

# **Advance Policy Gradient**

CMPT 729 G100

Jason Peng

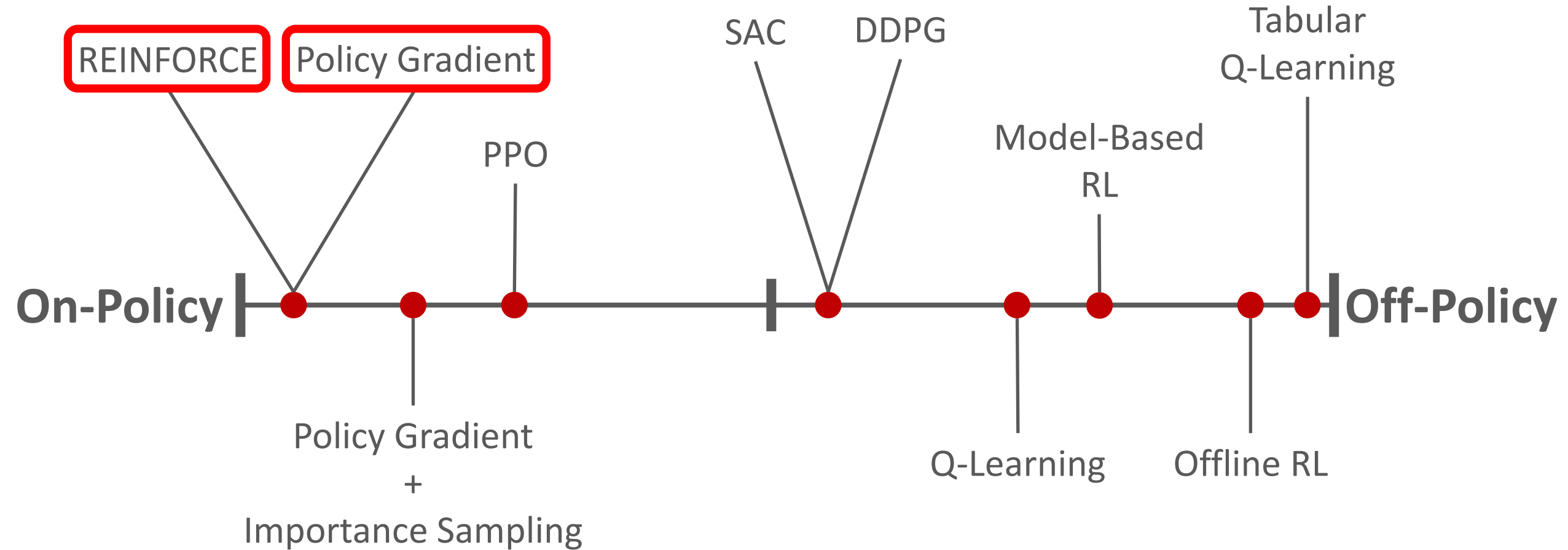
# Overview

---

- Off-Policy Policy Gradient
- Constrained Policy Optimization
- Proximal Policy Optimization

# On-Policy vs Off-Policy

---



# 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  $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$

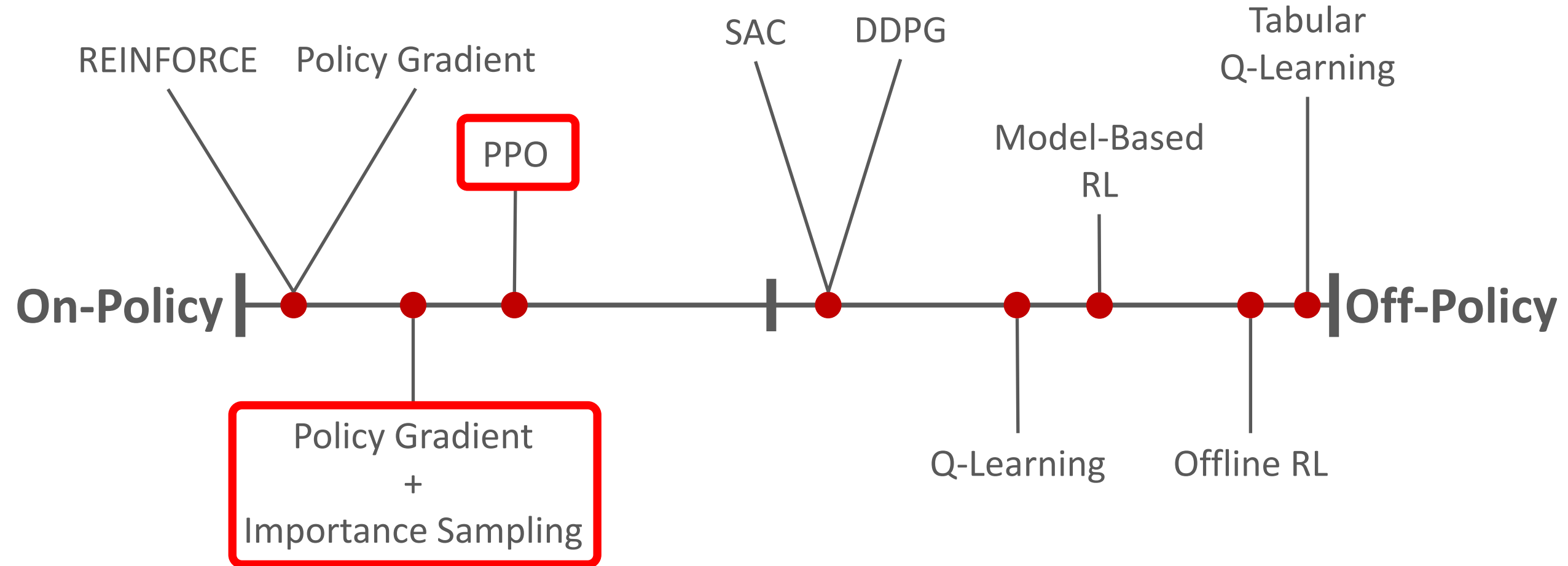
6: **end while**

7: return policy  $\pi_\theta$

---

Perform just one grad update,  
then throw out data

# On-Policy vs Off-Policy



# Off-Policy REINFORCE

---

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

Must be from  
current policy

- Off-Policy Reinforce: can we estimate  $\nabla_{\pi} J(\pi)$  using data from another policy  $\mu(\mathbf{a}|\mathbf{s})$ ?

# Importance Sampling

---

- Want to estimate  $\mathbb{E}_{x \sim p(x)} [f(x)]$ , but only have data  $x \sim q(x)$

$$\begin{aligned}\mathbb{E}_{x \sim p(x)} [f(x)] &= \sum_x p(x) f(x) \\ &= \sum_x \frac{q(x)}{\underbrace{q(x)}_{=1}} p(x) f(x)\end{aligned}$$

# Importance Sampling

---

- Want to estimate  $\mathbb{E}_{x \sim p(x)} [f(x)]$ , but only have data  $x \sim q(x)$

$$\begin{aligned}\mathbb{E}_{x \sim p(x)} [f(x)] &= \sum_x p(x) f(x) \\ &= \sum_x \frac{q(x)}{q(x)} p(x) f(x) \\ &= \sum_x q(x) \frac{p(x)}{q(x)} f(x) = \mathbb{E}_{x \sim q(x)} \left[ \frac{p(x)}{q(x)} f(x) \right]\end{aligned}$$

“Importance Sampling”  
weight



# Off-Policy REINFORCE

---

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

$\mu(\mathbf{a}|\mathbf{s})$  : behavior policy

$$\begin{aligned}&= \sum_{\tau} \frac{p(\tau|\mu)}{\underbrace{p(\tau|\mu)}_{=1}} p(\tau|\pi) \nabla_{\pi} \log p(\tau|\pi) R(\tau)\end{aligned}$$

# Off-Policy REINFORCE

---

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

$\mu(\mathbf{a}|\mathbf{s})$  : behavior policy

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

# Off-Policy REINFORCE

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim \underline{p(\tau|\mu)}} \left[ \frac{p(\tau|\pi)}{p(\tau|\mu)} \nabla_{\pi} \log p(\tau|\pi) R(\tau) \right]$$

Data sampled  
according to  $\mu$

# Importance Sampling

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\mu)} \left[ \underbrace{\frac{p(\tau|\pi)}{p(\tau|\mu)}}_{\text{"Importance Sampling" weight}} \nabla_{\pi} \log p(\tau|\pi) R(\tau) \right]$$

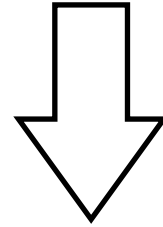
“Importance Sampling”  
weight

# Importance Sampling

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\mu)} \left[ \underbrace{\frac{p(\tau|\pi)}{p(\tau|\mu)}}_{=1} \nabla_{\pi} \log p(\tau|\pi) R(\tau) \right]$$

If  $p(\tau|\mu) = p(\tau|\pi)$ :



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

# Importance Sampling

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\mu)} \left[ \underbrace{\frac{p(\tau|\pi)}{p(\tau|\mu)}}_{< 1} \nabla_{\pi} \log p(\tau|\pi) R(\tau) \right]$$

If  $p(\tau|\pi) < p(\tau|\mu)$ :

- Down-weight likelihood of trajectory

# Importance Sampling

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\mu)} \left[ \underbrace{\frac{p(\tau|\pi)}{p(\tau|\mu)}}_{> 1} \nabla_{\pi} \log p(\tau|\pi) R(\tau) \right]$$

If  $p(\tau|\pi) > p(\tau|\mu)$ :

- Up-weight likelihood of trajectory

# Importance Sampling

---

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

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



# Importance Sampling

---

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

# Importance Sampling

---

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

# Importance Sampling

---

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

# Importance Sampling

$$\mathcal{T} = \left\{ \begin{array}{l} \boxed{s_0 \quad a_0 \quad r_0} + \boxed{\log \mu(a_0 | s_0)} \\ \boxed{s_1 \quad a_1 \quad r_1} + \boxed{\log \mu(a_1 | s_1)} \\ \vdots \\ \boxed{s_T} \end{array} \right.$$

$\left( \frac{\prod_{t=0}^{T-1} \pi(a_t | s_t)}{\prod_{t=0}^{T-1} \mu(a_t | s_t)} \right)$

# Importance Sampling

---

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

- Can estimate gradient from arbitrary distribution, as long as  $\mu(\mathbf{a}|\mathbf{s}) > 0$  for all actions (e.g. Gaussian distribution)
- Never used in practice

# Importance Sampling

---

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

- Can estimate gradient from arbitrary distribution, as long as  $\mu(\mathbf{a}|\mathbf{s}) > 0$  for all actions (e.g. Gaussian distribution)
- Never used in practice
  - Very high variance if  $\pi \neq \mu$
  - Importance sampling weights very quickly vanish or explode

# Reward-to-Go Policy Gradient

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \underbrace{(Q^{\pi}(\mathbf{s}, \mathbf{a}) - V^{\pi}(\mathbf{s}))}_{\text{“advantage”}}]$$

“advantage”

$$A^{\pi}(\mathbf{s}, \mathbf{a})$$

# Reward-to-Go Policy Gradient

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a})]$$

$\mu(\mathbf{a}|\mathbf{s})$  : behavior policy

$$\begin{aligned} \nabla_{\pi} J(\pi) &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \left[ \frac{\mu(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right] \\ &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \underline{\mu(\mathbf{a}|\mathbf{s})}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right] \end{aligned}$$



# Reward-to-Go Policy Gradient

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a})]$$

$\mu(\mathbf{a}|\mathbf{s})$  : behavior policy

$$\begin{aligned} \nabla_{\pi} J(\pi) &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \left[ \frac{\mu(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right] \\ &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right] \end{aligned}$$

single-step  
lower variance

# Reward-to-Go Policy Gradient

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a})]$$

$\mu(\mathbf{a}|\mathbf{s})$  : behavior policy

$$\begin{aligned} \nabla_{\pi} J(\pi) &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} \left[ \frac{\mu(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right] \\ &= \mathbb{E}_{\mathbf{s} \sim d_{\pi}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right] \end{aligned}$$

What about the  
state distribution?

# Reward-to-Go Policy Gradient

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\pi}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right]$$

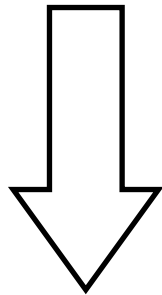
Computing the IS weights  
for  $d_{\pi}(\mathbf{s})$  is intractable.

$$\frac{d_{\pi}(\mathbf{s})}{d_{\mu}(\mathbf{s})}$$

# Reward-to-Go Policy Gradient

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\pi}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right]$$



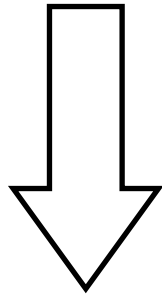
$$\nabla_{\pi} J(\pi) \approx \mathbb{E}_{\mathbf{s} \sim \underline{d_{\mu}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right]$$

Ok, if  $\mu \approx \pi$ ?

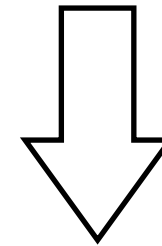
# Reward-to-Go Policy Gradient

---

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\pi}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\pi}(\mathbf{s}, \mathbf{a}) \right]$$



$$\nabla_{\pi} J(\pi) \approx \mathbb{E}_{\mathbf{s} \sim \underline{d_{\mu}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \underline{A^{\pi}(\mathbf{s}, \mathbf{a})} \right]$$



$$\approx \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \underline{A^{\mu}(\mathbf{s}, \mathbf{a})} \right]$$

# Policy Gradient + Importance Sampling

$$\nabla_{\pi} J^{\mu}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) A^{\mu}(\mathbf{s}, \mathbf{a}) \right]$$

Surrogate objective:

$$J^{\mu}(\pi) = \underline{\mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right]$$

# Surrogate Objective

---

Policy Gradient + Importance Sampling:

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

Soft Actor-Critic:

$$\hat{J}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q^\pi(\mathbf{s}, \mathbf{a})]$$

# Surrogate Objective

---

Policy Gradient + Importance Sampling:

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \underline{A^\mu(\mathbf{s}, \mathbf{a})} \right]$$

Soft Actor-Critic:

$$\hat{J}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [\underline{Q^\pi(\mathbf{s}, \mathbf{a})}]$$



# Surrogate Objective

---

Policy Gradient + Importance Sampling:

$$J^{\mu}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right]$$

Soft Actor-Critic:

$$\hat{J}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q^{\pi}(\mathbf{s}, \mathbf{a})]$$

# Surrogate Objective

---

Policy Gradient + Importance Sampling:

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

Soft Actor-Critic:

$$\hat{J}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q^\pi(\mathbf{s}, \mathbf{a})]$$

# Surrogate Objective

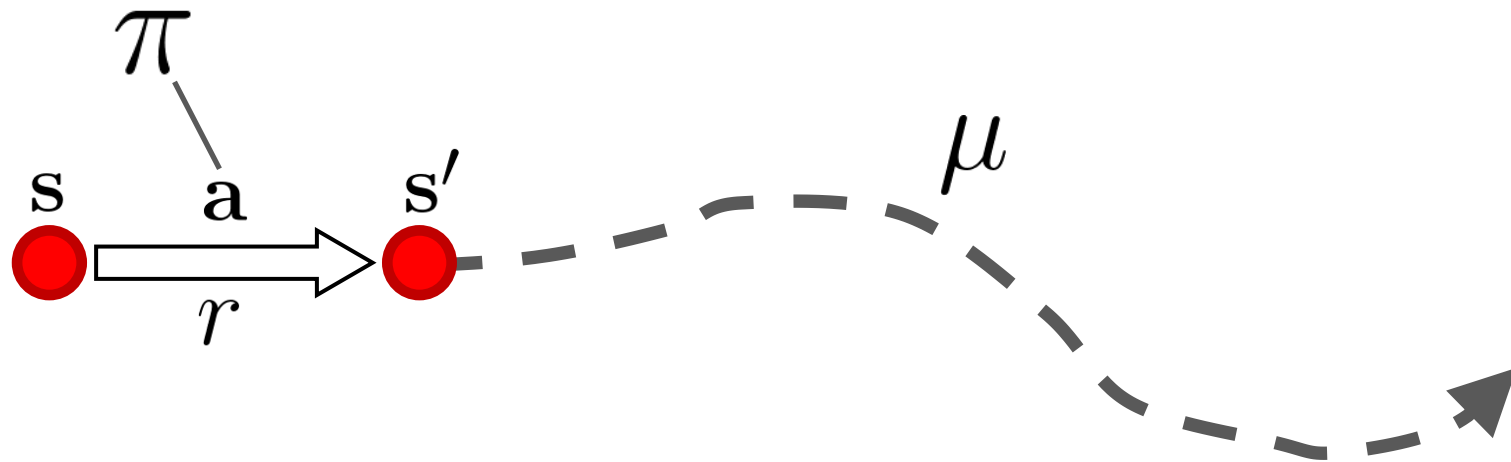
---

Policy Gradient + Importance Sampling:

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

Soft Actor-Critic:

$$\hat{J}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q^\pi(\mathbf{s}, \mathbf{a})]$$



# Surrogate Objective

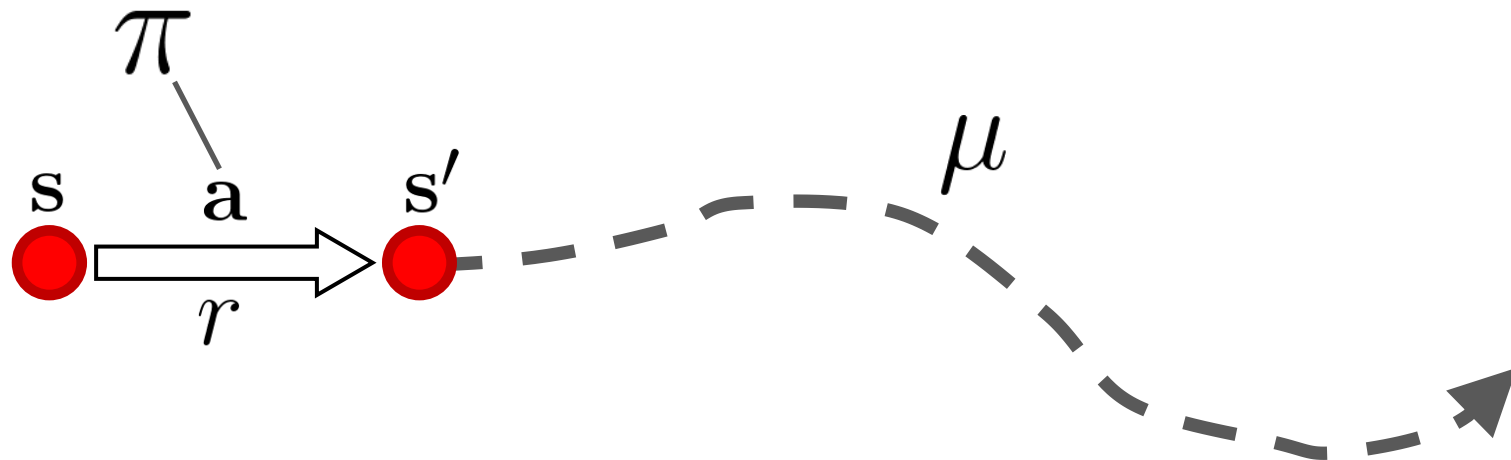
---

Policy Gradient + Importance Sampling:

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

Soft Actor-Critic:

$$\hat{J}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q^\pi(\mathbf{s}, \mathbf{a})]$$



# Surrogate Objective

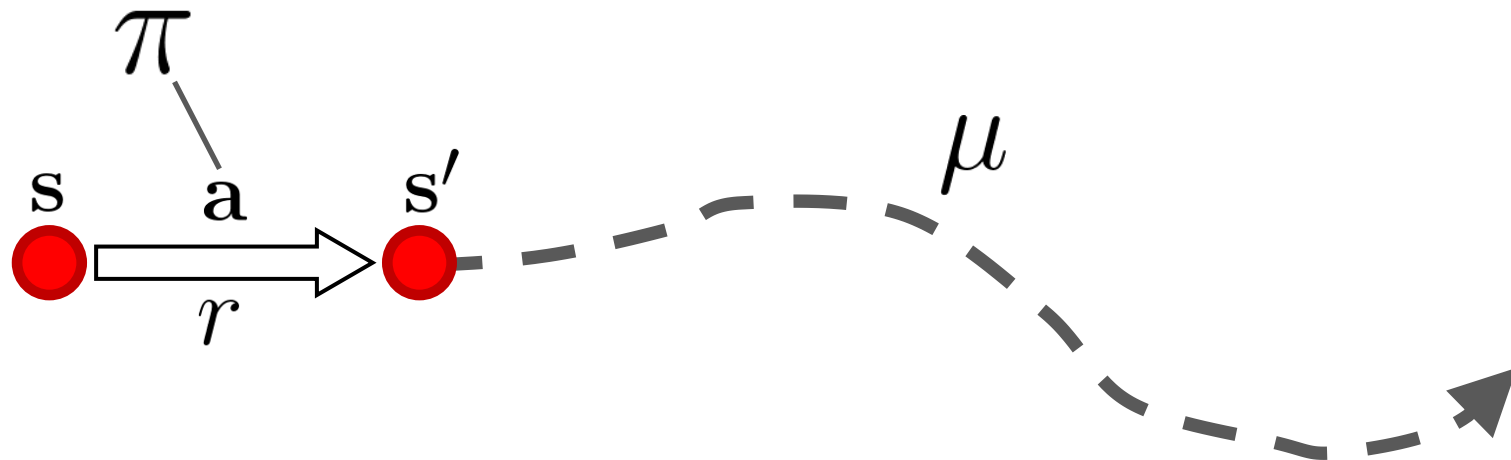
---

Policy Gradient + Importance Sampling:

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

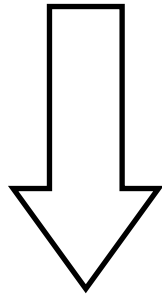
Soft Actor-Critic:

$$\hat{J}(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})} [Q^\pi(\mathbf{s}, \mathbf{a})]$$

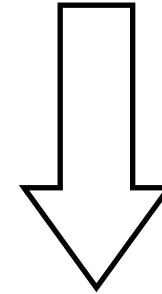


# Policy Gradient + Importance Sampling

$$\nabla_{\pi} J(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\pi}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \underline{A^{\pi}(\mathbf{s}, \mathbf{a})} \right]$$



$$\nabla_{\pi} J^{\mu}(\pi) = \mathbb{E}_{\mathbf{s} \sim \underline{d_{\mu}(\mathbf{s})}} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} \nabla_{\pi} \log \pi(\mathbf{a}|\mathbf{s}) \underline{A^{\mu}(\mathbf{s}, \mathbf{a})} \right]$$



Ok, if  $\mu \approx \pi$ ?

# Surrogate Objective

---

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

Reasonable if  $\pi$  is *close* to  $\mu$

$$D_{\text{KL}}^{\max}(\mu, \pi) = \max_{\mathbf{s}} D_{\text{KL}}(\mu(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))$$

# Surrogate Objective

---

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

If  $D_{\text{KL}}^{\max}(\mu, \pi) \leq \epsilon$ ,

$$J(\pi) \geq J^\mu(\pi) - \underline{C}\epsilon$$

constant



# Surrogate Objective

---

$$J^\mu(\pi) = \mathbb{E}_{\mathbf{s} \sim d_\mu(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^\mu(\mathbf{s}, \mathbf{a}) \right]$$

If  $D_{\text{KL}}^{\max}(\mu, \pi) \leq \epsilon$ ,

$$J(\pi) \geq J^\mu(\pi) - C_{\epsilon}$$

The surrogate objective is a lower bound on the real objective for sufficiently small  $\epsilon$ !

# Constrained Optimization

---

$$\arg \max_{\pi} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \underline{D_{\text{KL}}^{\max}(\mu, \pi)} \leq \epsilon \quad \text{“Trust region”}$$


$$D_{\text{KL}}^{\max}(\mu, \pi) = \max_{\mathbf{s}} D_{\text{KL}}(\mu(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))$$

# Constrained Optimization

---

$$\arg \max_{\pi} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \underline{D_{\text{KL}}^{\max}(\mu, \pi)} \leq \epsilon$$


$$D_{\text{KL}}^{\max}(\mu, \pi) = \max_{\underline{\mathbf{s}}} D_{\text{KL}}(\mu(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))$$

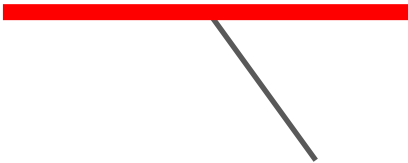
Hard to compute

# Constrained Optimization

---

$$\arg \max_{\pi} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \underline{D_{\text{KL}}^{\text{mean}}(\mu, \pi)} \leq \epsilon$$


$$D_{\text{KL}}^{\text{mean}}(\mu, \pi) = \mathbb{E}_{\mathbf{s} \sim d^{\mu}(\mathbf{s})} [D_{\text{KL}}(\mu(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))]$$

# Constrained Optimization

---

$$\begin{aligned} \arg \max_{\pi} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] \\ \text{s.t. } D_{\text{KL}}^{\text{mean}}(\mu, \pi) \leq \epsilon \end{aligned}$$

How do we pick  $\mu$ ?

- In practice, collect data using current policy  $\mu = \pi^k$

# Constrained Policy Optimization

---

## ALGORITHM: Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: **for** iteration  $k = 0, \dots, n - 1$  **do**
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: **end for**
  - 9: **return** policy  $\pi^n$
-

# Constrained Policy Optimization

---

---

**ALGORITHM:** Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: for iteration  $k = 0, \dots, n - 1$  do
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: end for
  - 9: return policy  $\pi^n$
-

# Constrained Policy Optimization

---

---

**ALGORITHM:** Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: **for** iteration  $k = 0, \dots, n - 1$  **do**
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: **end for**
  - 9: return policy  $\pi^n$
-



# Constrained Policy Optimization

---

---

**ALGORITHM:** Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: **for** iteration  $k = 0, \dots, n - 1$  **do**
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: **end for**
  - 9: return policy  $\pi^n$
-

# Constrained Policy Optimization

---

---

**ALGORITHM:** Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: **for** iteration  $k = 0, \dots, n - 1$  **do**
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: **end for**
  - 9: return policy  $\pi^n$
-

# Constrained Policy Optimization

---

---

**ALGORITHM:** Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: **for** iteration  $k = 0, \dots, n - 1$  **do**
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: **end for**
  - 9: return policy  $\pi^n$
-

# Constrained Policy Optimization

---

---

**ALGORITHM:** Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: **for** iteration  $k = 0, \dots, n - 1$  **do**
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: **end for**
  - 9: **return** policy  $\pi^n$
-

# Constrained Policy Optimization

---

---

**ALGORITHM:** Constrained Policy Optimization

---

- 1:  $\pi_0 \leftarrow$  initialize policy
  - 2: **for** iteration  $k = 0, \dots, n - 1$  **do**
  - 3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$
  - 4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$
  - 5:   Fit value function  $V^k(\mathbf{s})$
  - 6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$
  - 7:   Update policy:
$$\pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$
$$\text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon$$
  - 8: **end for**
  - 9: return policy  $\pi^n$
-

# Constrained Policy Optimization

---

## ALGORITHM: Constrained Policy Optimization

---

1:  $\pi_0 \leftarrow$  initialize policy

2: **for** iteration  $k = 0, \dots, n - 1$  **do**

3: Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$

4: Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$

5: Fit value function  $V^k(\mathbf{s})$

6: Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$

7: Update policy:

$$\begin{aligned} \pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right] \\ \text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon \end{aligned}$$

8: **end for**

9: return policy  $\pi^n$

---

Still need to collect a new batch of data every iteration

# Constrained Policy Optimization

---

**ALGORITHM:** Constrained Policy Optimization

---

1:  $\pi_0 \leftarrow$  initialize policy

2: **for** iteration  $k = 0, \dots, n - 1$  **do**

3:   Sample trajectories  $\tau^i$  from policy  $\pi^k(\mathbf{a}|\mathbf{s})$

4:   Store trajectories in dataset  $\mathcal{D} = \{\tau^i\}$

5:   Fit value function  $V^k(\mathbf{s})$

6:   Calculate advantage  $A^k(\mathbf{s}, \mathbf{a})$  for every  $(\mathbf{s}, \mathbf{a})$  in  $\mathcal{D}$

7:   Update policy:

$$\begin{aligned} \pi^{k+1} = \arg \max_{\pi} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right] \\ \text{s.t. } \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [D_{\text{KL}}(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}))] \leq \epsilon \end{aligned}$$

Update policy with  
multiple grad steps

8: **end for**

9: **return** policy  $\pi^n$

---

# Constrained Optimization

---

$$\begin{aligned} \arg \max_{\pi} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] \\ \text{s.t. } D_{\text{KL}}^{\text{mean}}(\mu, \pi) \leq \epsilon \end{aligned}$$

How do we solve this?


Trust Region Policy Optimization (TRPO):

- Linear approximation of objective
- Quadratic approximation of constraint
- Solve with conjugate gradient method



# Lagrangian

---

$$\begin{aligned} \arg \max_{\pi} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] \\ \text{s.t. } D_{\text{KL}}^{\text{mean}}(\mu, \pi) \leq \epsilon \end{aligned}$$


# Lagrangian

---

$$\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda \left( \underline{D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon} \right)$$

# Lagrangian

---

$$\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon)$$

“Lagrange multiplier”

# Lagrangian

---

$$\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda \underbrace{(D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon)}_{\substack{> 0 \\ \text{constraint violated}}}$$

$\lambda \rightarrow \infty$

# Lagrangian

---

$$\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda \underbrace{\left( D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon \right)}_{< 0}$$

$\lambda \rightarrow 0$  constraint satisfied

# Lagrangian

---

$$\underbrace{\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon)}_{\mathcal{L}(\pi, \lambda)}$$

Dual gradient descent:

- Maximize  $\mathcal{L}(\pi, \lambda)$  wrt  $\pi$
- Update  $\lambda$ :  $\lambda \leftarrow \max \left( 0, \lambda + \alpha \left( \underline{D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon} \right) \right)$   
 $= -\nabla_{\lambda} \mathcal{L}(\pi, \lambda)$

# Lagrangian

---

$$\underbrace{\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon)}_{\mathcal{L}(\pi, \lambda)}$$

Dual gradient descent:

- Maximize  $\mathcal{L}(\pi, \lambda)$  wrt  $\pi$
- Update  $\lambda$ :  $\lambda \leftarrow \max(0, \lambda + \underline{\alpha} (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon))$   
stepsize

# Lagrangian

---

$$\underbrace{\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon)}_{\mathcal{L}(\pi, \lambda)}$$

Dual gradient descent:

- Maximize  $\mathcal{L}(\pi, \lambda)$  wrt  $\pi$
- Update  $\lambda$  :  $\lambda \leftarrow \max \left( 0, \lambda + \alpha (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon) \right)$

gradient descent



# Lagrangian

---

$$\underbrace{\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon)}_{\mathcal{L}(\pi, \lambda)}$$

Dual gradient descent:

- Maximize  $\mathcal{L}(\pi, \lambda)$  wrt  $\pi$
- Update  $\lambda$  :  $\lambda \leftarrow \max(0, \lambda + \alpha (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon))$

# Lagrangian

---

$$\underbrace{\arg \max_{\pi} \min_{\lambda \geq 0} \mathbb{E}_{\mathbf{s} \sim d_{\mu}(\mathbf{s})} \mathbb{E}_{\mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})} \left[ \frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} A^{\mu}(\mathbf{s}, \mathbf{a}) \right] - \lambda (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon)}_{\mathcal{L}(\pi, \lambda)}$$

Dual gradient descent:

- Maximize  $\mathcal{L}(\pi, \lambda)$  wrt  $\pi$
  - Update  $\lambda : \lambda \leftarrow \max(0, \lambda + \alpha (D_{\text{KL}}^{\text{mean}}(\mu, \pi) - \epsilon))$
- } Proximal Policy Optimization (PPO)

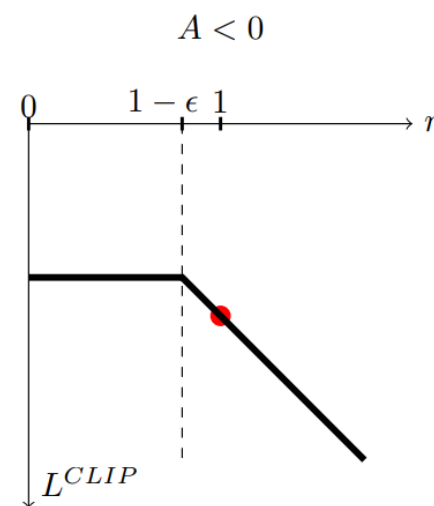
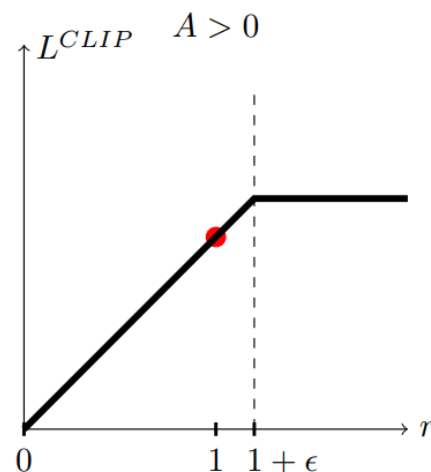
# PPO

---

In practice:

- Most PPO implementations use a clipping objective:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right]$$



# Robotic Locomotion

---



Learning Robust Perceptive Locomotion for Quadrupedal Robots in the Wild  
[Miki et al. 2022]



# Dota



Dota 2 with Large Scale Deep Reinforcement Learning  
[OpenAI et al. 2019]

# ChatGPT

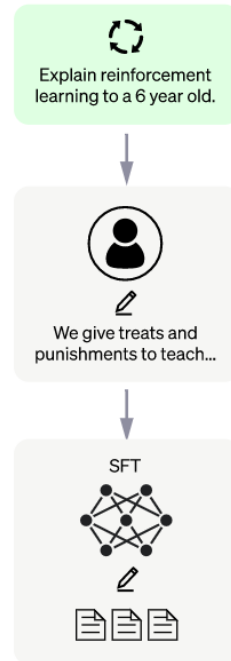
## Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



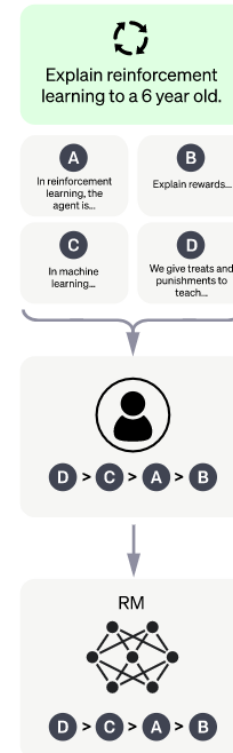
## Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



## Step 3

Optimize a policy against the reward model using the **PPO** reinforcement learning algorithm.

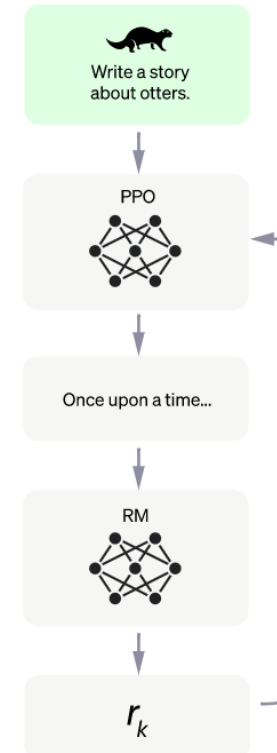
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



[OpenAI 2022]

# Summary

---

- Off-Policy Policy Gradient
- Constrained Policy Optimization
- Proximal Policy Optimization