

Policy Gradient

CMPT 729 G100

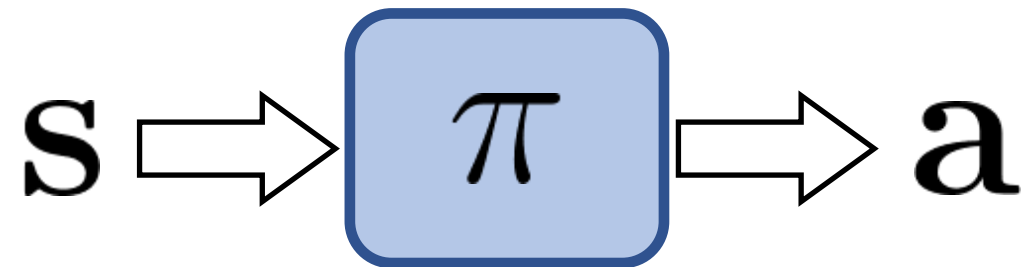
Jason Peng

Overview

- Taxonomy of RL Algorithms
- Policy Gradient
- Derivation
- Variance Reduction
- Applications
- General View of PG

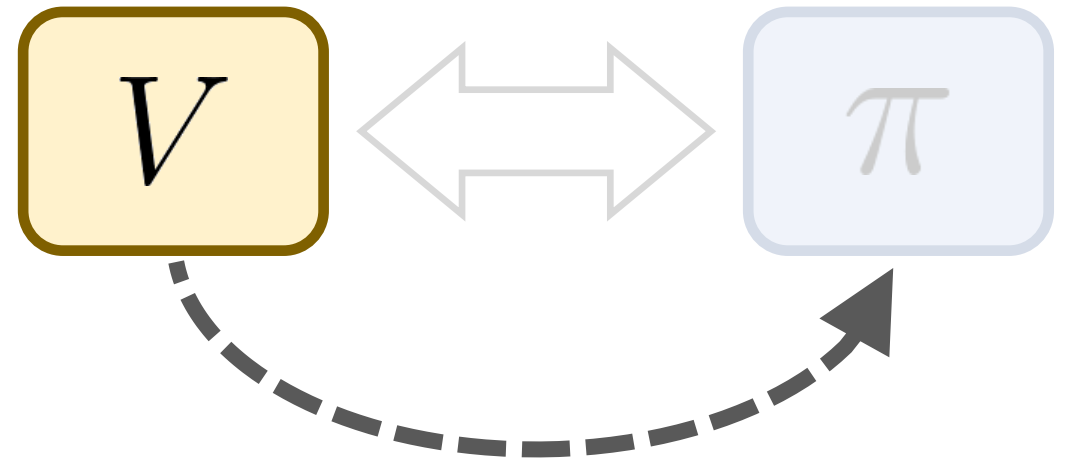
Taxonomy of RL Algorithms

- Policy-Based Methods
- Value-Based Methods
- Actor-Critic Methods
- Model-Based Methods



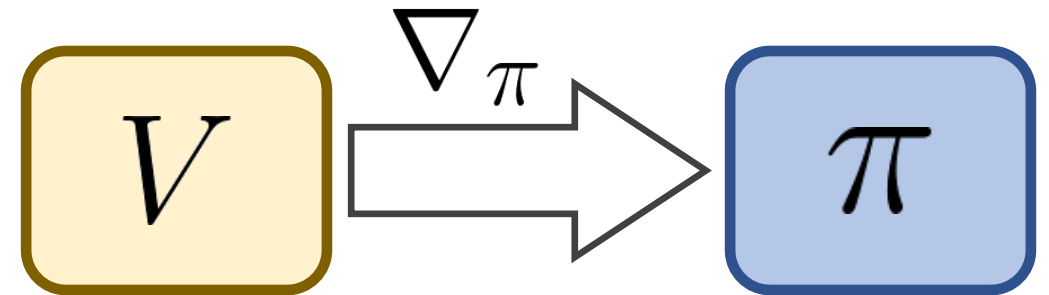
Taxonomy of RL Algorithms

- Policy-Based Methods
- Value-Based Methods
- Actor-Critic Methods
- Model-Based Methods



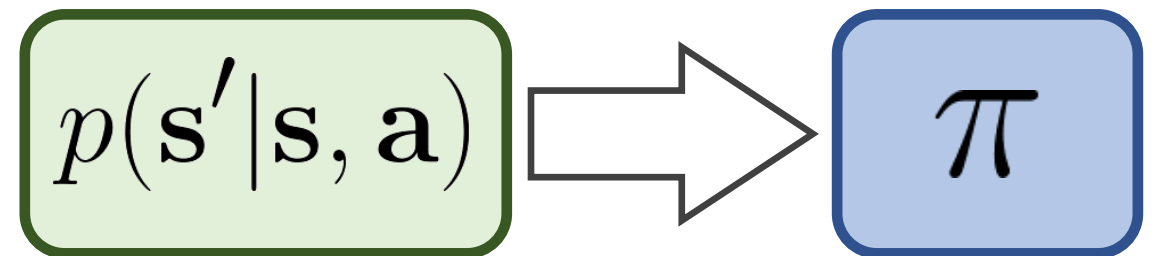
Taxonomy of RL Algorithms

- Policy-Based Methods
- Value-Based Methods
- Actor-Critic Methods
- Model-Based Methods



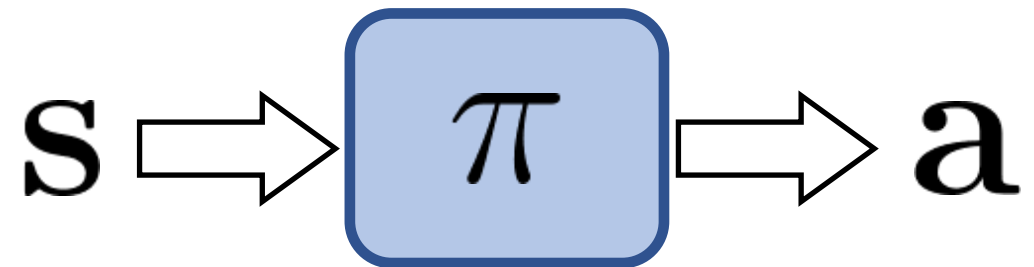
Taxonomy of RL Algorithms

- Policy-Based Methods
- Value-Based Methods
- Actor-Critic Methods
- Model-Based Methods



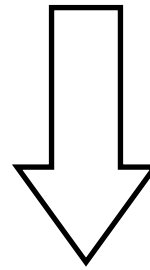
Taxonomy of RL Algorithms

- Policy-Based Methods
- Value-Based Methods
- Actor-Critic Methods
- Model-Based Methods



Nondifferentiable Objective

$$\theta^* = \arg \max_{\theta} J(\pi_{\theta})$$



Just use gradient ascent!

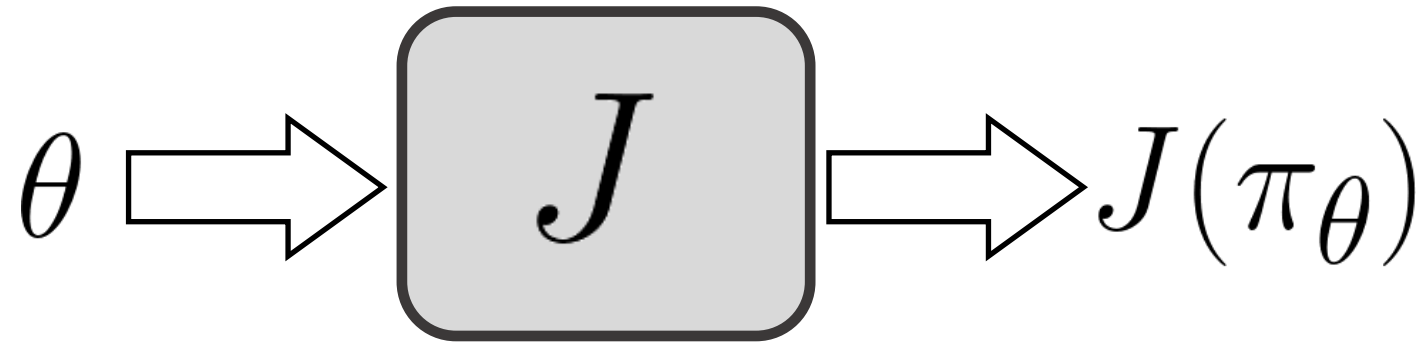
Objective is often
NOT differentiable

~~$$\nabla_{\theta} J(\pi_{\theta})$$~~

Black Box Optimization

$$\theta^* = \arg \max_{\theta} \underline{J(\pi_{\theta})}$$

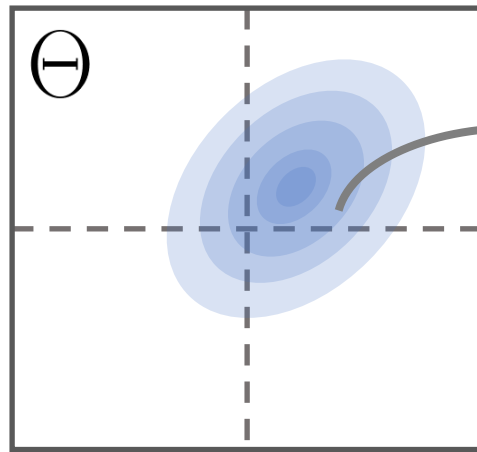
black box



Black Box Optimization

- Adapt search samples base on objective

search distribution

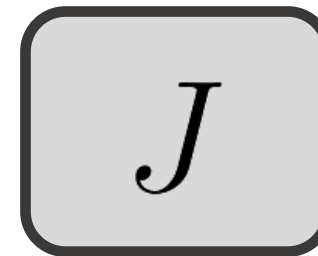


sample

θ^j



evaluate



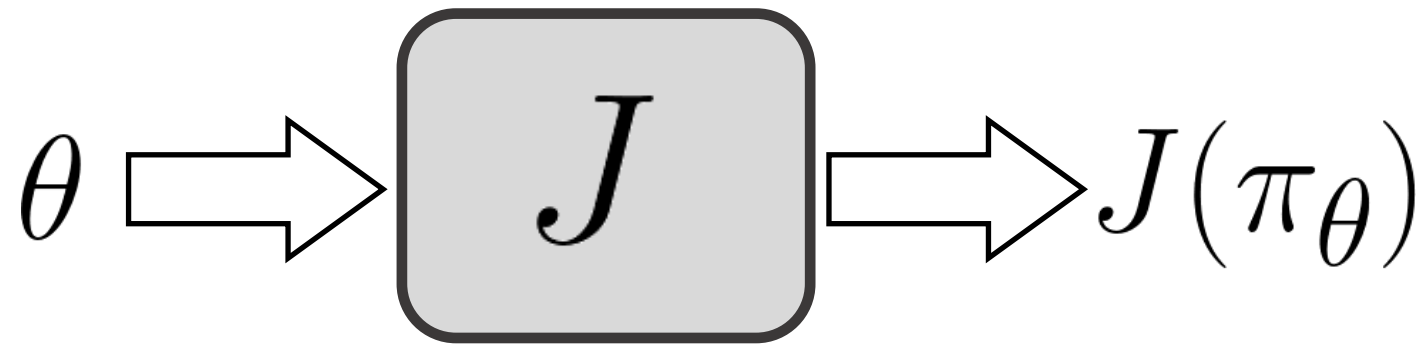
$J(\pi_{\theta^j})$

update

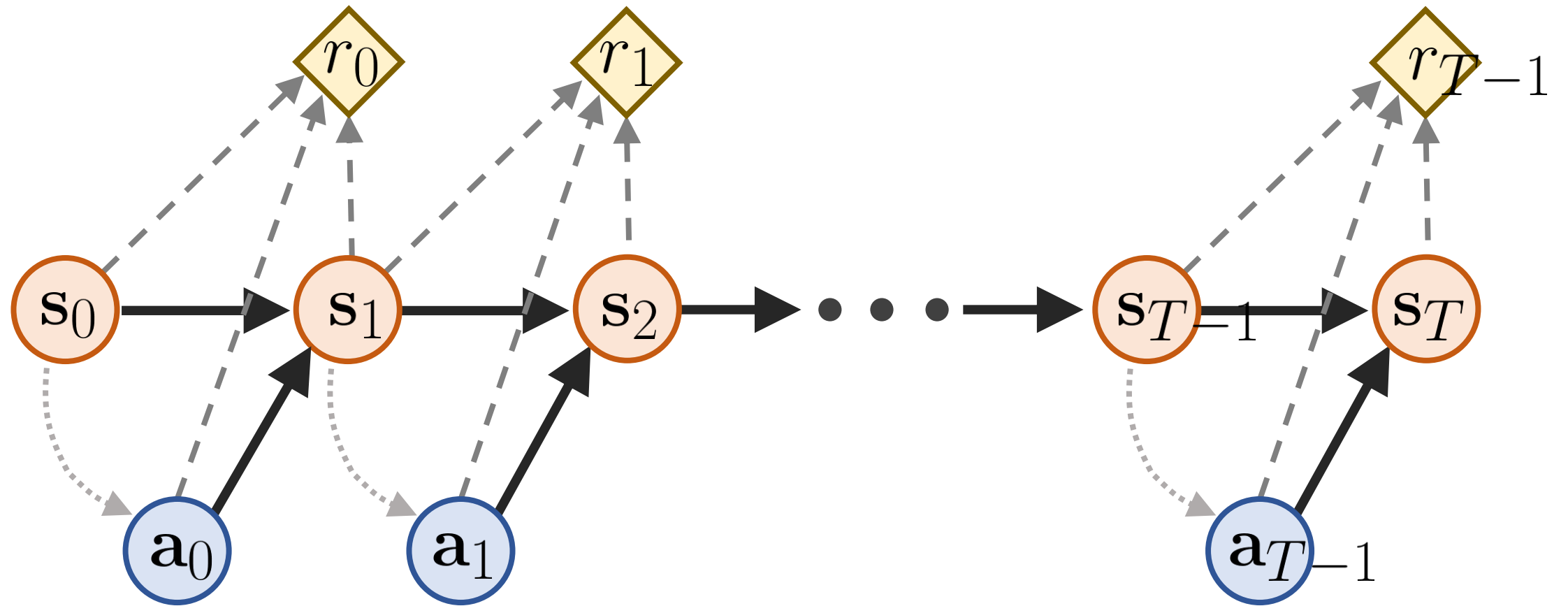


Black Box Optimization

$$\theta^* = \arg \max_{\theta} J(\pi_{\theta})$$



MDP



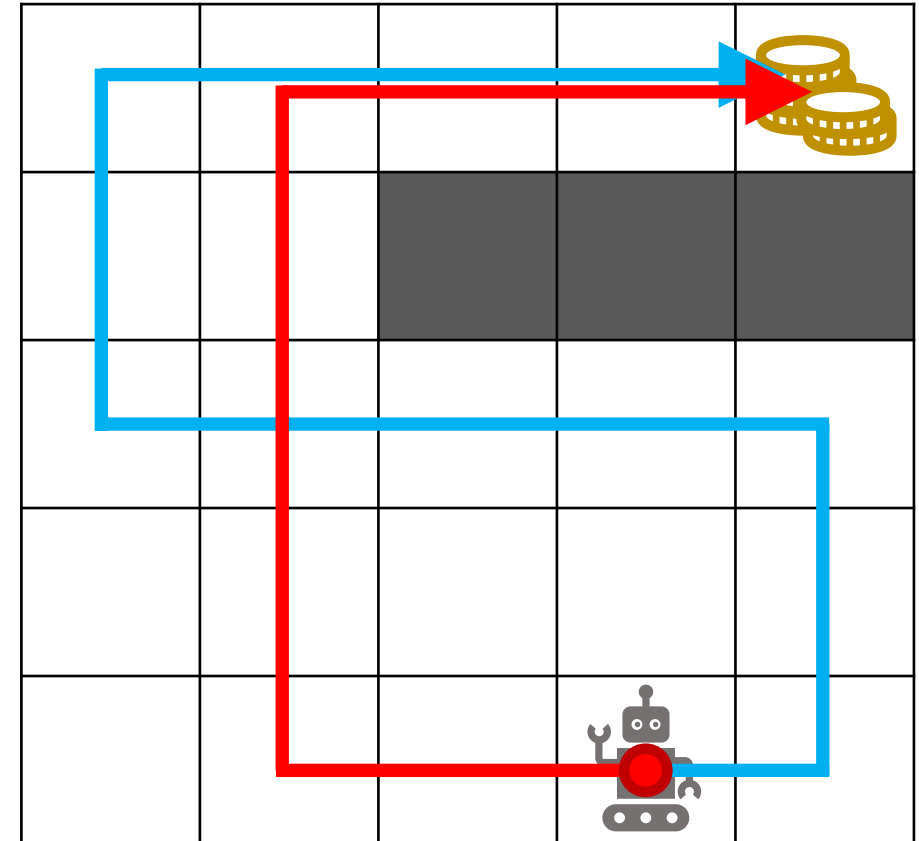
Behavioral Timescales

- Lifetime

Behavioral Timescales

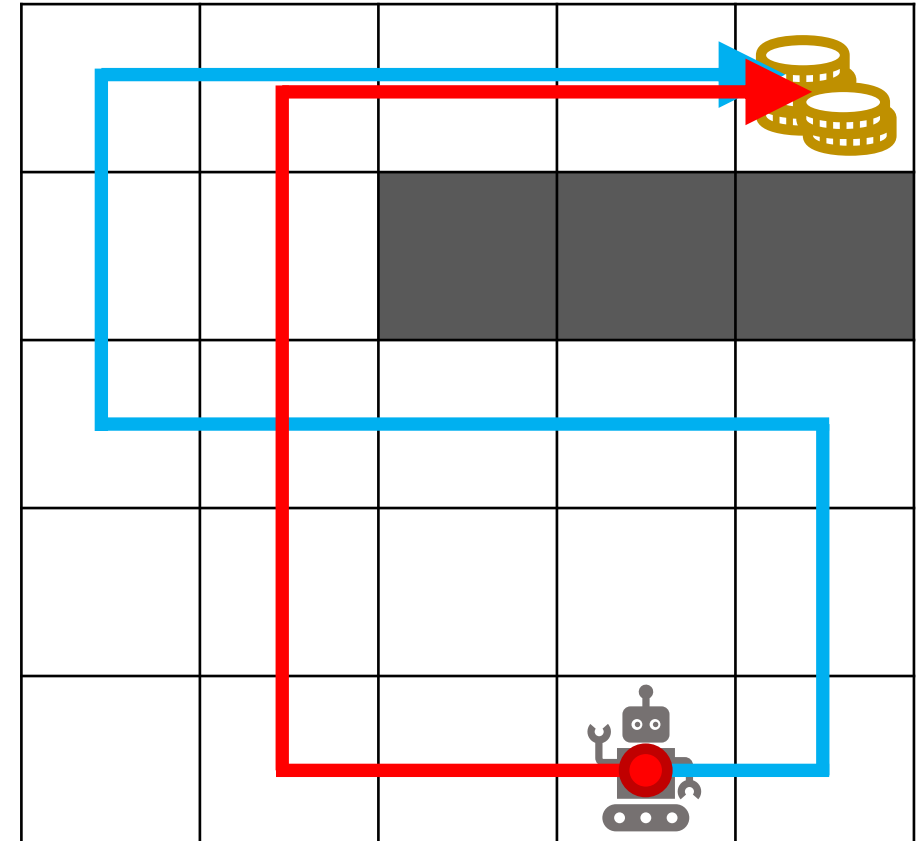
- Lifetime

Evolutionary Methods



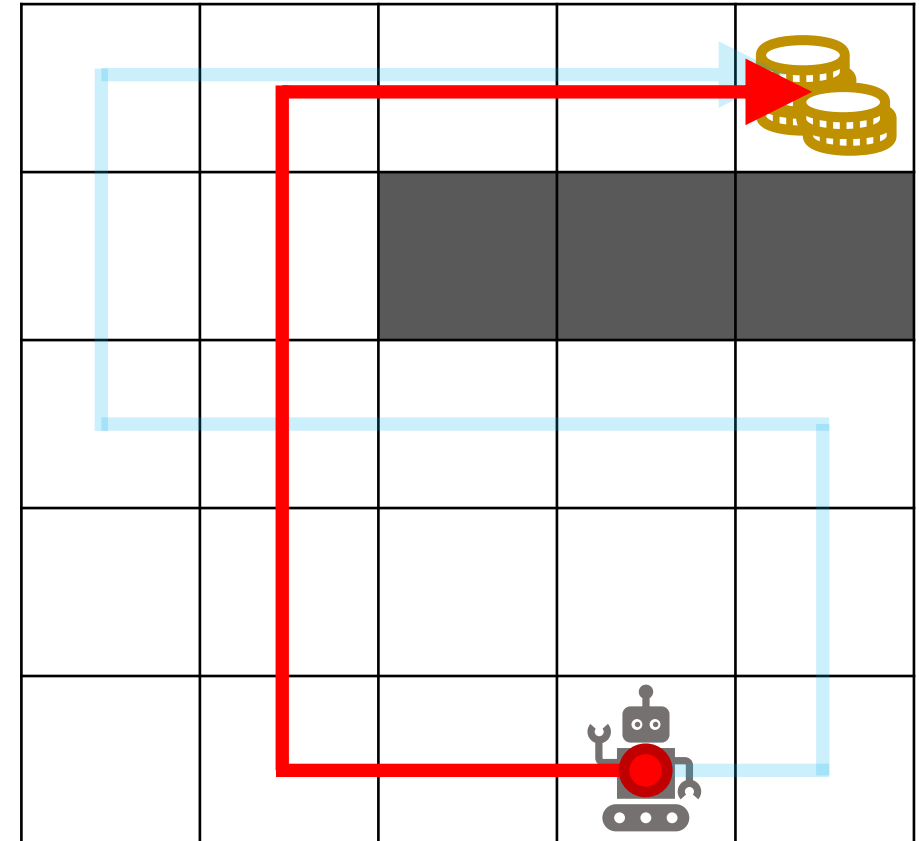
Behavioral Timescales

- Lifetime
- Trajectories



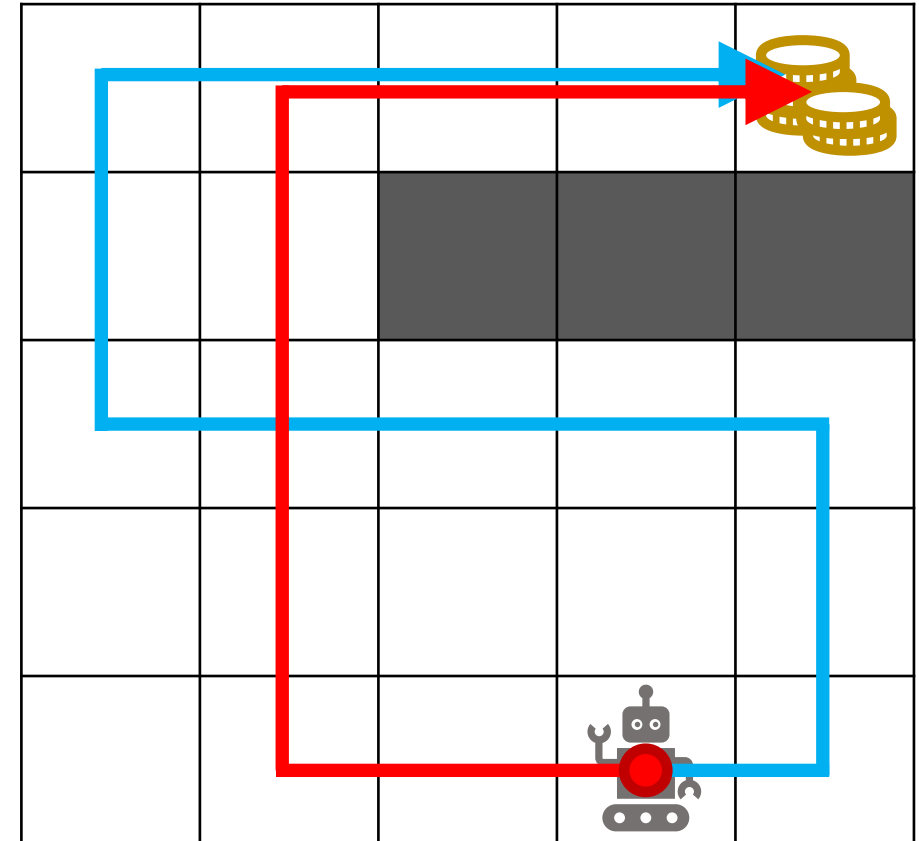
Behavioral Timescales

- Lifetime
- Trajectories

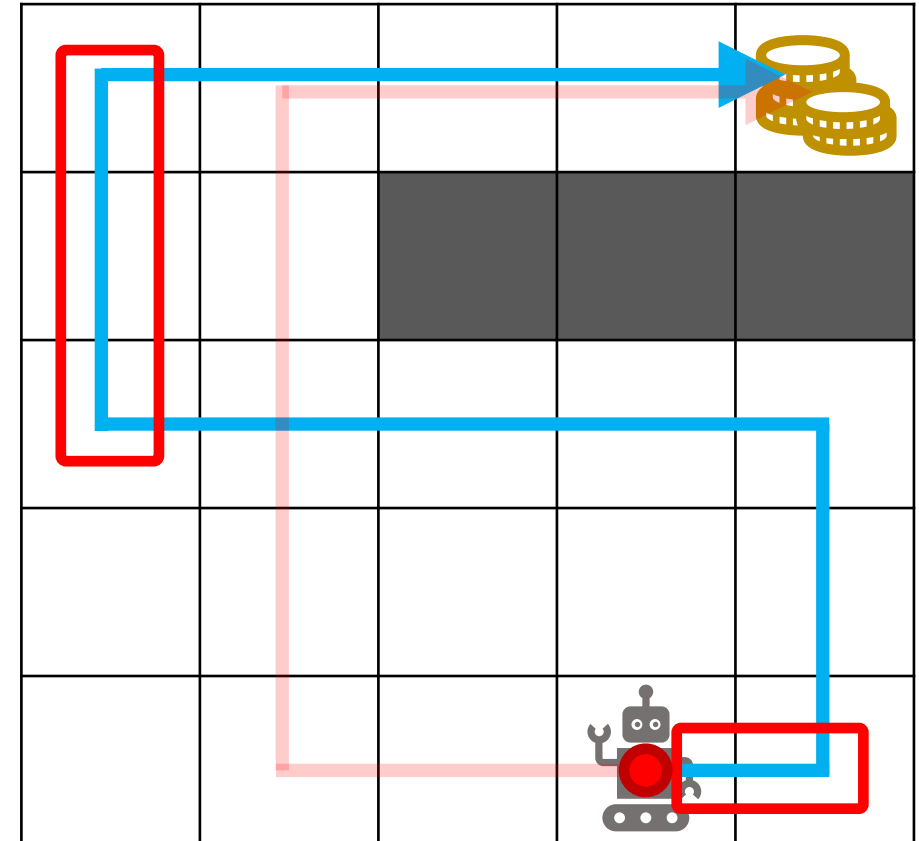


Behavioral Timescales

- Lifetime
- Trajectories
- Actions



- Lifetime
- Trajectories
- Actions

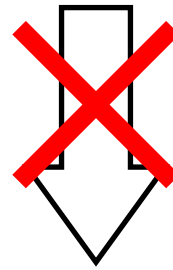


Nondifferentiable Objective

$$\theta^* = \arg \max_{\theta} J(\pi_{\theta})$$

$J(\pi_{\theta})$

nondifferentiable

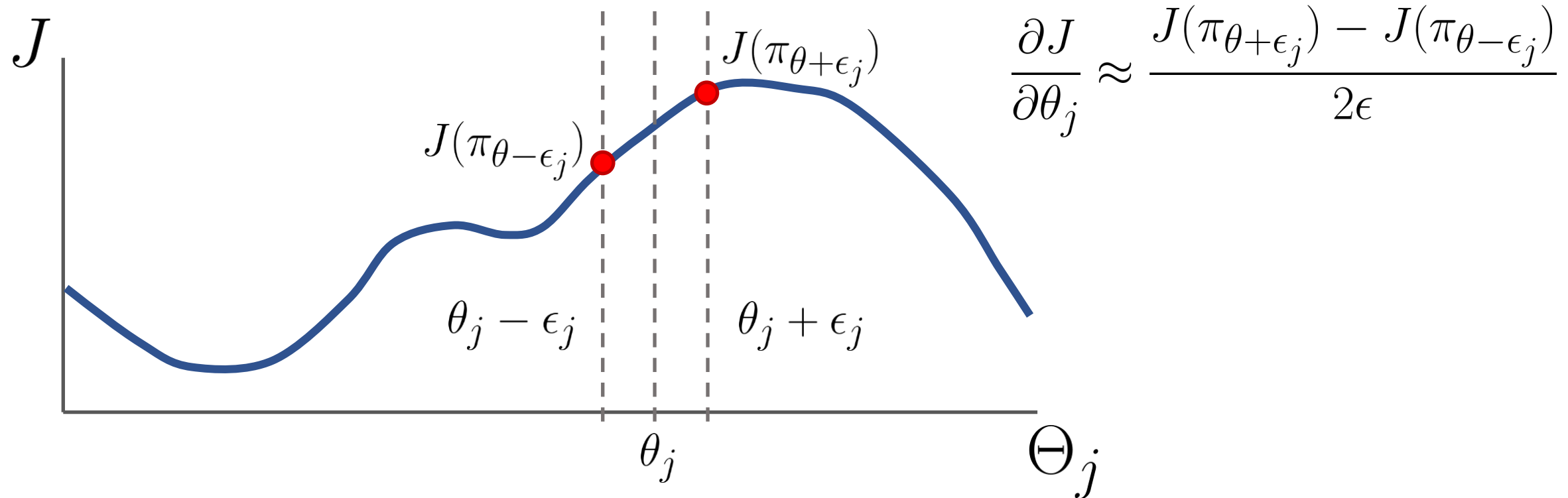


$$\nabla_{\theta} J(\pi_{\theta})$$

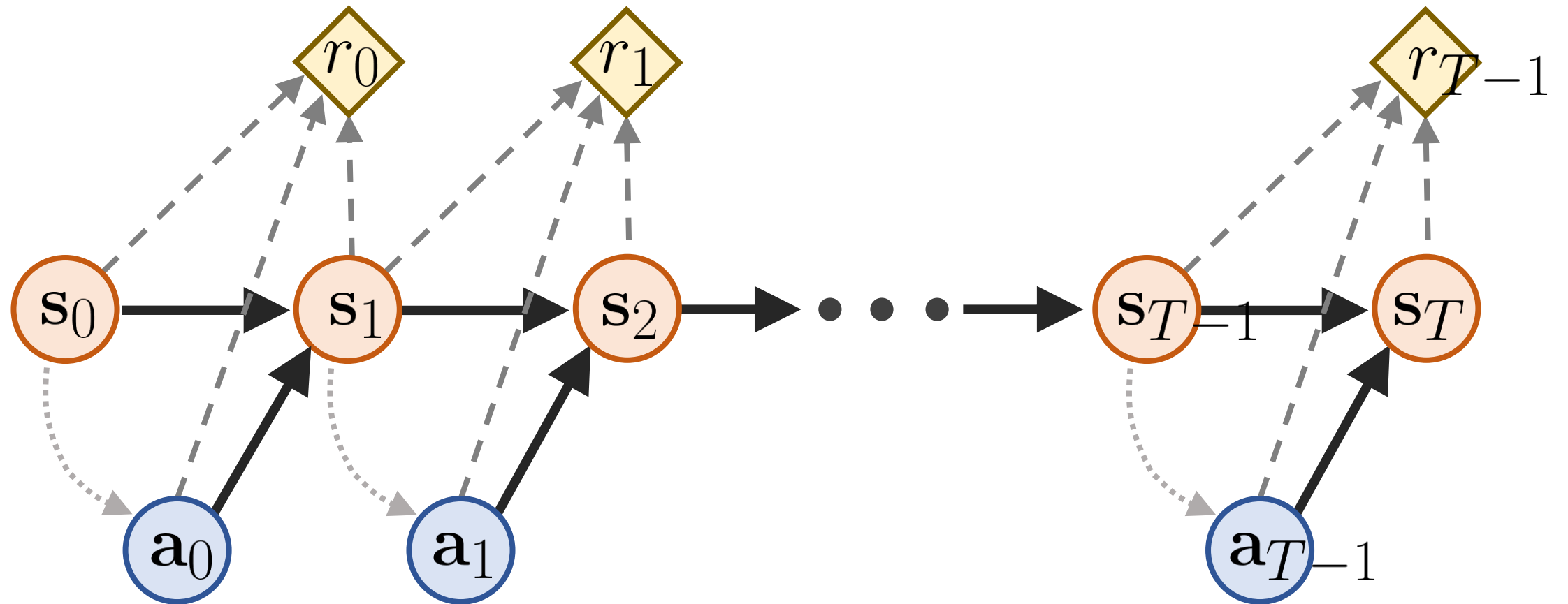
Can we approximate?

Finite-Differences

- Approximate gradient using finite-differences

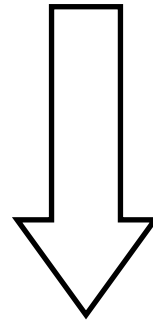


MDP

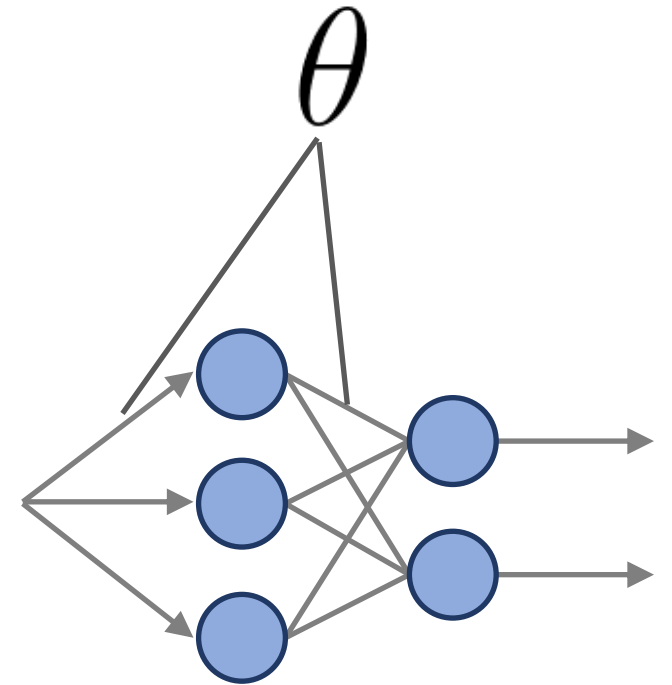


Notation

$$\nabla_{\theta} J(\pi_{\theta})$$



$$\underline{\nabla_{\pi}} J(\pi)$$



Policy Gradients

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right] = \mathbb{E}_{\tau \sim p(\tau|\pi)} [\underline{R(\tau)}]$$

return of a trajectory

Policy Gradients

$$\begin{aligned} J(\pi) &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right] = \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] \\ &= \sum_{\tau} \underline{p(\tau|\pi)} R(\tau) \end{aligned}$$

Policy Gradients

$$\begin{aligned} J(\pi) &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right] = \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] \\ &= \sum_{\tau} p(\tau|\pi) \underline{R(\tau)} \end{aligned}$$

Policy Gradients

$$\begin{aligned} J(\pi) &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right] = \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] \\ &= \sum_{\tau} p(\tau|\pi) R(\tau) \end{aligned}$$

$$\nabla_{\pi} J(\pi) = \sum_{\tau} \nabla_{\pi} p(\tau|\pi) R(\tau)$$

completely intractable

Policy Gradients

$$\nabla_{\pi} J(\pi) = \sum_{\tau} \nabla_{\pi} p(\tau|\pi) R(\tau)$$

Score Function

$$\nabla_p p(x) = p(x) \frac{\nabla_p p(x)}{p(x)} = \underline{p(x)} \nabla_p \log p(x)$$

Policy Gradients

$$\begin{aligned}\nabla_{\pi} J(\pi) &= \sum_{\tau} \nabla_{\pi} p(\tau|\pi) R(\tau) \\ &= \sum_{\tau} \underline{p(\tau|\pi)} \nabla_{\pi} \log p(\tau|\pi) R(\tau)\end{aligned}$$

Score Function

$$\nabla_p p(x) = p(x) \frac{\nabla_p p(x)}{p(x)} = p(x) \nabla_p \log p(x)$$

Policy Gradients

$$\begin{aligned}\nabla_{\pi} J(\pi) &= \sum_{\tau} \nabla_{\pi} p(\tau|\pi) R(\tau) \\ &= \sum_{\tau} \underline{p(\tau|\pi)} \nabla_{\pi} \log p(\tau|\pi) R(\tau)\end{aligned}$$

Score Function

$$\nabla_p p(x) = p(x) \frac{\nabla_p p(x)}{p(x)} = p(x) \nabla_p \log p(x)$$

Policy Gradients

$$\begin{aligned}\nabla_{\pi} J(\pi) &= \sum_{\tau} \nabla_{\pi} p(\tau|\pi) R(\tau) \\ &= \sum_{\tau} p(\tau|\pi) \nabla_{\pi} \log p(\tau|\pi) R(\tau) \\ &= \mathbb{E}_{\tau \sim p(\tau|\pi)} [\nabla_{\pi} \log p(\tau|\pi) R(\tau)]\end{aligned}$$

Score Function

$$\nabla_p p(x) = p(x) \frac{\nabla_p p(x)}{p(x)} = p(x) \nabla_p \log p(x)$$

Policy Gradients

$$\begin{aligned}\nabla_{\pi} J(\pi) &= \sum_{\tau} \nabla_{\pi} p(\tau|\pi) R(\tau) \\ &= \sum_{\tau} p(\tau|\pi) \nabla_{\pi} \log p(\tau|\pi) R(\tau) \\ &= \mathbb{E}_{\tau \sim p(\tau|\pi)} [\nabla_{\pi} \log p(\tau|\pi) R(\tau)]\end{aligned}$$

Score Function

$$\nabla_p p(x) = p(x) \frac{\nabla_p p(x)}{p(x)} = p(x) \nabla_p \log p(x)$$

Policy Gradients

$$\nabla_{\pi} \log p(\tau|\pi) = \nabla_{\pi} \log \underbrace{\left(p(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right)}_{= p(\tau|\pi)}$$

Policy Gradients

$$\nabla_{\pi} \log p(\tau|\pi) = \nabla_{\pi} \log \left(\underline{p(\mathbf{s}_0)} \prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right)$$

Policy Gradients

$$\nabla_{\pi} \log p(\tau|\pi) = \nabla_{\pi} \log \left(p(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right)$$

Policy Gradients

$$\nabla_{\pi} \log p(\tau|\pi) = \nabla_{\pi} \log \left(p(\mathbf{s}_0) \prod_{t=0}^{T-1} \underline{\pi(\mathbf{a}_t|\mathbf{s}_t)} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right)$$

Policy Gradients

$$\nabla_{\pi} \log p(\tau|\pi) = \nabla_{\pi} \log \left(p(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t) \underline{p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)} \right)$$

Policy Gradients

$$\begin{aligned}\nabla_{\pi} \log p(\tau|\pi) &= \nabla_{\pi} \log \left(p(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right) \\ &= \nabla_{\pi} \left(\log p(\mathbf{s}_0) + \sum_{t=0}^{T-1} \log \pi(\mathbf{a}_t|\mathbf{s}_t) + \log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right)\end{aligned}$$

Policy Gradients

$$\begin{aligned}\nabla_{\pi} \log p(\tau|\pi) &= \nabla_{\pi} \log \left(p(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right) \\ &= \nabla_{\pi} \left(\underbrace{\log p(\mathbf{s}_0)}_{\text{Independent of } \pi} + \sum_{t=0}^{T-1} \log \pi(\mathbf{a}_t|\mathbf{s}_t) + \underbrace{\log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}_{\text{Independent of } \pi} \right)\end{aligned}$$

Independent of π

Policy Gradients

$$\begin{aligned}\nabla_{\pi} \log p(\tau|\pi) &= \nabla_{\pi} \log \left(p(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right) \\ &= \nabla_{\pi} \left(\cancel{\log p(\mathbf{s}_0)} + \sum_{t=0}^{T-1} \log \pi(\mathbf{a}_t|\mathbf{s}_t) + \cancel{\log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)} \right) \\ &= \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t|\mathbf{s}_t)\end{aligned}$$

Policy Gradients

$$\begin{aligned}\nabla_{\pi} J(\pi) &= \sum_{\tau} \nabla_{\pi} p(\tau|\pi) R(\tau) \\ &= \sum_{\tau} p(\tau|\pi) \nabla_{\pi} \log p(\tau|\pi) R(\tau) \\ &= \mathbb{E}_{\tau \sim p(\tau|\pi)} [\nabla_{\pi} \log p(\tau|\pi) R(\tau)] \\ &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]\end{aligned}$$

Score Function

$$\nabla_{\pi} \pi(\tau) = \pi(\tau) \frac{\nabla_{\pi} \pi(\tau)}{\pi(\tau)} = \pi(\tau) \nabla_{\pi} \log \pi(\tau)$$

policy gradient
AKA. REINFORCE [Williams 1992]

REINFORCE

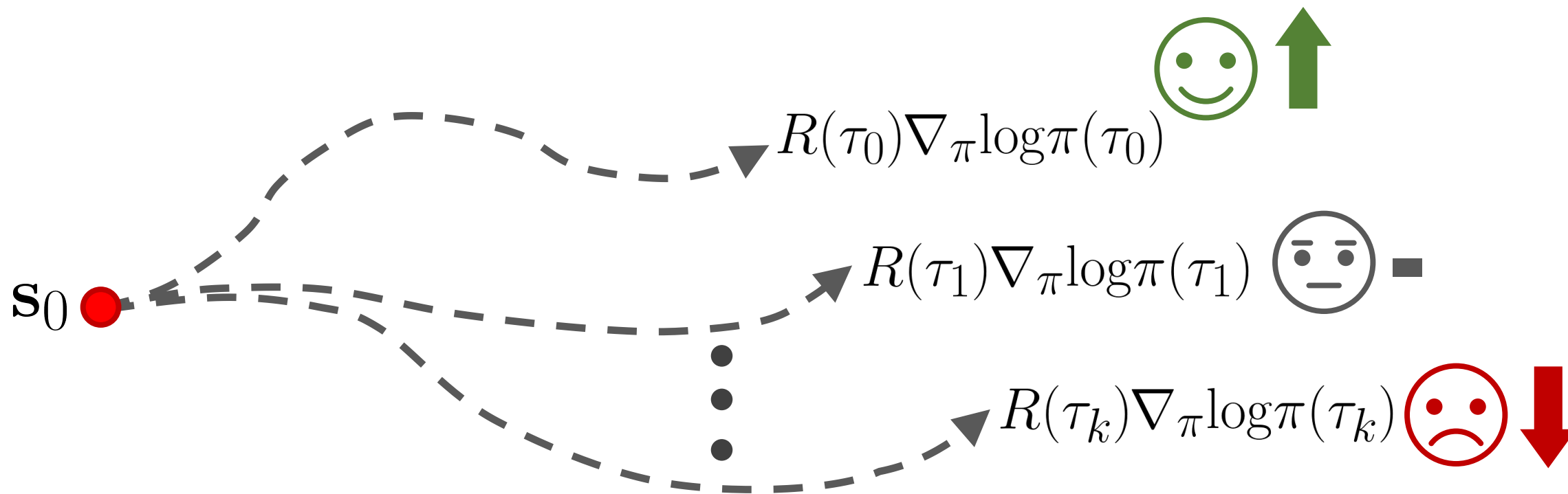
$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

REINFORCE

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\underline{R(\tau)} \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

REINFORCE

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



REINFORCE

ALGORITHM: REINFORCE

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate policy gradient
$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i R(\tau^i) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$$
 - 5: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 6: **end while**
 - 7: return policy π_θ
-

REINFORCE

ALGORITHM: REINFORCE

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate policy gradient
$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i R(\tau^i) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$$
 - 5: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 6: **end while**
 - 7: return policy π_θ
-

REINFORCE

ALGORITHM: REINFORCE

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate policy gradient
$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i R(\tau^i) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$$
 - 5: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 6: **end while**
 - 7: return policy π_θ
-

REINFORCE

ALGORITHM: REINFORCE

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate policy gradient
$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i R(\tau^i) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$$
 - 5: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 6: **end while**
 - 7: return policy π_θ
-

REINFORCE

ALGORITHM: REINFORCE

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate policy gradient
$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i R(\tau^i) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$$
 - 5: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 6: **end while**
 - 7: return policy π_θ
-

REINFORCE

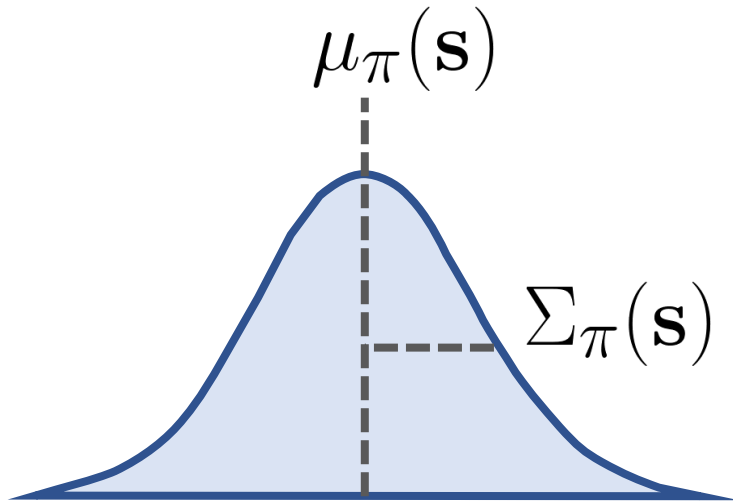
ALGORITHM: REINFORCE

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate policy gradient
$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i R(\tau^i) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$$
 - 5: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 6: **end while**
 - 7: return policy π_θ
-

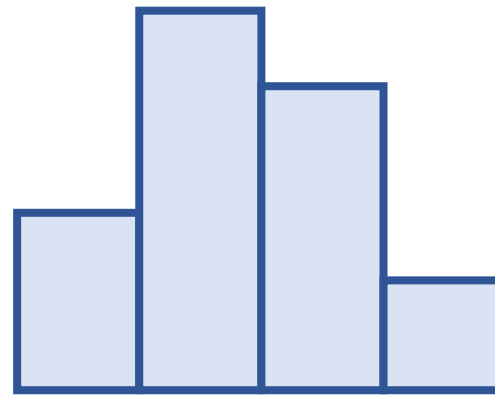
Action Distribution

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

must be differentiable



Gaussian Distribution
(Continuous Actions)

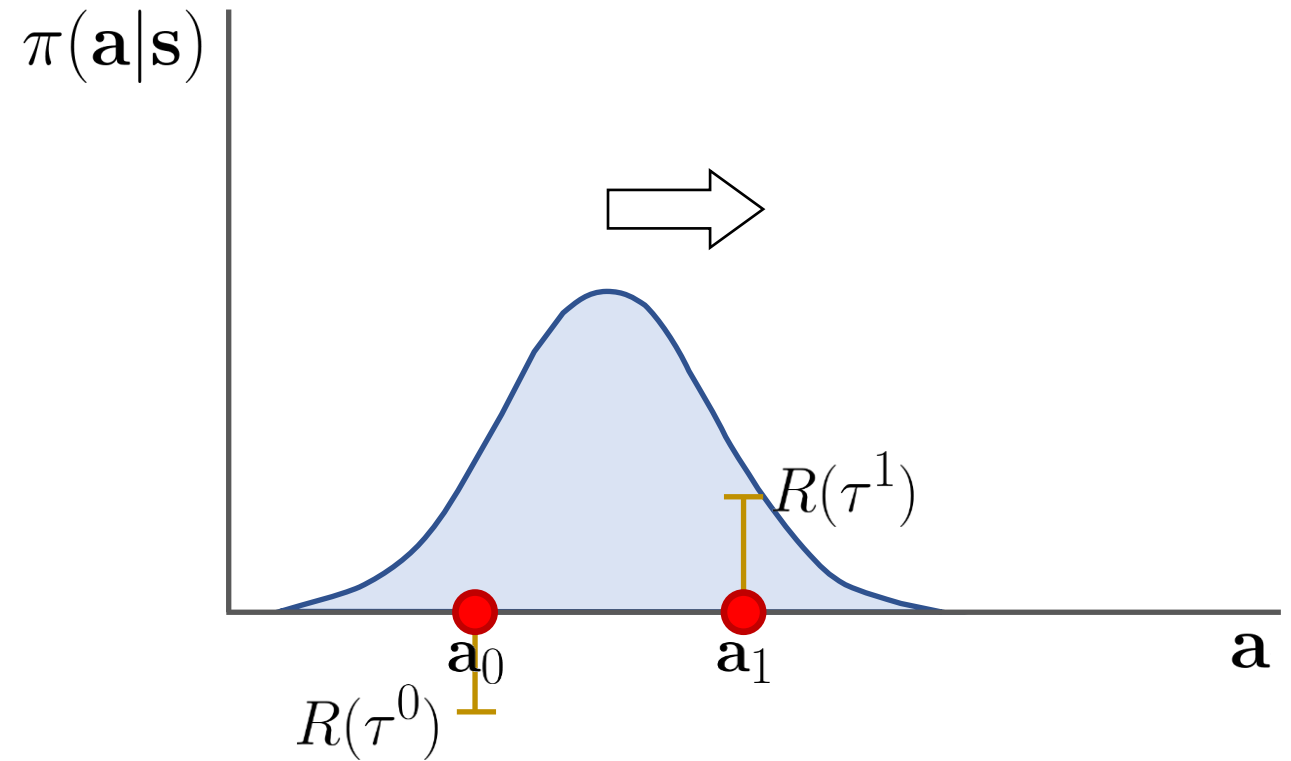


Categorical Distribution
(Discrete Actions)

Etc...

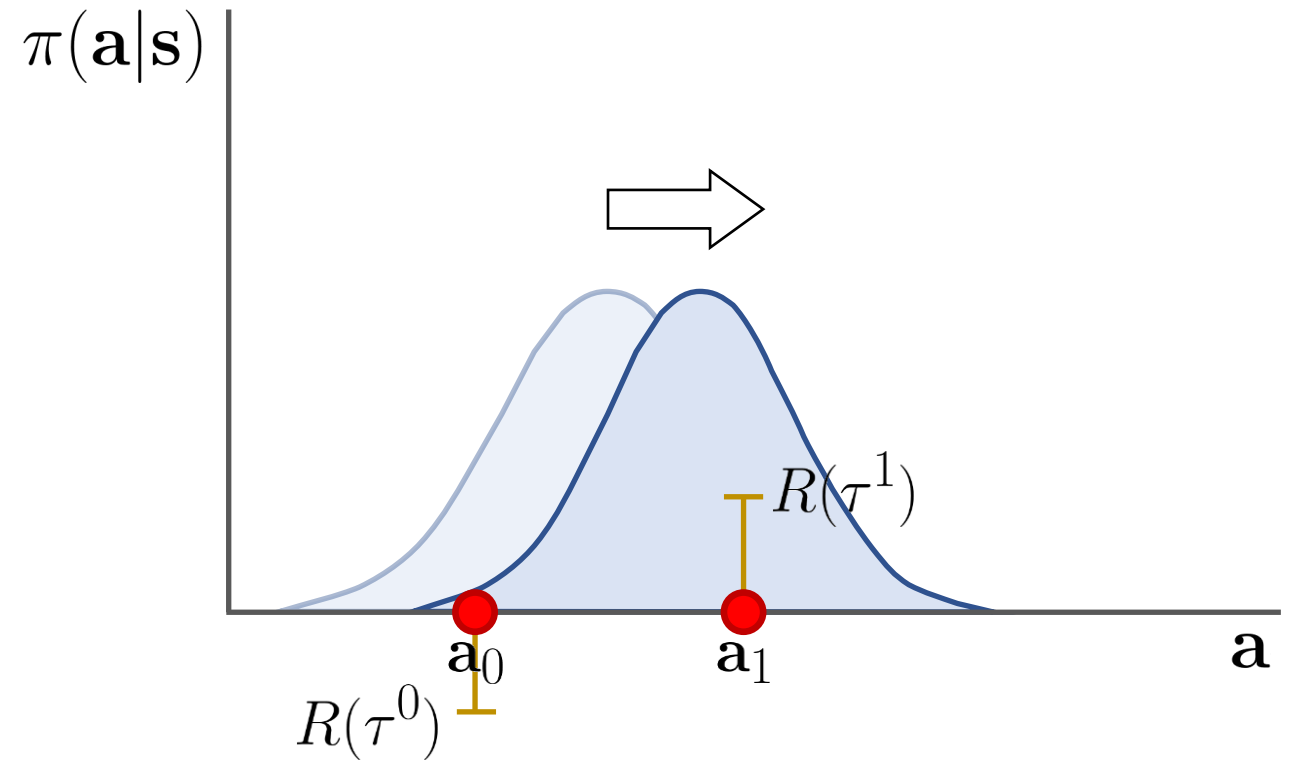
Problems

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



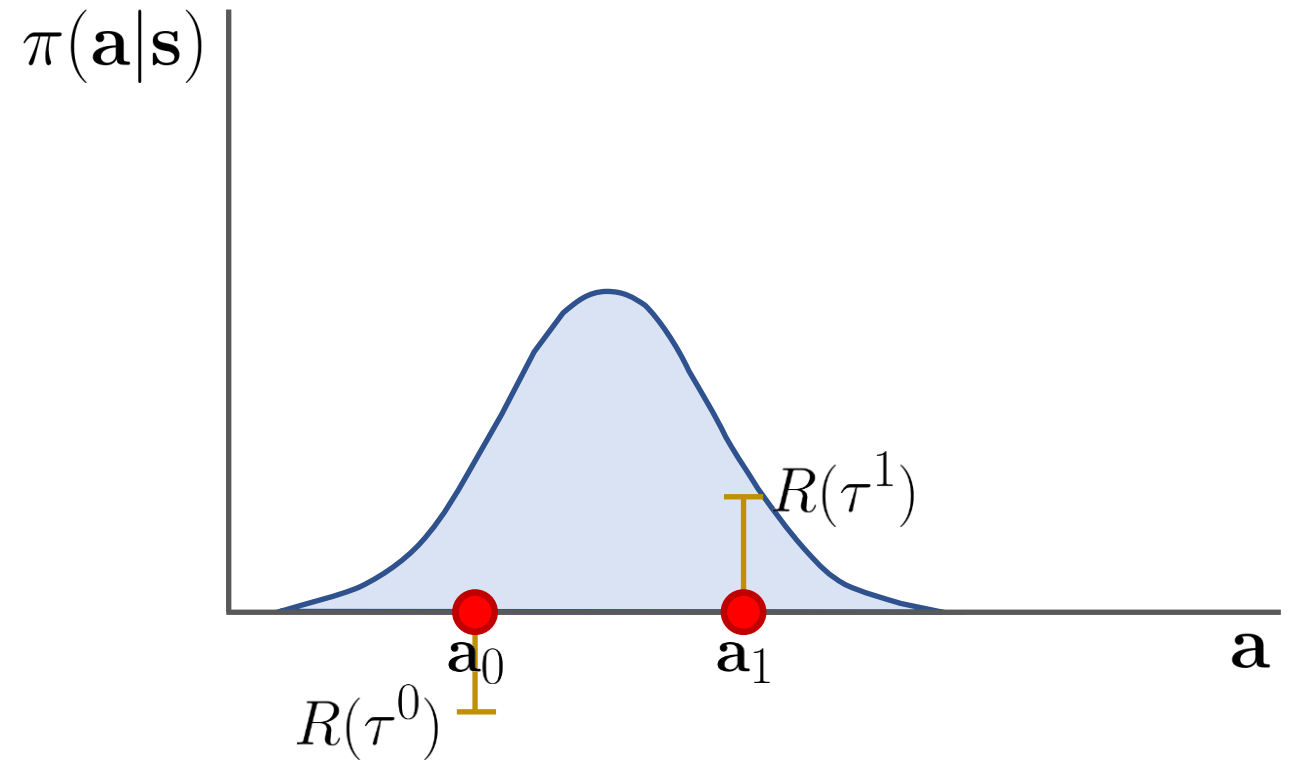
Problems

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



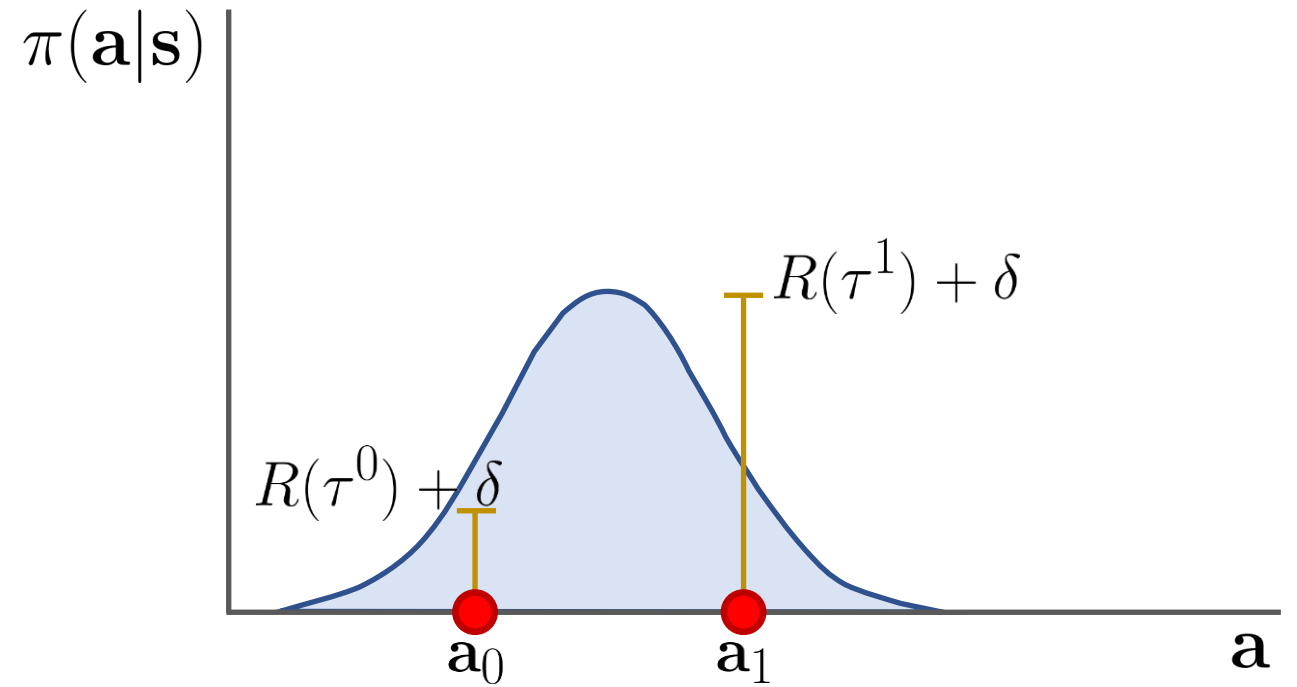
Problems

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



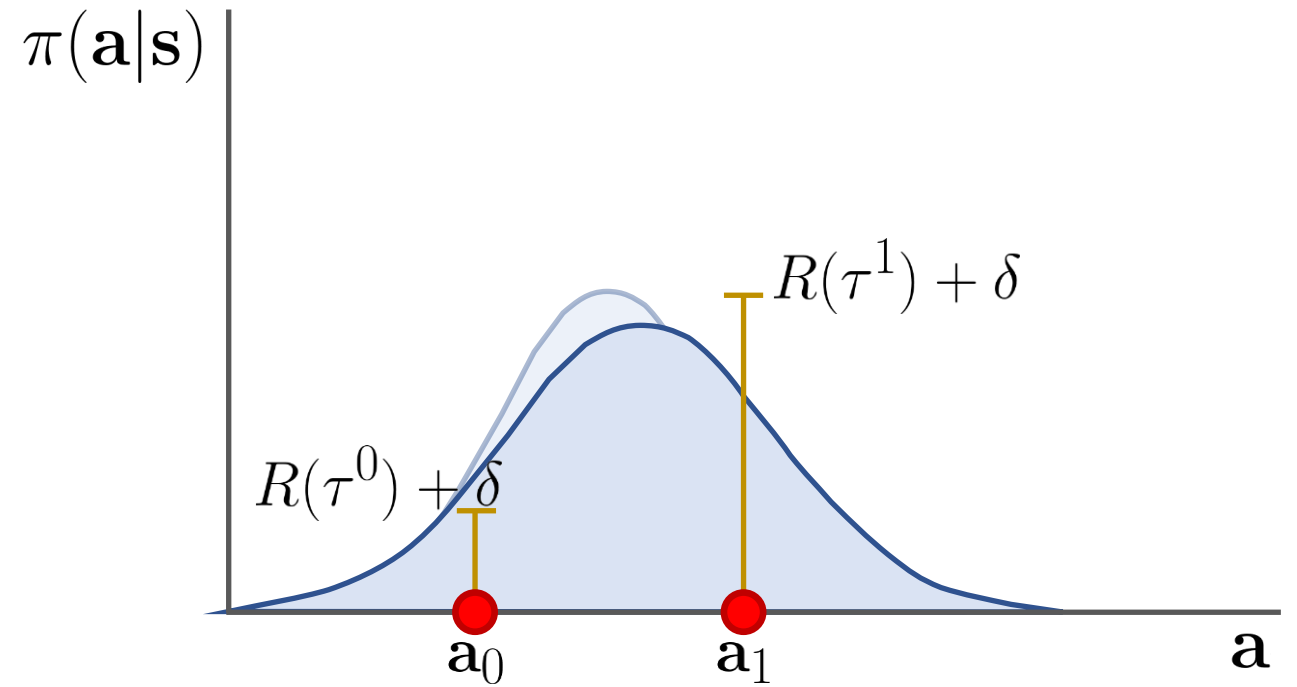
Problems

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



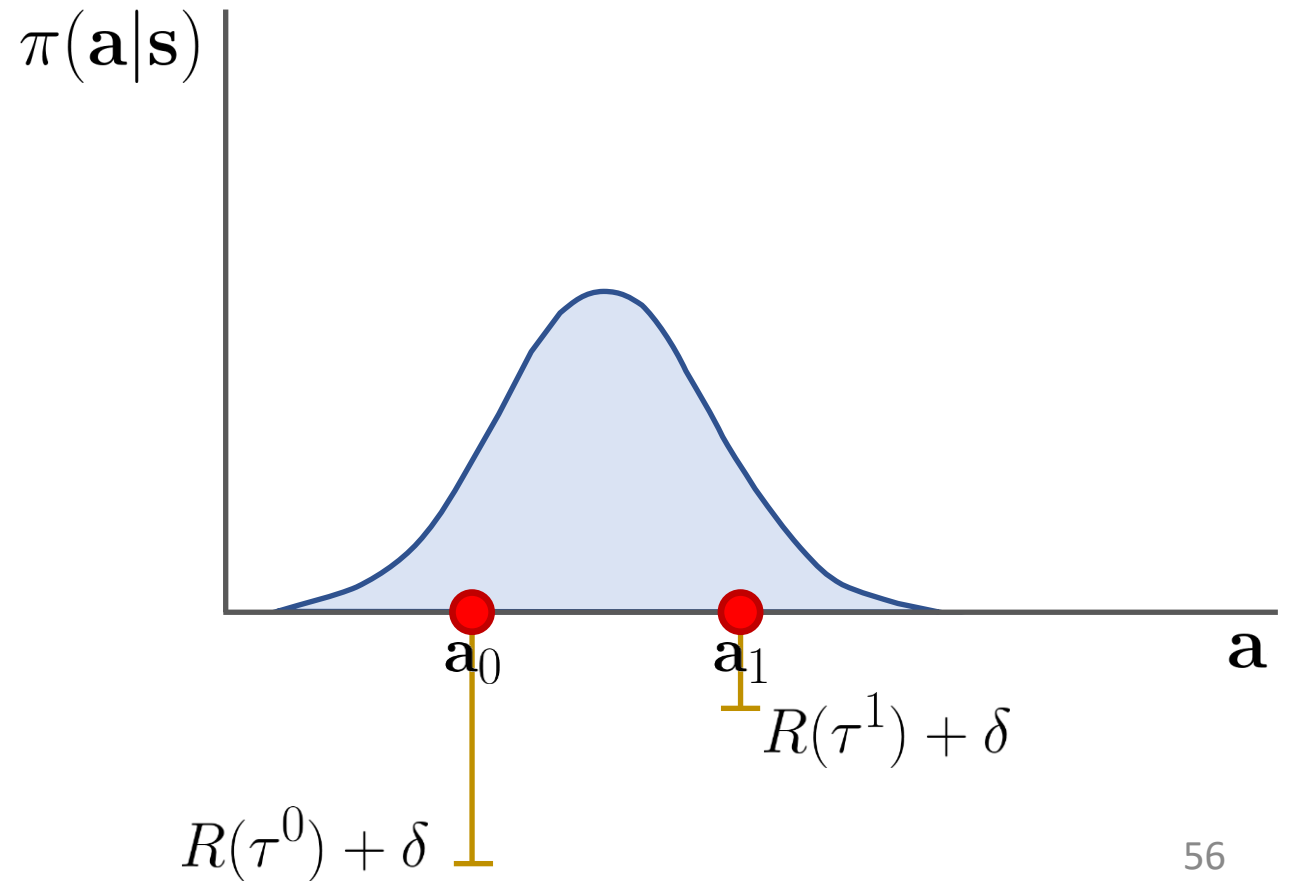
Problems

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



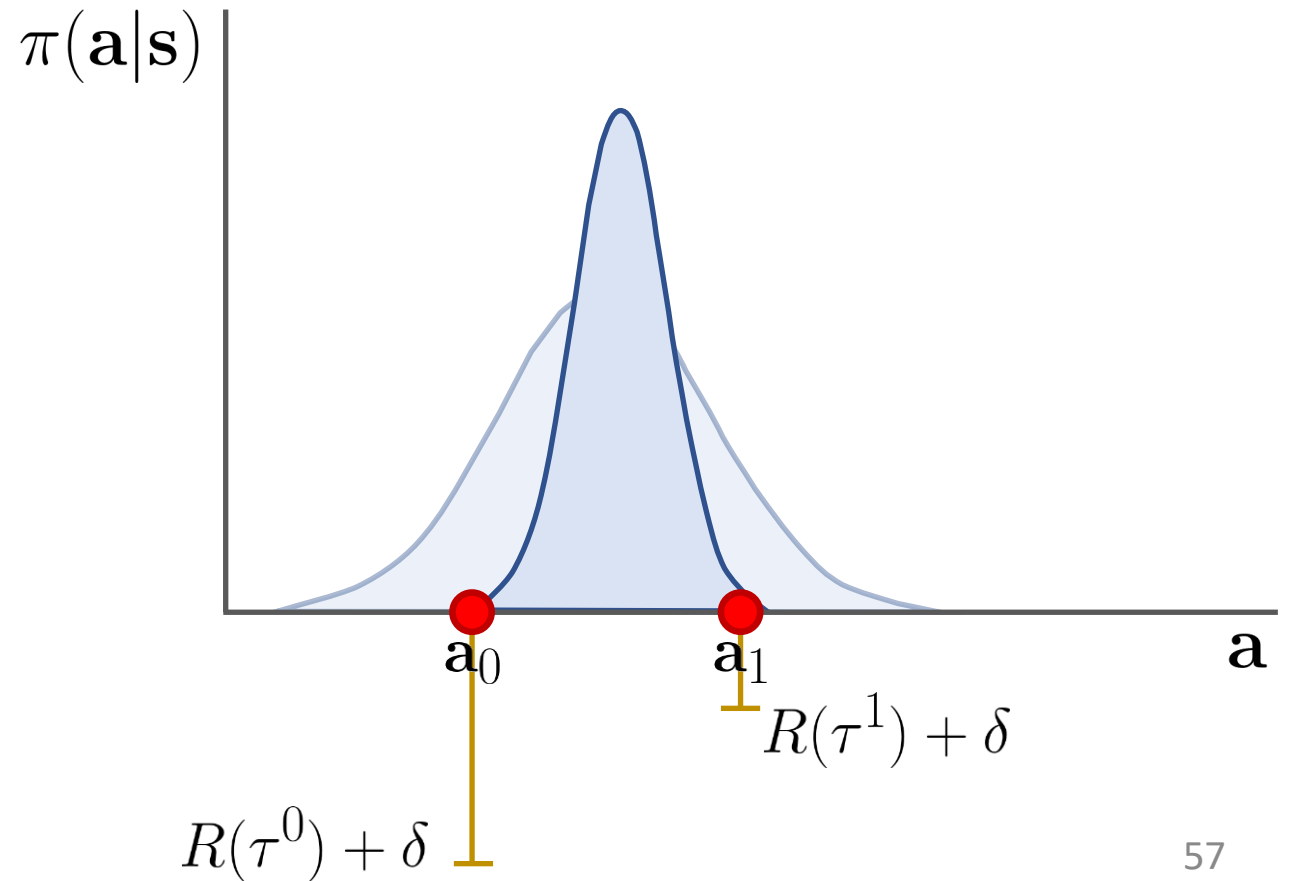
Problems

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



Problems

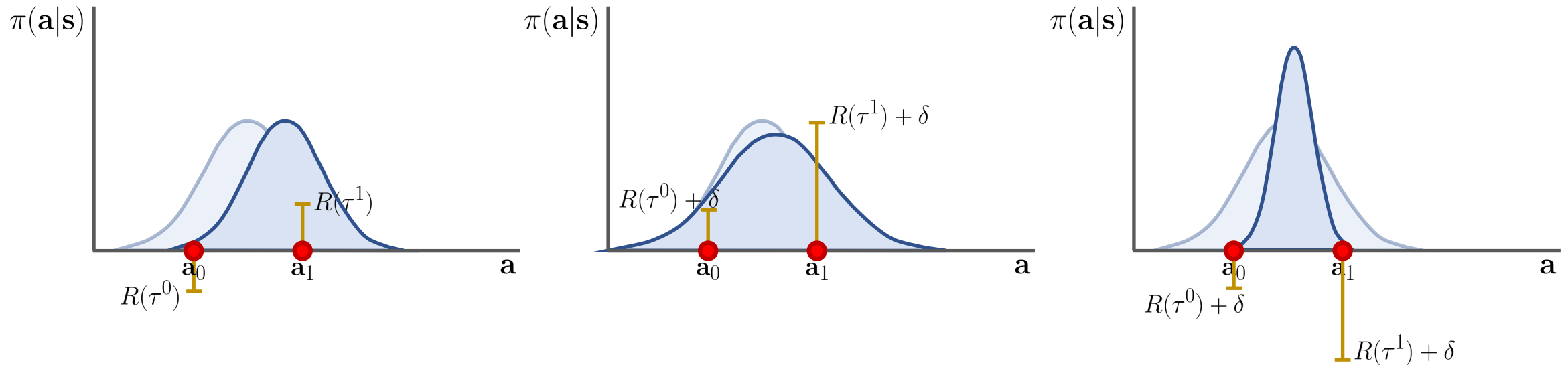
$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



Problems

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

Problem: Not invariant to reward translations



Reward Translation

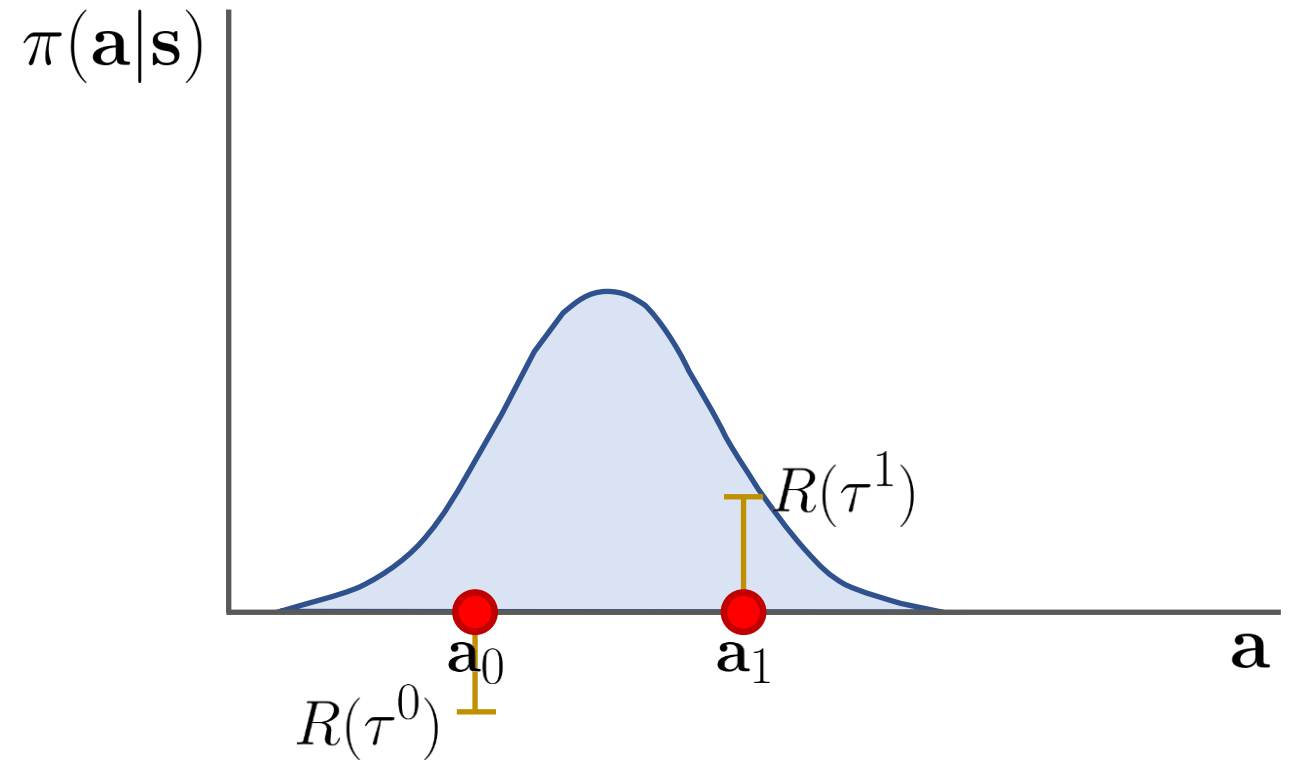
- Optimal policy is invariant to reward translation
- Gradient estimator is *not* invariant to reward translation
- Problem: Variance
 - Monte-Carlo estimate with finite samples
 - Goes away in expectation with infinite samples

Variance Reduction

- Baselines
- Causality
- Bootstrapping

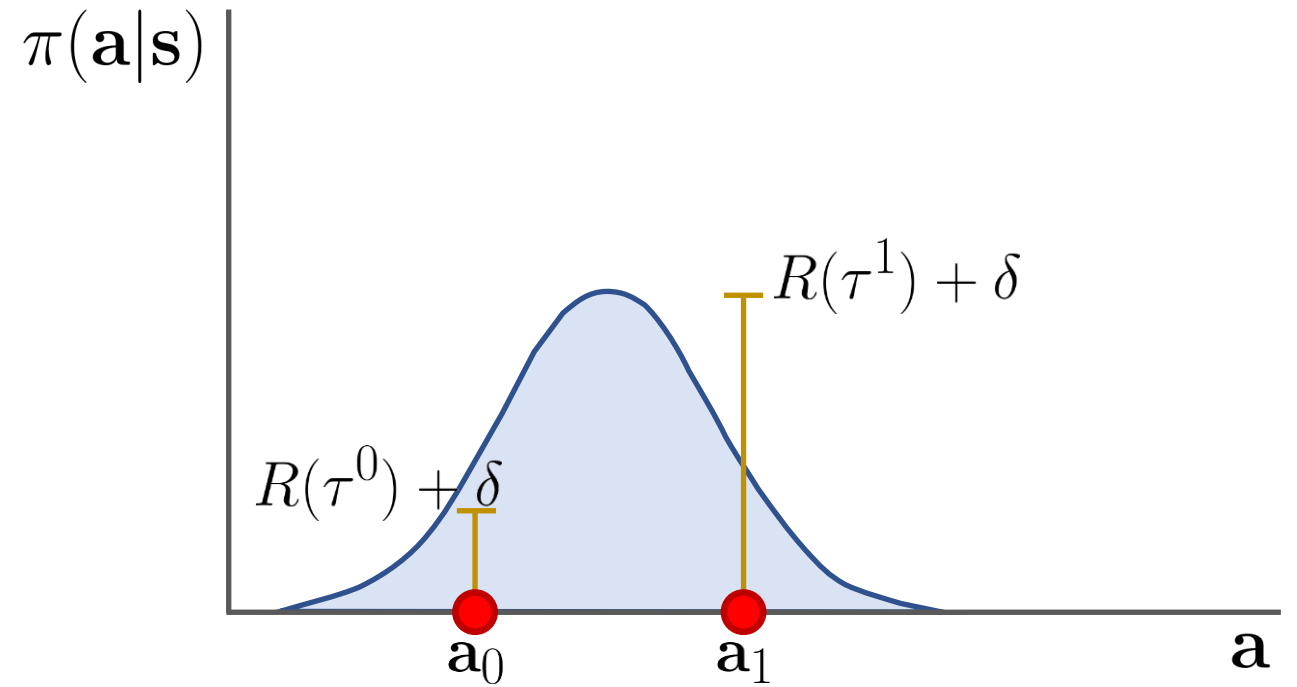
Variance Reduction: Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



Variance Reduction: Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$



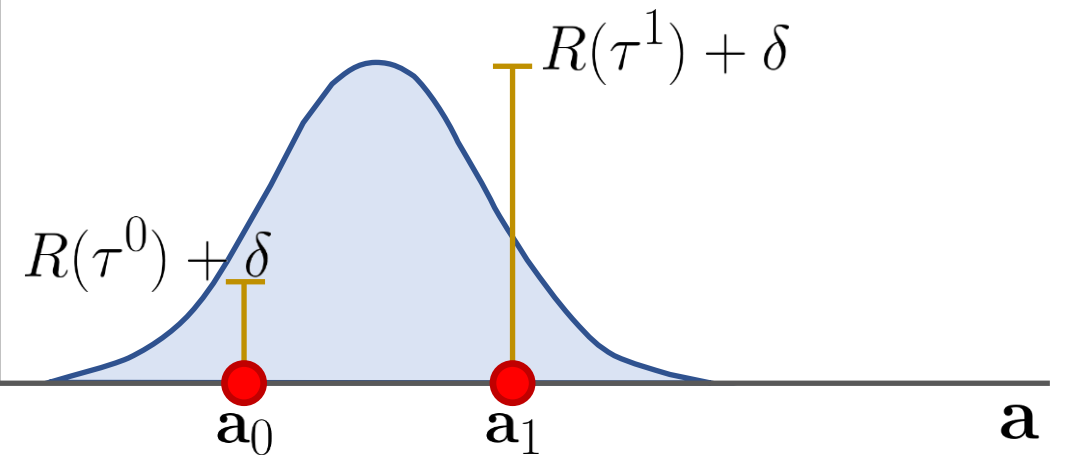
Variance Reduction: Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(R(\tau) - \underline{b}) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

baseline

e.g. $b = \delta$

$\pi(\mathbf{a}|\mathbf{s})$



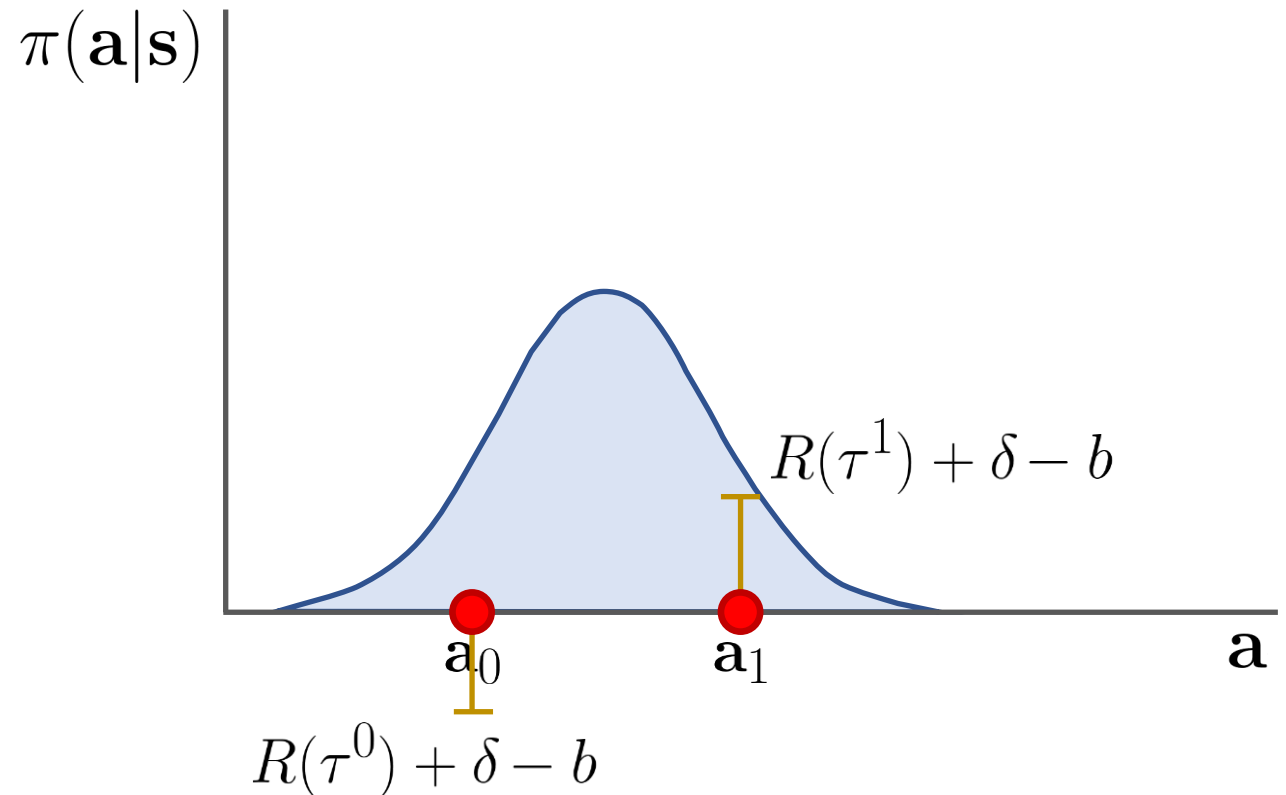
Variance Reduction: Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(R(\tau) - \underline{b}) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

baseline

e.g. $b = \delta$

- Baseline reduces variance
- Is this allowed?
- What is the optimal baseline?



Variance Reduction: Baseline

- How does the baseline effect the gradient?

$$R(\tau) \Longrightarrow \hat{R}(\tau) = R(\tau) - b$$

$$\nabla_{\pi} \hat{J}(\pi) = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [\hat{R}(\tau)] = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) - b]$$

$$= \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

$$= \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) \underbrace{\nabla_{\pi} \log p(\tau|\pi)}_{\text{score function}}] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

score function

Variance Reduction: Baseline

- How does the baseline effect the gradient?

$$R(\tau) \Longrightarrow \hat{R}(\tau) = R(\tau) - b$$

$$\nabla_{\pi} \hat{J}(\pi) = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [\hat{R}(\tau)] = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) - b]$$

$$= \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

$$= \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) \nabla_{\pi} \log p(\tau|\pi)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

Variance Reduction: Baseline

- How does the baseline effect the gradient?

$$R(\tau) \Longrightarrow \hat{R}(\tau) = R(\tau) - b$$

$$\nabla_{\pi} \hat{J}(\pi) = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [\hat{R}(\tau)] = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) - b]$$

$$= \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

$$= \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) \nabla_{\pi} \log p(\tau|\pi)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

$$\nabla_{\pi} b = 0$$

Variance Reduction: Baseline

- How does the baseline effect the gradient?

$$R(\tau) \implies \hat{R}(\tau) = R(\tau) - b$$

$$\nabla_{\pi} \hat{J}(\pi) = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [\hat{R}(\tau)] = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) - b]$$

$$= \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

$$= \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) \nabla_{\pi} \log p(\tau|\pi)] - \cancel{\nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]} 0$$

Variance Reduction: Baseline

- How does the baseline effect the gradient?

$$R(\tau) \Longrightarrow \hat{R}(\tau) = R(\tau) - b$$

$$\nabla_{\pi} \hat{J}(\pi) = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [\hat{R}(\tau)] = \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) - b]$$

$$= \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

$$= \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) \nabla_{\pi} \log p(\tau|\pi)] - \nabla_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} [b]$$

$$= \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) \nabla_{\pi} \log p(\tau|\pi)]$$

$$= \nabla_{\pi} J(\pi)$$

Variance Reduction: Baseline

- How does the baseline effect the gradient?

$$R(\tau) \Longrightarrow \hat{R}(\tau) = R(\tau) - b$$

- Baseline does not change the gradient!
- Reduces variance without introducing bias
- Any constant value for the baseline will preserve policy gradient

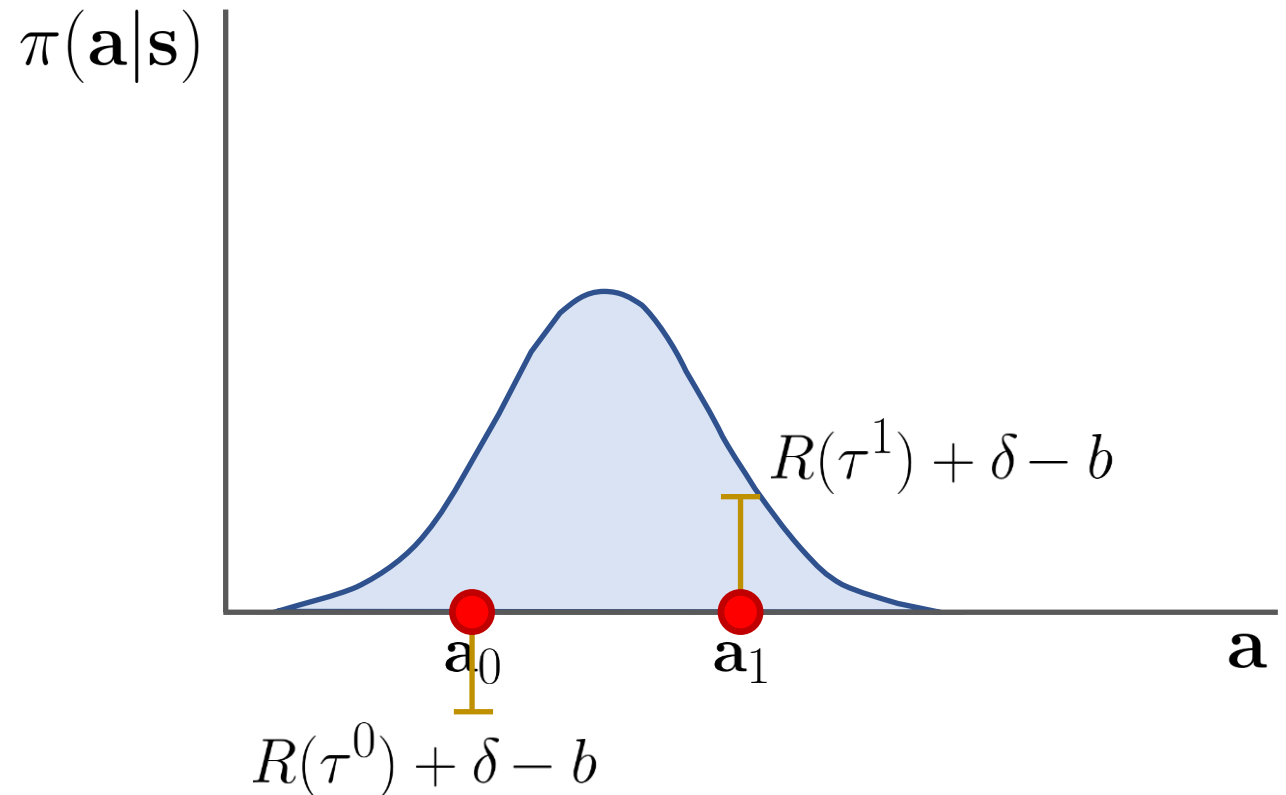
Variance Reduction: Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(R(\tau) - \underline{b}) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

baseline

e.g. $b = \delta$

- Baseline reduces variance
- Is this allowed?
- What is the optimal baseline?



Optimal Baseline

- Minimize variance of gradient estimator

$$\text{Var} [x] = \mathbb{E}[x^2] - (\mathbb{E} [x])^2$$

$$\begin{aligned}\text{Var} [\nabla_{\pi} J(\pi)] &= \text{Var} [(R(\tau) - b) \nabla_{\pi} \log p(\tau|\pi)] \\ &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[((R(\tau) - b) \nabla_{\pi} \log p(\tau|\pi))^2 \right] - \underbrace{\left(\mathbb{E}_{\tau \sim p(\tau|\pi)} [(R(\tau) - b) \nabla_{\pi} \log p(\tau|\pi)] \right)^2}_{\begin{aligned} &= \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau) \nabla_{\pi} \log p(\tau|\pi)] \\ &= \nabla_{\pi} J(\pi) \\ &\text{independent of baseline} \end{aligned}}\end{aligned}$$

Optimal Baseline

- Minimize variance of gradient estimator

$$\text{Var} [x] = \mathbb{E}[x^2] - (\mathbb{E} [x])^2$$

$$\begin{aligned} \text{Var} [\nabla_{\pi} J(\pi)] &= \text{Var} [(R(\tau) - b) \nabla_{\pi} \log p(\tau|\pi)] \\ &= \underline{\mathbb{E}_{\tau \sim p(\tau|\pi)} \left[((R(\tau) - b) \nabla_{\pi} \log p(\tau|\pi))^2 \right]} - \left(\mathbb{E}_{\tau \sim p(\tau|\pi)} [(R(\tau) - b) \nabla_{\pi} \log p(\tau|\pi)] \right)^2 \end{aligned}$$

Optimal Baseline

$$\begin{aligned}\frac{d\text{Var}}{db} &= \frac{d}{db} \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(R(\tau) - b)^2 (\nabla_{\pi} \log p(\tau|\pi))^2 \right] = 0 \\ &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[-2 (R(\tau) - b) (\nabla_{\pi} \log p(\tau|\pi))^2 \right] \\ &= -2 \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) (\nabla_{\pi} \log p(\tau|\pi))^2 \right] + 2b \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(\nabla_{\pi} \log p(\tau|\pi))^2 \right]\end{aligned}$$

$$2b \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(\nabla_{\pi} \log p(\tau|\pi))^2 \right] = 2 \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) (\nabla_{\pi} \log p(\tau|\pi))^2 \right]$$

$$b = \frac{\mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) (\nabla_{\pi} \log p(\tau|\pi))^2 \right]}{\mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(\nabla_{\pi} \log p(\tau|\pi))^2 \right]} \xrightarrow{w(\tau) = (\nabla_{\pi} \log p(\tau|\pi))^2} b = \frac{\mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)w(\tau)]}{\mathbb{E}_{\tau \sim p(\tau|\pi)} [w(\tau)]}$$

Optimal Baseline

$$R(\tau) \Longrightarrow \hat{R}(\tau) = R(\tau) - b$$

- Optimal baseline:

$$b = \frac{\mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)w(\tau)]}{\mathbb{E}_{\tau \sim p(\tau|\pi)} [w(\tau)]} \quad \text{where} \quad w(\tau) = (\nabla_{\pi} \log p(\tau|\pi))^2$$

- In practice:

$$b = \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)]$$

easier to estimate

Optimal Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[(R(\tau) - b) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

where

$$b = \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)]$$

- Interpretation:
 - Increase likelihood of trajectories that do *better* than average
 - Decrease likelihood of trajectories that do *worse* than average

Optimal Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\underbrace{(R(\tau) - b)}_{> 0} \sum_{t=0}^{T-1} \underbrace{\nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t)}_{\text{increase likelihood}} \right]$$

where

$$b = \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)]$$

- Interpretation:
 - Increase likelihood of trajectories that do *better* than average
 - Decrease likelihood of trajectories that do *worse* than average

Optimal Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\underbrace{(R(\tau) - b)}_{< 0} \sum_{t=0}^{T-1} \underbrace{\nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t)}_{\text{decrease likelihood}} \right]$$

where

$$b = \mathbb{E}_{\tau \sim p(\tau|\pi)} [R(\tau)]$$

- Interpretation:
 - Increase likelihood of trajectories that do *better* than average
 - Decrease likelihood of trajectories that do *worse* than average

Policy Gradient

ALGORITHM: Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate baseline $b = \frac{1}{N} R(\tau^i)$
 - 5: Estimate policy gradient
$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i (R(\tau^i) - b) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$$
 - 6: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 7: **end while**
 - 8: return policy π_θ
-

Policy Gradient

ALGORITHM: Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: **while** not done **do**
 - 3: Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 4: Estimate baseline $b = \frac{1}{N} R(\tau^i)$
 - 5: Estimate policy gradient
 $\nabla_\theta J(\pi_\theta) \approx \frac{1}{N} \sum_i (R(\tau^i) - b) \sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i | \mathbf{s}_t^i)$
 - 6: Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$
 - 7: **end while**
 - 8: return policy π_θ
-

Variance Reduction

- Baselines
- Causality
- Bootstrapping

Variance Reduction: Causality

$$\begin{aligned}\nabla_{\pi} J(\pi) &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\underline{(R(\tau) - b)} \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right] \\ &= \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\underline{\left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right)} \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]\end{aligned}$$

rewards across all timesteps

Variance Reduction: Causality

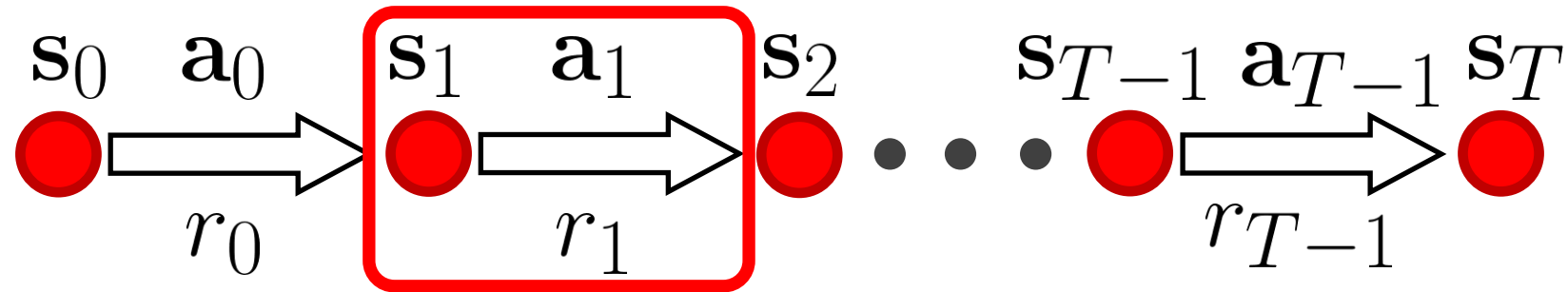
- Gradient at single timestep t

$$\left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t)$$

sum rewards across all timesteps

- Current action *does not* affect past rewards
- $r_{t'}$ is independent of \mathbf{a}_t for all $t' < t$

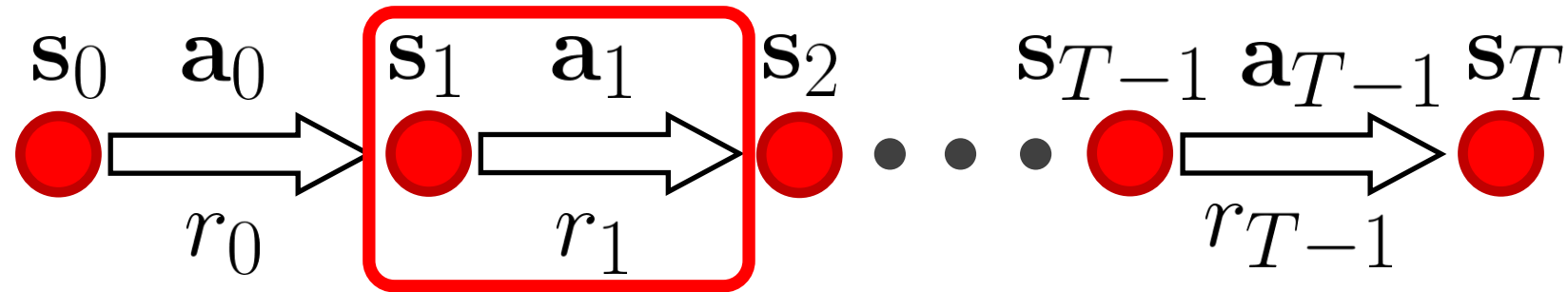
Variance Reduction: Causality



$$(\underline{r_0} + r_1 + r_2 + \dots + r_{T-1}) \nabla_{\pi} \log \pi(\mathbf{a}_1 | \mathbf{s}_1)$$

Not affected by \mathbf{a}_1

Variance Reduction: Causality



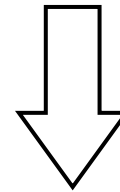
$$(r_1 + r_2 + \dots + r_{T-1}) \nabla_{\pi} \log \pi(\mathbf{a}_1 | \mathbf{s}_1)$$

Generally:

$$(r_t + r_{t+1} + \dots + r_{T-1}) \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t)$$

Variance Reduction: Causality

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \left(\sum_{\underline{t'=0}}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$



$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \left(\sum_{\underline{t'=t}}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

“reward-to-go”

fewer reward terms → lower variance

Variance Reduction: Causality

- Trajectory-based estimator:

$$\nabla_{\pi} J(\pi) = \underline{\mathbb{E}_{\tau \sim p(\tau|\pi)}} \left[\sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \left(\underline{\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b} \right) \right]$$

- Reward-to-Go estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

Variance Reduction: Causality

- Trajectory-based estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

- Reward-to-Go estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

treat every state as start of
a new trajectory

Variance Reduction: Causality

- Trajectory-based estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

- Reward-to-Go estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

“discounted” state distribution
of the policy π

sum future rewards


Discounted State Distribution

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$


“discounted” state distribution

$$d_{\pi}(\mathbf{s}) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t p(\mathbf{s}_t = \mathbf{s} | \pi)$$

Discounted State Distribution

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$


“discounted” state distribution

$$d_{\pi}(\mathbf{s}) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(\mathbf{s}_t = \mathbf{s} | \pi)$$


probability of being in \mathbf{S} after
following π for t timesteps

Discounted State Distribution

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\pi}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

“discounted” state distribution

$$d_{\pi}(\mathbf{s}) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \underline{p(\mathbf{s}_t = \mathbf{s}|\pi)}$$

Discounted State Distribution

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\pi}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

“discounted” state distribution

$$d_{\pi}(\mathbf{s}) = \underline{(1 - \gamma)} \sum_{t=0}^{\infty} \gamma^t p(\mathbf{s}_t = \mathbf{s} | \pi)$$

Discounted State Distribution

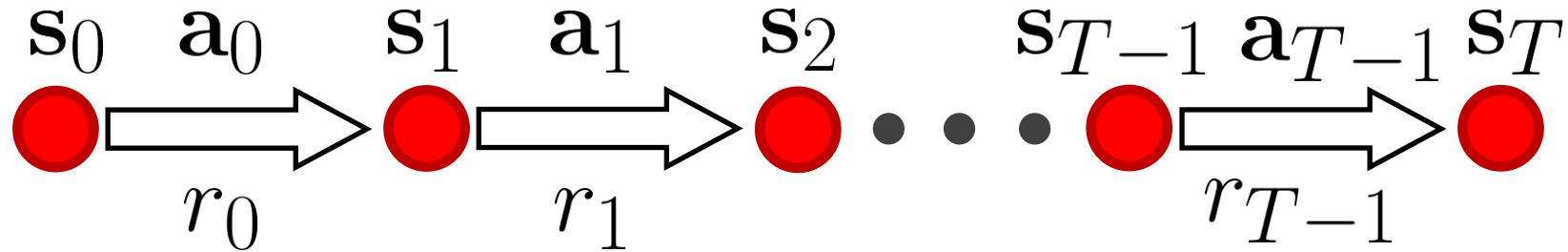
$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\pi}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$

“discounted” state distribution

$$d_{\pi}(\mathbf{s}) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t p(\mathbf{s}_t = \mathbf{s} | \pi) \quad \Longrightarrow \quad p(\mathbf{s} | \pi)$$

In practice, just use the marginal
state distribution instead

Reward-to-Go Gradient Estimator



$$\left. \begin{aligned} \nabla_0 &= \left(r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{T-1} r_{T-1} \right) \nabla_{\pi} \log \pi(\mathbf{a}_0 | \mathbf{s}_0) \\ \nabla_1 &= \left(r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{T-2} r_{T-1} \right) \nabla_{\pi} \log \pi(\mathbf{a}_1 | \mathbf{s}_1) \\ &\vdots \\ \nabla_{T-1} &= (r_{T-1}) \nabla_{\pi} \log \pi(\mathbf{a}_{T-1} | \mathbf{s}_{T-1}) \end{aligned} \right\} \begin{array}{l} \text{average grads} \\ \approx \nabla_{\pi} J(\pi) \end{array}$$

State-Based Baseline

- Reward-to-Go estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - \underline{b} \right) \right]$$

Can use a better baseline for
even lower variance

State-Based Baseline

- Reward-to-Go estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$
$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - \underline{V^{\pi}(\mathbf{s})} \right) \right]$$

Value function baseline

State-Based Baseline

- Reward-to-Go estimator:

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - b \right) \right]$$
$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \underbrace{\left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right)}_{\text{“Advantage”}} \right]$$

- Advantage > 0: Action is better than average
- Advantage < 0: Action is worse than average


Value Function Baseline

$$\begin{aligned} & \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right] \\ &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\ & \quad - \underbrace{\mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim \pi(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) V^{\pi}(\mathbf{s})]} \end{aligned}$$

Value Function Baseline

$$\begin{aligned} & \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right] \\ &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\ &\quad - \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim \pi(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) V^{\pi}(\mathbf{s})] \end{aligned}$$

Value Function Baseline

$$\begin{aligned}
 & \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right] \\
 &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\
 &\quad - \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \underline{V^{\pi}(\mathbf{s})}]
 \end{aligned}$$


Value Function Baseline

$$\begin{aligned} & \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right] \\ &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\ & \quad - \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim \pi(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) V^{\pi}(\mathbf{s})] \\ &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\ & \quad - \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \left[\underline{V^{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s})] \right] \end{aligned}$$

Value Function Baseline

$$\begin{aligned} & \mathbb{E}_{\mathbf{s} \sim d_\pi(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^\pi(\mathbf{s}) \right) \right] \\ &= \mathbb{E}_{\mathbf{s} \sim d_\pi(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\ &\quad - \mathbb{E}_{\mathbf{s} \sim d_\pi(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim \pi(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) V^\pi(\mathbf{s})] \\ &= \mathbb{E}_{\mathbf{s} \sim d_\pi(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\ &\quad - \mathbb{E}_{\mathbf{s} \sim d_\pi(\mathbf{s})} \left[V^\pi(\mathbf{s}) \underbrace{\mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s})]} \right] \\ &\quad = \nabla_{\pi} \sum_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{s}) = \nabla_{\pi} 1 \end{aligned}$$

Value Function Baseline

$$\begin{aligned}
 & \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right] \\
 &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\
 &\quad - \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim \pi(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) V^{\pi}(\mathbf{s})] \\
 &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} \right] \\
 &\quad - \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \left[V^{\pi}(\mathbf{s}) \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s})] \right]
 \end{aligned}$$

Value function baseline is unbiased!

Value Function Baseline

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right]$$

- Value function baseline is unbiased!
 - Substantial variance reduction
 - Any baseline that is only a function of the state is unbiased
- [Sutton et al. 1990]

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_{\theta}) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_{\theta})$
 - 11: **end while**
 - 12: return policy π_{θ}
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_{\theta}) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_{\theta})$
 - 11: **end while**
 - 12: return policy π_{θ}
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - \underline{V(\mathbf{s}_t)} \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters

 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$

 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**

 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**

 - 12: return policy π_θ
-

Reward-to-Go Policy Gradient

ALGORITHM: Reward-to-Go Policy Gradient

- 1: $\theta \leftarrow$ initialize policy parameters
 - 2: $V \leftarrow$ initialize value function parameters
 - 3: **while** not done **do**
 - 4: Sample trajectory τ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$
 - 5: Fit value function $V(\mathbf{s})$
 - 6: **for** every timestep t **do**
 - 7: $\nabla_t \leftarrow \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - V(\mathbf{s}_t) \right) \nabla_{\theta} \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$
 - 8: **end for**
 - 9: $\nabla_{\theta} J(\pi_\theta) \approx \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t$
 - 10: Update policy $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\pi_\theta)$
 - 11: **end while**
 - 12: return policy π_θ
-

Variance Reduction

- Baselines
- Causality
- Bootstrapping

Variance Reduction: Bootstrapping

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right]$$

sum of random variables \rightarrow high variance

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \dots$$

Variance Reduction: Bootstrapping

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right]$$

sum of random variables \rightarrow high variance

n-step return: $r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{k-1} r_{k-1}$

Variance Reduction: Bootstrapping

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \mathbb{E}_{\tau \sim p(\tau|\pi, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a})} \left[\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t'=0}^{T-1} \gamma^{t'} r_{t'} - V^{\pi}(\mathbf{s}) \right) \right]$$

sum of random variables \rightarrow high variance

n-step return: $r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{k-1} r_{k-1} + \underbrace{\gamma^k V^{\pi}(\mathbf{s}_k)}_{\text{bootstrap}}$

N-Step Bootstrapping

1-step bootstrap: $y = r_0 + \gamma \hat{V}^\pi(\mathbf{s}_1)$

2-step bootstrap: $y = r_0 + \gamma r_1 + \gamma^2 \hat{V}^\pi(\mathbf{s}_2)$

3-step bootstrap: $y = r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 \hat{V}^\pi(\mathbf{s}_3)$

•
•
•

n-step bootstrap: $y = \sum_{t=0}^{n-1} \gamma^t r_t + \gamma^n \hat{V}^\pi(\mathbf{s}_n)$

High variance

Biased

$$\text{TD}(\lambda)$$

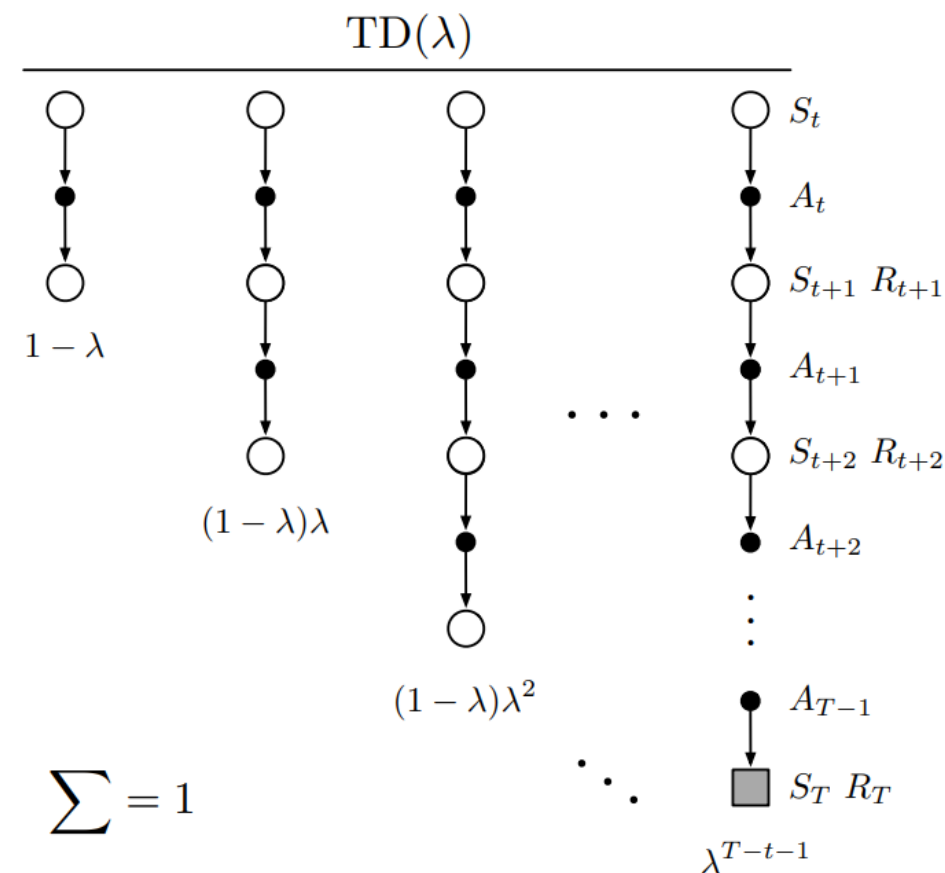
- Use TD(λ) to estimate return

$$\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t=0}^{\infty} \gamma^t r_t - V^{\pi}(\mathbf{s}) \right)$$

↓

$$\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\underline{R^{\lambda}(\mathbf{s}, \mathbf{a})} - V^{\pi}(\mathbf{s}) \right)$$

λ -return



Reinforcement Learning: An Introduction

[Sutton and Barto 1998]

TD(λ)

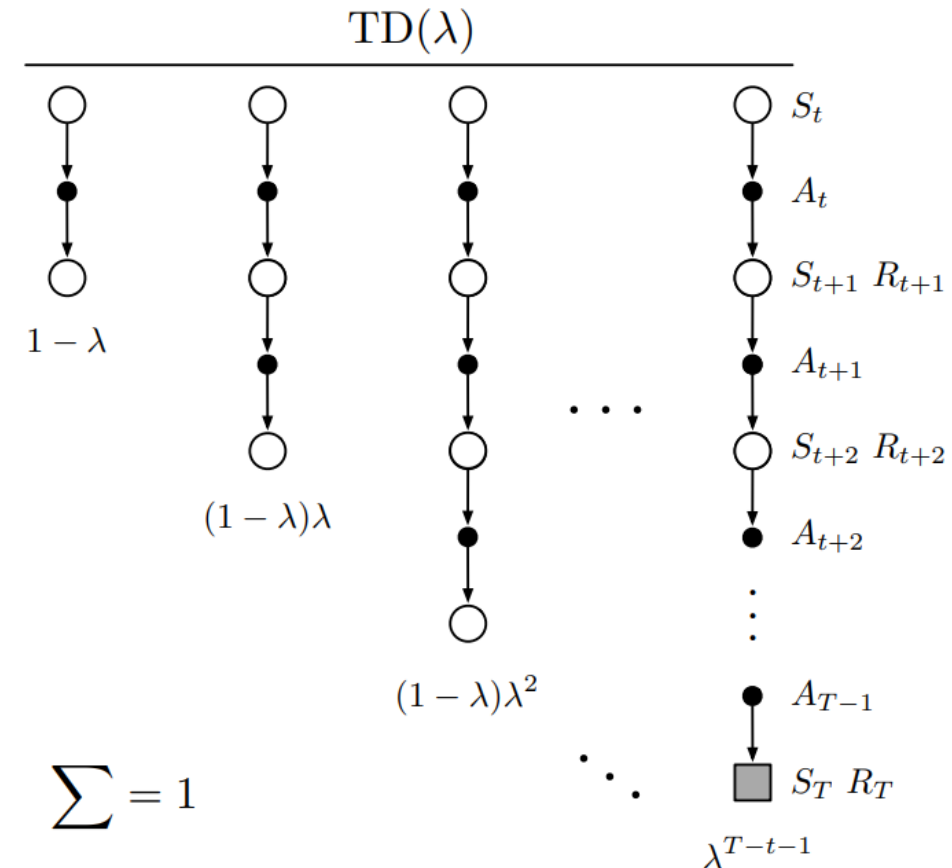
- Use TD(λ) to estimate return

$$\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\sum_{t=0}^{\infty} \gamma^t r_t - V^{\pi}(\mathbf{s}) \right)$$

$$\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \left(\underbrace{R^{\lambda}(\mathbf{s}, \mathbf{a}) - V^{\pi}(\mathbf{s})}_{\text{Generalized Advantage Estimation (GAE)}} \right)$$

Generalized Advantage Estimation (GAE)

High-Dimensional Continuous Control Using
Generalized Advantage Estimation
[Schulman et al. 2016]



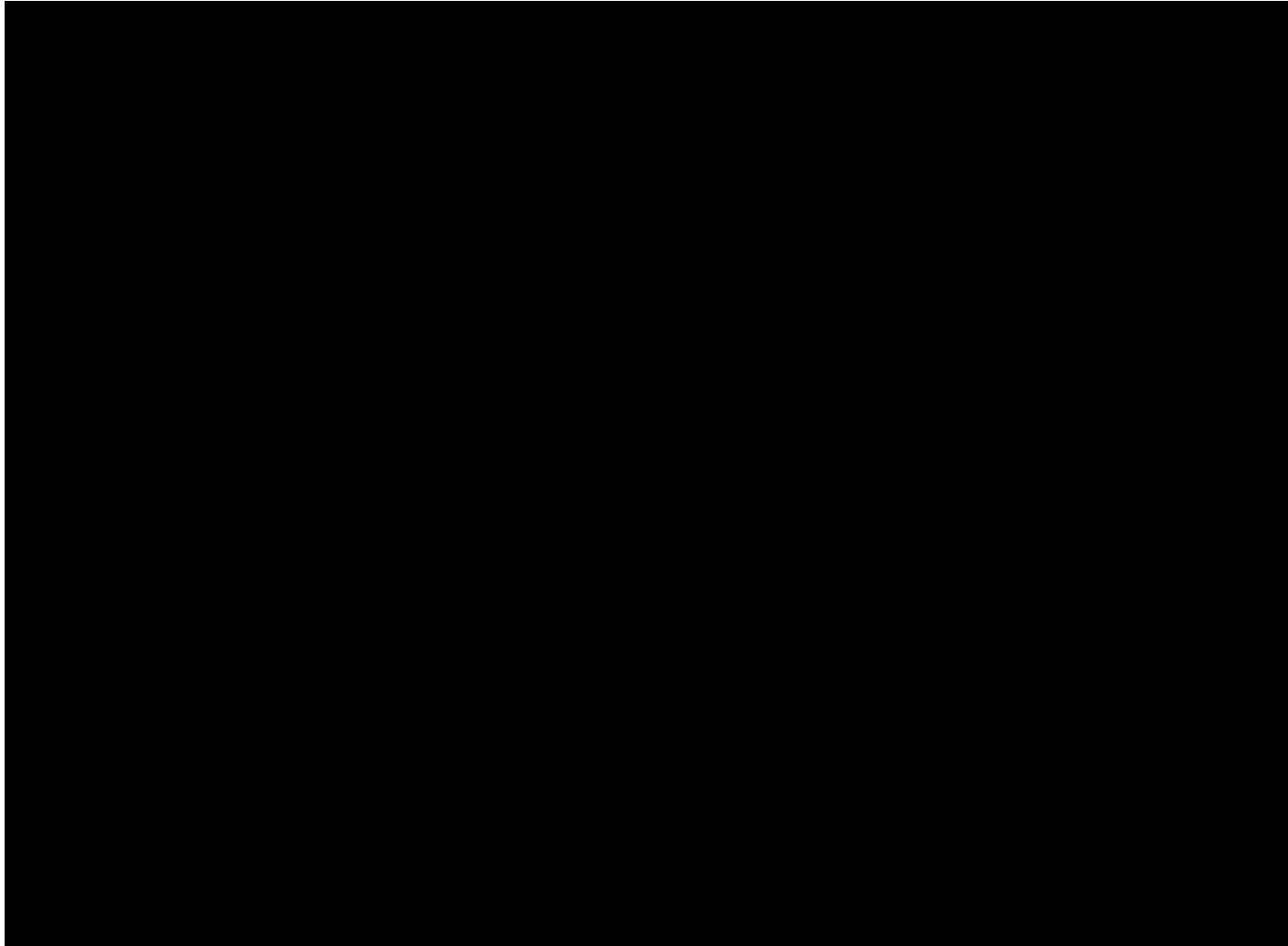
Reinforcement Learning: An Introduction
[Sutton and Barto 1998]

Variance Reduction

- Baselines
- Causality
- Bootstrapping

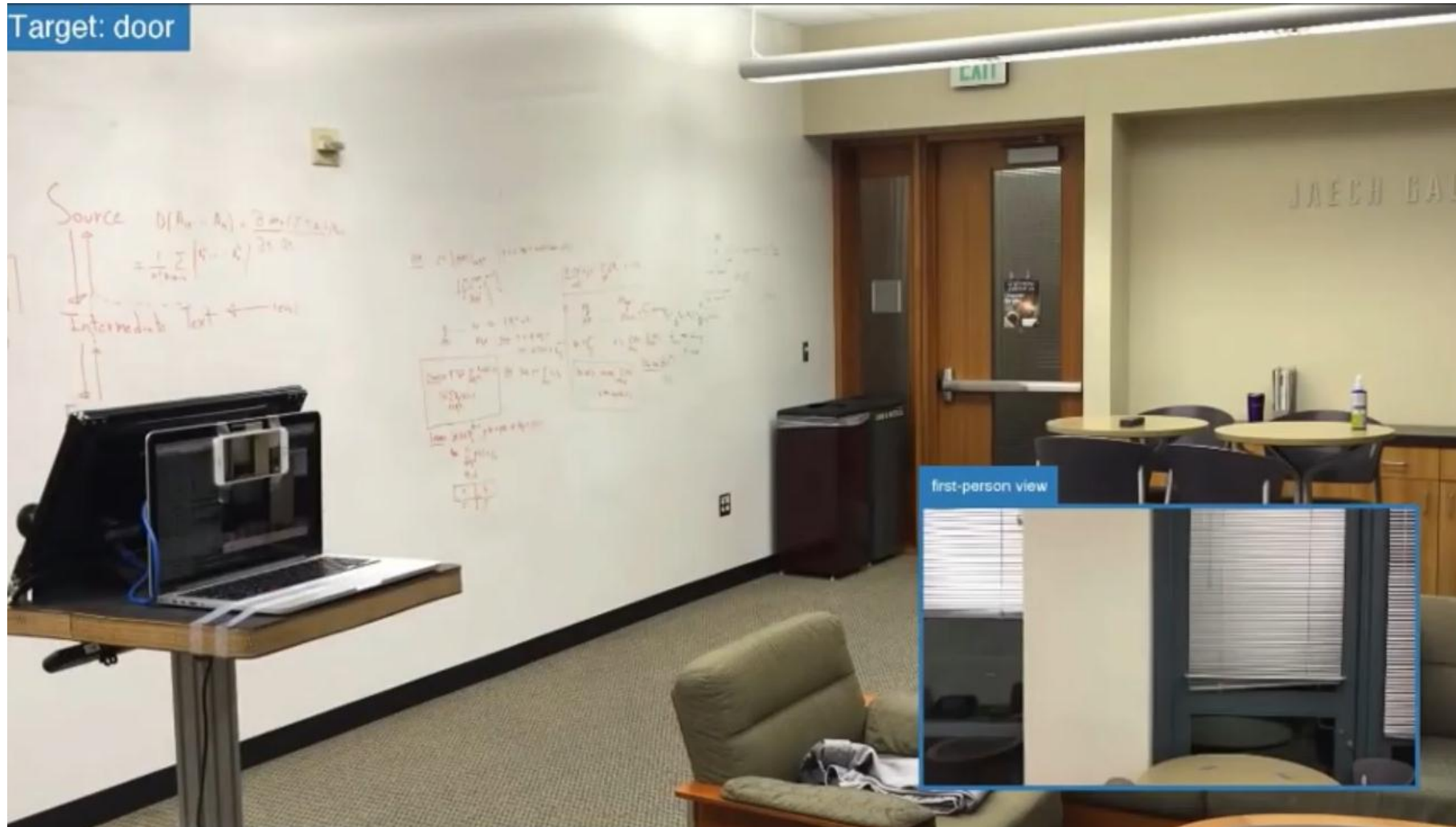
Applications

Visual Navigation



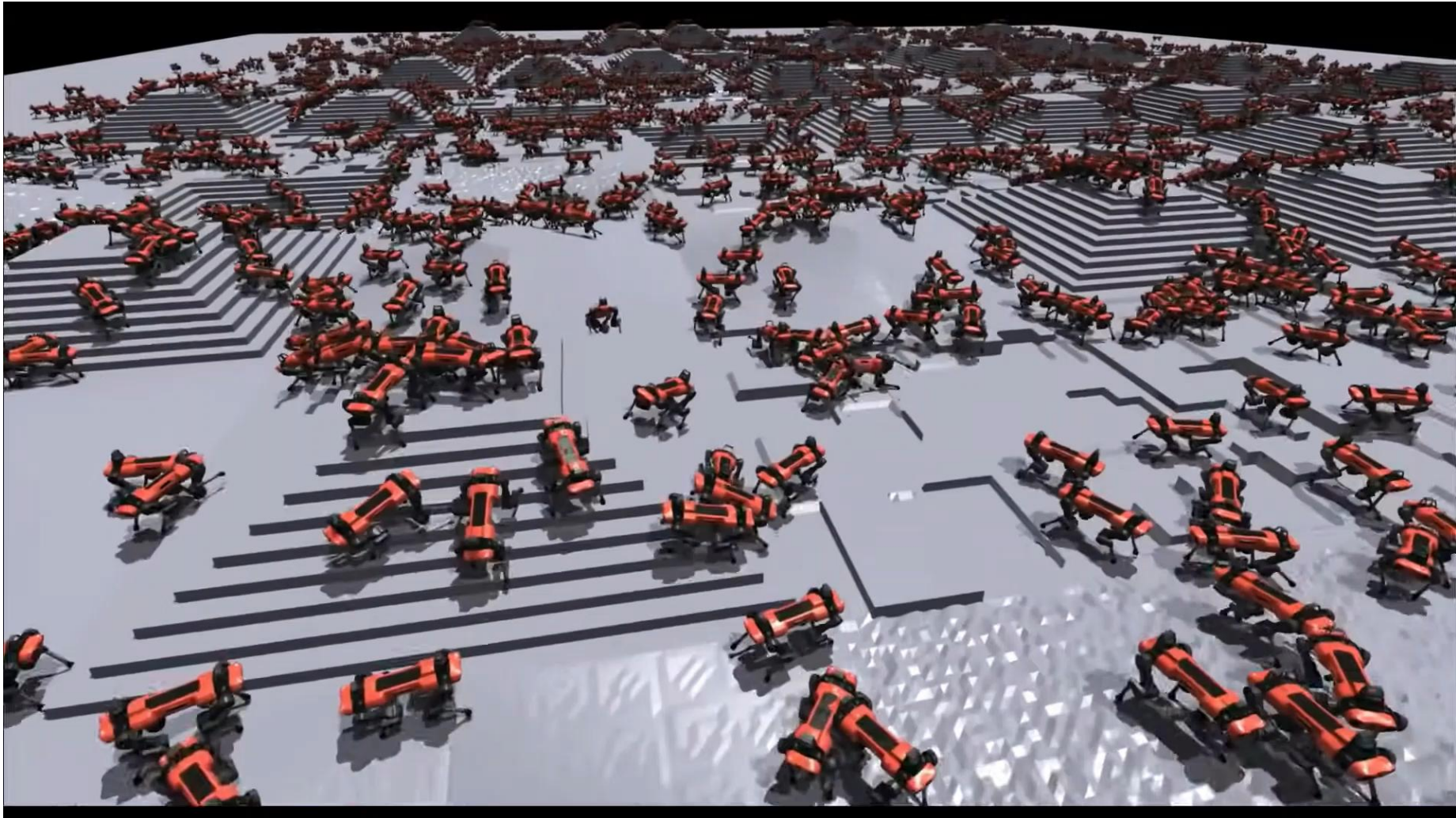
Asynchronous Methods for Deep Reinforcement Learning
[Mnih et al. 2016]

Visual Navigation



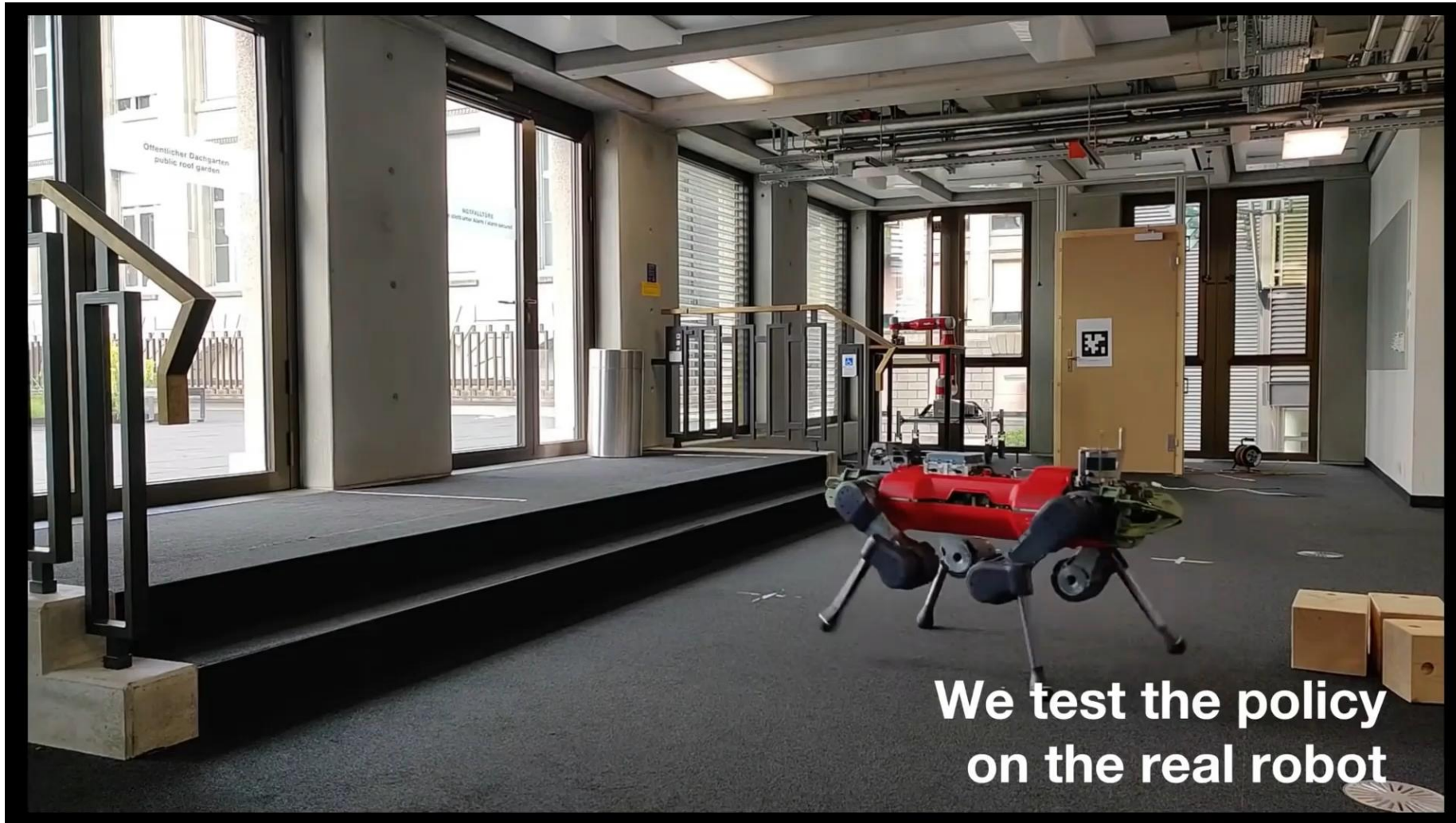
Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning
[Zhu et al. 2017]

Robotic Locomotion



Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning
[Rudin et al. 2022]

Robotic Locomotion



Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning
[Rudin et al. 2022]

Policy Gradient

- ✓ Directly optimize $J(\pi)$ by estimating gradient $\nabla_{\pi} J(\pi)$
- ✓ General: can be applied to continuous and discrete states and actions
- ✗ High-variance gradient estimator \rightarrow unstable/slow convergence
- ✗ Very sample inefficient

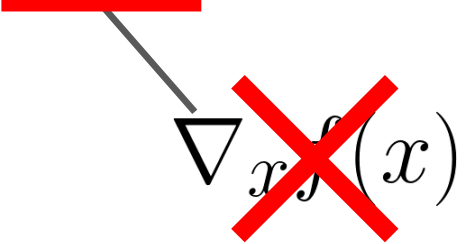
General View of PG

Nondifferentiable Functions

- Why does PG allow us to calculate gradients for a nondifferentiable function?
 - Gradient exists but unknown
 - Gradient does not exist

$$\underline{\nabla_{\pi} J(\pi)} = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_{\pi} \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

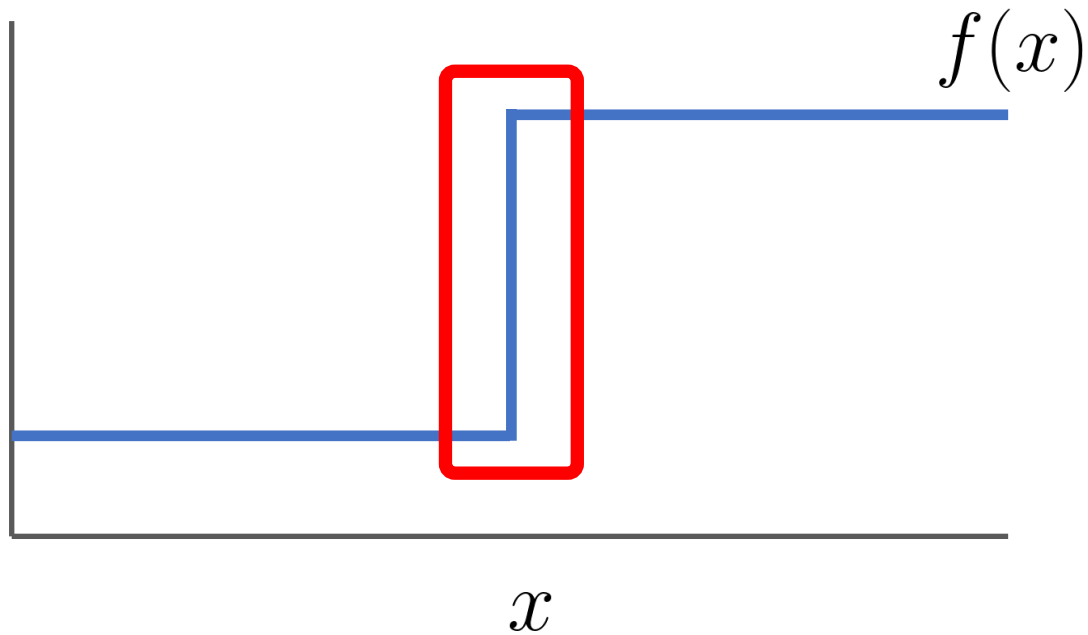
Nondifferentiable Functions

$$\arg \max_x \underline{f(x)}$$

$$\nabla_x f(x)$$

The diagram illustrates that for a nondifferentiable function, the gradient $\nabla_x f(x)$ does not exist. A red underline is placed under $f(x)$ in the maximization expression. A grey arrow points from this underlined $f(x)$ down to the expression $\nabla_x f(x)$, which is then crossed out with a large red 'X'.

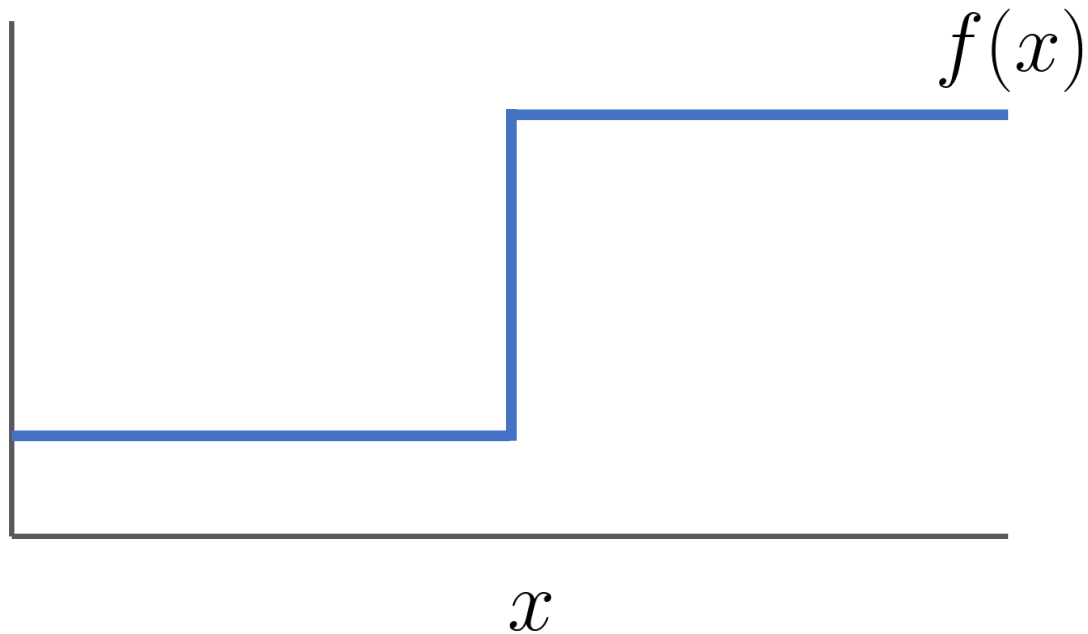
Nondifferentiable Functions

$$\arg \max_x f(x)$$



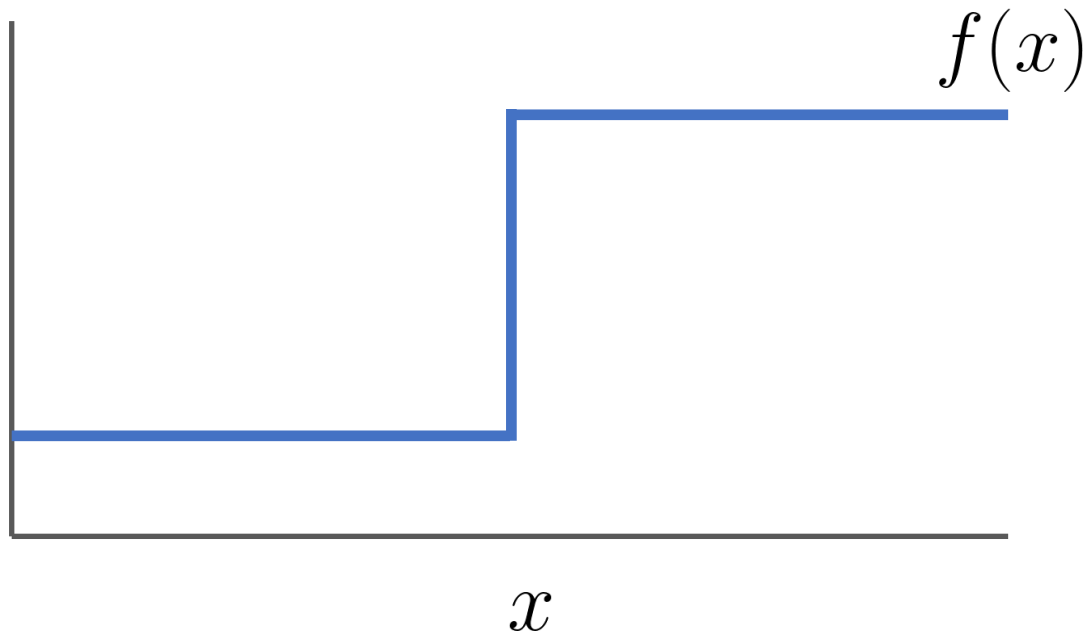
Nondifferentiable Functions

$$\arg \max_x f(x) \longrightarrow \arg \max_{\underline{p}} \mathbb{E}_{x \sim p(x)} [f(x)]$$



Nondifferentiable Functions

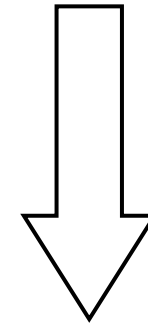
$$\arg \max_x f(x) \longrightarrow \arg \max_p \mathbb{E}_{x \sim \underline{p(x)}} [f(x)]$$



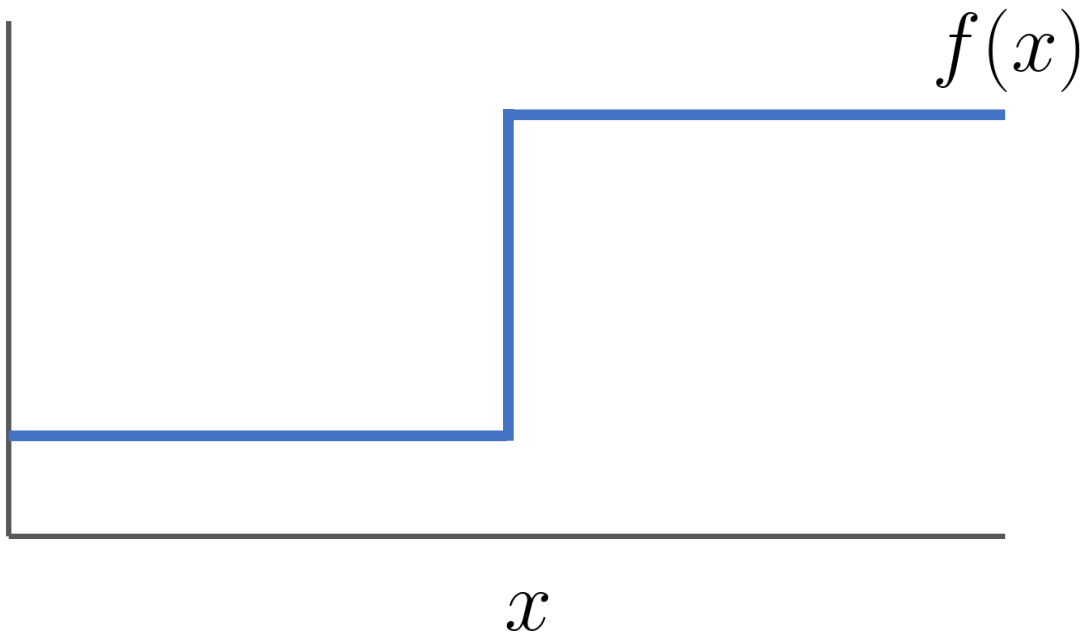
Nondifferentiable Functions

$$\arg \max_x f(x) \longrightarrow \arg \max_p \mathbb{E}_{x \sim p(x)} [f(x)]$$

Score
Function



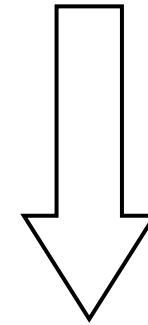
$$\nabla_p = \mathbb{E}_{x \sim p(x)} [\underline{f(x)} \nabla_p \log p(x)]$$



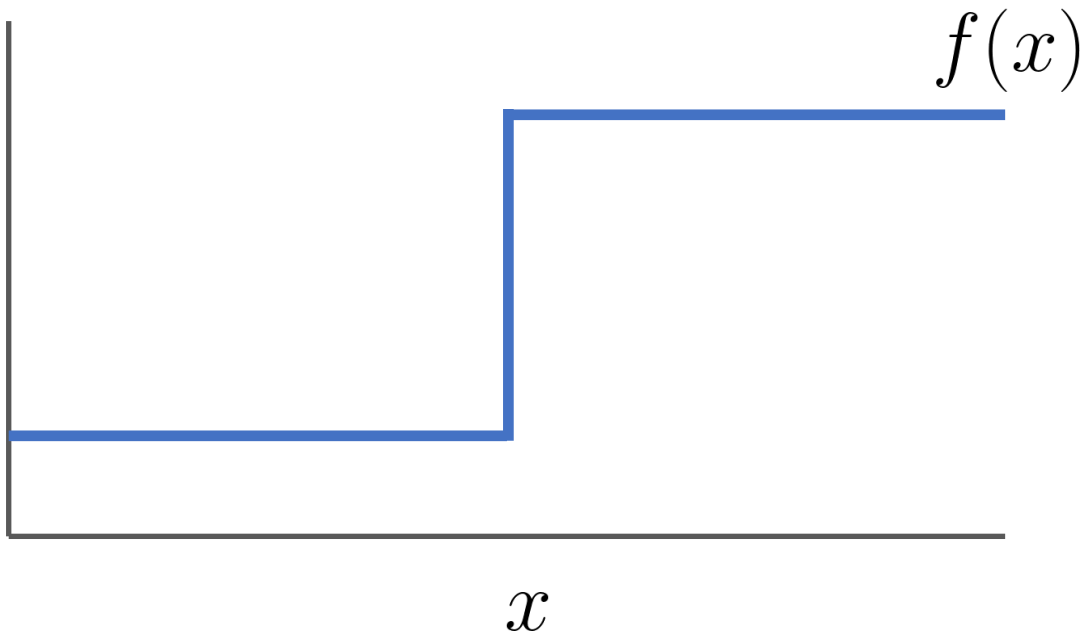
Nondifferentiable Functions

$$\arg \max_x f(x) \longrightarrow \arg \max_p \mathbb{E}_{x \sim p(x)} [f(x)]$$

Score
Function

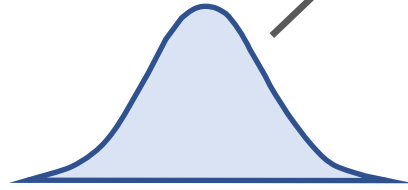


$$\nabla_p = \mathbb{E}_{x \sim p(x)} [f(x) \nabla_p \log p(x)]$$



Nondifferentiable Functions

$$\arg \max_p \mathbb{E}_{x \sim \underline{p(x)}} [f(x)]$$



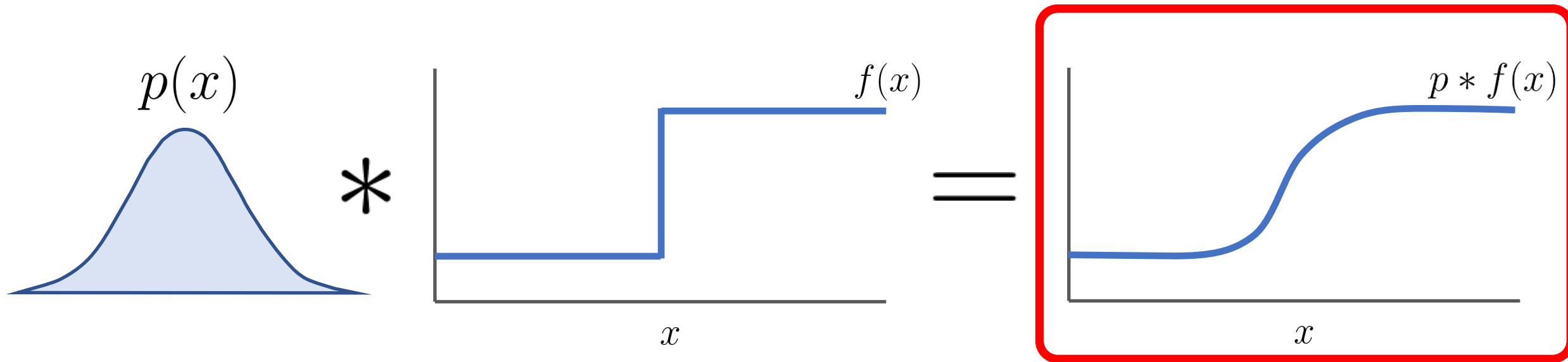
Nondifferentiable Functions

$$\arg \max_p \quad \underline{\mathbb{E}_{x \sim p(x)} [f(x)]}$$
$$= \sum_x \underbrace{p(x) f(x)}$$

This is a convolution!

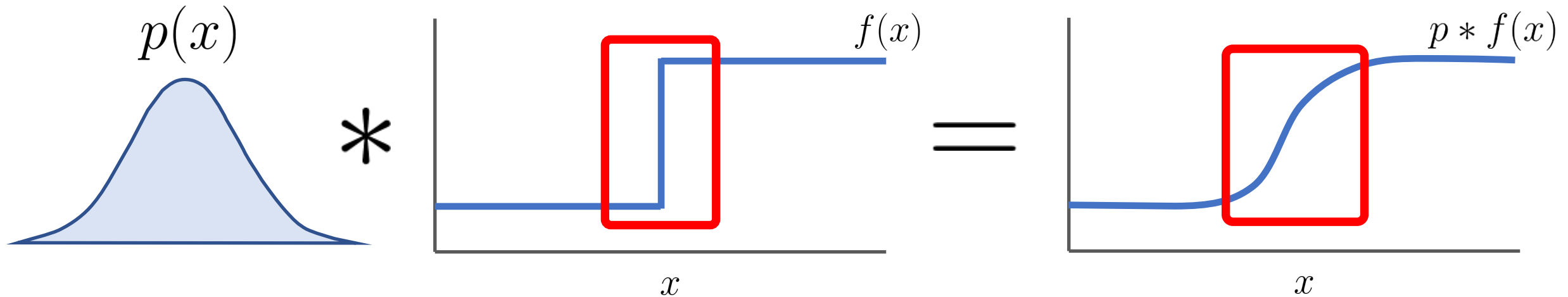
Nondifferentiable Functions

$$\arg \max_p \mathbb{E}_{x \sim p(x)} [f(x)]$$



Nondifferentiable Functions

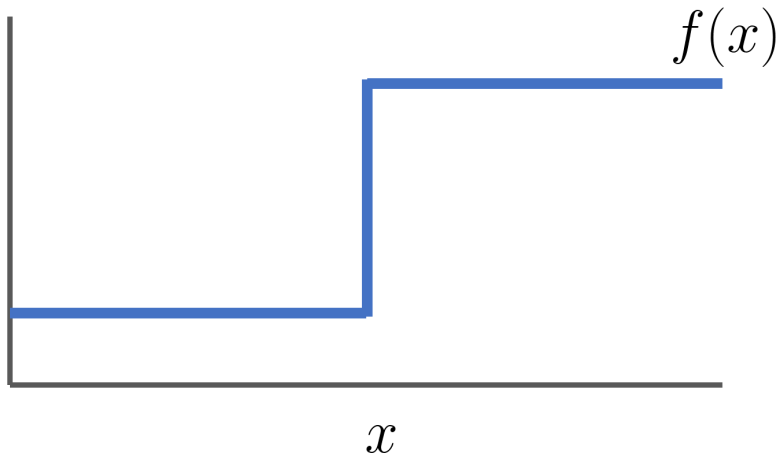
$$\arg \max_p \mathbb{E}_{x \sim p(x)} [f(x)]$$



Nondifferentiable Functions

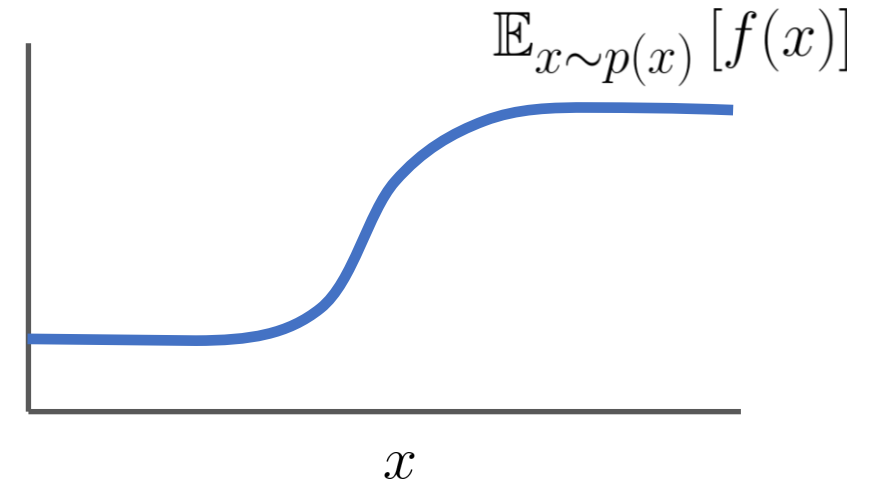
not differentiable

$$\arg \max_x \overbrace{f(x)}$$



differentiable!

$$\arg \max_p \overbrace{\mathbb{E}_{x \sim p(x)} [f(x)]}$$



Key Idea: Expectation of a **nondifferentiable** function is **differentiable**

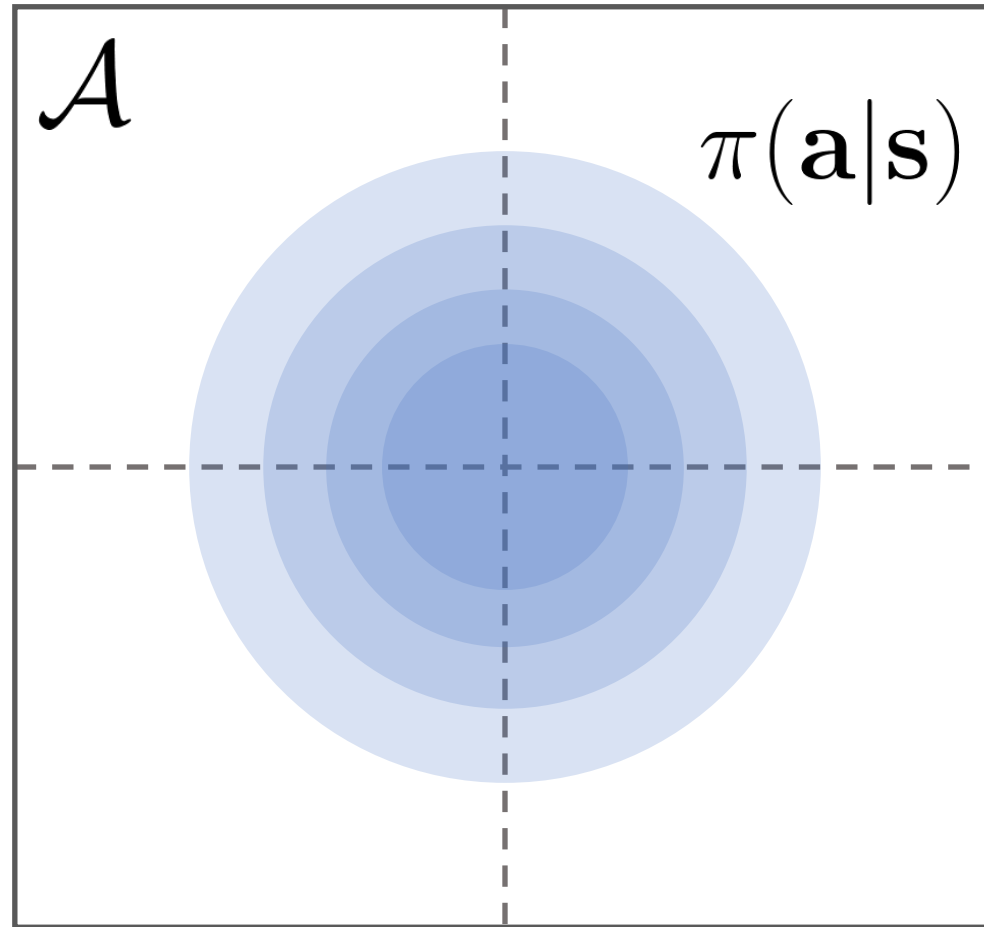
Score Function

- Score function can be applied to estimate gradients for any nondifferentiable function
 - Converts an optimization of a deterministic variable into an optimization of a stochastic distribution
- Policy gradient only works for stochastic policies

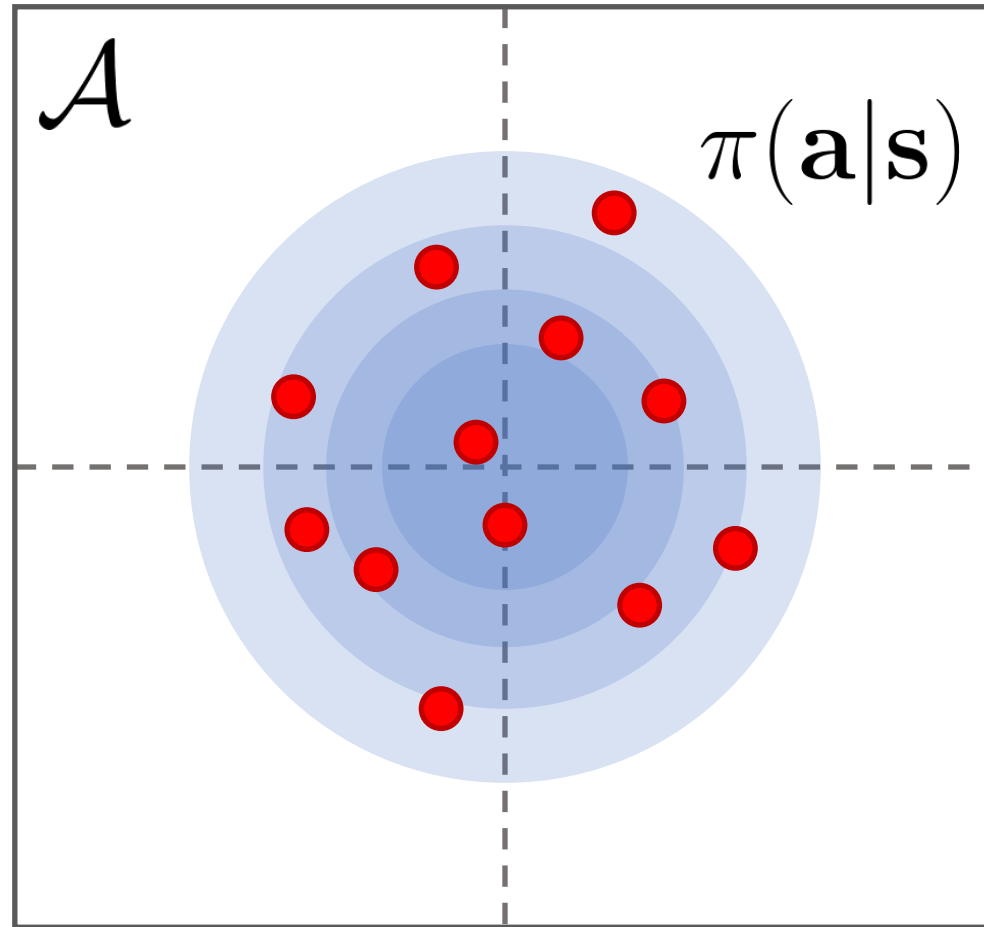
$$\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})$$

$$\mathbf{a} = \pi(\mathbf{s})$$

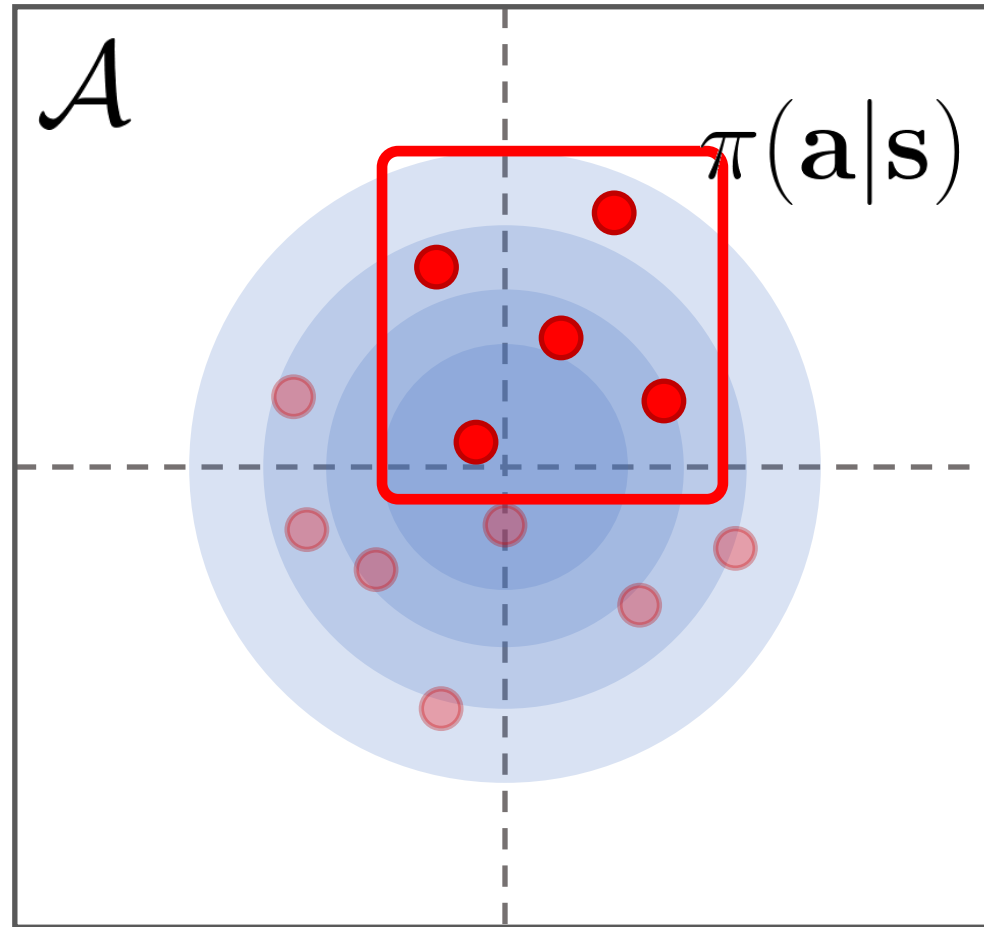
Evolutionary Strategies



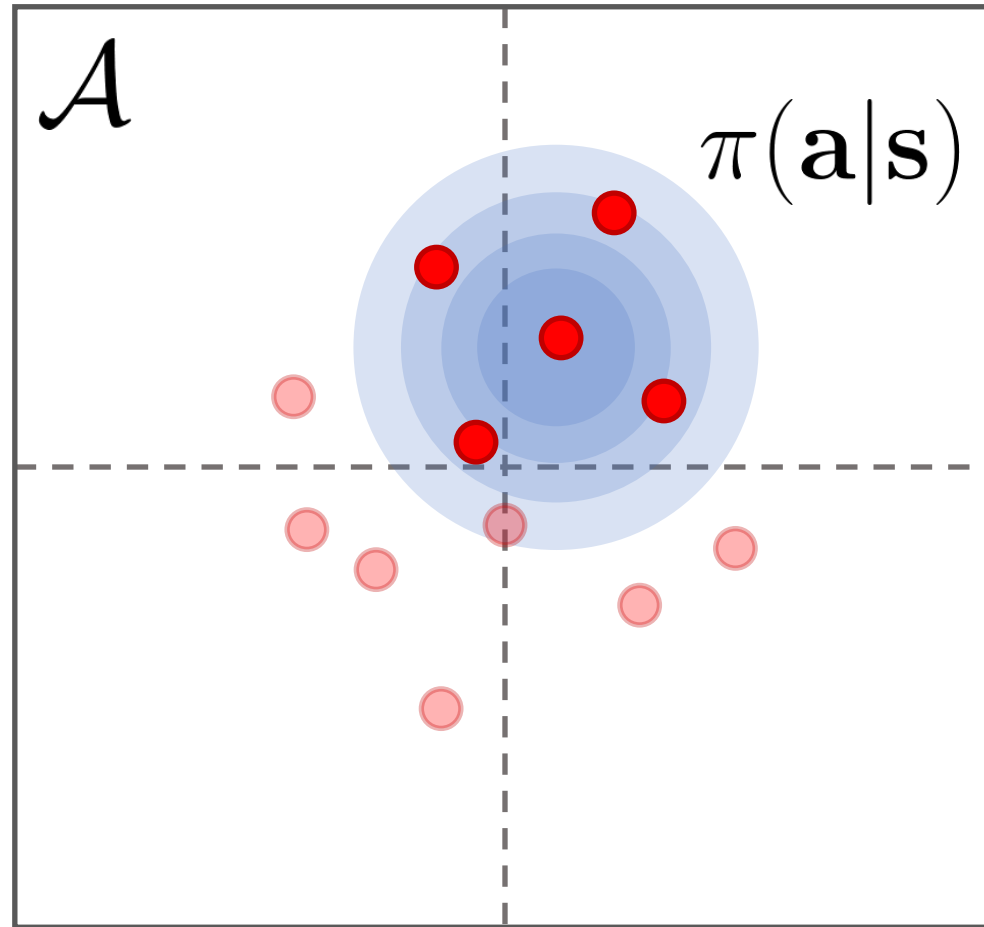
Evolutionary Strategies



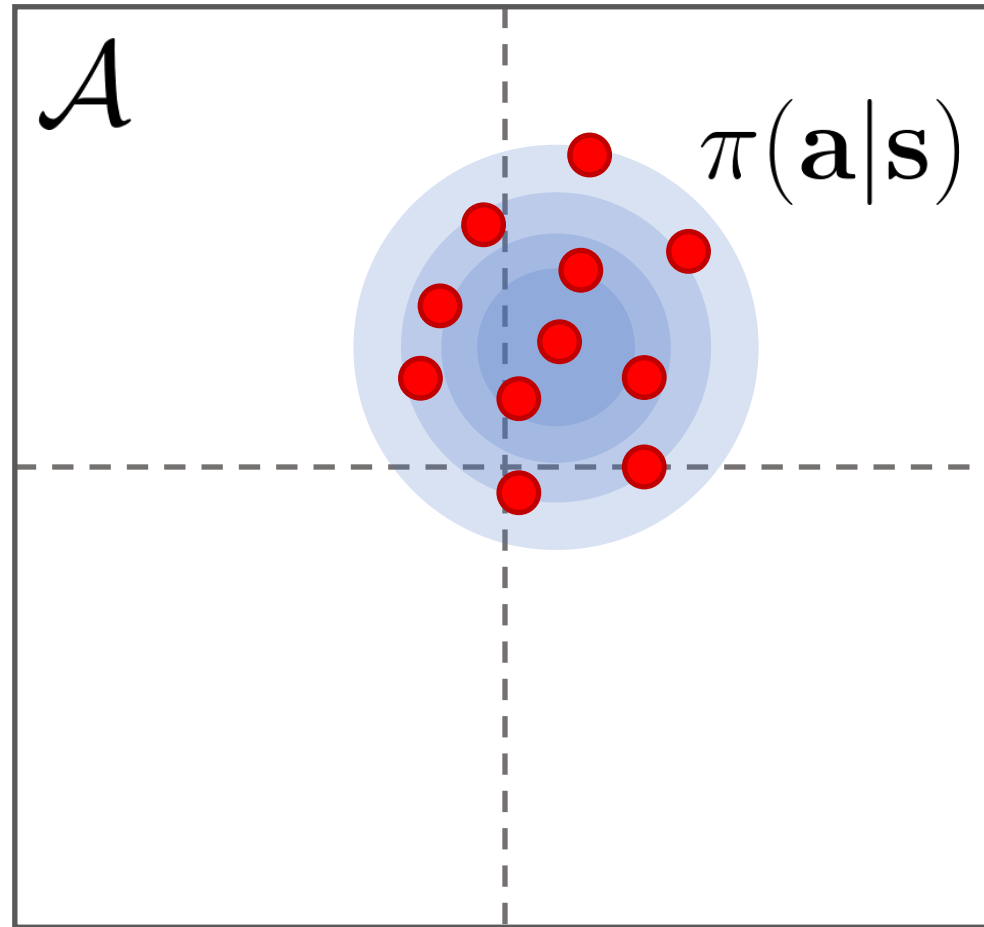
Evolutionary Strategies



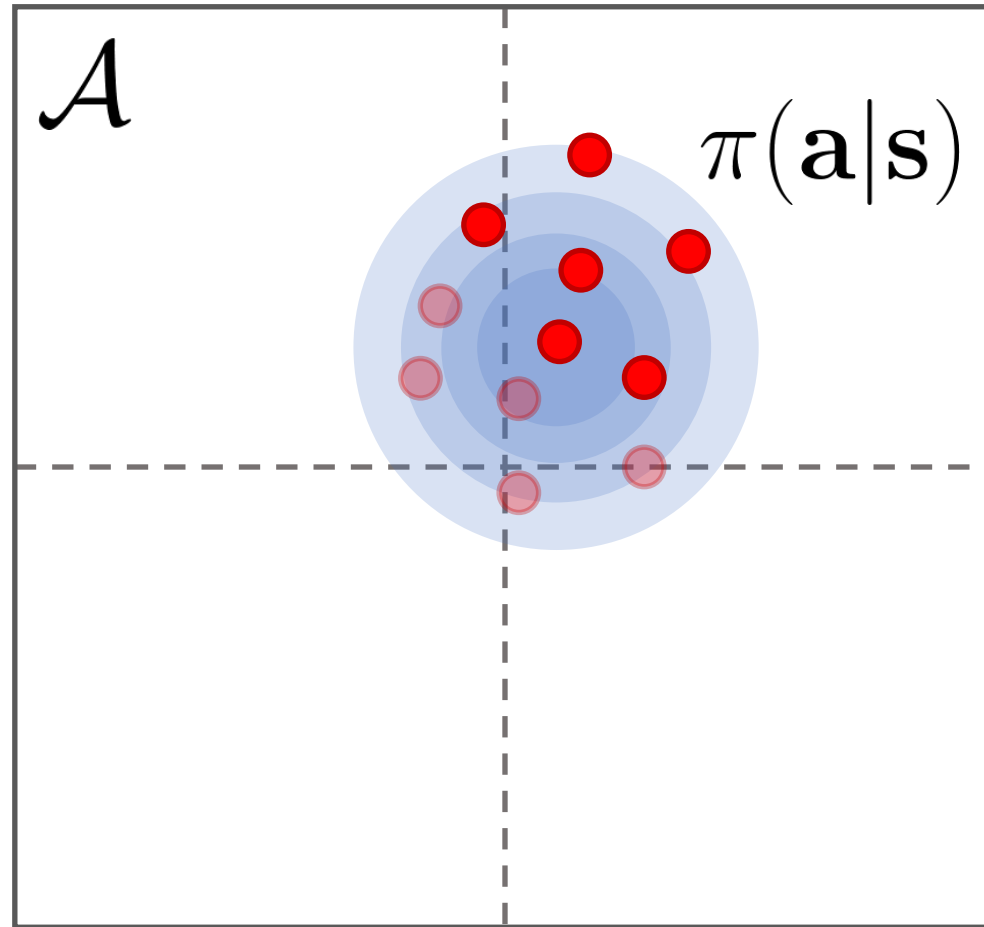
Evolutionary Strategies



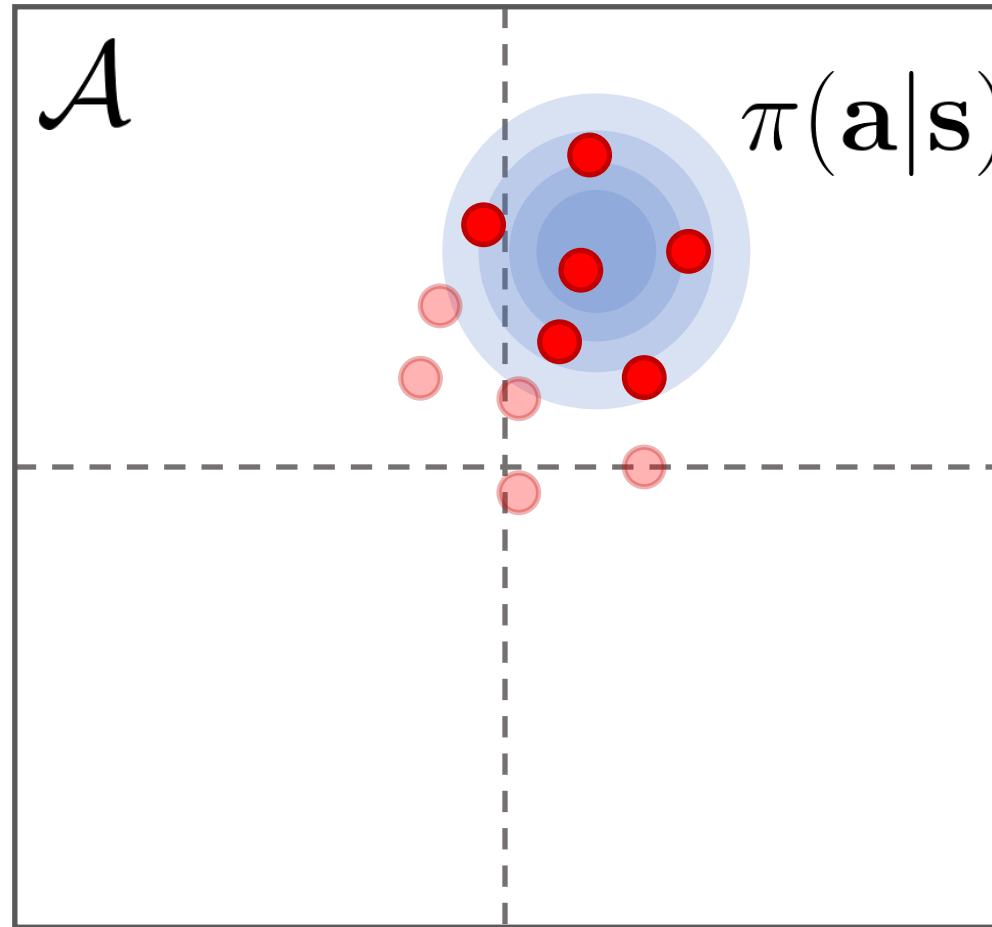
Evolutionary Strategies



Evolutionary Strategies

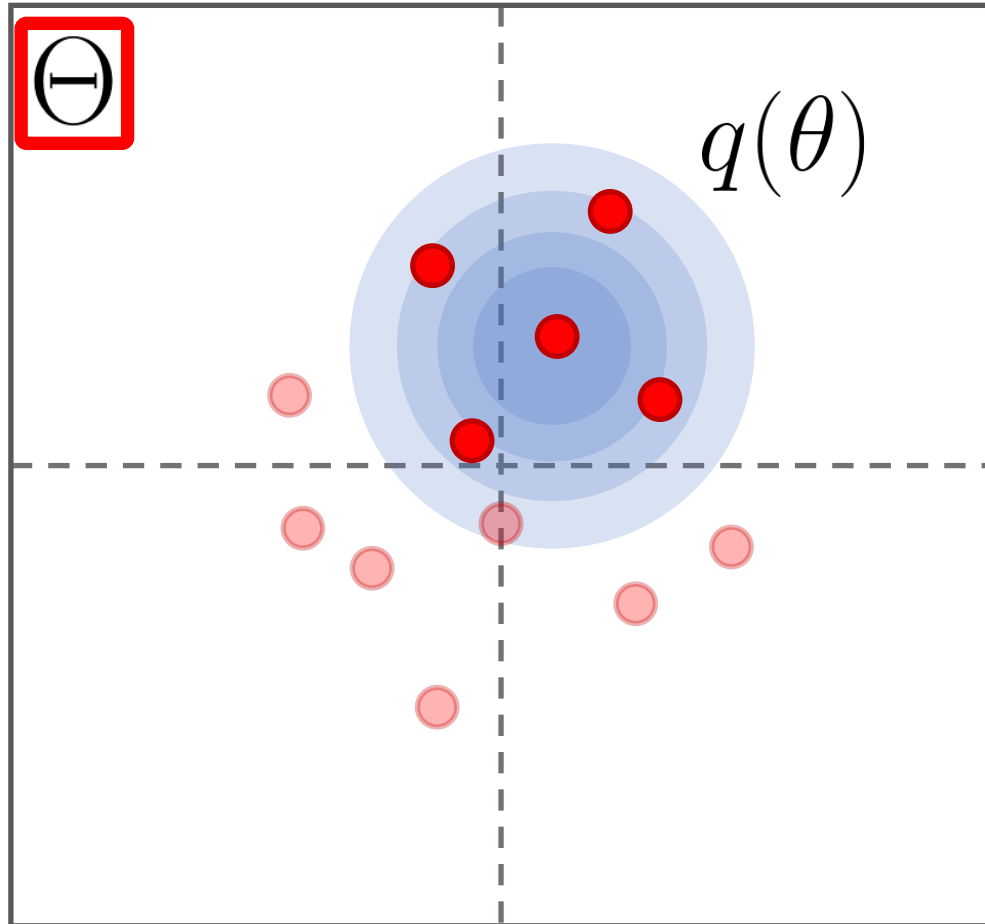


Evolutionary Strategies

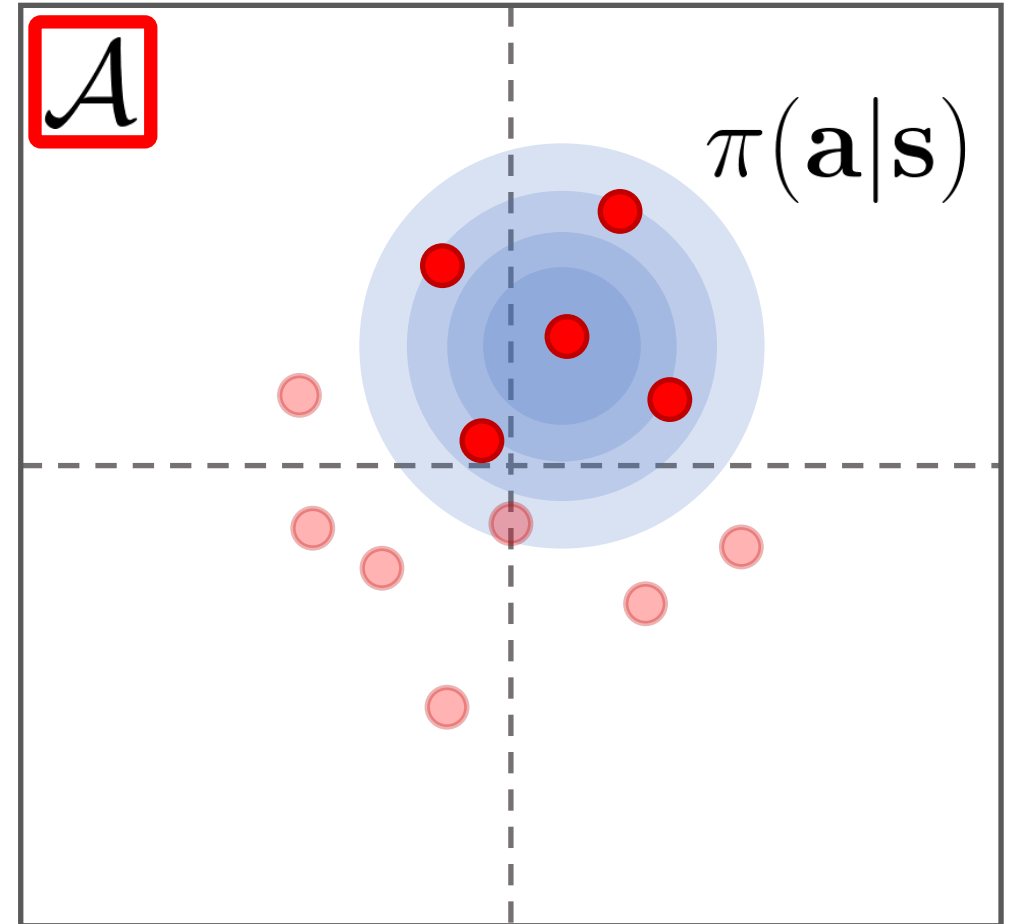


Evolutionary Strategies

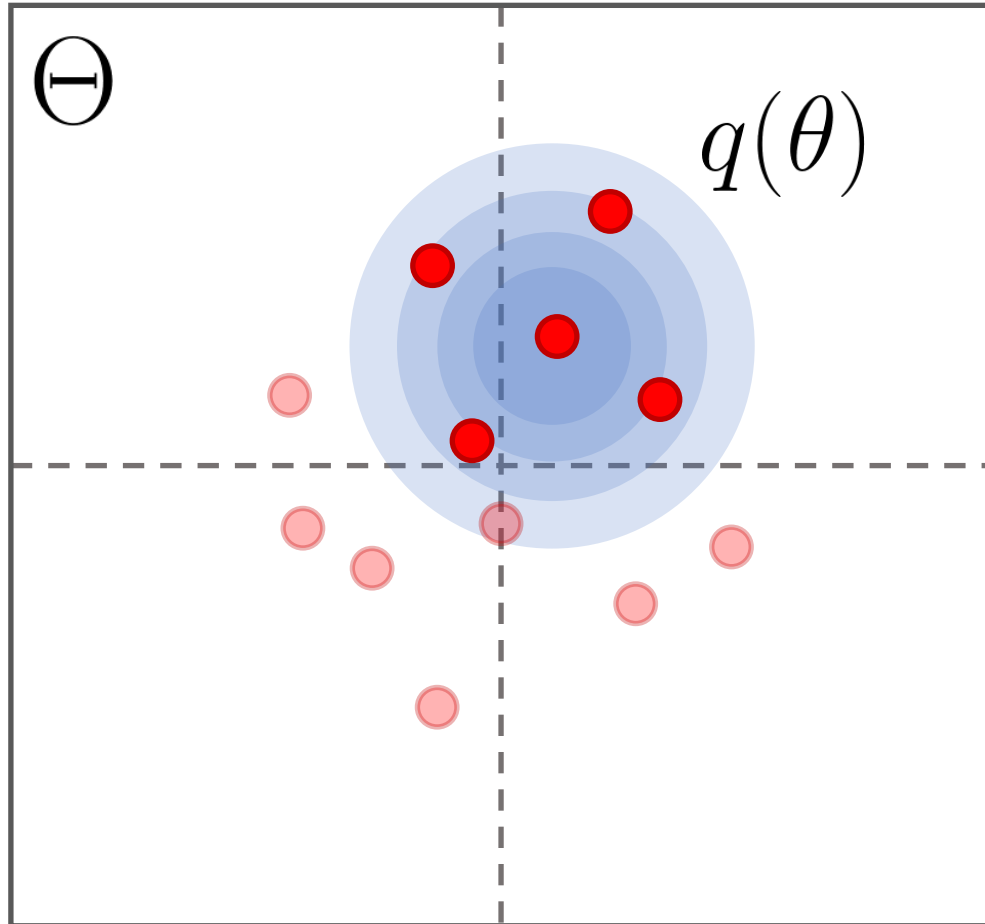
Cross-Entropy Method



Policy Gradient

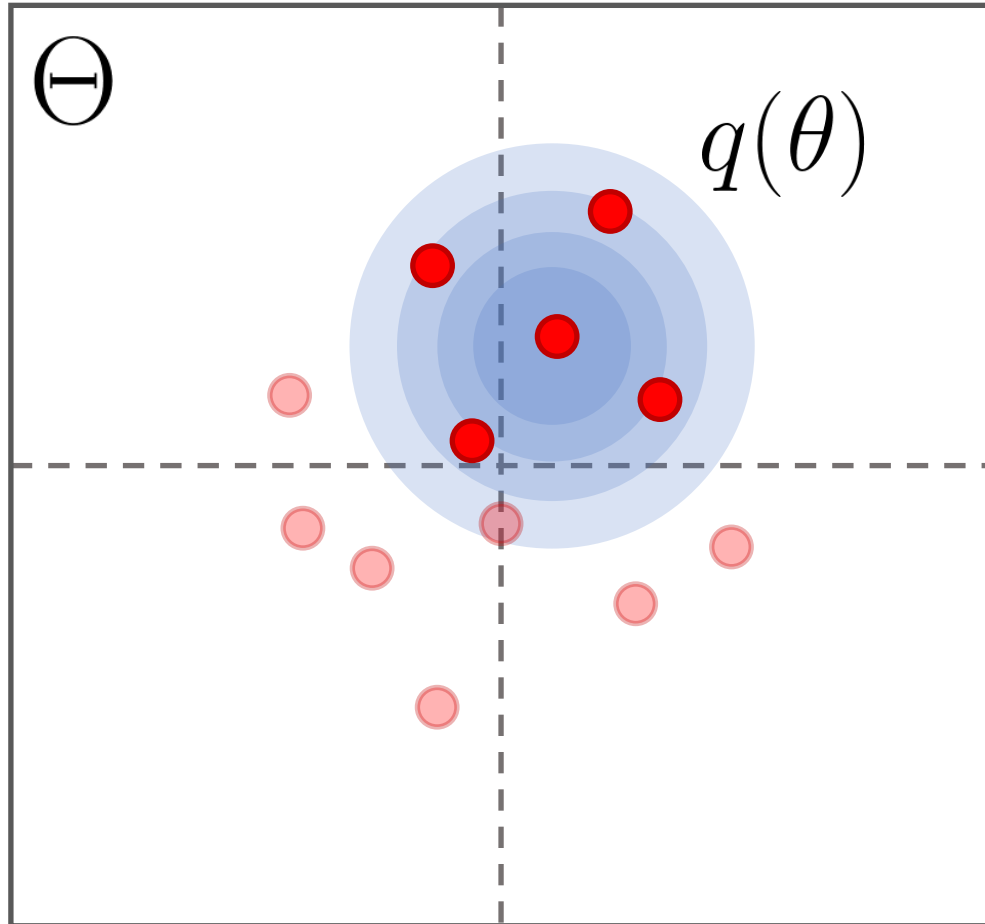


Evolutionary Strategies

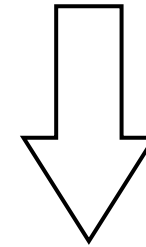


$$J(q) = \mathbb{E}_{\theta \sim \underline{q(\theta)}} [J(\pi_{\theta})]$$

Evolutionary Strategies



$$J(q) = \mathbb{E}_{\theta \sim q(\theta)} [J(\pi_{\theta})]$$



$$\nabla_q J(q) = \mathbb{E}_{\theta \sim q(\theta)} [J(\pi_{\theta}) \nabla_q \log q(\theta)]$$

evolutionary strategy

Evolution is doing gradient ascent!

Evolutionary Strategies

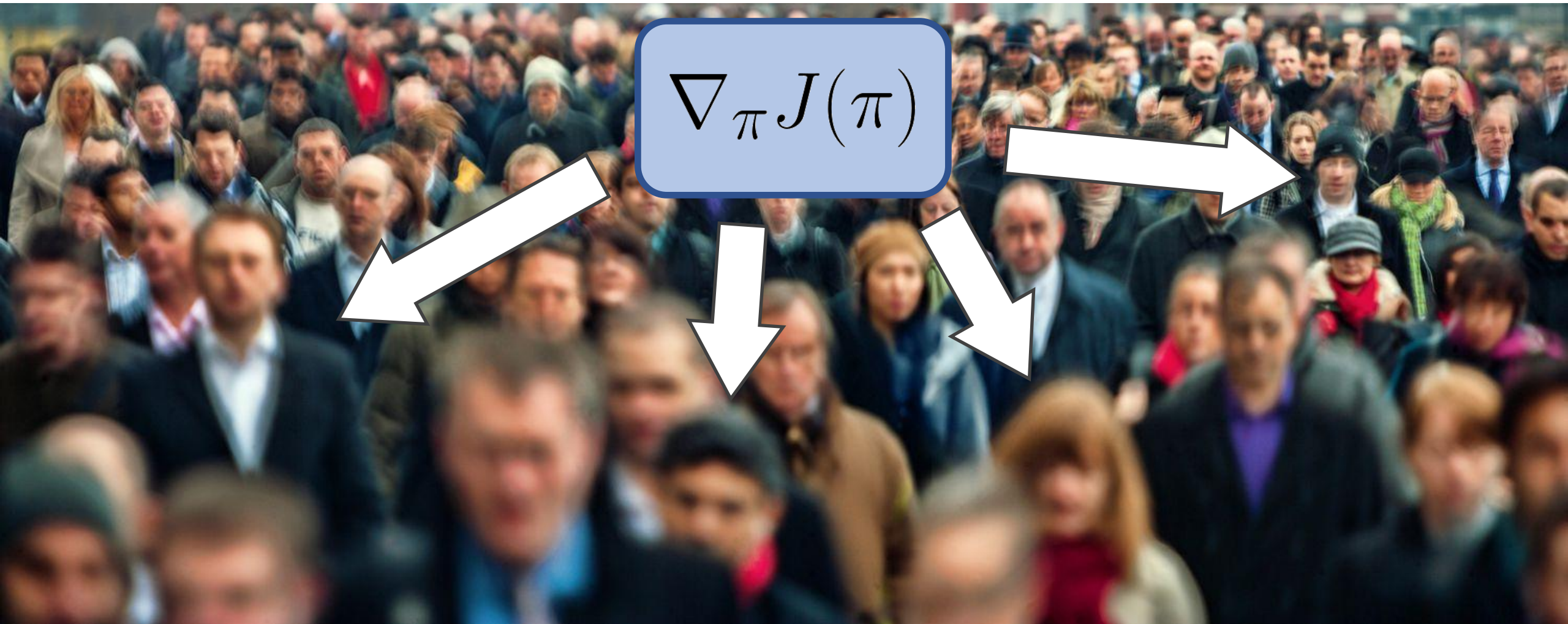
Cross-Entropy Method:

- Optimize distribution over parameters

Policy Gradient:

- Optimize distribution over actions

Evolution



Summary

- Policy Gradient
- Derivation
- Variance Reduction
- Applications
- General View of PG