Advance Policy Gradient

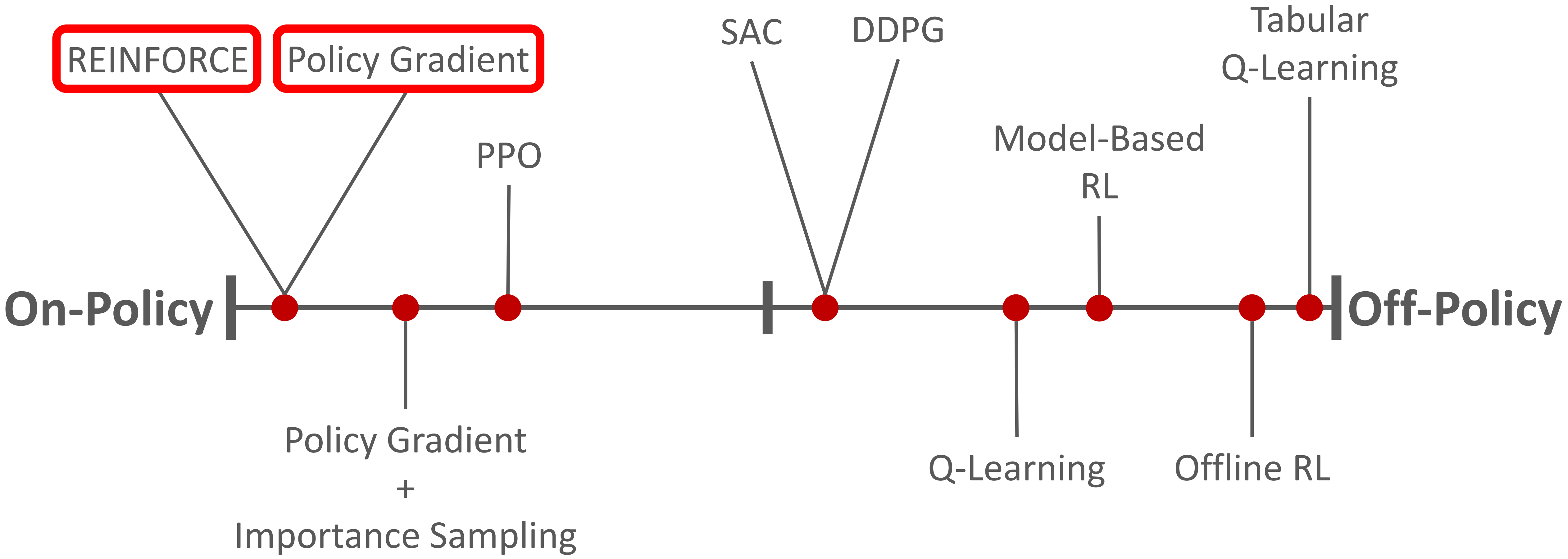
CMPT 729 G100

Jason Peng

Overview

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

On-Policy vs Off-Policy



REINFORCE

ALGORITHM: REINFORCE

$\theta \leftarrow$ initialize policy parameters

while not done **do**

  Sample trajectories $\{\tau^i\}$ from policy $\pi_\theta(\mathbf{a}|\mathbf{s})$

  Estimate policy gradient $\nabla_\theta J(\pi_\theta) \approx \frac{1}{N}\sum_i R(\tau^i)\sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i)$

  Update policy $\theta \leftarrow \theta + \alpha \nabla_\theta J(\pi_\theta)$

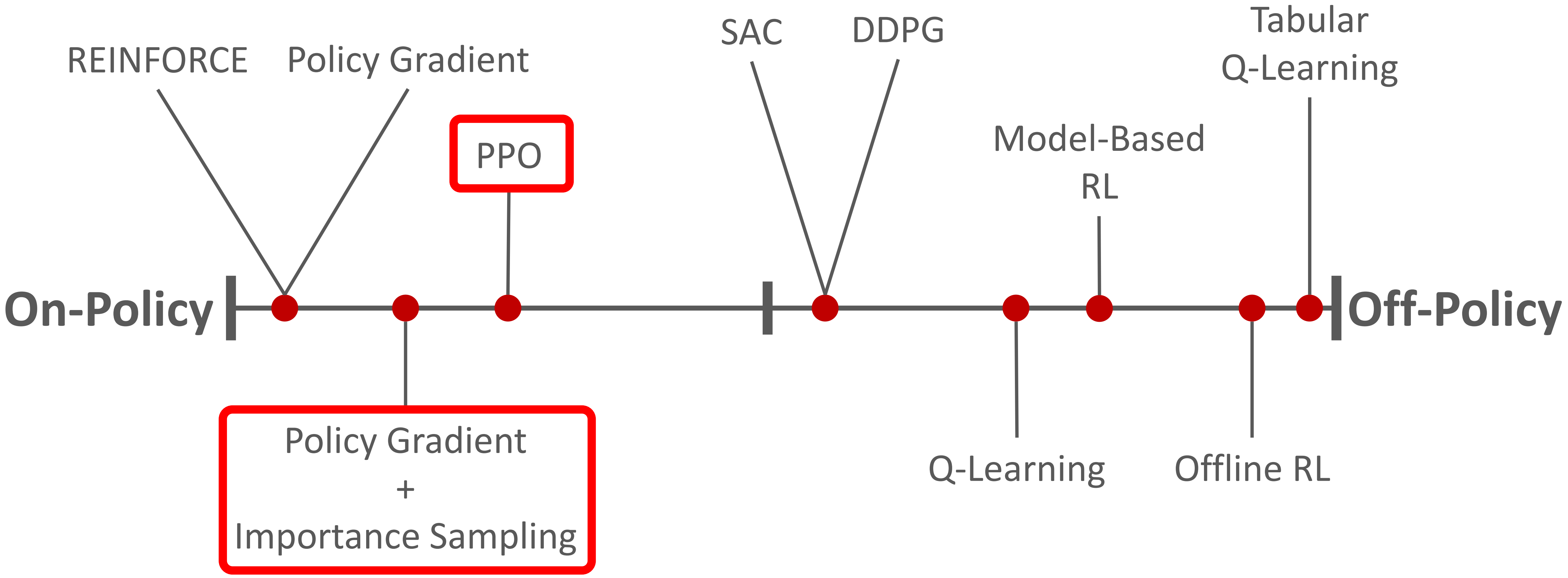
end while

return policy π_θ

Perform just one grad update, then throw out data

Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning [Williams 1992]

On-Policy vs Off-Policy



Off-Policy REINFORCE

$$\nabla_\pi J(\pi) = \underbrace{\mathbb{E}_{\tau \sim p(\tau|\pi)}}_{\text{Must be from current policy}} \left[\nabla_\pi \log p(\tau|\pi) R(\tau)\right]$$

- 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 $x \sim q(x)$

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

$$= \sum_x \underbrace{\frac{q(x)}{q(x)}}_{=1} p(x) f(x)$$

Importance Sampling

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

$$\mathbb{E}_{x\sim p(x)}\left[f(x)\right] = \sum_x p(x)f(x)$$

$$= \sum_x \frac{q(x)}{q(x)}p(x)f(x)$$

$$= \sum_x q(x)\frac{p(x)}{q(x)}f(x) = \mathbb{E}_{x\sim q(x)}\left[\underbrace{\frac{p(x)}{q(x)}}_{\text{"Importance Sampling" weight}}f(x)\right]$$

Off-Policy REINFORCE

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

$$= \sum_\tau p(\tau|\pi) \nabla_\pi \log\, p(\tau|\pi) R(\tau)$$

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

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

Off-Policy REINFORCE

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

$$= \sum_\tau p(\tau|\pi) \nabla_\pi \log p(\tau|\pi) R(\tau)$$

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

$$= \sum_\tau \frac{p(\tau|\mu)}{p(\tau|\mu)} p(\tau|\pi) \nabla_\pi \log p(\tau|\pi) R(\tau)$$

$$= \sum_\tau p(\tau|\mu) \frac{p(\tau|\pi)}{p(\tau|\mu)} \nabla_\pi \log p(\tau|\pi) R(\tau)$$

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

Off-Policy REINFORCE

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

Importance Sampling

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

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

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

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

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

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

$$= \frac{\prod_{t=0}^{T-1} \pi(\mathbf{a}_t|\mathbf{s}_t)}{\prod_{t=0}^{T-1} \mu(\mathbf{a}_t|\mathbf{s}_t)}$$

Importance Sampling

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

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

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

Importance Sampling

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

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

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

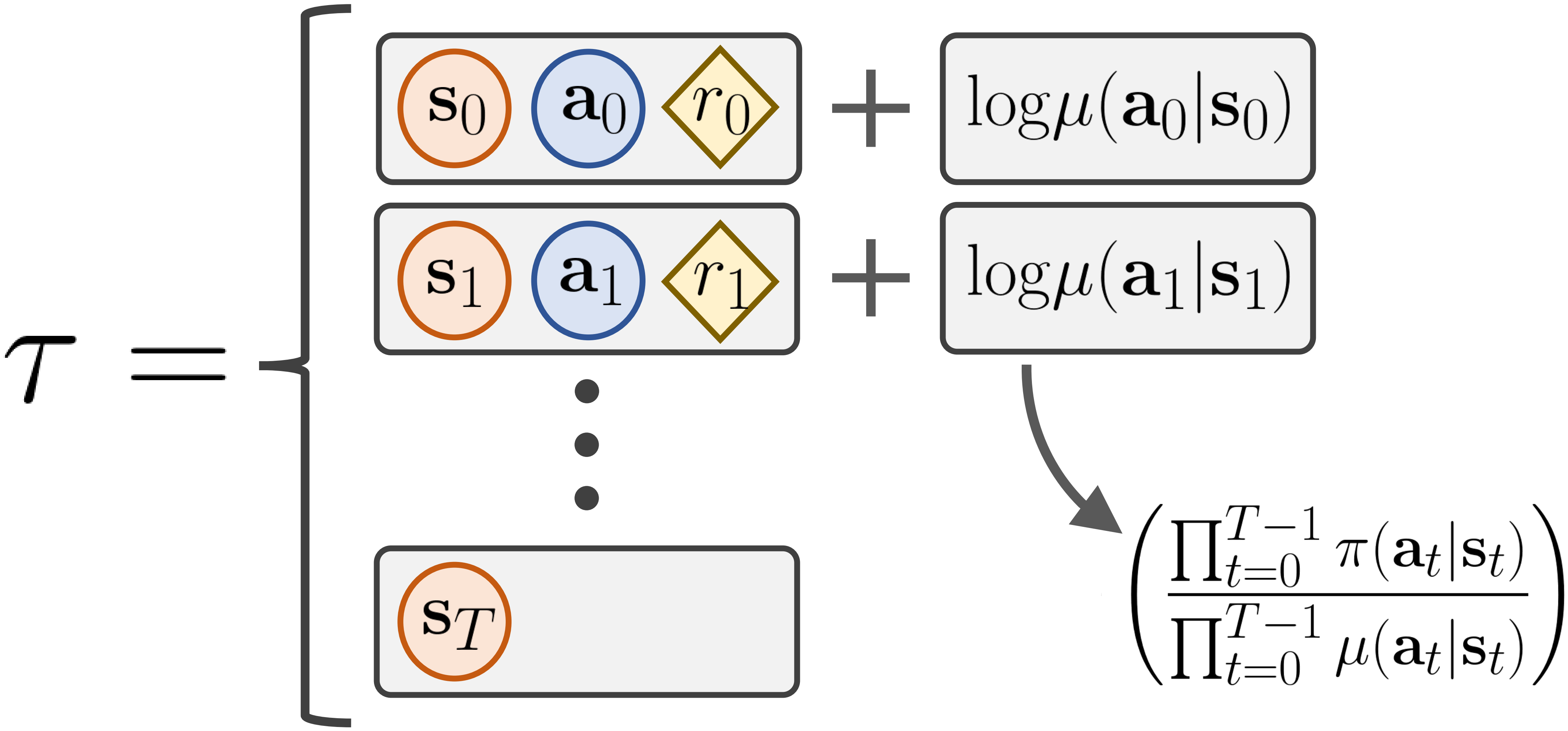
Importance Sampling

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

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

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

Importance Sampling



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

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

- 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

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

Reward-to-Go Policy Gradient

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

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

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

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

Reward-to-Go Policy Gradient

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

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

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

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

single-step
lower variance

Reward-to-Go Policy Gradient

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

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

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

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

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.

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

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

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

$$\Downarrow$$

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

$$\Downarrow$$

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

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:

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

Soft Actor-Critic:

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

Surrogate Objective

Policy Gradient + Importance Sampling:

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

Soft Actor-Critic:

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

Surrogate Objective

Policy Gradient + Importance Sampling:

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

Soft Actor-Critic:

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

Surrogate Objective

Policy Gradient + Importance Sampling:

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

Soft Actor-Critic:

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

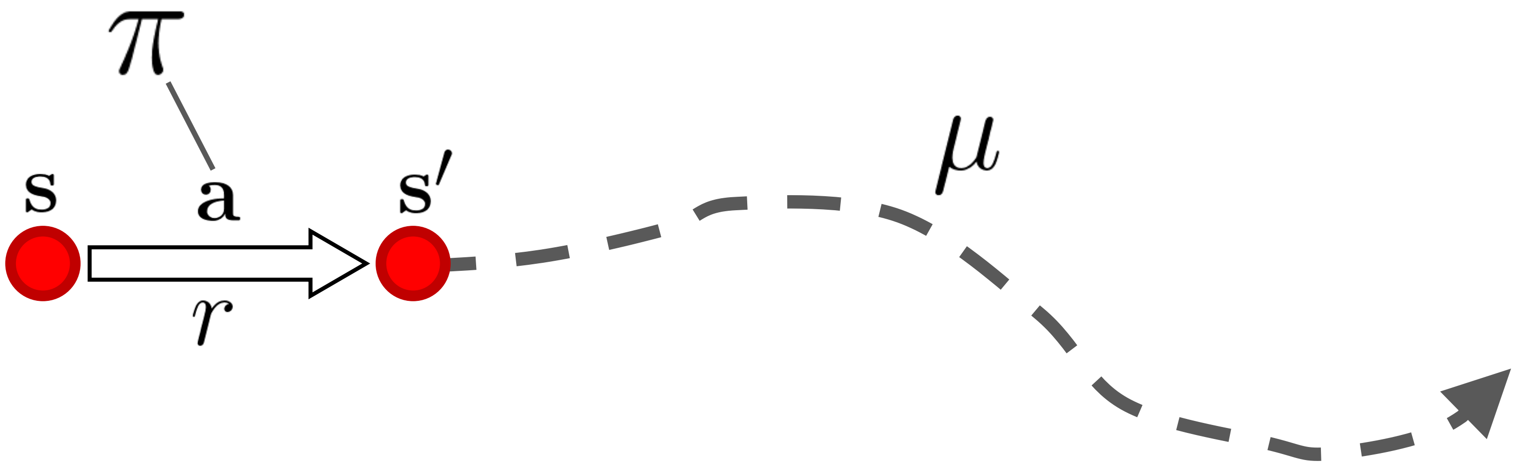
Surrogate Objective

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:

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



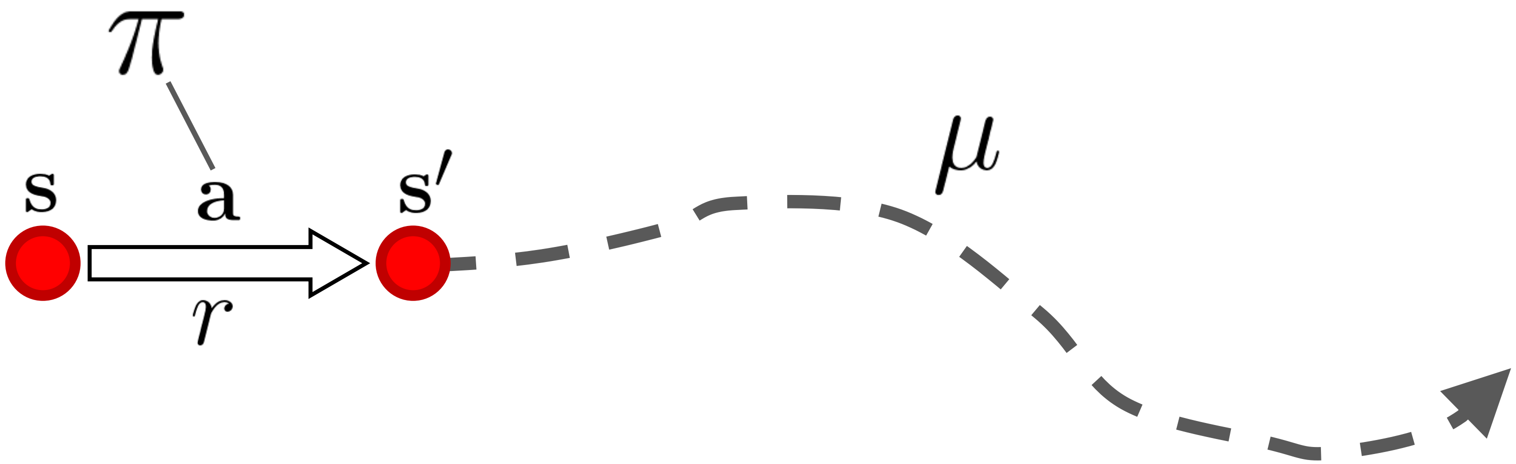
Surrogate Objective

Policy Gradient + Importance Sampling:

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

Soft Actor-Critic:

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



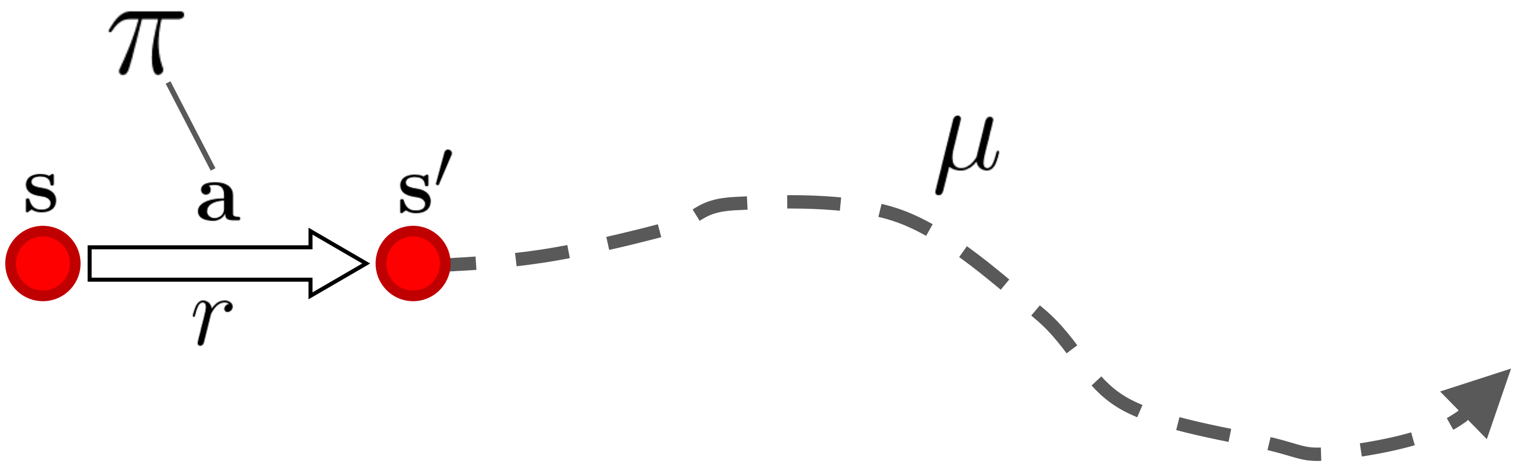
Surrogate Objective

Policy Gradient + Importance Sampling:

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

Soft Actor-Critic:

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



Policy Gradient + Importance Sampling

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

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

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

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 π is *close* to μ

$$D_{\mathrm{KL}}^{\max}(\mu,\pi) = \max_{\mathbf{s}}\; D_{\mathrm{KL}}\left(\mu(\cdot|\mathbf{s})||\pi(\cdot|\mathbf{s})\right)$$

Trust Region Policy Optimization
[Schulman et al. 2015]

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 $D_{\mathrm{KL}}^{\max}(\mu,\pi) \le \epsilon$,

$$J(\pi) \ge J^{\mu}(\pi) - \underbrace{C}_{\text{constant}}\epsilon$$

Trust Region Policy Optimization
[Schulman et al. 2015]

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

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

The surrogate objective is a lower bound on the real objective for sufficiently small ϵ!

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Optimization

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

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

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

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Optimization

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

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

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

Hard to compute

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Optimization

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

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

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

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Optimization

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

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

How do we pick μ?

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

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

$\pi_0 \leftarrow$ initialize policy

for iteration $k = 0, \ldots, n-1$ **do**

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

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

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

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

  Update policy:

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}} \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi^k(\cdot|\mathbf{s}) \| \pi(\cdot|\mathbf{s}) \right) \right] \leq \epsilon$$

end for

return policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

$\pi_0 \leftarrow$ initialize policy

for iteration $k = 0, ..., n-1$ **do**

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

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

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

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

  Update policy:

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}} \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}) \right) \right] \leq \epsilon$$

end for

return policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

1: $\pi_0 \leftarrow$ initialize policy

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

3: Sample trajectories τ^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:

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

8: **end for**

9: return policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

1: $\pi_0 \leftarrow$ initialize policy

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

3: Sample trajectories τ^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:

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}} \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi^k(\cdot|\mathbf{s}) \| \pi(\cdot|\mathbf{s}) \right) \right] \leq \epsilon$$

8: **end for**

9: return policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

$\pi_0 \leftarrow$ initialize policy

for iteration $k = 0, ..., n-1$ **do**

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

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

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

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

  Update policy:

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}} \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}) \right) \right] \leq \epsilon$$

end for

return policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

1: $\pi_0 \leftarrow$ initialize policy

2: **for** iteration $k = 0, ..., n-1$ **do**
3: Sample trajectories τ^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:

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

8: **end for**

9: **return** policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

1: $\pi_0 \leftarrow$ initialize policy

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

3: Sample trajectories τ^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:

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}}\left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a})\right]$$

$$\text{s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}}\left[D_{\mathrm{KL}}\left(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s})\right)\right] \leq \epsilon$$

8: **end for**

9: return policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

1: $\pi_0 \leftarrow$ initialize policy

2: **for** iteration $k = 0, ..., n-1$ **do**
3: Sample trajectories τ^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:

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

8: **end for**

9: return policy π^n

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

$\pi_0 \leftarrow$ initialize policy

for iteration $k = 0, ..., n-1$ **do**

  Sample trajectories τ^i from policy $\pi^k(\mathbf{a}|\mathbf{s})$
  Store trajectories in dataset $\mathcal{D} = \{\tau^i\}$
  Fit value function $V^k(\mathbf{s})$

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

  Update policy:

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}} \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}) \right) \right] \le \epsilon$$

end for

return policy π^n

Still need to collect a new batch of data every iteration

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Policy Optimization

ALGORITHM: Constrained Policy Optimization

1: $\pi_0 \leftarrow$ initialize policy

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

3: Sample trajectories τ^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:

$$\pi^{k+1} = \arg\max_{\pi} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}} \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\pi^k(\mathbf{a}|\mathbf{s})} A^k(\mathbf{s}, \mathbf{a}) \right]$$

$$\text{s.t. } \mathbb{E}_{\mathbf{s}\sim\mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi^k(\cdot|\mathbf{s}) || \pi(\cdot|\mathbf{s}) \right) \right] \leq \epsilon$$

8: **end for**

9: return policy π^n

Update policy with multiple grad steps

Trust Region Policy Optimization
[Schulman et al. 2015]

Constrained Optimization

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

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

How do we solve this?

Trust Region Policy Optimization (TRPO):

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

Trust Region Policy Optimization
[Schulman et al. 2015]

Lagrangian

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

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

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

"Lagrange multiplier"

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

$$\lambda \rightarrow \infty$$

constraint violated

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

constraint satisfied

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

Dual gradient descent:

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

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

Dual gradient descent:

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

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

Dual gradient descent:

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

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

Dual gradient descent:

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

Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

Lagrangian

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

Dual gradient descent:

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

Proximal Policy Optimization (PPO)

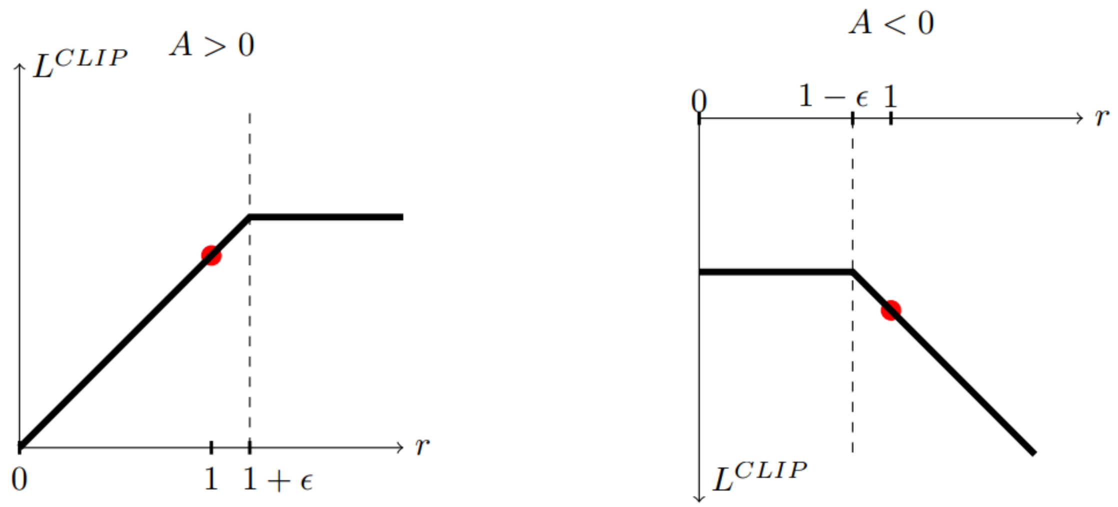
Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

PPO

In practice:

- Most PPO implementations use a clipping objective:

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



Proximal Policy Optimization Algorithms
[Schulman et al. 2017]

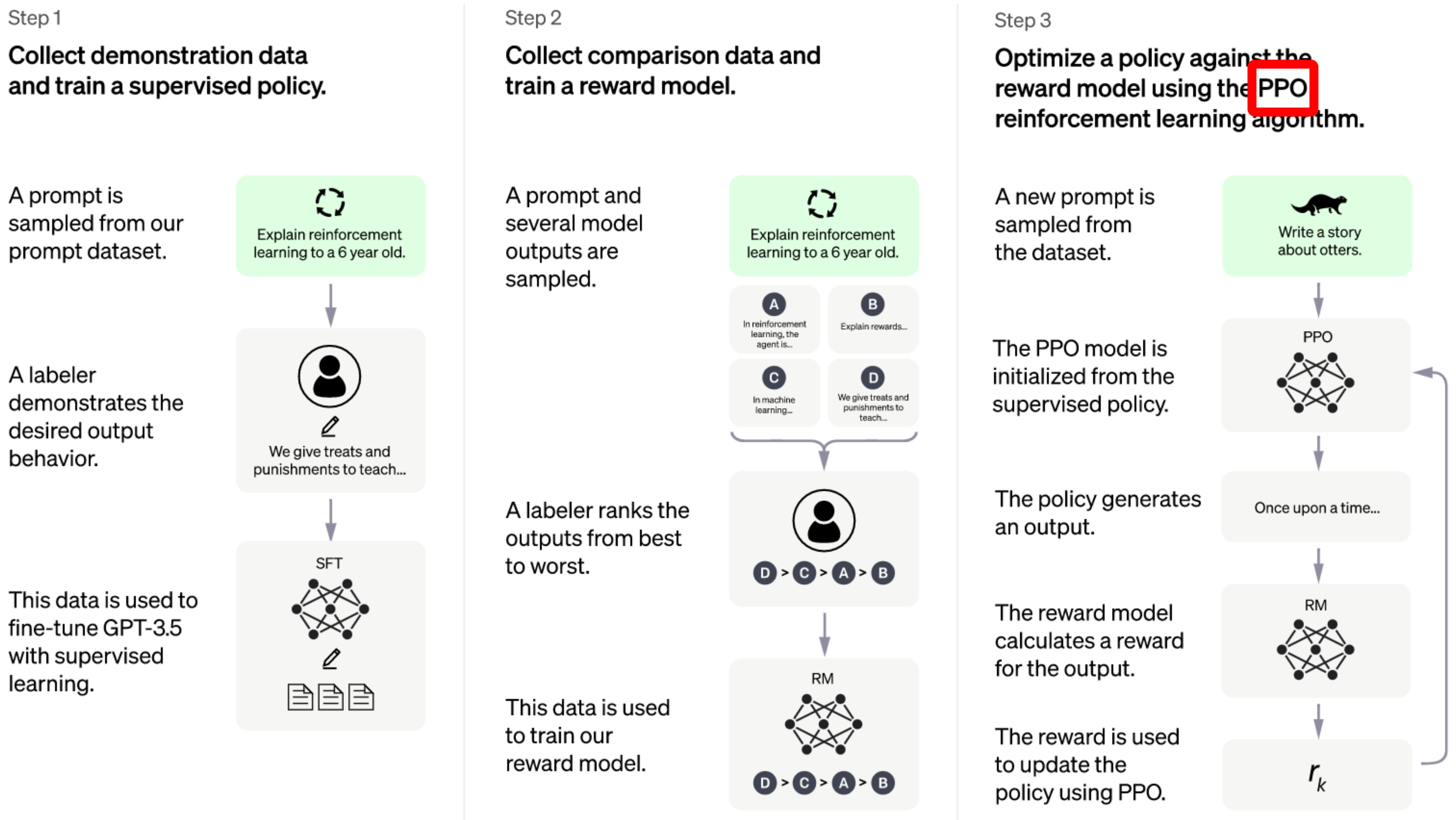
Robotic Locomotion

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



[OpenAI 2022]

Summary

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