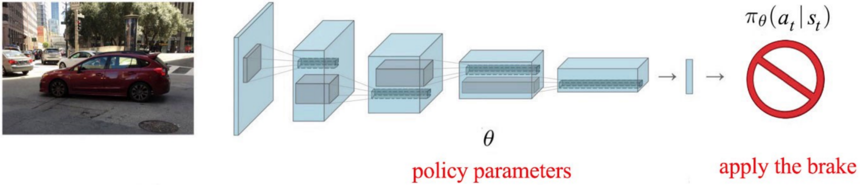
# Actor-Critic Methods

**Insoon Yang**

Department of Electrical and Computer Engineering
Seoul National University

CORE
Control + Optimization Research Lab

# Recap: Parameterizing Policy

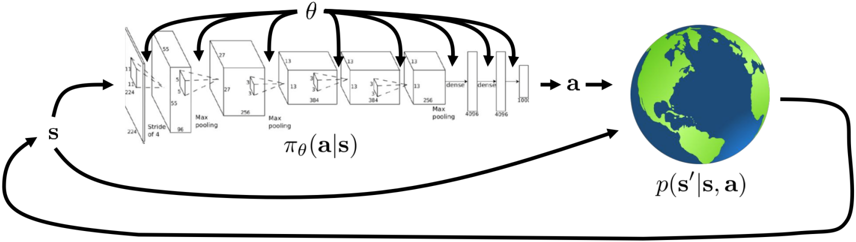



Q) How can we find a good $\pi(a|s)$, which is a **function**?

Idea:

- Parameterize policy by a parameter vector $\theta \in \mathbb{R}^\ell$: $\pi_\theta(a|s)$
- Find an optimal $\theta$

# Recap: How to find optimal parameters $\theta$?




- Let $\tau := (s_0, a_0, \dots, s_T, a_T)$ denote the state-action trajectory
- By Markov property,

$$p_\theta(\tau) = p(s_0) \prod_{t=0}^{T} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

- Approximate MDP problem:

$$\max_\theta \; \mathbb{E}_{\tau\sim p_\theta(\tau)} \left[ \sum_t r(s_t, a_t) \right] =: J(\theta)$$

# Recap: Policy Gradient Theorem & REINFORCE

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[ \left( \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \left( \sum_{t=0}^{T} r(s_t, a_t) \right) \right]$$

Initialize $\theta$;

1. Sample $\{\tau^i\}_{i=1}^N := \{(s_0^i, a_0^i, \ldots, s_T^i, a_T^i)\}_{i=1}^N$ using the current policy $\pi_\theta(a_t|s_t)$
2. Estimate the gradient

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left( \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) \right) \left( \sum_{t=0}^{T} r(s_t^i, a_t^i) \right)$$

3. Perform gradient ascent:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta);$$

# Recap: Policy Gradient with Baselines

- Policy gradient

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i | s_t^i)[Q^{\pi_\theta}(s_t^i, a_t^i) - b]$$

# Recap: Policy Gradient with Baselines

- Policy gradient

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i)[Q^{\pi_\theta}(s_t^i, a_t^i) - b]$$

- Good baseline: $b = v^{\pi_\theta}(s_t, a_t)$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) A^{\pi_\theta}(s_t^i, a_t^i),$$

where $A^{\pi_\theta}(s_t^i, a_t^i) := Q^{\pi_\theta}(s_t^i, a_t^i) - v^{\pi_\theta}(s_t^i)$ is called the advantage function

# How can we compute a good gradient?

Good gradient:

1. Unbiased (Ok!)
2. Low variance (How?)

# How can we compute a good gradient?

Good gradient:

1. Unbiased (Ok!)
2. Low variance (How?)

- Use baseline $b = v^{\pi_\theta}(s_t, a_t)$:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i | s_t^i) \underbrace{(Q^{\pi_\theta}(s_t^i, a_t^i) - v^{\pi_\theta}(s_t^i))}_{A^{\pi_\theta}(s_t^i, a_t^i)}$$

- Need to accurately estimate $v^\pi, Q^\pi$ or $A^\pi$ (**Policy evaluation**)

# Policy evaluation

Q) How to evaluate $v^{\pi}(s_t)$?

# Policy evaluation

Q) How to evaluate $v^\pi(s_t)$?

- Monte Carlo:

$$v^\pi(s_t) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t'=t}^{T} r_{t'}^{i}$$

# Policy evaluation with function approximation

Want to fit the value function by

$$v^{\pi}(s_t) \approx v_{\phi}^{\pi}(s_t) \quad \phi : \text{parameter vector}$$

Q) What are the training data?

# Policy evaluation with function approximation

Want to fit the value function by

$$v^{\pi}(s_t) \approx v^{\pi}_{\phi}(s_t) \quad \phi : \text{parameter vector}$$

Q) What are the training data?

- $\{(\underbrace{s^i_t}_{s^i}, \underbrace{\sum_{t'=t}^{T} r^i_{t'}}_{y^i})\} =: \{(s^i, y^i)\}$

# Policy evaluation with function approximation

Want to fit the value function by

$$v^\pi(s_t) \approx v^\pi_\phi(s_t) \quad \phi : \text{parameter vector}$$

Q) What are the training data?

- $\{(\underbrace{s_t^i}_{s^i}, \underbrace{\sum_{t'=t}^{T} r_{t'}^i}_{y^i})\} =: \{(s^i, y^i)\}$

Q) How can we find the best $\phi$?

# Policy evaluation with function approximation

Want to fit the value function by

$$v^\pi(s_t) \approx v^\pi_\phi(s_t) \quad \phi : \text{parameter vector}$$

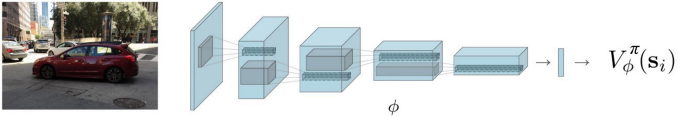
Q) What are the training data?

- $\{(\underbrace{s^i_t}_{s^i}, \underbrace{\sum_{t'=t}^{T} r^i_{t'}}_{y^i})\} =: \{(s^i, y^i)\}$

Q) How can we find the best $\phi$?

- (supervised) regression:

$$\min_\phi \; \mathcal{L}(\phi) := \frac{1}{2}\sum_i \|v^\pi_\phi(s_i) - y_i\|^2$$



# Better version?

Recap) Monte Carlo target: $y_i := \sum_{t'=t}^{T} r_{t'}^{i}$

Q) Can we choose a better target?

## Better version?

Recap) Monte Carlo target: $y_i := \sum_{t'=t}^{T} r_{t'}^i$

Q) Can we choose a better target?

- Use previous fitted value function:

$$y_i := r(s_t^i, a_t^i) + \underbrace{v_\phi^\pi(s_{t+1}^i)}_{\text{fitted value ftn}}$$

# Better version?

Recap) Monte Carlo target: $y_i := \sum_{t'=t}^{T} r_{t'}^i$

Q) Can we choose a better target?

- Use previous fitted value function:

$$y_i := r(s_t^i, a_t^i) + \underbrace{v_\phi^\pi(s_{t+1}^i)}_{\text{fitted value ftn}}$$

- Training data:
  - $\{(\underbrace{s_t^i}_{s^i}, \underbrace{r_t^i + v_\phi^\pi(s_{t+1}^i)}_{y^i})\} =: \{(s^i, y^i)\}$

# Better version?

Recap) Monte Carlo target: $y_i := \sum_{t'=t}^{T} r_{t'}^i$

Q) Can we choose a better target?

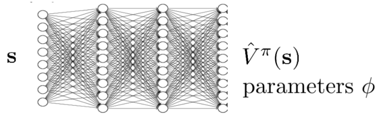
- Use previous fitted value function:

$$y_i := r(s_t^i, a_t^i) + \underbrace{v_\phi^\pi(s_{t+1}^i)}_{\text{fitted value ftn}}$$

- Training data:
  - $\{(\underbrace{s_t^i}_{s^i}, \underbrace{r_t^i + v_\phi^\pi(s_{t+1}^i)}_{y^i})\} =: \{(s^i, y^i)\}$

- Regression:

$$\min_\phi \mathcal{L}(\phi) := \frac{1}{2}\sum_i \|v_\phi^\pi(s_i) - y_i\|^2$$



# Actor-Critic algorithm

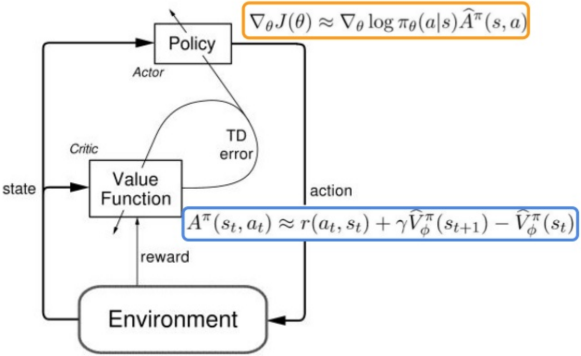
Batch version:

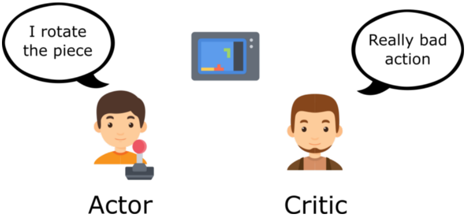
1. Sample $\{(s_t^i, a_t^i, s_{t+1}^i, r_t^i)\}$ using $\pi_\theta(a|s)$;
2. Fit $v_\phi^\pi(s)$ by solving the regression problem $\min_\phi \ \mathcal{L}(\phi) := \frac{1}{2}\sum_i \|v_\phi^\pi(s_i) - y_i\|^2$;
3. Evaluate Advantage $A^\pi(s_t^i, a_t^i) = r_t^i + v_\phi^\pi(s_{t+1}^i) - v_\phi^\pi(s_t^i)$;
4. Estimate SG $\nabla_\theta J(\theta) \approx \sum_i \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) A^\pi(s_t^i, a_t^i)$;
5. Update $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$;

Note:

- Critic: Step 2, 3
- Actor: Step 4, 5

# Intuition behind actor-critic



# Actor-Critic algorithm with discount factor

With discount factor $\gamma$:

1. Sample $\{(s_t^i, a_t^i, s_{t+1}^i, r_t^i)\}$ using $\pi_\theta(a|s)$;
2. Fit $v_\phi^\pi(s)$ by solving the regression problem $\min_\phi \ \mathcal{L}(\phi) := \frac{1}{2}\sum_i \|v_\phi^\pi(s_i) - y_i\|^2$;
3. Evaluate Advantage $A^\pi(s_t^i, a_t^i) = r_t^i + \gamma v_\phi^\pi(s_{t+1}^i) - v_\phi^\pi(s_t^i)$;
4. Estimate SG $\nabla_\theta J(\theta) \approx \sum_i \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) A^\pi(s_t^i, a_t^i)$;
5. Update $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$;

# Online Actor-Critic algorithm

Fully incremental online version:

1. Take action $a \sim \pi_\theta(a|s)$, and observe $(s, a, s', r)$;
2. Fit $v_\phi^\pi(s)$ using target $r + \gamma v_\phi^\pi(s')$;
3. Estimate Advantage $A^\pi(s, a) = r + \gamma v_\phi^\pi(s') - v_\phi^\pi(s)$;
4. Estimate SG $\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(a|s) A^\pi(s, a)$;
5. Update $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$;

# Online Actor-Critic algorithm

Fully incremental online version:

1. Take action $a \sim \pi_\theta(a|s)$, and observe $(s, a, s', r)$;
2. Fit $v_\phi^\pi(s)$ using target $r + \gamma v_\phi^\pi(s')$;
3. Estimate Advantage $A^\pi(s, a) = r + \gamma v_\phi^\pi(s') - v_\phi^\pi(s)$;
4. Estimate SG $\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(a|s) A^\pi(s, a)$;
5. Update $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$;

Q) What's an issue in actor-critic?

# Online Actor-Critic algorithm

Fully incremental online version:

1. Take action $a \sim \pi_\theta(a|s)$, and observe $(s, a, s', r)$;
2. Fit $v_\phi^\pi(s)$ using target $r + \gamma v_\phi^\pi(s')$;
3. Estimate Advantage $A^\pi(s, a) = r + \gamma v_\phi^\pi(s') - v_\phi^\pi(s)$;
4. Estimate SG $\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(a|s) A^\pi(s, a)$;
5. Update $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$;

Q) What's an issue in actor-critic?

- On-policy: sample inefficient

# Off-Policy Actor-Critic

## Off-Policy Actor-Critic

**Thomas Degris** THOMAS.DEGRIS@INRIA.FR
Flowers Team, INRIA, Talence, ENSTA-ParisTech, Paris, France

**Martha White** WHITEM@CS.UALBERTA.CA
**Richard S. Sutton** SUTTON@CS.UALBERTA.CA
RLAI Laboratory, Department of Computing Science, University of Alberta, Edmonton, Canada

Key idea: Use behavior policy $\beta(a|s) \neq \pi_\theta(a|s)$

# Off-Policy Actor-Critic

## Off-Policy Actor-Critic

**Thomas Degris** THOMAS.DEGRIS@INRIA.FR
Flowers Team, INRIA, Talence, ENSTA-ParisTech, Paris, France

**Martha White** WHITEM@CS.UALBERTA.CA
**Richard S. Sutton** SUTTON@CS.UALBERTA.CA
RLAI Laboratory, Department of Computing Science, University of Alberta, Edmonton, Canada

Key idea: Use behavior policy $\beta(a|s) \neq \pi_\theta(a|s)$

- Changes in objective: $J(\theta) \to J_\beta(\theta)$, where

$$J_\beta(\theta) := \mathbb{E}_{\tau \sim p^\beta}\left[\sum_t r(s_t, a_t)\right] = \int p^\beta(\tau) r(\tau) d\tau$$

# Off-Policy Actor-Critic

**Off-Policy Actor-Critic**

**Thomas Degris** THOMAS.DEGRIS@INRIA.FR
Flowers Team, INRIA, Talence, ENSTA-ParisTech, Paris, France

**Martha White** WHITEM@CS.UALBERTA.CA
**Richard S. Sutton** SUTTON@CS.UALBERTA.CA
RLAI Laboratory, Department of Computing Science, University of Alberta, Edmonton, Canada

Key idea: Use behavior policy $\beta(a|s) \neq \pi_\theta(a|s)$

- Changes in objective: $J(\theta) \to J_\beta(\theta)$, where

$$J_\beta(\theta) := \mathbb{E}_{\tau \sim p^\beta}\left[\sum_t r(s_t, a_t)\right] = \int p^\beta(\tau) r(\tau) d\tau$$

- Approximate gradient:

$$\nabla_\theta J_\beta(\theta) \approx \mathbb{E}_{\tau \sim p^\beta}\left[\frac{\pi_\theta(a|s)}{\beta(a|s)} \nabla_\theta \log \pi_\theta(a|s) r(\tau)\right]$$

# Advantages and Disadvantages

Advantages:

- Lower variance (thanks to critic)
- (Often) Fast and stable convergence

# Advantages and Disadvantages

Advantages:

- Lower variance (thanks to critic)
- (Often) Fast and stable convergence

Disadvantages:

- Not unbiased (because the critic is not perfect)
- Training two networks required (for actor and critic)

# Will learn actor-critic deep RL methods

Published as a conference paper at ICLR 2016

## HIGH-DIMENSIONAL CONTINUOUS CONTROL USING GENERALIZED ADVANTAGE ESTIMATION

**John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan and Pieter Abbeel**
Department of Electrical Engineering and Computer Science
University of California, Berkeley
{joschu,pcmoritz,levine,jordan,pabbeel}@eecs.berkeley.edu