Actor Critic I

CISC 7404 - Decision Making

Steven Morad

University of Macau

Admin

Admin

How is homework 2?

Admin

How is homework 2?

Deadlines

Admin

How is homework 2?

Deadlines

- 1 Quiz next week

Admin

How is homework 2?

Deadlines

- 1 Quiz next week
- Final project proposal due day after quiz

Admin

How is homework 2?

Deadlines

- 1 Quiz next week
- Final project proposal due day after quiz
- Homework 2 due in 2 weeks

Admin

How is homework 2?

Deadlines

- 1 Quiz next week
- Final project proposal due day after quiz
- Homework 2 due in 2 weeks
- Last quiz in ~ 1 month
- Final project ~ 6 weeks

Admin

Quiz next week, topics:

Admin

Quiz next week, topics:

- Actor critic
  - Basic algorithm
  - Advantages
  - Off-policy gradient
  - Trust regions
  - Will not cover PPO

Admin

Quiz next week, topics:

- Actor critic
  - Basic algorithm
  - Advantages
  - Off-policy gradient
  - Trust regions
  - Will not cover PPO
- Policy gradient
  - Expectations, returns, notation
  - Optimization objective
  - REINFORCE

Admin

Quiz next week, topics:

- Actor critic
  - Basic algorithm
  - Advantages
  - Off-policy gradient
  - Trust regions
  - Will not cover PPO
- Policy gradient
  - Expectations, returns, notation
  - Optimization objective
  - REINFORCE
- Q learning
  - Expectations, returns, notation
  - Different objectives
  - Relationships between V and Q

Admin

Quiz next week, topics:

- Actor critic
  - Basic algorithm
  - Advantages
  - Off-policy gradient
  - Trust regions
  - Will not cover PPO
- Policy gradient
  - Expectations, returns, notation
  - Optimization objective
  - REINFORCE
- Q learning
  - Expectations, returns, notation
  - Different objectives
  - Relationships between V and Q

Format/difficulty will be similar to last time (3-4 questions, 75 mins)

Admin

Quiz next week, topics:

- Actor critic
  - Basic algorithm
  - Advantages
  - Off-policy gradient
  - Trust regions
  - Will not cover PPO
- Policy gradient
  - Expectations, returns, notation
  - Optimization objective
  - REINFORCE
- Q learning
  - Expectations, returns, notation
  - Different objectives
  - Relationships between V and Q

Format/difficulty will be similar to last time (3-4 questions, 75 mins)

Continue lecture after quiz next week? Will you be too tired?

Final Project

Final Project

Final project information is released

Final Project

Final project information is released

Suggest project and group members by next Friday (28th)

Final Project

Final project information is released

Suggest project and group members by next Friday (28th)

Find (or create) a gymnasium environment

Final Project

Final project information is released

Suggest project and group members by next Friday (28th)

Find (or create) a gymnasium environment

- Ensure your task is MDP

Final Project

Final project information is released

Suggest project and group members by next Friday (28th)

Find (or create) a gymnasium environment

- Ensure your task is MDP
- Can also try POMDP, but make sure you are prepared!

Final Project

Final project information is released

Suggest project and group members by next Friday (28th)

Find (or create) a gymnasium environment

- Ensure your task is MDP
- Can also try POMDP, but make sure you are prepared!
- Groups of 5, results should be impressive

Final Project

Final project information is released

Suggest project and group members by next Friday (28th)

Find (or create) a gymnasium environment

- Ensure your task is MDP
- Can also try POMDP, but make sure you are prepared!
- Groups of 5, results should be impressive
- Due just before final exam study week

Final Project

Final project information is released

Suggest project and group members by next Friday (28th)

Find (or create) a gymnasium environment
- Ensure your task is MDP
- Can also try POMDP, but make sure you are prepared!
- Groups of 5, results should be impressive
- Due just before final exam study week

https://ummoodle.um.edu.mo/pluginfile.php/6900679/mod_resource/content/6/project.pdf

Review

Actor Critic

Actor Critic

Today, we will investigate modern forms of policy gradient

Actor Critic

Today, we will investigate modern forms of policy gradient

These forms of policy gradient also learn Q or V functions jointly

Actor Critic

Today, we will investigate modern forms of policy gradient

These forms of policy gradient also learn Q or V functions jointly

We will learn the prerequisites to implement PPO, the most popular RL algorithm

Actor Critic

Today, we will investigate modern forms of policy gradient

These forms of policy gradient also learn Q or V functions jointly

We will learn the prerequisites to implement PPO, the most popular RL algorithm

Actor Critic

Recall the policy gradient

Actor Critic

Recall the policy gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Actor Critic

Recall the policy gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

We previously computed the Monte Carlo policy gradient (REINFORCE)

Actor Critic

Recall the policy gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

We previously computed the Monte Carlo policy gradient (REINFORCE)

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Actor Critic

Recall the policy gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

We previously computed the Monte Carlo policy gradient (REINFORCE)

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Question: Why don't we always use Monte Carlo?

Actor Critic

Recall the policy gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

We previously computed the Monte Carlo policy gradient (REINFORCE)

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Question: Why don't we always use Monte Carlo?

Answer: Requires collecting an infinite sequence of rewards!

Actor Critic

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Actor Critic

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Question: Other way to compute $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$?

Actor Critic

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Question: Other way to compute $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$?

Answer: Can use Q or V function with TD objective

Actor Critic

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Question: Other way to compute $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$?

Answer: Can use Q or V function with TD objective

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

Actor Critic

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Question: Other way to compute $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$?

Answer: Can use Q or V function with TD objective

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Actor Critic

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{t=0}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi]$$

Question: Other way to compute $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$?

Answer: Can use Q or V function with TD objective

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

I want to quickly repeat the relationship between V and Q

Actor Critic

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Actor Critic

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Q and V are closely related, we derived Q from V

Actor Critic

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Q and V are closely related, we derived Q from V

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

Actor Critic

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Q and V are closely related, we derived Q from V

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

This means we can convert Q to V or V to Q

Actor Critic

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Q and V are closely related, we derived Q from V

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

This means we can convert Q to V or V to Q

If you choose $a_0 \sim \pi(\cdot \mid s_0; \theta_\pi)$ for Q

Actor Critic

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Q and V are closely related, we derived Q from V

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

This means we can convert Q to V or V to Q

If you choose $a_0 \sim \pi(\cdot \mid s_0; \theta_\pi)$ for Q

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \cancel{a_0}, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

Actor Critic

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma Q(s_1, a_1, \theta_\pi)$$

Q and V are closely related, we derived Q from V

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0, \theta_\pi] + \gamma V(s_1, \theta_\pi)$$

This means we can convert Q to V or V to Q

If you choose $a_0 \sim \pi(\cdot \mid s_0; \theta_\pi)$ for Q

$$\begin{aligned} Q(s_0, a_0, \theta_\pi) &= \mathbb{E}[\mathcal{R}(s_1) \mid s_0, \cancel{a_0}, \theta_\pi] + \gamma V(s_1, \theta_\pi) \\ &= V(s_0, \theta_\pi) \end{aligned}$$

Actor Critic

Now that V and Q are clear, back to policy gradient

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Represent expected policy-conditioned return using value function

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Represent expected policy-conditioned return using value function

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Represent expected policy-conditioned return using value function

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Replace MC return with V / Q in policy gradient, call it **actor-critic**

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Represent expected policy-conditioned return using value function

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Replace MC return with V / Q in policy gradient, call it **actor-critic**

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Represent expected policy-conditioned return using value function

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Replace MC return with V / Q in policy gradient, call it **actor-critic**

Actor pick action

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \boxed{\pi(a_0 \mid s_0; \theta_\pi)}$$

Actor Critic

Now that V and Q are clear, back to policy gradient

Policy gradient objective uses the expected policy-conditioned return

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Represent expected policy-conditioned return using value function

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Replace MC return with V/Q in policy gradient, call it **actor-critic**

Actor pick action

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \boxed{V(s_0, \theta_\pi)} \cdot \nabla_{\theta_\pi} \log \boxed{\pi(a_0 \mid s_0; \theta_\pi)}$$

Critic gives actor score

Actor Critic

Definition: The actor-critic algorithm that jointly trains a policy network and value function

Actor Critic

Definition: The actor-critic algorithm that jointly trains a policy network and value function

$$\theta_{\pi, i+1} = \theta_{\pi, i} + \alpha \cdot \underbrace{V(s_0, \theta_{\pi, i}, \theta_{V, i})}_{\text{Expected return}} \cdot \nabla_{\theta_{\pi, i}} \log \pi(a_0 \mid s_0; \theta_{\pi, i})$$

Actor Critic

Definition: The actor-critic algorithm that jointly trains a policy network and value function

$$\theta_{\pi, i+1} = \theta_{\pi, i} + \alpha \cdot \underbrace{V(s_0, \theta_{\pi, i}, \theta_{V, i})}_{\text{Expected return}} \cdot \nabla_{\theta_{\pi, i}} \log \pi(a_0 \mid s_0; \theta_{\pi, i})$$

$$\theta_{V, i+1} =$$

$$\arg\min_{\theta_{V, i}} \underbrace{\left(V(s_0, \theta_{\pi, i}, \theta_{V, i}) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] + \neg d \gamma V(s_0, \theta_{\pi, i}, \theta_{V, i})\right)\right)^2}_{\text{TD error}}$$

Actor Critic

Definition: The actor-critic algorithm that jointly trains a policy network and value function

$$\theta_{\pi, i+1} = \theta_{\pi, i} + \alpha \cdot \underbrace{V(s_0, \theta_{\pi, i}, \theta_{V, i})}_{\text{Expected return}} \cdot \nabla_{\theta_{\pi, i}} \log \pi(a_0 \mid s_0; \theta_{\pi, i})$$

$$\theta_{V, i+1} =$$

$$\arg \min_{\theta_{V, i}} \underbrace{\left(V(s_0, \theta_{\pi, i}, \theta_{V, i}) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] + \neg d \gamma V(s_0, \theta_{\pi, i}, \theta_{V, i}) \right) \right)^2}_{\text{TD error}}$$

Repeat process until convergence: $\theta_{\pi, i+1} = \theta_{\pi, i}, \quad \theta_{V, i+1} = \theta_{V, i}$

Actor Critic

Definition: The actor-critic algorithm that jointly trains a policy network and value function

$$\theta_{\pi, i+1} = \theta_{\pi, i} + \alpha \cdot \underbrace{V(s_0, \theta_{\pi, i}, \theta_{V, i})}_{\text{Expected return}} \cdot \nabla_{\theta_{\pi, i}} \log \pi(a_0 \mid s_0; \theta_{\pi, i})$$

$$\theta_{V, i+1} =$$

$$\arg\min_{\theta_{V, i}} \underbrace{\left(V(s_0, \theta_{\pi, i}, \theta_{V, i}) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] + \neg d \gamma V(s_0, \theta_{\pi, i}, \theta_{V, i})\right)\right)^2}_{\text{TD error}}$$

Repeat process until convergence: $\theta_{\pi, i+1} = \theta_{\pi, i}, \quad \theta_{V, i+1} = \theta_{V, i}$

Can train policy with single transition s_0, a_0, s_1, r_0, d_0

Advantage Actor Critic

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Any scenarios where reward is always negative?

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Any scenarios where reward is always negative?

Answer: Distance to goal, $\mathcal{R}(s_{t+1}) = -(s_{t+1} - s_g)^2$

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Any scenarios where reward is always negative?

Answer: Distance to goal, $\mathcal{R}(s_{t+1}) = -(s_{t+1} - s_g)^2$

Question: What happens to return if reward is always negative?

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Any scenarios where reward is always negative?

Answer: Distance to goal, $\mathcal{R}(s_{t+1}) = -(s_{t+1} - s_g)^2$

Question: What happens to return if reward is always negative?

Answer: Return always negative

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Any scenarios where reward is always negative?

Answer: Distance to goal, $\mathcal{R}(s_{t+1}) = -(s_{t+1} - s_g)^2$

Question: What happens to return if reward is always negative?

Answer: Return always negative

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Any scenarios where reward is always negative?

Answer: Distance to goal, $\mathcal{R}(s_{t+1}) = -(s_{t+1} - s_g)^2$

Question: What happens to return if reward is always negative?

Answer: Return always negative

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Similar results if reward is always positive

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Any scenarios where reward is always negative?

Answer: Distance to goal, $\mathcal{R}(s_{t+1}) = -(s_{t+1} - s_g)^2$

Question: What happens to return if reward is always negative?

Answer: Return always negative

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Similar results if reward is always positive

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Example: MDP with one state and continuous actions, negative reward

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Example: MDP with one state and continuous actions, negative reward

Sample k transitions $(s_0, a_0, r_0, d_0, s_1)$ for each policy update

Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Example: MDP with one state and continuous actions, negative reward

Sample k transitions $(s_0, a_0, r_0, d_0, s_1)$ for each policy update

What if we cannot sample all possible actions?

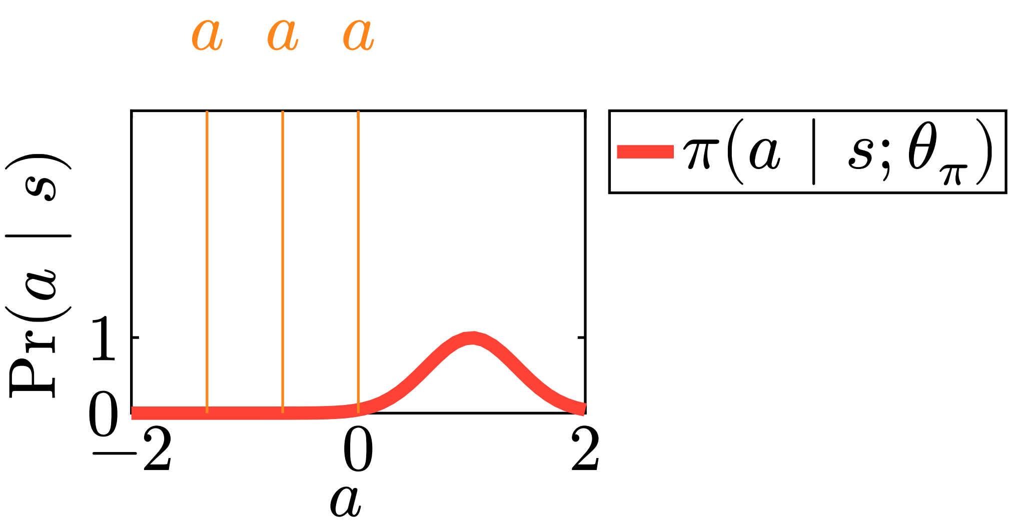
Advantage Actor Critic

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Example: MDP with one state and continuous actions, negative reward

Sample k transitions $(s_0, a_0, r_0, d_0, s_1)$ for each policy update

What if we cannot sample all possible actions?



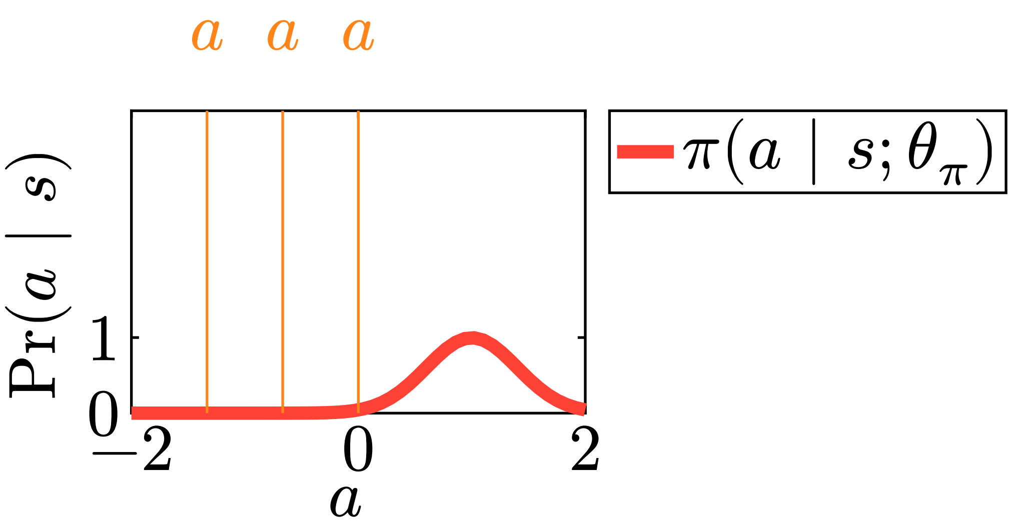
Advantage Actor Critic

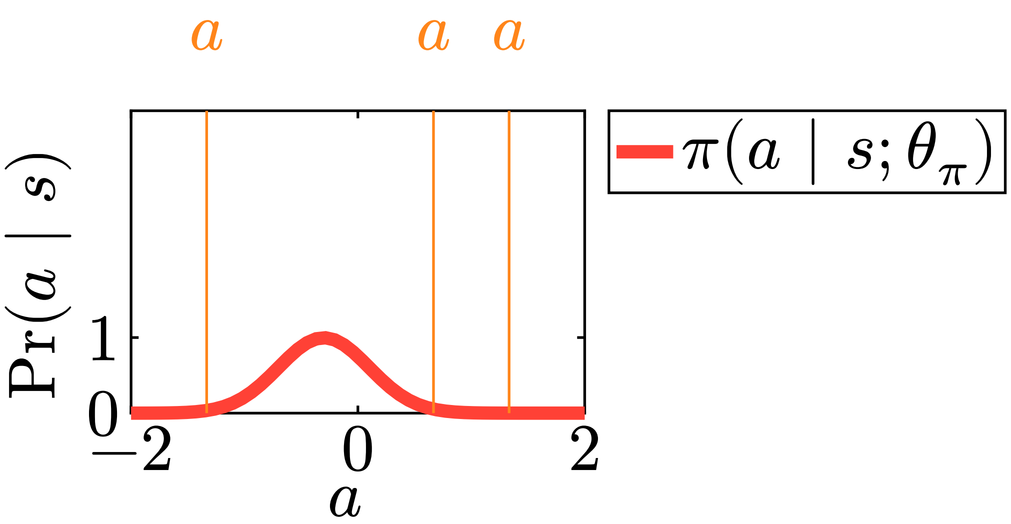
$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = - \mid V(s_0, \theta_\pi) \mid \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Example: MDP with one state and continuous actions, negative reward

Sample k transitions $(s_0, a_0, r_0, d_0, s_1)$ for each policy update

What if we cannot sample all possible actions?

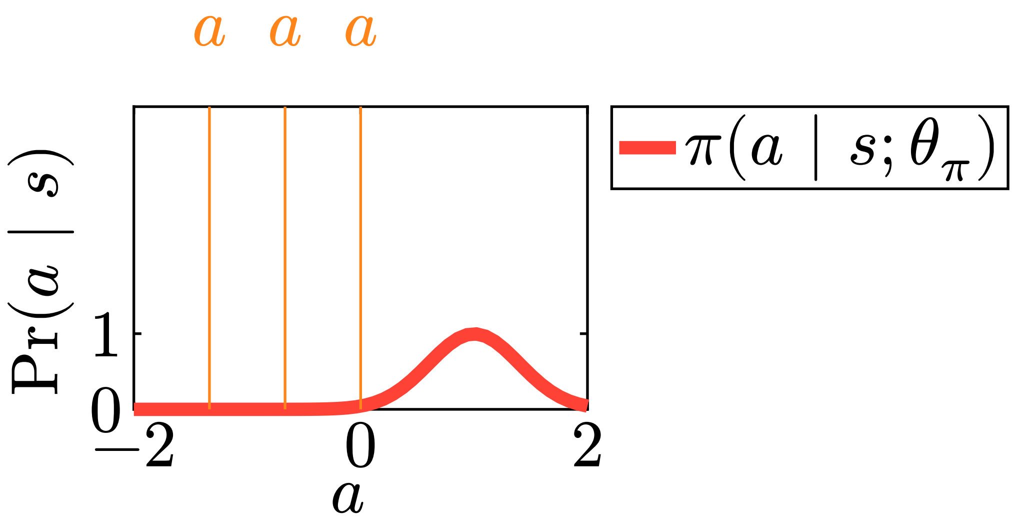


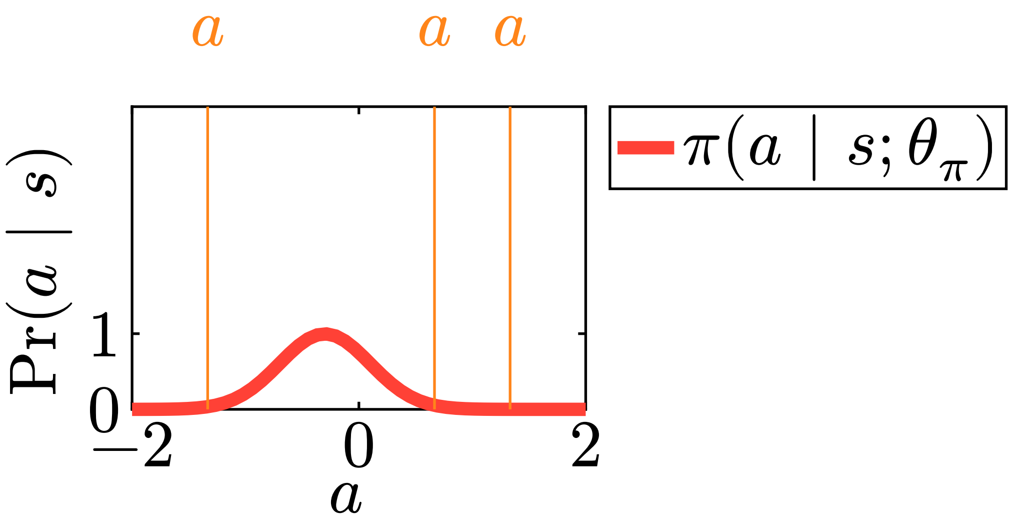


Advantage Actor Critic

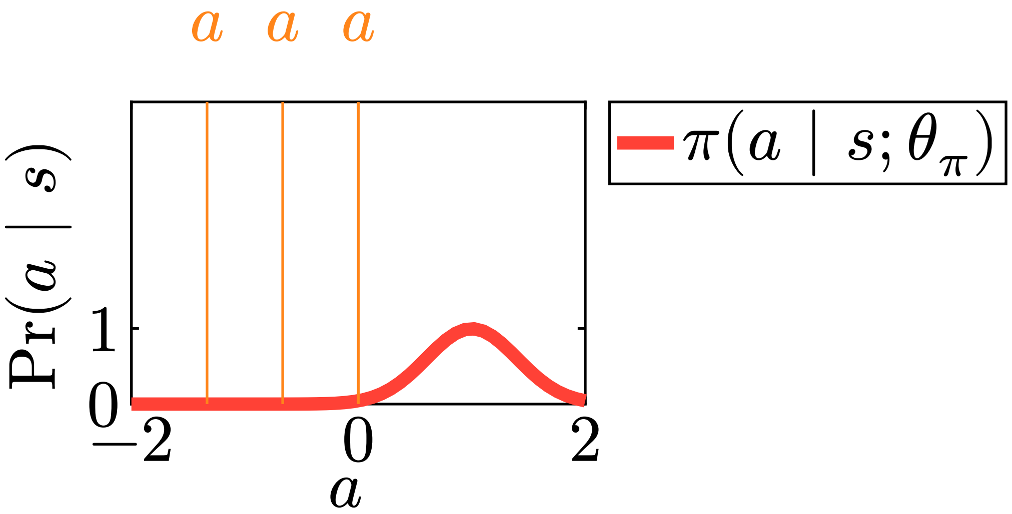
https://media0.giphy.com/media/v1.Y2lkPTc5MGI3NjExeGdqZm56NDgzcmY2Ym95dG13Ynczdm9lbDY0cGpjczdtMHBmcnJmMSZlcD12MV9pbnRlcm5hbF9naWZfYnlfaWQmY3Q9Zw/MVUyVpyjakkRW/giphy.gif

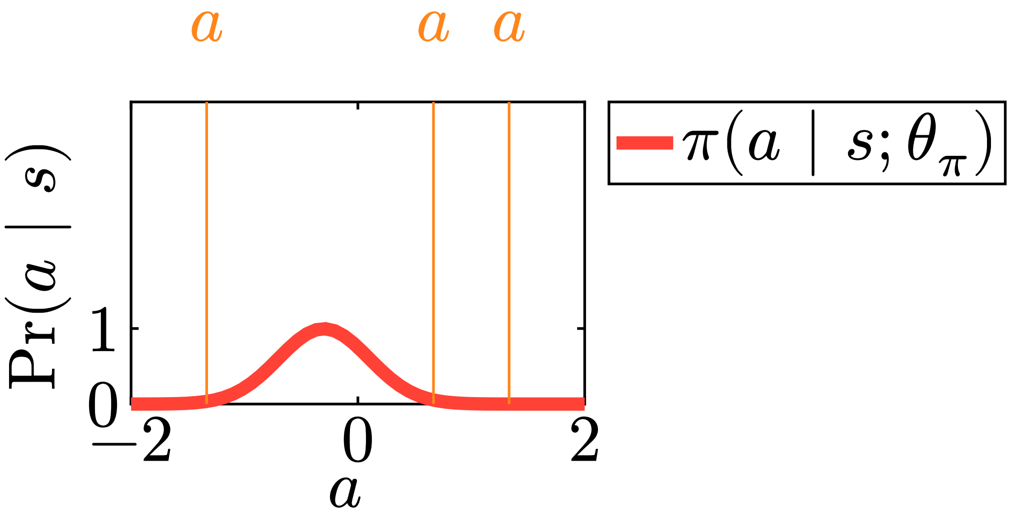
Advantage Actor Critic





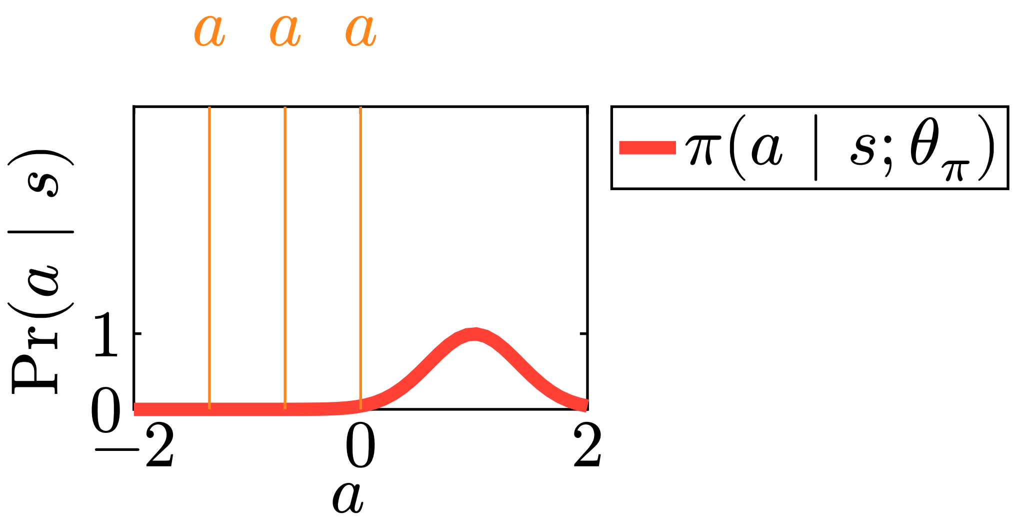
Advantage Actor Critic

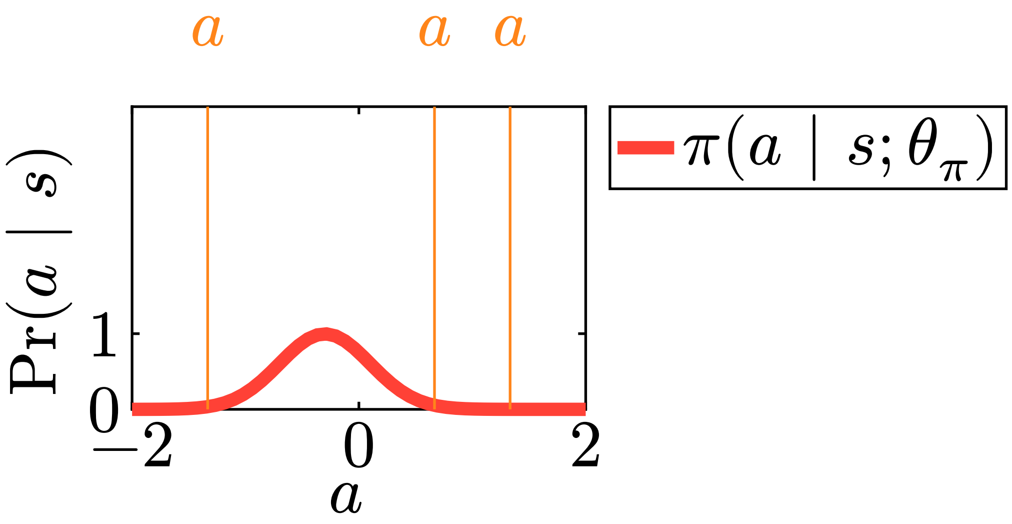




Policy keeps oscillating, can destabilize learning

Advantage Actor Critic

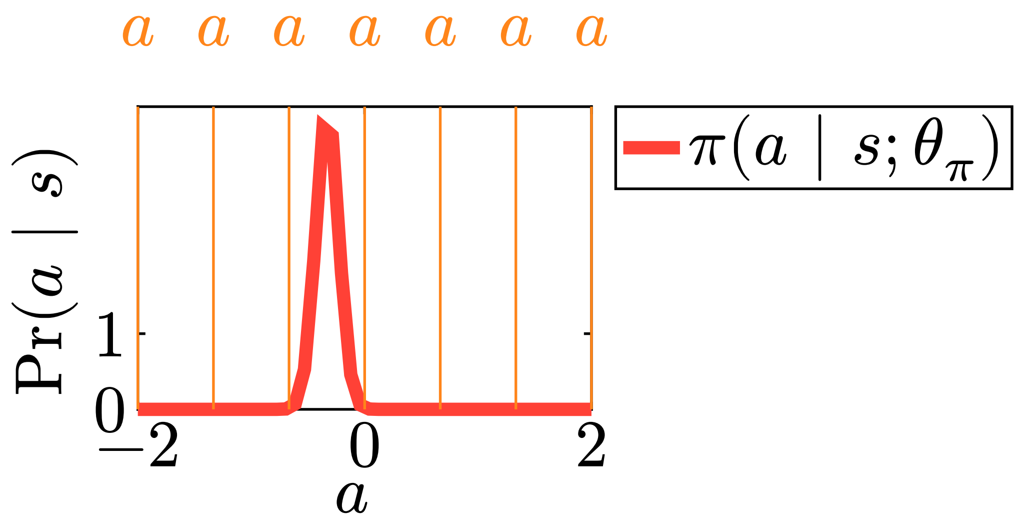




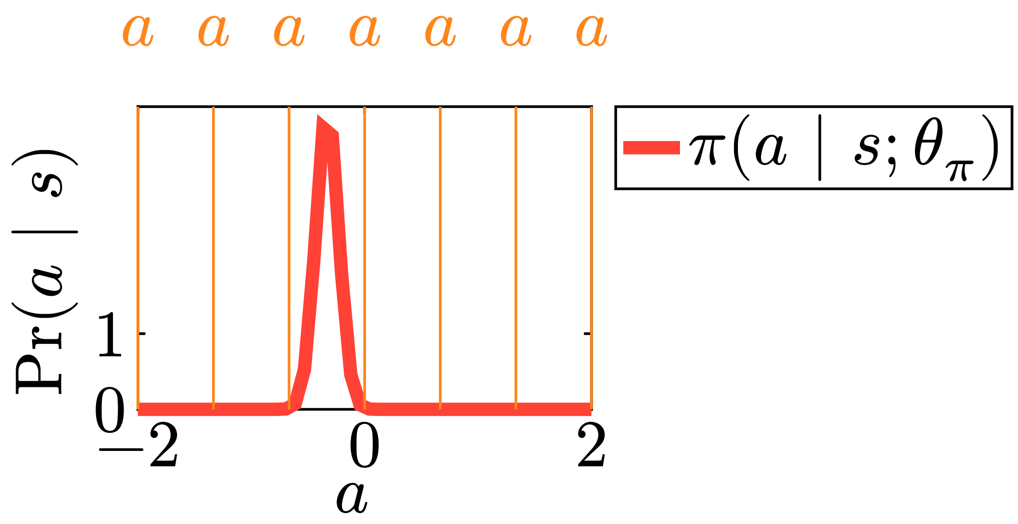
Policy keeps oscillating, can destabilize learning

Question: If we take 8 actions, will this fix it?

Advantage Actor Critic

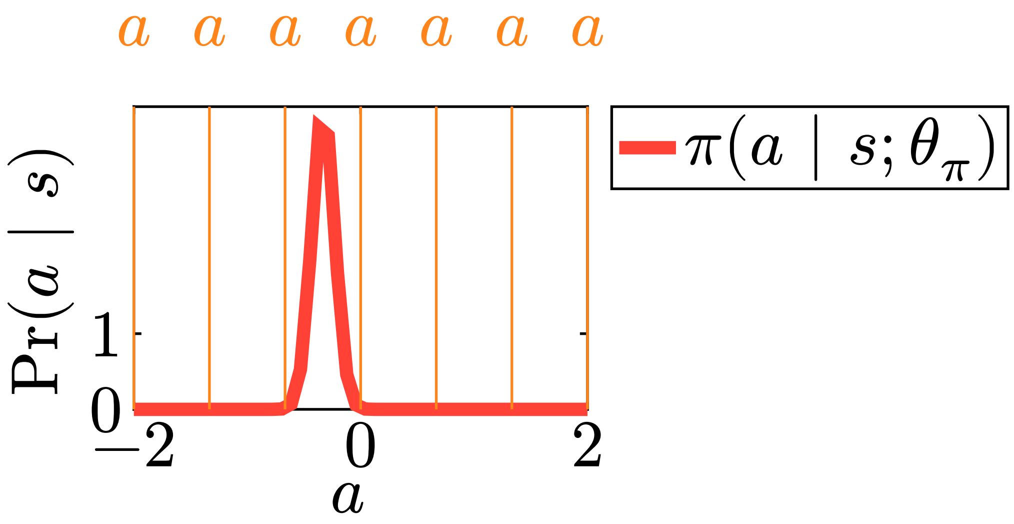


Advantage Actor Critic



Question: Any solutions?

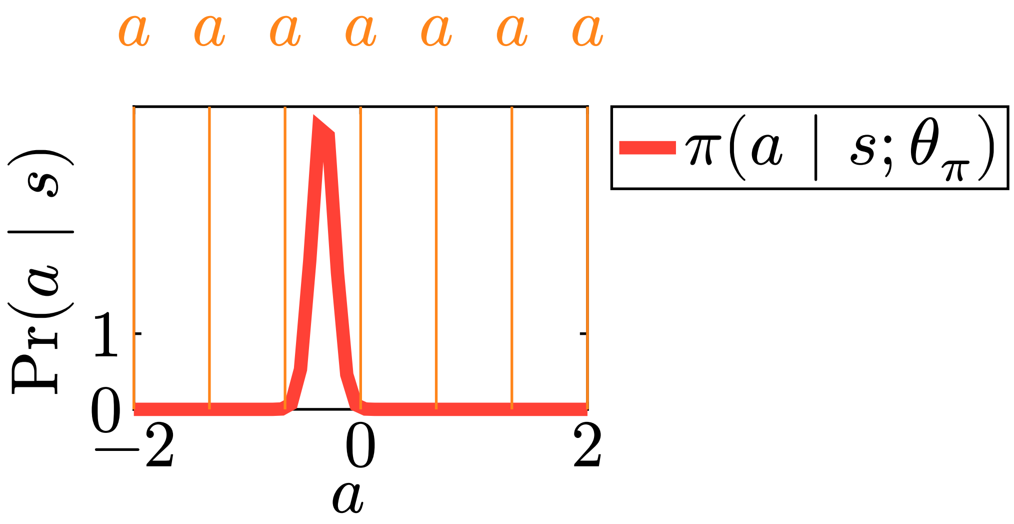
Advantage Actor Critic



Question: Any solutions?

Hint: Think about the mean of the return

Advantage Actor Critic

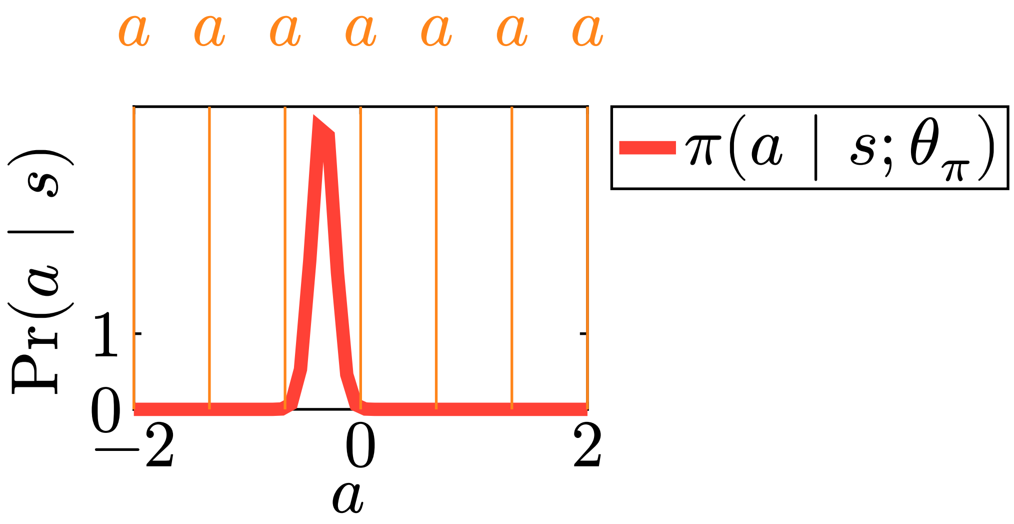


Question: Any solutions?

Hint: Think about the mean of the return

Answer: Recenter return such that mean is zero

Advantage Actor Critic



What if we:

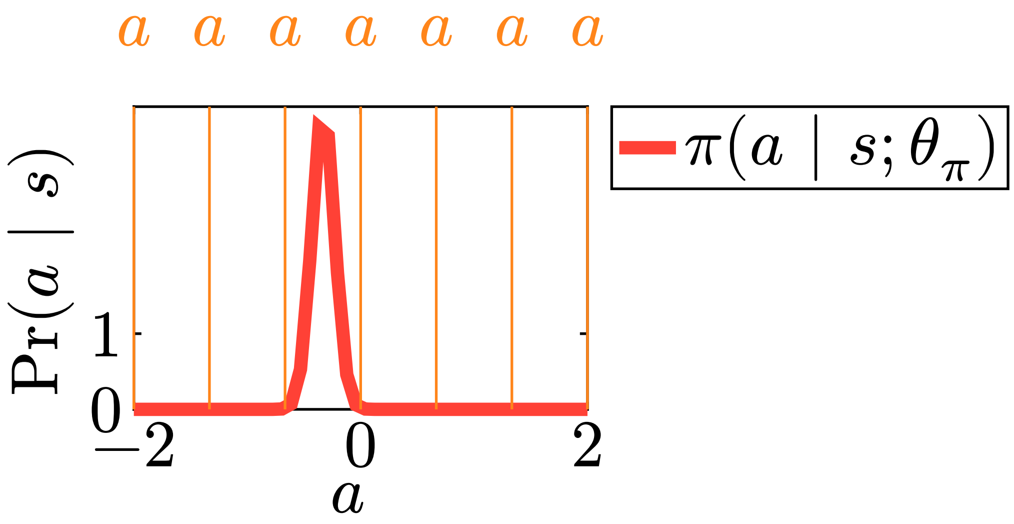
Question: Any solutions?

Hint: Think about the mean of the return

Answer: Recenter return such that mean is zero

Does not completely solve issue, maybe $\mathcal{R}(s_A) < 0$

Advantage Actor Critic



Question: Any solutions?

Hint: Think about the mean of the return

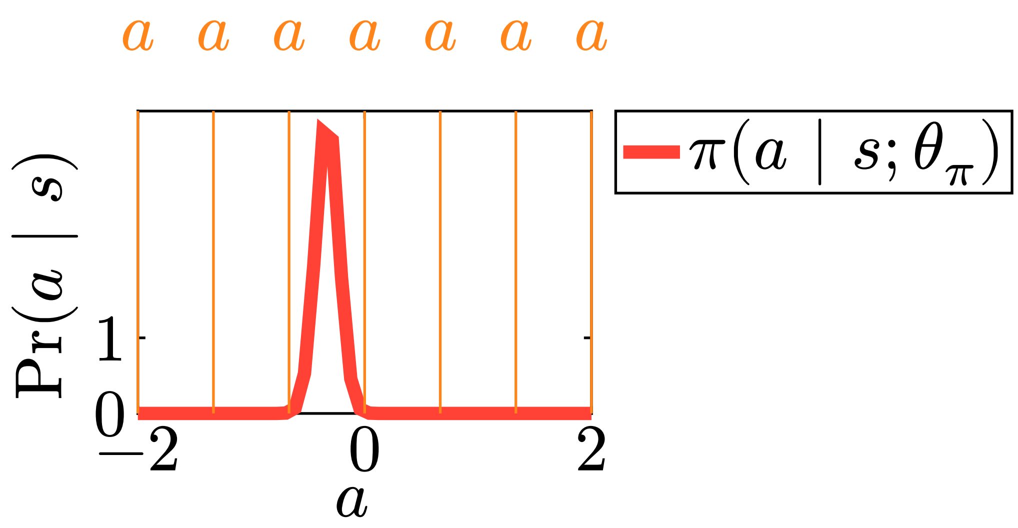
Answer: Recenter return such that mean is zero

Does not completely solve issue, maybe $\mathcal{R}(s_A) < 0$

What if we:

- Almost never update policy

Advantage Actor Critic



Question: Any solutions?

Hint: Think about the mean of the return

Answer: Recenter return such that mean is zero

Does not completely solve issue, maybe $\mathcal{R}(s_A) < 0$

What if we:

- Almost never update policy
- Update the policy **only** if action is better/worse than expected

Advantage Actor Critic

Question: What is the expected performance of the policy?

Advantage Actor Critic

Question: What is the expected performance of the policy?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Advantage Actor Critic

Question: What is the expected performance of the policy?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Question: What is the expected performance of a specific action?

Advantage Actor Critic

Question: What is the expected performance of the policy?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Question: What is the expected performance of a specific action?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0; \theta_\pi] = Q(s_0, a_0, \theta_\pi)$$

Advantage Actor Critic

Question: What is the expected performance of the policy?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Question: What is the expected performance of a specific action?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0; \theta_\pi] = Q(s_0, a_0, \theta_\pi)$$

Question: How can we tell if an action is better/worse than expected?

Advantage Actor Critic

Question: What is the expected performance of the policy?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Question: What is the expected performance of a specific action?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0; \theta_\pi] = Q(s_0, a_0, \theta_\pi)$$

Question: How can we tell if an action is better/worse than expected?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0; \theta_\pi] - \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$$

Advantage Actor Critic

Question: What is the expected performance of the policy?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = V(s_0, \theta_\pi)$$

Question: What is the expected performance of a specific action?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0; \theta_\pi] = Q(s_0, a_0, \theta_\pi)$$

Question: How can we tell if an action is better/worse than expected?

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0; \theta_\pi] - \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

We call this the **advantage**, tells us if we should change policy

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

We call this the **advantage**, tells us if we should change policy

If a_0 produces better than expected return, increase policy probability

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

We call this the **advantage**, tells us if we should change policy

If a_0 produces better than expected return, increase policy probability

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \mid A(s_0, a_0, \theta_{\pi,i}) \mid \cdot \nabla_{\theta_{\pi,i}} \log \pi(a_0 \mid s_0; \theta_{\pi,i})$$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

We call this the **advantage**, tells us if we should change policy

If a_0 produces better than expected return, increase policy probability

$$\theta_{\pi,i+1} = \theta_{\pi,i} + | \ A(s_0, a_0, \theta_{\pi,i}) \ | \cdot \nabla_{\theta_{\pi,i}} \log \pi(a_0 \mid s_0; \theta_{\pi,i})$$

If action a_0 produces worse return than expected, reduce probability

$$\theta_{\pi,i+1} = \theta_{\pi,i} - | \ A(s_0, a_0, \theta_{\pi,i}) \ | \cdot \nabla_{\theta_{\pi,i}} \log \pi(a_0 \mid s_0; \theta_{\pi,i})$$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

We call this the **advantage**, tells us if we should change policy

If a_0 produces better than expected return, increase policy probability

$$\theta_{\pi,i+1} = \theta_{\pi,i} + | \ A\left(s_0, a_0, \theta_{\pi,i}\right) \ | \cdot \nabla_{\theta_{\pi,i}} \log \pi\left(a_0 \mid s_0; \theta_{\pi,i}\right)$$

If action a_0 produces worse return than expected, reduce probability

$$\theta_{\pi,i+1} = \theta_{\pi,i} - | \ A\left(s_0, a_0, \theta_{\pi,i}\right) \ | \cdot \nabla_{\theta_{\pi,i}} \log \pi\left(a_0 \mid s_0; \theta_{\pi,i}\right)$$

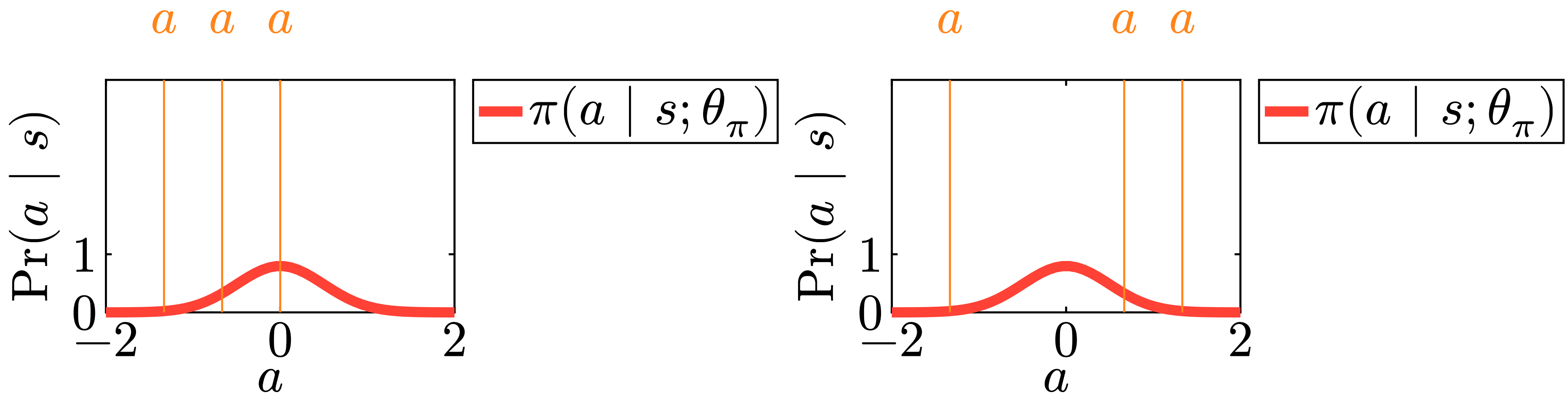
If action a_0 produced expected return, do nothing $\theta_{\pi,i+1} = \theta_{\pi,i} + 0$

Advantage Actor Critic

The policy will not oscillate – policy only changes if it improves return

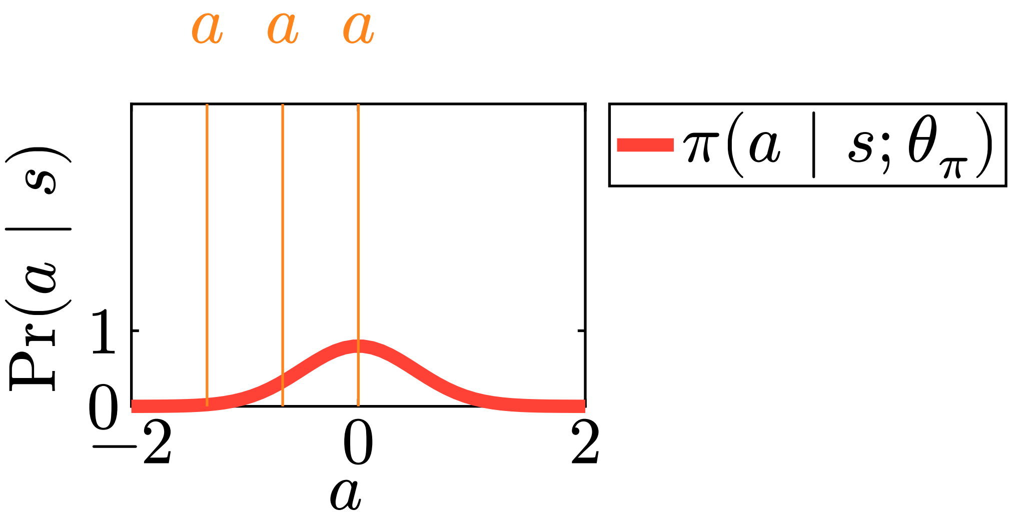
Advantage Actor Critic

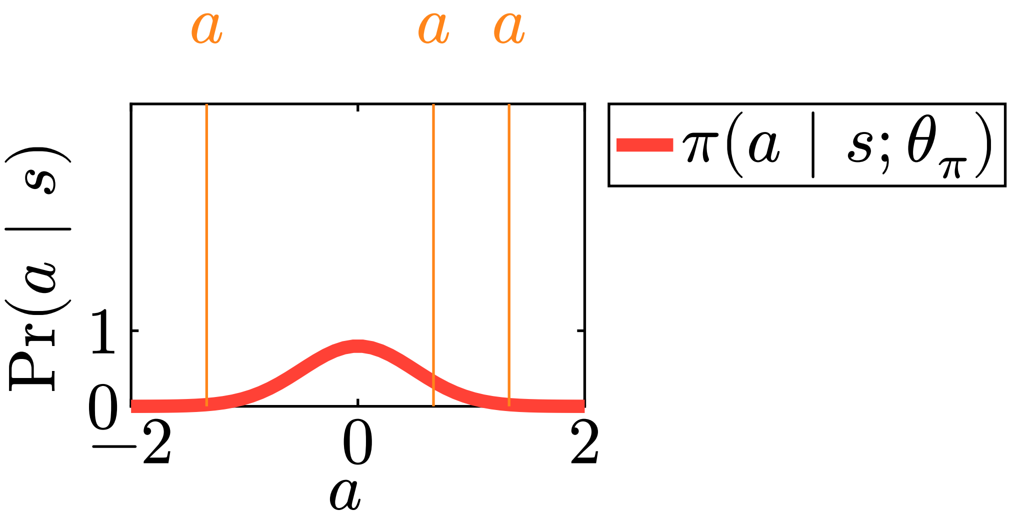
The policy will not oscillate – policy only changes if it improves return



Advantage Actor Critic

The policy will not oscillate – policy only changes if it improves return





Results in more stable training and faster convergence

Advantage Actor Critic

Definition: The advantage A determines the relative advantage/ disadvantage of taking an action a_0 in state s_0 for a policy θ_π

Advantage Actor Critic

Definition: The advantage A determines the relative advantage/ disadvantage of taking an action a_0 in state s_0 for a policy θ_π

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage requires both Q and V

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage requires both Q and V

But earlier, we saw $Q = V$ in some circumstances

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage requires both Q and V

But earlier, we saw $Q = V$ in some circumstances

Question: Can we replace Q with V? How?

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage requires both Q and V

But earlier, we saw $Q = V$ in some circumstances

Question: Can we replace Q with V? How?

HINT: Think about TD error, will write as $A(s_0, s_1, r_0, \theta_\pi)$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage requires both Q and V

But earlier, we saw $Q = V$ in some circumstances

Question: Can we replace Q with V? How?

HINT: Think about TD error, will write as $A(s_0, s_1, r_0, \theta_\pi)$

$$A(s_0, s_1, r_0, \theta_\pi) = - \underbrace{V(s_0, \theta_\pi)}_{\text{What we expect}} + \underbrace{(\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] + \neg d \gamma V(s_1, \theta_\pi))}_{\text{What happens}}$$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage requires both Q and V

But earlier, we saw $Q = V$ in some circumstances

Question: Can we replace Q with V? How?

HINT: Think about TD error, will write as $A(s_0, s_1, r_0, \theta_\pi)$

$$A(s_0, s_1, r_0, \theta_\pi) = - \underbrace{V(s_0, \theta_\pi)}_{\text{What we expect}} + \underbrace{(\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] + \neg d \gamma V(s_1, \theta_\pi))}_{\text{What happens}}$$

Better than expected: $|A| > 0$

Advantage Actor Critic

$$A(s_0, a_0, \theta_\pi) = Q(s_0, a_0, \theta_\pi) - V(s_0, \theta_\pi)$$

Advantage requires both Q and V

But earlier, we saw $Q = V$ in some circumstances

Question: Can we replace Q with V? How?

HINT: Think about TD error, will write as $A(s_0, s_1, r_0, \theta_\pi)$

$$A(s_0, s_1, r_0, \theta_\pi) = -\underbrace{V(s_0, \theta_\pi)}_{\text{What we expect}} + \underbrace{(\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] + \neg d \gamma V(s_1, \theta_\pi))}_{\text{What happens}}$$

Better than expected: $|A| > 0$, worse than expected $|A| < 0$

Advantage Actor Critic

Definition: Advantage actor critic (A2C) updates the policy using the advantage, and repeats until convergence

Advantage Actor Critic

Definition: Advantage actor critic (A2C) updates the policy using the advantage, and repeats until convergence

$$A(s_0, s_1, r_0, \theta_\pi, \theta_V) = -V(s_0, \theta_\pi, \theta_V) + \underbrace{\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\pi]}_{r_0} + \neg d \gamma V(s_1, \theta_\pi, \theta_V)$$

Advantage Actor Critic

Definition: Advantage actor critic (A2C) updates the policy using the advantage, and repeats until convergence

$$A(s_0, s_1, r_0, \theta_\pi, \theta_V) = -V(s_0, \theta_\pi, \theta_V) + \underbrace{\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\pi]}_{r_0} + \neg d \gamma V(s_1, \theta_\pi, \theta_V)$$

$$\theta_{\pi, i+1} = \theta_{\pi, i} + \alpha \cdot \underbrace{A(s_0, \theta_{\pi, i}, \theta_{V, i})}_{\text{Advantage}} \cdot \underbrace{\nabla_{\theta_{\pi, i}} \log \pi(a_0 \mid s_0; \theta_{\pi, i})}_{\text{Policy gradient}}$$

Advantage Actor Critic

Definition: Advantage actor critic (A2C) updates the policy using the advantage, and repeats until convergence

$$A(s_0, s_1, r_0, \theta_\pi, \theta_V) = -V(s_0, \theta_\pi, \theta_V) + \underbrace{\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\pi]}_{r_0} + \neg d \gamma V(s_1, \theta_\pi, \theta_V)$$

$$\theta_{\pi, i+1} = \theta_{\pi, i} + \alpha \cdot \underbrace{A(s_0, \theta_{\pi, i}, \theta_{V, i})}_{\text{Advantage}} \cdot \underbrace{\nabla_{\theta_{\pi, i}} \log \pi(a_0 \mid s_0; \theta_{\pi, i})}_{\text{Policy gradient}}$$

$$\theta_{V, i+1} =$$

$$\arg\min_{\theta_{V, i}} \underbrace{\left(V(s_0, \theta_{\pi, i}, \theta_{V, i}) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] + \neg d \gamma V(s_0, \theta_{\pi, i}, \theta_{V, i})\right)\right)^2}_{\text{TD error}}$$

Off-Policy Gradient

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Answer: On-policy, expected return depends on θ_π

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Answer: On-policy, expected return depends on θ_π

Question: Why do we care about being off-policy?

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Answer: On-policy, expected return depends on θ_π

Question: Why do we care about being off-policy?

Answer: Algorithm can reuse data, much more efficient

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Answer: On-policy, expected return depends on θ_π

Question: Why do we care about being off-policy?

Answer: Algorithm can reuse data, much more efficient

Question: What do we need to make policy gradient off-policy?

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Answer: On-policy, expected return depends on θ_π

Question: Why do we care about being off-policy?

Answer: Algorithm can reuse data, much more efficient

Question: What do we need to make policy gradient off-policy?

Must approximate $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$ using $\mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\beta\right]$

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Answer: On-policy, expected return depends on θ_π

Question: Why do we care about being off-policy?

Answer: Algorithm can reuse data, much more efficient

Question: What do we need to make policy gradient off-policy?

Must approximate $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \underbrace{\theta_\pi}_{\text{Training policy}}]$ using $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \underbrace{\theta_\beta}_{\text{Behavior policy}}]$

Off-Policy Gradient

$$\nabla_{\theta_\pi} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_\pi)$$

Question: Is policy gradient off-policy or on-policy?

Answer: On-policy, expected return depends on θ_π

Question: Why do we care about being off-policy?

Answer: Algorithm can reuse data, much more efficient

Question: What do we need to make policy gradient off-policy?

Must approximate $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \underbrace{\theta_\pi}_{\text{Training policy}}]$ using $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \underbrace{\theta_\beta}_{\text{Behavior policy}}]$

Question: Any statistics students know how to do this?

Off-Policy Gradient

In **importance sampling**, we want to estimate

Off-Policy Gradient

In **importance sampling**, we want to estimate

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)]$$

Off-Policy Gradient

In **importance sampling**, we want to estimate

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)]$$

Unfortunately, we only have data from

Off-Policy Gradient

In **importance sampling**, we want to estimate

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)]$$

Unfortunately, we only have data from

$$\mathbb{E}[f(x) \mid x \sim \Pr(Y)]$$

Off-Policy Gradient

In **importance sampling**, we want to estimate

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)]$$

Unfortunately, we only have data from

$$\mathbb{E}[f(x) \mid x \sim \Pr(Y)]$$

We can use their ratio to approximate the expectation

Off-Policy Gradient

In **importance sampling**, we want to estimate

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)]$$

Unfortunately, we only have data from

$$\mathbb{E}[f(x) \mid x \sim \Pr(Y)]$$

We can use their ratio to approximate the expectation

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y)\right]$$

Off-Policy Gradient

In **importance sampling**, we want to estimate

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)]$$

Unfortunately, we only have data from

$$\mathbb{E}[f(x) \mid x \sim \Pr(Y)]$$

We can use their ratio to approximate the expectation

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y) \right]$$

Question: How can we use this to make policy gradient off-policy?

Off-Policy Gradient

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \,\middle|\, x \sim \Pr(Y)\right]$$

Off-Policy Gradient

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y)\right]$$

Consider our current policy is θ_π, we want $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

Off-Policy Gradient

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y)\right]$$

Consider our current policy is θ_π, we want $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

A **behavior policy** θ_β collected the data $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\beta]$

Off-Policy Gradient

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y)\right]$$

Consider our current policy is θ_π, we want $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

A **behavior policy** θ_β collected the data $\mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\beta\right]$

θ_β can be an old policy or some other policy

Off-Policy Gradient

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y)\right]$$

Consider our current policy is θ_π, we want $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

A **behavior policy** θ_β collected the data $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\beta]$

θ_β can be an old policy or some other policy

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y)\right]$$

Consider our current policy is θ_π, we want $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

A **behavior policy** θ_β collected the data $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\beta]$

θ_β can be an old policy or some other policy

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Reward following θ_β

Off-Policy Gradient

$$\mathbb{E}[f(x) \mid x \sim \Pr(X)] = \mathbb{E}\left[f(x) \cdot \frac{\Pr(X)}{\Pr(Y)} \middle| x \sim \Pr(Y)\right]$$

Consider our current policy is θ_π, we want $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

A **behavior policy** θ_β collected the data $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\beta]$

θ_β can be an old policy or some other policy

Reward following θ_π

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Reward following θ_β

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Seems like magic, how does this actually work?

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Seems like magic, how does this actually work?

Let us find out, start with expected reward from behavior policy

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \,\middle|\, s_0; \theta_\beta\right]$$

Seems like magic, how does this actually work?

Let us find out, start with expected reward from behavior policy

$$\mathbb{E}\left[\mathcal{R}(s_1) \mid s_0; \theta_\beta\right] = \sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi\left(a_0 \mid s_0; \theta_\beta\right)$$

Off-Policy Gradient

$$\mathbb{E}\left[\mathcal{R}(s_1) \mid s_0; \theta_\beta\right] = \underbrace{\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi\left(a_0 \mid s_0; \theta_\beta\right)}_{\text{Expected reward}} \underbrace{\frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi\left(a_0 \mid s_0; \theta_\beta\right)}}_{\text{Correction}}$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\beta] = \underbrace{\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta)}_{\text{Expected reward}} \underbrace{\frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}}_{\text{Correction}}$$

$$\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cancel{\pi(a_0 \mid s_0; \theta_\beta)} \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\cancel{\pi(a_0 \mid s_0; \theta_\beta)}}$$

$$\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\pi)$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\beta] = \underbrace{\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta)}_{\text{Expected reward}} \underbrace{\frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}}_{\text{Correction}}$$

$$\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cancel{\pi(a_0 \mid s_0; \theta_\beta)} \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\cancel{\pi(a_0 \mid s_0; \theta_\beta)}}$$

$$\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi]$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\beta] = \underbrace{\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta)}_{\text{Expected reward}} \underbrace{\frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}}_{\text{Correction}}$$

$$\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cancel{\pi(a_0 \mid s_0; \theta_\beta)} \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\cancel{\pi(a_0 \mid s_0; \theta_\beta)}}$$

$$\sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi]$$

Left with expected reward following θ_π

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \;\middle|\; s_0; \theta_\beta\right]$$

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta) \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \;\middle|\; s_0; \theta_\beta\right]$$

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta) \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}$$

We found a way to estimate the off-policy reward

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta) \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}$$

We found a way to estimate the off-policy reward

Apply the same approach to find the off-policy return (won't derive, trust me)

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta) \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}$$

We found a way to estimate the off-policy reward

Apply the same approach to find the off-policy return (won't derive, trust me)

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta) \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}$$

We found a way to estimate the off-policy reward

Apply the same approach to find the off-policy return (won't derive, trust me)

Return following θ_β

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{R}(s_1) \cdot \frac{\pi(a \mid s_0; \theta_\pi)}{\pi(a \mid s_0; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

$$\mathbb{E}[\mathcal{R}(s_1) \mid s_0; \theta_\pi] = \sum_{s_1 \in S} \mathcal{R}(s_1) \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \pi(a_0 \mid s_0; \theta_\beta) \frac{\pi(a_0 \mid s_0; \theta_\pi)}{\pi(a_0 \mid s_0; \theta_\beta)}$$

We found a way to estimate the off-policy reward

Apply the same approach to find the off-policy return (won't derive, trust me)

Return following θ_β

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

Definition: Off-policy gradient uses importance sampling to learn from off-policy data

Off-Policy Gradient

Definition: Off-policy gradient uses importance sampling to learn from off-policy data

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

Definition: Off-policy gradient uses importance sampling to learn from off-policy data

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_{\pi,i}} \log \pi(a_0 \mid s_0; \theta_{\pi,i})$$

Off-Policy Gradient

Definition: Off-policy gradient uses importance sampling to learn from off-policy data

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_{\pi,i}} \log \pi\left(a_0 \mid s_0; \theta_{\pi,i}\right)$$

Note: Wrote MC version for clarity, but you can use V too

Off-Policy Gradient

Definition: Off-policy gradient uses importance sampling to learn from off-policy data

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

$$\theta_{\pi, i+1} = \theta_{\pi, i} + \alpha \cdot \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] \cdot \nabla_{\theta_{\pi, i}} \log \pi(a_0 \mid s_0; \theta_{\pi, i})$$

Note: Wrote MC version for clarity, but you can use V too

$$V(s_0, \theta_\pi, \theta_\beta) = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta \right]$$

Why did I tell you policy gradient is on policy?

Off-Policy Gradient

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Why did I tell you policy gradient is on policy?

Off-policy gradient does not work in most cases

Off-Policy Gradient

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Why did I tell you policy gradient is on policy?

Off-policy gradient does not work in most cases

Question: Why?

Off-Policy Gradient

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Why did I tell you policy gradient is on policy?

Off-policy gradient does not work in most cases

Question: Why? HINT: What happens to $\prod$?

Off-Policy Gradient

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \middle| s_0; \theta_\beta\right]$$

Why did I tell you policy gradient is on policy?

Off-policy gradient does not work in most cases

Question: Why? HINT: What happens to $\prod$?

$$\prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi(a_t \mid s_t; \theta_\beta)} \to 0, \infty$$

Off-Policy Gradient

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \mathbb{E}\left[\mathcal{G}(\boldsymbol{\tau}) \prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi\left(a_t \mid s_t; \theta_\beta\right)} \middle| s_0; \theta_\beta\right]$$

Why did I tell you policy gradient is on policy?

Off-policy gradient does not work in most cases

Question: Why? HINT: What happens to $\prod$?

$$\prod_{t=0}^{\infty} \frac{\pi(a_t \mid s_t; \theta_\pi)}{\pi\left(a_t \mid s_t; \theta_\beta\right)} \rightarrow 0, \infty$$

Only works if $\pi(a_t \mid s_t; \theta_\pi) \approx \pi\left(a_t \mid s_t; \theta_\beta\right) \quad \forall t$

Trust Regions

Trust Regions

Training policies in RL is difficult

Trust Regions

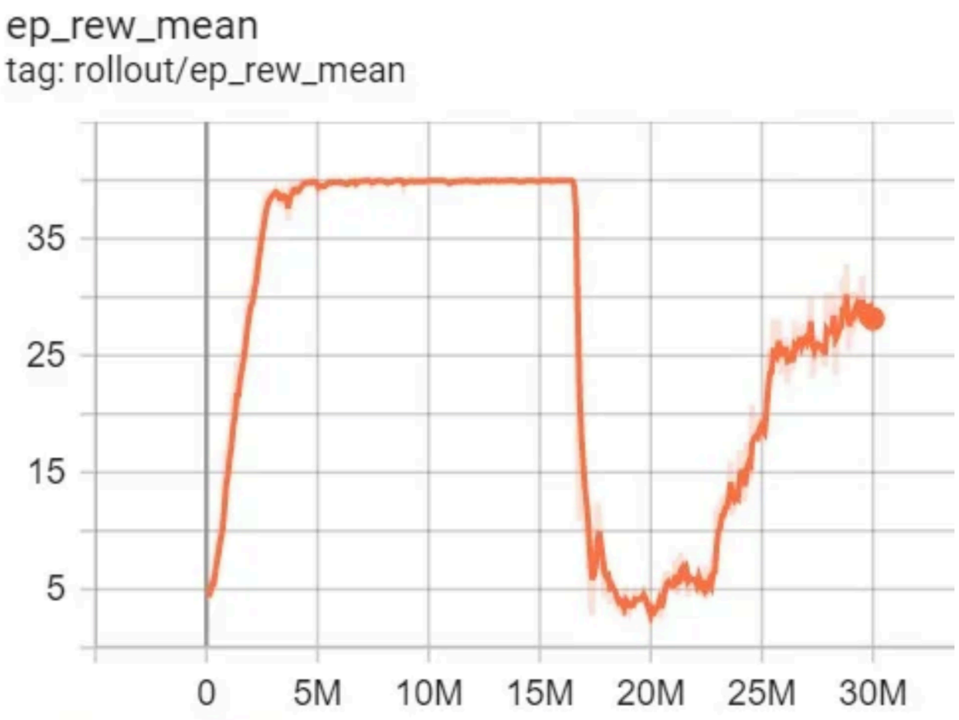
Training policies in RL is difficult

We often see strange results during training

Trust Regions

Training policies in RL is difficult

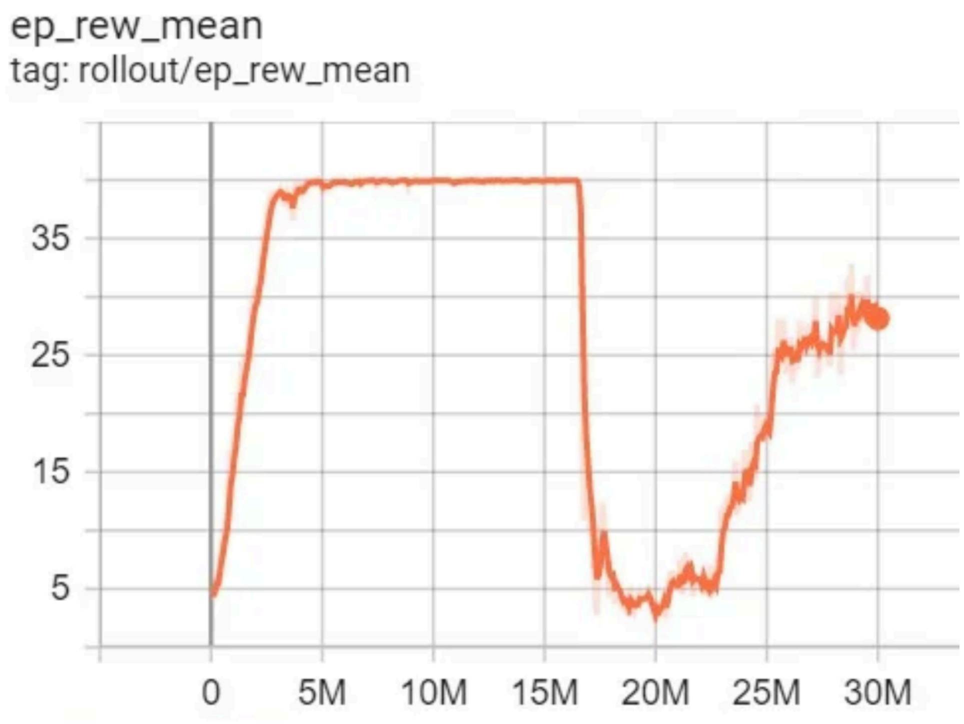
We often see strange results during training



Trust Regions

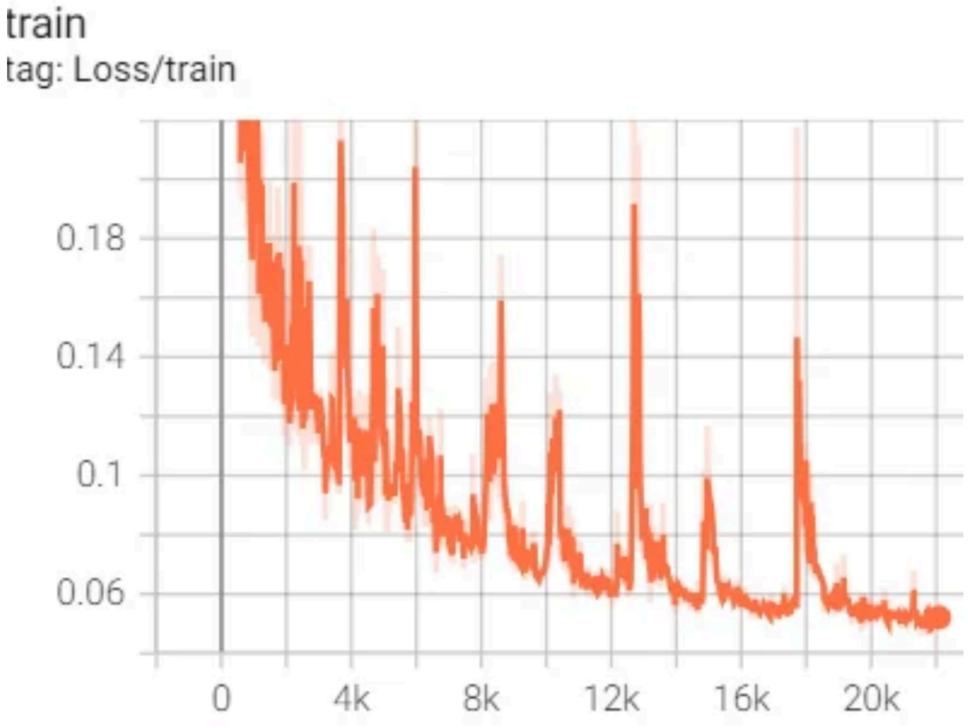
Training policies in RL is difficult

We often see strange results during training

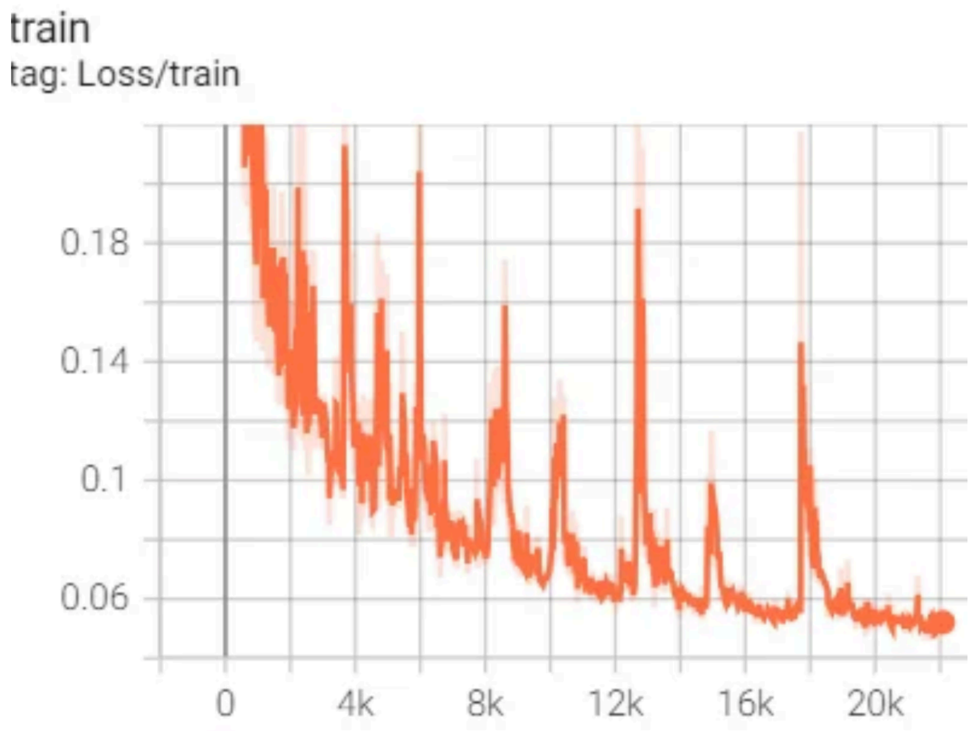


Question: Any idea why?

Trust Regions

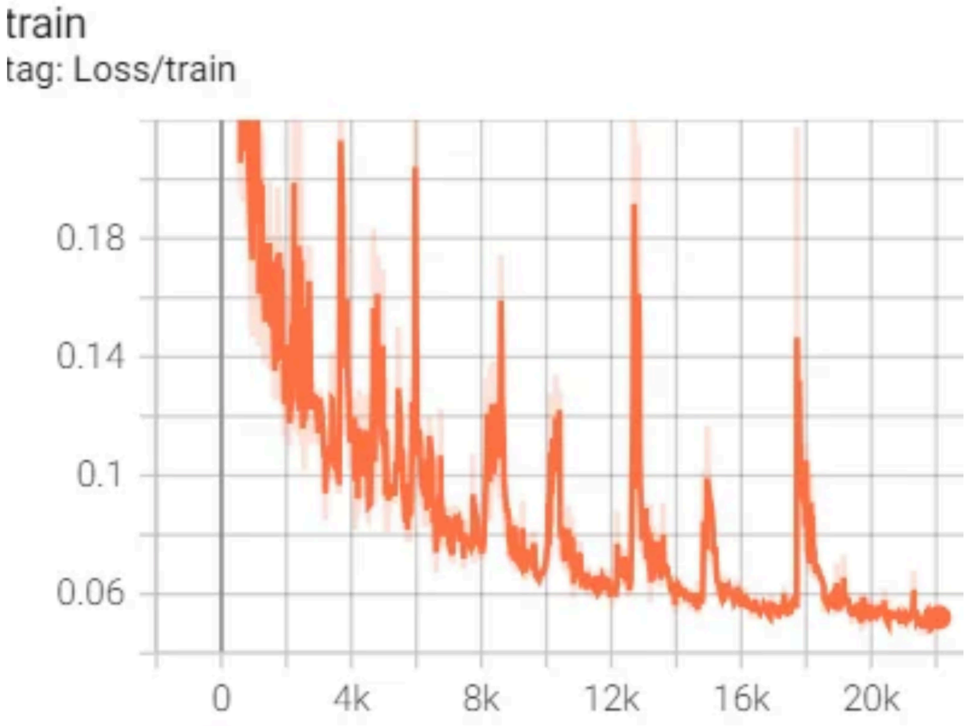


Trust Regions



See it in supervised learning too

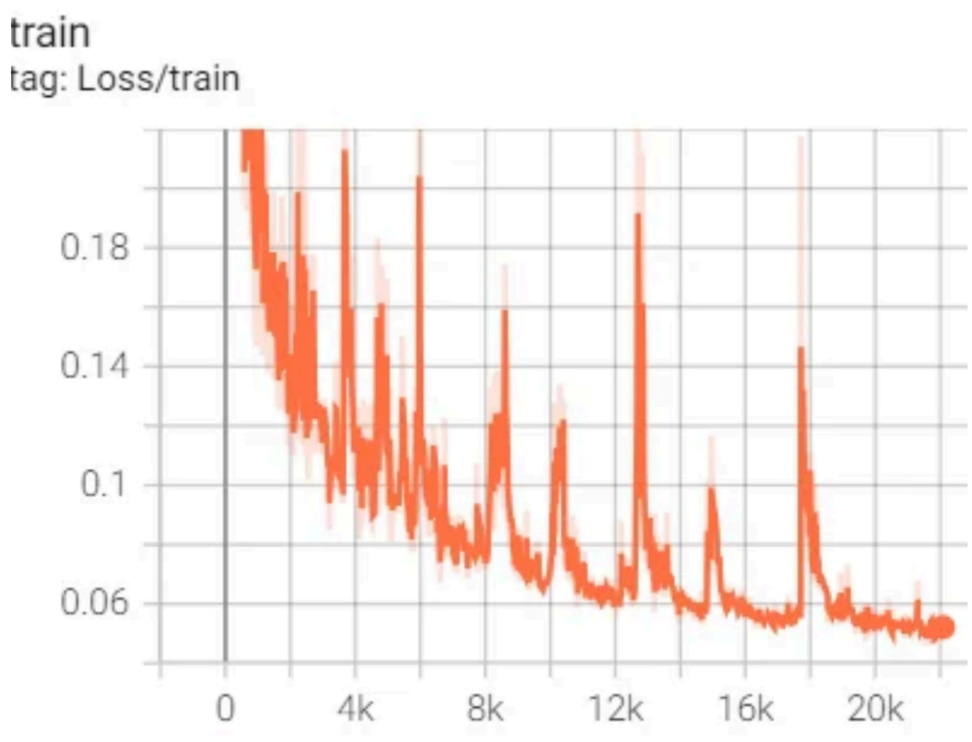
Trust Regions



See it in supervised learning too

Sometimes, the gradient is inaccurate producing a bad update

Trust Regions

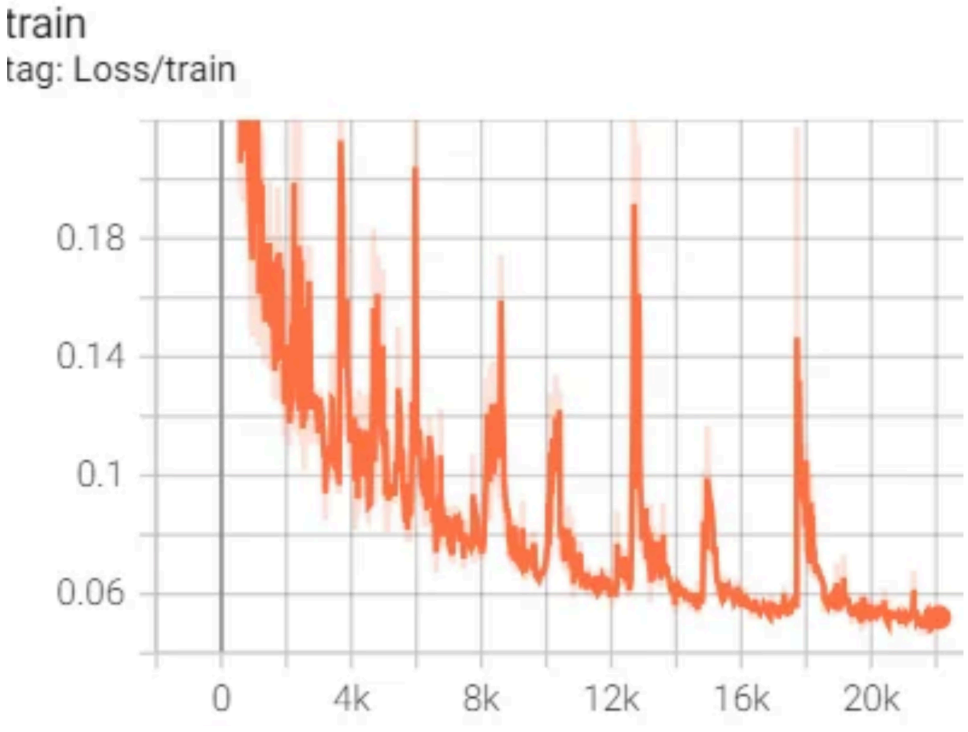


See it in supervised learning too

Sometimes, the gradient is inaccurate producing a bad update

In supervised learning, the network can easily recover

Trust Regions



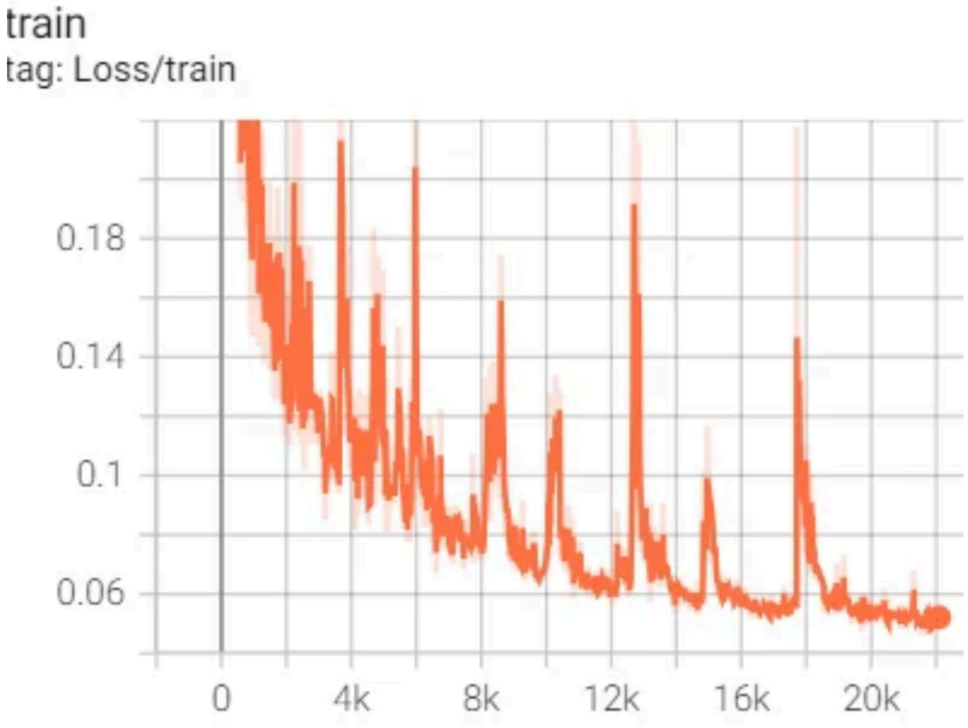
See it in supervised learning too

Sometimes, the gradient is inaccurate producing a bad update

In supervised learning, the network can easily recover

With policy gradient, it is much harder to recover

Trust Regions



See it in supervised learning too

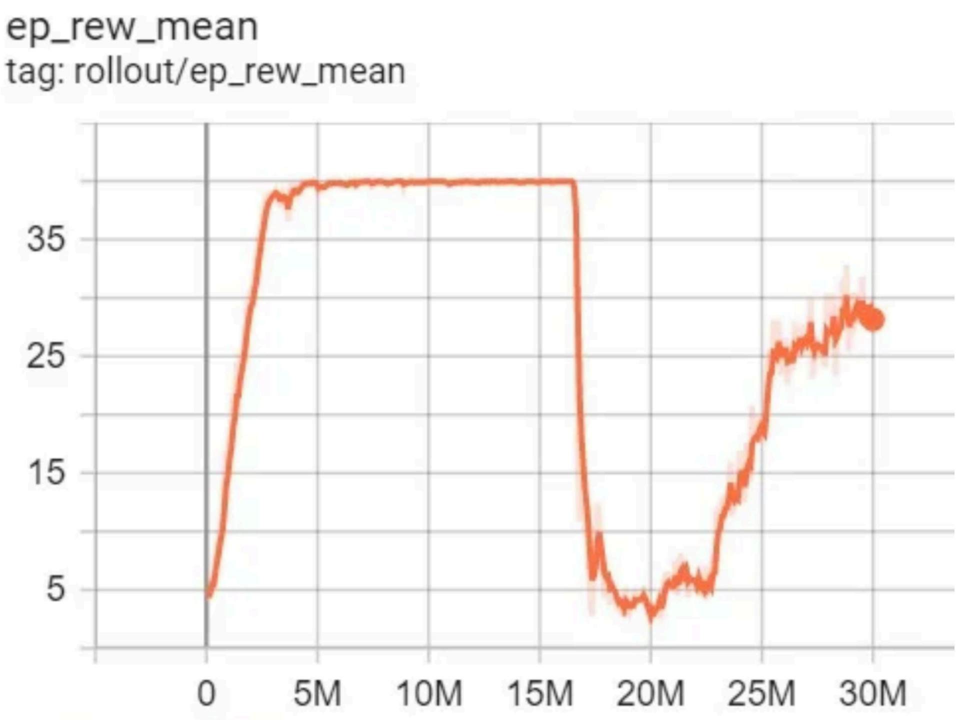
Sometimes, the gradient is inaccurate producing a bad update

In supervised learning, the network can easily recover

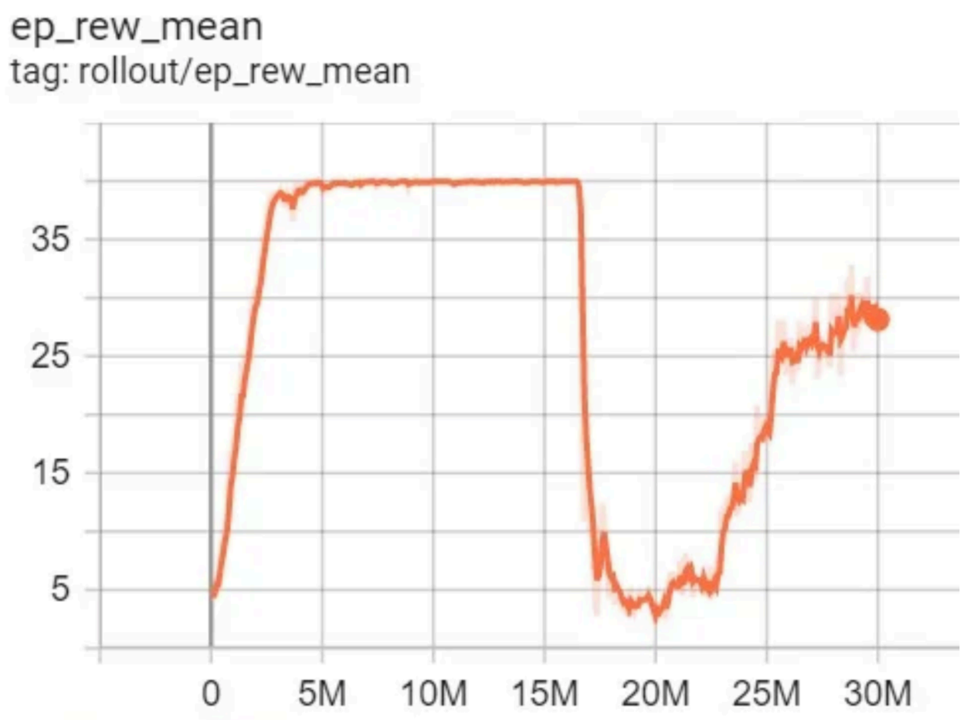
With policy gradient, it is much harder to recover

Question: Why is it harder to recover with policy gradient?

Trust Regions

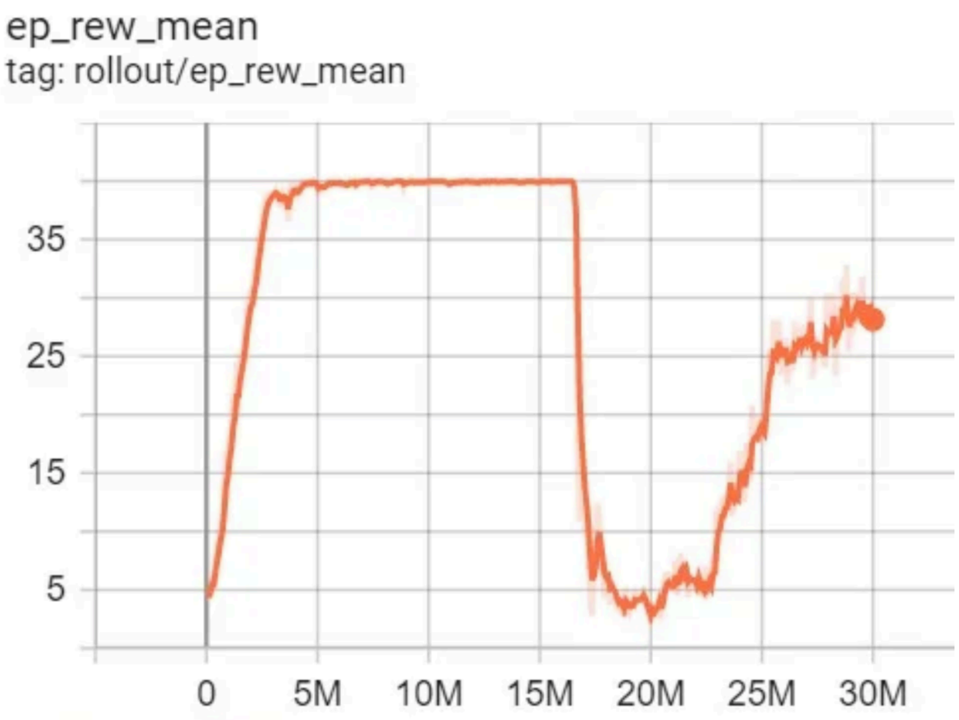


Trust Regions



Our policy provides the training data $a \sim \pi(\cdot \mid s; \theta_\pi)$

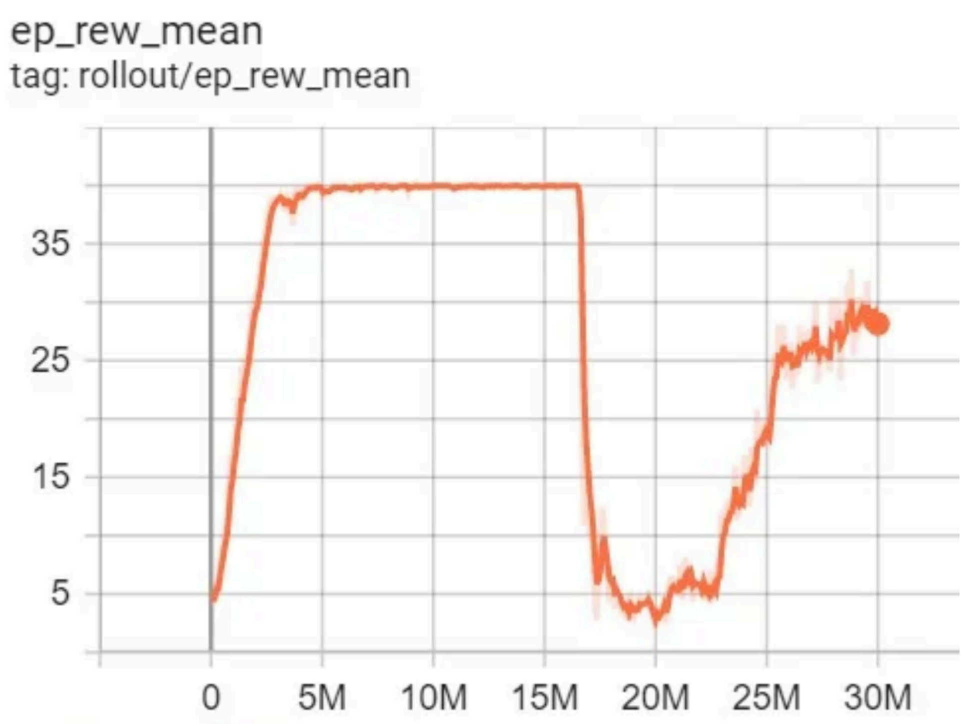
Trust Regions



Our policy provides the training data $a \sim \pi(\cdot \mid s; \theta_\pi)$

One bad update breaks the policy

Trust Regions

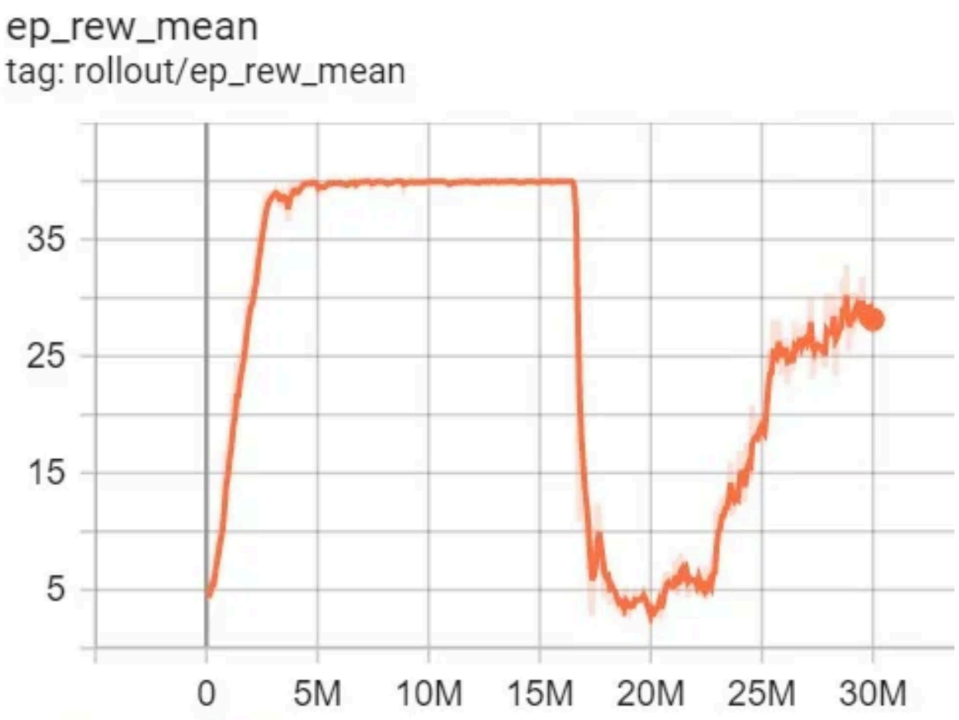


Our policy provides the training data $a \sim \pi(\cdot \mid s; \theta_\pi)$

One bad update breaks the policy

Policy collects useless data

Trust Regions



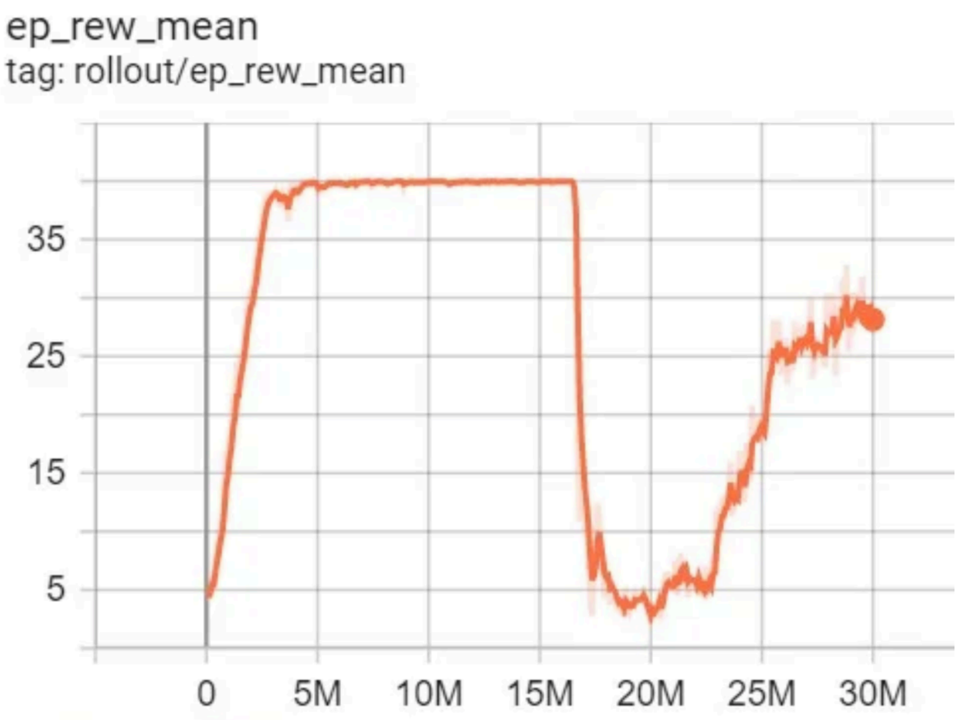
Our policy provides the training data $a \sim \pi(\cdot \mid s; \theta_\pi)$

One bad update breaks the policy

Policy collects useless data

Bad data = no policy recovery!

Trust Regions



Our policy provides the training data $a \sim \pi(\cdot \mid s; \theta_\pi)$

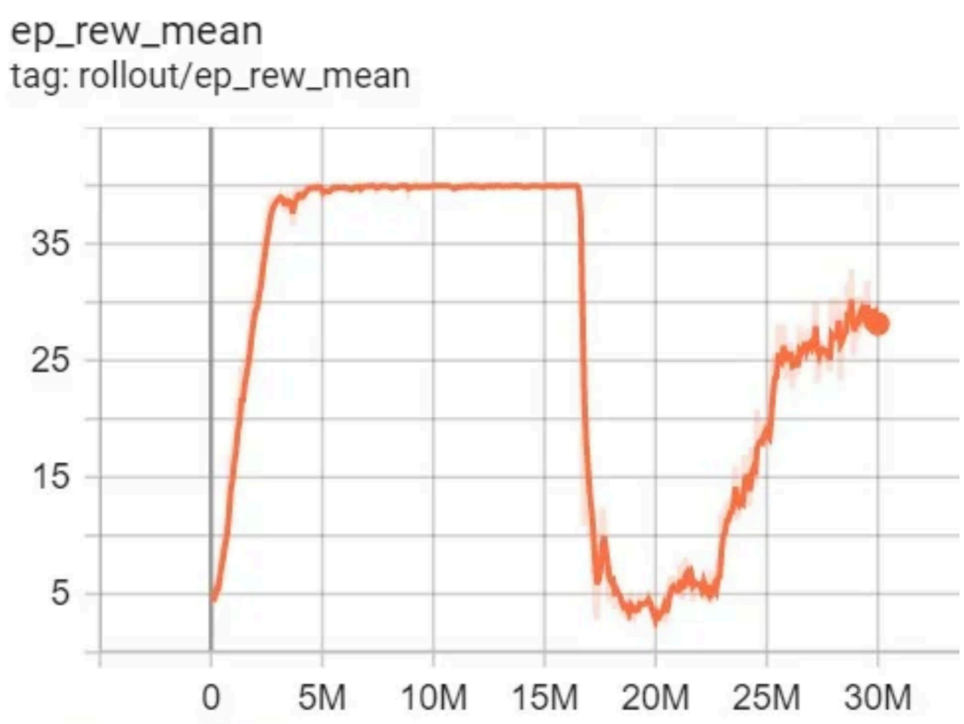
One bad update breaks the policy

Policy collects useless data

Bad data = no policy recovery!

Off-policy methods recover with "good" data from replay buffer, on-policy cannot

Trust Regions



Our policy provides the training data $a \sim \pi(\cdot \mid s; \theta_\pi)$

One bad update breaks the policy

Policy collects useless data

Bad data = no policy recovery!

Off-policy methods recover with "good" data from replay buffer, on-policy cannot

Must be very careful when updating policy using on-policy algorithms

Trust Regions

We can fix this issue with small changes to the policy

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Lower learning rate? Can help a little

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Lower learning rate? Can help a little

Small parameter change can cause large changes in deep networks

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Lower learning rate? Can help a little

Small parameter change can cause large changes in deep networks

$$\pi(a \mid s_A; \theta_{\pi, i}) = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$$

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Lower learning rate? Can help a little

Small parameter change can cause large changes in deep networks

$$\pi(a \mid s_A; \theta_{\pi,i}) = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \qquad \pi(a \mid s_A; \theta_{\pi,i+1}) = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Lower learning rate? Can help a little

Small parameter change can cause large changes in deep networks

$$\pi(a \mid s_A; \theta_{\pi,i}) = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \qquad \pi(a \mid s_A; \theta_{\pi,i+1}) = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Parameter-space constraints (learning rate) does not work very well!

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Lower learning rate? Can help a little

Small parameter change can cause large changes in deep networks

$$\pi(a \mid s_A; \theta_{\pi,i}) = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \qquad \pi(a \mid s_A; \theta_{\pi,i+1}) = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Parameter-space constraints (learning rate) does not work very well!

Question: What else can we constrain?

Trust Regions

We can fix this issue with small changes to the policy

Question: How can we make policy changes small?

Lower learning rate? Can help a little

Small parameter change can cause large changes in deep networks

$$\pi(a \mid s_A; \theta_{\pi,i}) = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \qquad \pi(a \mid s_A; \theta_{\pi,i+1}) = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Parameter-space constraints (learning rate) does not work very well!

Question: What else can we constrain?

Answer: The action distributions

Trust Regions

Can measure the difference in distributions using KL divergence

Trust Regions

Can measure the difference in distributions using KL divergence

$$\mathrm{KL}[\Pr(X), \Pr(Y)] \in [0, \infty]$$

Trust Regions

Can measure the difference in distributions using KL divergence

$$\mathrm{KL}[\Pr(X), \Pr(Y)] \in [0, \infty]$$

Policies are just action distributions

Trust Regions

Can measure the difference in distributions using KL divergence

$$\mathrm{KL}[\Pr(X), \Pr(Y)] \in [0, \infty]$$

Policies are just action distributions

$$\mathrm{KL}\left[\pi\left(a \mid s; \theta_{\pi,i}\right), \pi\left(a \mid s; \theta_{\pi,i+1}\right)\right]$$

Trust Regions

Can measure the difference in distributions using KL divergence

$$\mathrm{KL}[\Pr(X), \Pr(Y)] \in [0, \infty]$$

Policies are just action distributions

$$\mathrm{KL}\left[\pi\left(a \mid s; \theta_{\pi,i}\right), \pi\left(a \mid s; \theta_{\pi,i+1}\right)\right]$$

Introduce **trust region** k to prevent large policy changes

Trust Regions

Can measure the difference in distributions using KL divergence

$$\mathrm{KL}[\Pr(X), \Pr(Y)] \in [0, \infty]$$

Policies are just action distributions

$$\mathrm{KL}\left[\pi\left(a \mid s; \theta_{\pi,i}\right), \pi\left(a \mid s; \theta_{\pi,i+1}\right)\right]$$

Introduce **trust region** k to prevent large policy changes

$$\theta_{\pi,i+1} = V\left(s_0, \theta_{\pi,i}\right) \cdot \nabla_{\theta_\pi} \log \pi\left(a_0 \mid s_0; \theta_{\pi,i}\right)$$

Trust Regions

Can measure the difference in distributions using KL divergence

$$\mathrm{KL}[\Pr(X), \Pr(Y)] \in [0, \infty]$$

Policies are just action distributions

$$\mathrm{KL}\left[\pi\left(a \mid s; \theta_{\pi,i}\right), \pi\left(a \mid s; \theta_{\pi,i+1}\right)\right]$$

Introduce **trust region** k to prevent large policy changes

$$\theta_{\pi,i+1} = V\left(s_0, \theta_{\pi,i}\right) \cdot \nabla_{\theta_\pi} \log \pi\left(a_0 \mid s_0; \theta_{\pi,i}\right)$$
$$s.t.\ \ \mathrm{KL}\left[\pi\left(a \mid s; \theta_{\pi,i}\right), \pi\left(a \mid s; \theta_{\pi,i+1}\right)\right] < k$$

Trust Regions

Can measure the difference in distributions using KL divergence

$$\mathrm{KL}[\Pr(X), \Pr(Y)] \in [0, \infty]$$

Policies are just action distributions

$$\mathrm{KL}\left[\pi\left(a \mid s; \theta_{\pi,i}\right), \pi\left(a \mid s; \theta_{\pi,i+1}\right)\right]$$

Introduce **trust region** k to prevent large policy changes

$$\theta_{\pi,i+1} = V\left(s_0, \theta_{\pi,i}\right) \cdot \nabla_{\theta_\pi} \log \pi\left(a_0 \mid s_0; \theta_{\pi,i}\right)$$
$$s.t.\ \ \mathrm{KL}\left[\pi\left(a \mid s; \theta_{\pi,i}\right), \pi\left(a \mid s; \theta_{\pi,i+1}\right)\right] < k$$

See Trust Region Policy Optimization (TRPO), Natural Policy Gradient

Trust Regions

$$\theta_{\pi, i+1} = V(s_0, \theta_{\pi, i}) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_{\pi, i})$$

$$s.t. \; \mathrm{KL}[\pi(a \mid s; \theta_{\pi, i}), \pi(a \mid s; \theta_{\pi, i+1})] < k$$

Trust Regions

$$\theta_{\pi,i+1} = V(s_0, \theta_{\pi,i}) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_{\pi,i})$$

$$s.t. \;\; \mathrm{KL}[\pi(a \mid s; \theta_{\pi,i}), \pi(a \mid s; \theta_{\pi,i+1})] < k$$

Constrained optimization can be expensive and tricky to implement

Trust Regions

$$\theta_{\pi, i+1} = V(s_0, \theta_{\pi, i}) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_{\pi, i})$$

$$s.t. \ \ \mathrm{KL}[\pi(a \mid s; \theta_{\pi, i}), \pi(a \mid s; \theta_{\pi, i+1})] < k$$

Constrained optimization can be expensive and tricky to implement

Often requires inverting the gradient or computing Hessian

Trust Regions

$$\theta_{\pi,i+1} = V(s_0, \theta_{\pi,i}) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_{\pi,i})$$
$$s.t. \;\; \mathrm{KL}[\pi(a \mid s; \theta_{\pi,i}), \pi(a \mid s; \theta_{\pi,i+1})] < k$$

Constrained optimization can be expensive and tricky to implement

Often requires inverting the gradient or computing Hessian

Hack: Add KL term to the objective (soft constraint)

Trust Regions

$$\theta_{\pi,i+1} = V(s_0, \theta_{\pi,i}) \cdot \nabla_{\theta_\pi} \log \pi(a_0 \mid s_0; \theta_{\pi,i})$$
$$s.t. \;\; \mathrm{KL}[\pi(a \mid s; \theta_{\pi,i}), \pi(a \mid s; \theta_{\pi,i+1})] < k$$

Constrained optimization can be expensive and tricky to implement

Often requires inverting the gradient or computing Hessian

Hack: Add KL term to the objective (soft constraint)

$$\theta_{\pi,i+1} = V(s_0, \theta_{\pi,i}) \cdot \nabla_{\theta_{\pi,i}} [\log \pi(a_0 \mid s_0; \theta_{\pi,i})]$$
$$-\rho \nabla_{\theta_{\pi,i+1}} [\mathrm{KL}[\pi(a \mid s; \theta_{\pi,i}), \pi(a \mid s; \theta_{\pi,i+1})]]$$

Proximal Policy Optimization

Proximal Policy Optimization

Proximal policy optimization (PPO) combines all we learned today

Proximal Policy Optimization

Proximal policy optimization (PPO) combines all we learned today

- Value function for policy gradient (actor critic)

Proximal Policy Optimization

Proximal policy optimization (PPO) combines all we learned today

- Value function for policy gradient (actor critic)
- Advantage (stable updates, faster convergence)

Proximal Policy Optimization

Proximal policy optimization (PPO) combines all we learned today

- Value function for policy gradient (actor critic)
- Advantage (stable updates, faster convergence)
- Off-policy gradient (data efficiency)

Proximal Policy Optimization

Proximal policy optimization (PPO) combines all we learned today

- Value function for policy gradient (actor critic)
- Advantage (stable updates, faster convergence)
- Off-policy gradient (data efficiency)
- Trust regions (stable updates, prevents policy collapse)

Proximal Policy Optimization

Proximal policy optimization (PPO) combines all we learned today

- Value function for policy gradient (actor critic)
- Advantage (stable updates, faster convergence)
- Off-policy gradient (data efficiency)
- Trust regions (stable updates, prevents policy collapse)

Let us see a pseudocode PPO update

Proximal Policy Optimization

```python
for epoch in range(epochs):
    batch = collect_rollout(theta_beta)
    # Minibatching learns much faster
    # but is very slightly off-policy!
    for minibatch in batch:
        theta_pi = update_pi(
            theta_pi, theta_beta, theta_V, batch
        )
        theta_V = update_V(theta_V, batch)
    theta_beta = theta_pi
```

Proximal Policy Optimization

There are different variations of PPO

Proximal Policy Optimization

There are different variations of PPO

- PPO clip
- PPO KL penalty
- PPO clip + KL penalty
- PPO clip + KL penalty + entropy bonus

Proximal Policy Optimization

There are different variations of PPO

- PPO clip
- PPO KL penalty
- PPO clip + KL penalty
- PPO clip + KL penalty + entropy bonus

We will focus on the simplest version (PPO KL penalty)

Proximal Policy Optimization

$$\theta_{\pi, i+1} =$$

Proximal Policy Optimization

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \underbrace{\left(\frac{\pi(a \mid s; \theta_{\pi,i})}{\pi(a \mid s; \theta_\beta)} \; A(s_0, s_1, r_0, \theta_\beta, \theta_V) \right)}_{\text{Value}}$$

$$\cdot \left(\nabla_{\theta_{\pi,i}} \left[\log \pi(a_0 \mid s_0; \theta_{\pi,i}) \right] - \rho \nabla_{\theta_{\pi,i+1}} \left[\mathrm{KL}\left[\pi(a_0 \mid s_0; \theta_\beta), \pi(a_0 \mid s_0; \theta_{\pi,i+1}) \right] \right] \right)$$

Proximal Policy Optimization

Off-policy correction for minibatch

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \underbrace{\left(\frac{\pi(a \mid s; \theta_{\pi,i})}{\pi(a \mid s; \theta_{\beta})} A(s_0, s_1, r_0, \theta_\beta, \theta_V) \right)}_{\text{Value}}$$

$$\cdot \left(\nabla_{\theta_{\pi,i}} \left[\log \pi(a_0 \mid s_0; \theta_{\pi,i}) \right] - \rho \nabla_{\theta_{\pi,i+1}} \left[\mathrm{KL}\left[\pi(a_0 \mid s_0; \theta_\beta), \pi(a_0 \mid s_0; \theta_{\pi,i+1}) \right] \right] \right)$$

Proximal Policy Optimization

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \underbrace{\left(\overbrace{\frac{\pi(a \mid s; \theta_{\pi,i})}{\pi(a \mid s; \theta_\beta)}}^{\text{Off-policy correction for minibatch}} \overbrace{A(s_0, s_1, r_0, \theta_\beta, \theta_V)}^{\text{Advantage}} \right)}_{\text{Value}}$$

$$\cdot \left(\nabla_{\theta_{\pi,i}} \left[\log \pi(a_0 \mid s_0; \theta_{\pi,i}) \right] - \rho \nabla_{\theta_{\pi,i+1}} \left[\mathrm{KL}\left[\pi(a_0 \mid s_0; \theta_\beta), \pi(a_0 \mid s_0; \theta_{\pi,i+1}) \right] \right] \right)$$

Proximal Policy Optimization

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \underbrace{\left(\overbrace{\frac{\pi(a \mid s; \theta_{\pi,i})}{\pi(a \mid s; \theta_\beta)}}^{\text{Off-policy correction for minibatch}} \overbrace{A(s_0, s_1, r_0, \theta_\beta, \theta_V)}^{\text{Advantage}} \right)}_{\text{Value}}$$

$$\cdot \left(\underbrace{\nabla_{\theta_{\pi,i}} \left[\log \pi(a_0 \mid s_0; \theta_{\pi,i})\right]}_{\text{Policy gradient}} - \rho \nabla_{\theta_{\pi,i+1}} \left[\mathrm{KL}\left[\pi(a_0 \mid s_0; \theta_\beta), \pi(a_0 \mid s_0; \theta_{\pi,i+1})\right]\right] \right)$$

Proximal Policy Optimization

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \underbrace{\left(\overbrace{\frac{\pi(a \mid s; \theta_{\pi,i})}{\pi(a \mid s; \theta_\beta)}}^{\text{Off-policy correction for minibatch}} \overbrace{A(s_0, s_1, r_0, \theta_\beta, \theta_V)}^{\text{Advantage}} \right)}_{\text{Value}}$$

$$\cdot \left(\underbrace{\nabla_{\theta_{\pi,i}} [\log \pi(a_0 \mid s_0; \theta_{\pi,i})]}_{\text{Policy gradient}} - \rho \nabla_{\theta_{\pi,i+1}} \left[\underbrace{\mathrm{KL}[\pi(a_0 \mid s_0; \theta_\beta), \pi(a_0 \mid s_0; \theta_{\pi,i+1})]}_{\text{Trust region}} \right] \right)$$

Proximal Policy Optimization

Off-policy correction for minibatch

Advantage

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \underbrace{\left(\frac{\pi(a \mid s; \theta_{\pi,i})}{\pi(a \mid s; \theta_{\beta})} A(s_0, s_1, r_0, \theta_\beta, \theta_V) \right)}_{\text{Value}}$$

$$\cdot \left(\nabla_{\theta_{\pi,i}} [\log \pi(a_0 \mid s_0; \theta_{\pi,i})] - \rho \nabla_{\theta_{\pi,i+1}} [\mathrm{KL}[\pi(a_0 \mid s_0; \theta_\beta), \pi(a_0 \mid s_0; \theta_{\pi,i+1})]] \right)$$

Policy gradient

Trust region

$$A(s_0, s_1, r_0, \theta_\beta, \theta_V) = -V(s_0, \theta_\beta, \theta_V) + \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\beta] + \neg d \gamma V(s_1, \theta_\beta, \theta_V) \right)$$

Proximal Policy Optimization

Off-policy correction for minibatch

Advantage

$$\theta_{\pi,i+1} = \theta_{\pi,i} + \alpha \cdot \underbrace{\left(\frac{\pi(a \mid s; \theta_{\pi,i})}{\pi(a \mid s; \theta_\beta)} A(s_0, s_1, r_0, \theta_\beta, \theta_V) \right)}_{\text{Value}}$$

$$\cdot \left(\nabla_{\theta_{\pi,i}} \left[\log \pi(a_0 \mid s_0; \theta_{\pi,i}) \right] - \rho \nabla_{\theta_{\pi,i+1}} \left[\mathrm{KL}\left[\pi(a_0 \mid s_0; \theta_\beta), \pi(a_0 \mid s_0; \theta_{\pi,i+1}) \right] \right] \right)$$

Policy gradient

Trust region

$$A(s_0, s_1, r_0, \theta_\beta, \theta_V) = -V(s_0, \theta_\beta, \theta_V) + \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\beta] + \neg d \gamma V(s_1, \theta_\beta, \theta_V) \right)$$

$$\theta_{V,i+1} = \arg\min_{\theta_{V,i}} \left(V(s_0, \theta_\beta, \theta_{V,i}) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0; \theta_\beta] + \neg d \gamma V(s_0, \theta_\beta, \theta_{V,i}) \right) \right)^2$$

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Cohere finds REINFORCE better than PPO for LLM training

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Cohere finds REINFORCE better than PPO for LLM training

https://arxiv.org/pdf/2402.14740v1

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Cohere finds REINFORCE better than PPO for LLM training

https://arxiv.org/pdf/2402.14740v1

Our experiments find that Q learning outperforms PPO

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Cohere finds REINFORCE better than PPO for LLM training

https://arxiv.org/pdf/2402.14740v1

Our experiments find that Q learning outperforms PPO

My suggestions:

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Cohere finds REINFORCE better than PPO for LLM training

https://arxiv.org/pdf/2402.14740v1

Our experiments find that Q learning outperforms PPO

My suggestions:

- Try A2C first, solid actor-critic method, easy to implement

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Cohere finds REINFORCE better than PPO for LLM training

https://arxiv.org/pdf/2402.14740v1

Our experiments find that Q learning outperforms PPO

My suggestions:

- Try A2C first, solid actor-critic method, easy to implement
- Large batches and regularization (weight decay, layer norm) helpful

Proximal Policy Optimization

Personal opinion: PPO is overrated, for some reason very popular

Many hyperparameters, hard to implement, computationally expensive

Cohere finds REINFORCE better than PPO for LLM training

https://arxiv.org/pdf/2402.14740v1

Our experiments find that Q learning outperforms PPO

My suggestions:
- Try A2C first, solid actor-critic method, easy to implement
- Large batches and regularization (weight decay, layer norm) helpful
- You can make any algorithm work with enough effort!

Proximal Policy Optimization

PPO plays Pokemon!

Video describes the RL experiment process, helpful for your final project

https://youtu.be/DcYLT37ImBY?si=jJfZyYwFkPYMJYMy