gradient

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t\sim\pi}\left[\frac{\partial\, \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - V_\pi(s_t)\right)\right].$$

$$= \mathbf{g}(A_t)$$

Policy Gradient

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \left(Q_\pi(s_t, A_t) - V_\pi(s_t)\right)\right].$$

$$= \mathbf{g}(A_t)$$

- Randomly sample $a_t \sim \pi(\cdot \mid s_t; \boldsymbol{\theta})$.
- Then $\mathbf{g}(a_t)$ is an unbiased estimation of the policy gradient.

Approximations

Policy gradient:

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \big(Q_\pi(s_t, A_t) - V_\pi(s_t)\big)\right].$$

$$= \mathbf{g}(A_t)$$

Stochastic policy gradient:

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

Approximations

Stochastic policy gradient:

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

Approximations

Stochastic policy gradient:

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

Approximations

Stochastic policy gradient:

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

- Recall that $Q_\pi(s_t,a_t) = \mathbb{E}[U_t \mid s_t, a_t]$.
- Monte Carlo approximation to $Q_\pi(s_t,a_t) \approx u_t$ (REINFORCE):

Approximations

Stochastic policy gradient:

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

- Recall that $Q_\pi(s_t, a_t) = \mathbb{E}[U_t \mid s_t, a_t]$.
- Monte Carlo approximation to $Q_\pi(s_t, a_t) \approx u_t$ (REINFORCE):
  - Observing the trajectory: $s_t, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1}, \cdots, s_n, a_n, r_n$.
  - Compute return: $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$.
  - u_t is an unbiased estimate of $Q_\pi(s_t, a_t)$.

Approximations

Stochastic policy gradient:

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

- Approximate $V(s;\boldsymbol{\theta})$ by the value network, $v(s;\mathbf{w})$.

Approximations

Approximate policy gradient:

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

Summary of Approximations

Approximate policy gradient:

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

- Three approximations:

1. Approximate expectation using one sample, a_t. (Monte Carlo.)
2. Approximate $Q_\pi(s_t, a_t)$ by u_t. (Another Monte Carlo.)
3. Approximate $V_\pi(s)$ by the value network, $v(s;\mathbf{w})$.

Summary of Approximations

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\frac{\partial \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \big(Q_\pi(s_t, A_t) - V_\pi(s_t)\big)\right].$$

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

Summary of Approximations

$$\frac{\partial\, V_\pi(s_t)}{\partial\, \boldsymbol{\theta}} = \mathbb{E}_{A_t \sim \pi}\left[\frac{\partial\, \ln \pi(A_t \mid s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \big(Q_\pi(s_t, A_t) - V_\pi(s_t)\big)\right].$$

$$\downarrow$$

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

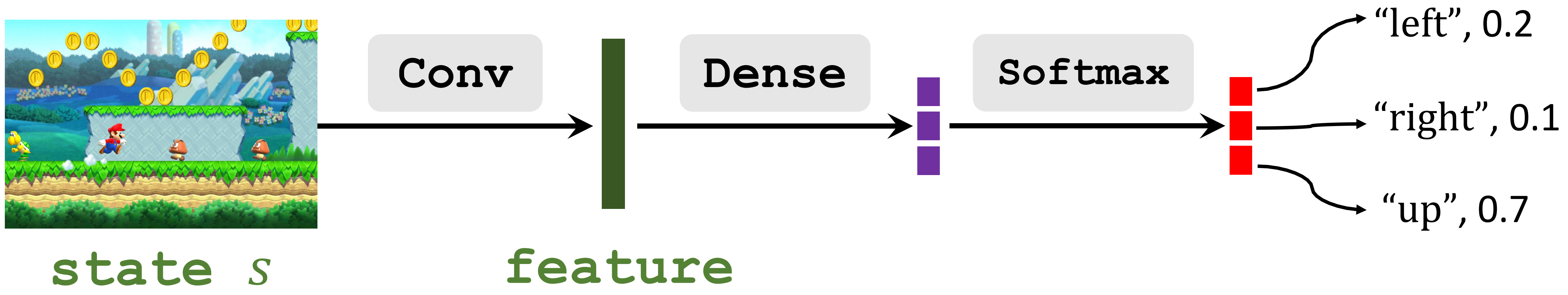
$$\downarrow$$

$$\mathbf{g}(a_t) \approx \frac{\partial\, \ln \pi(a_t | s_t; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}} \cdot \big(u_t - v(s_t; \mathbf{w})\big).$$

Policy and Value Networks

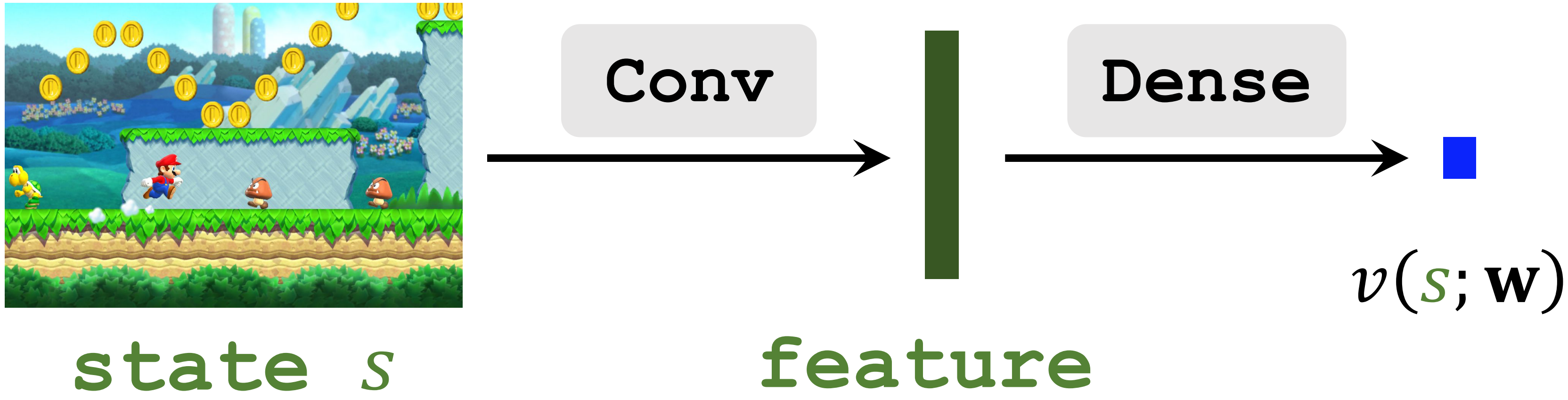
Policy Network

Approximate policy function, $\pi(a|s)$, by policy network, $\pi(a|s;\boldsymbol{\theta})$.

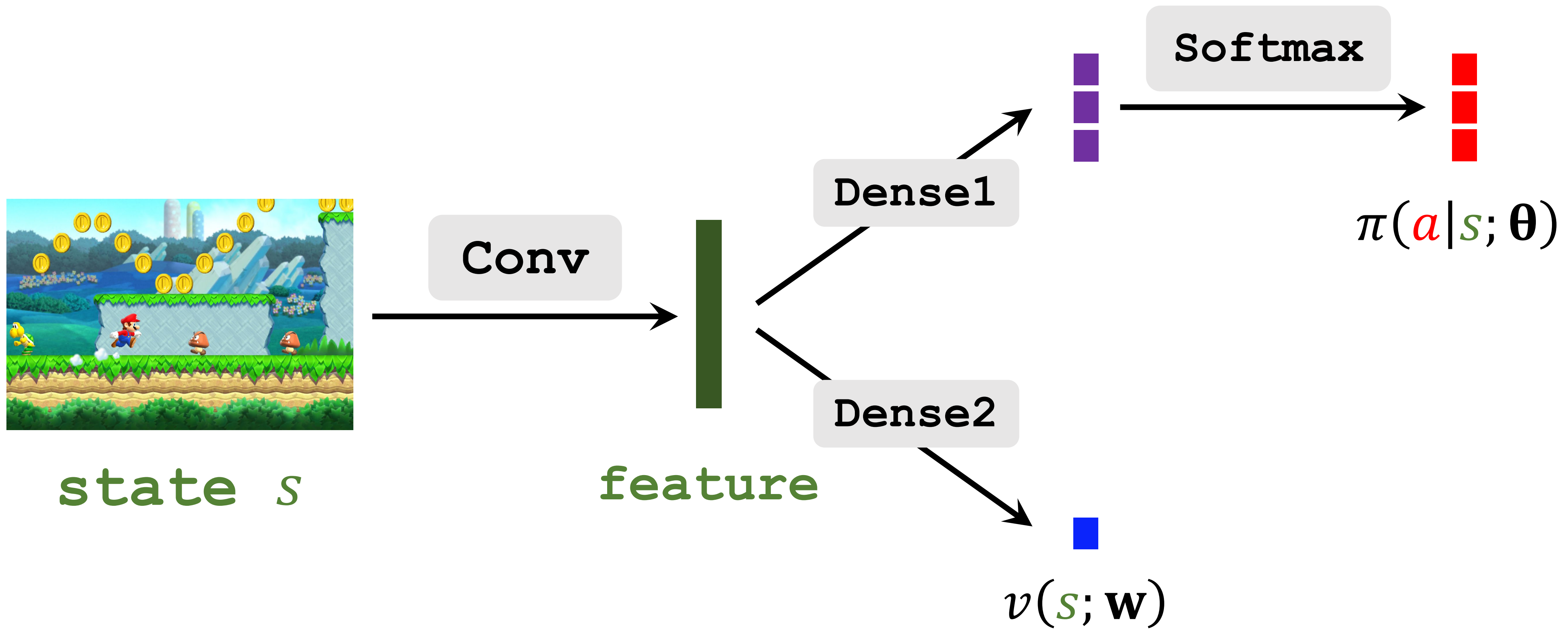


Value Network

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



Parameter Sharing



REINFORCE with Baseline

Updating the policy network

Approximate policy gradient:

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

- Update policy network by policy gradient ascent:

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

Updating the policy network

Approximate policy gradient:

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

- Update policy network by policy gradient ascent:

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

Updating the policy network

Approximate policy gradient:

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

- Update policy network by policy gradient ascent:

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

Updating the value network

- Recall $v(s_t; \mathbf{w})$ is an approximation to $V_\pi(s_t) = \mathbb{E}[U_t \mid s_t]$.

Updating the value network

- Recall $v(s_t; \mathbf{w})$ is an approximation to $V_\pi(s_t) = \mathbb{E}[U_t \mid s_t]$.
- Prediction error: $\delta_t = v(s_t; \mathbf{w}) - u_t$.

Updating the value network

- Recall $v(s_t; \mathbf{w})$ is an approximation to $V_\pi(s_t) = \mathbb{E}[U_t \mid s_t]$.
- Prediction error: $\delta_t = v(s_t; \mathbf{w}) - u_t$.
- Gradient: $\frac{\partial\, \delta_t^2/2}{\partial\, \mathbf{w}} = \delta_t \cdot \frac{\partial\, v(s_t;\mathbf{w})}{\partial\, \mathbf{w}}$.

Updating the value network

- Recall $v(s_t; \mathbf{w})$ is an approximation to $V_\pi(s_t) = \mathbb{E}[U_t \mid s_t]$.
- Prediction error: $\delta_t = v(s_t; \mathbf{w}) - u_t$.
- Gradient: $\frac{\partial\, \delta_t^2/2}{\partial\, \mathbf{w}} = \delta_t \cdot \frac{\partial\, v(s_t;\mathbf{w})}{\partial\, \mathbf{w}}$.
- Gradient descent:

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

Summary of Algorithm

- Play a game to the end and observe the trajectory:

$$s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_n, a_n, r_n \ .$$

- Compute $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$ and $\delta_t = v(s_t; \mathbf{w}) - u_t$.

Summary of Algorithm

- Play a game to the end and observe the trajectory:

$$s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_n, a_n, r_n \,.$$

- Compute $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$ and $\delta_t = v(s_t; \mathbf{w}) - u_t$.
- 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:

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

Summary of Algorithm

- Play a game to the end and observe the trajectory:

$$s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_n, a_n, r_n \; .$$

- Compute $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$ and $\delta_t = v(s_t; \mathbf{w}) - u_t$.
- 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:

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

Summary of Algorithm

- Play a game to the end and observe the trajectory:

$$s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_n, a_n, r_n \ .$$

- Compute $u_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$ and $\delta_t = v(s_t; \mathbf{w}) - u_t$.
- 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:

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

Repeat this procedure for $t = 1, \cdots, n$.

Thank you!