11-695: Competitive Engineering
Reinforcement Learning

Spring 2018

Outline

Notations

Carnegie Mellon

Name	Notation	Meaning
Environment	$\mathcal{S}$	Set of states
	$\mathcal{A}$	Set of actions
	$p(s'\|s,a)$	State transition probabilities
	$s_0 \sim \rho_0$	Initial state and initial distribution
	$r(s,a)$	Reward function
	$\gamma \in [0,1]$	Discount factor
Policy	$\pi(a \in \mathcal{A}\|s)$	Distribution of actions given a state

Trajectory

$$\boxed{s_0 \sim \rho_0} \rightarrow \cdots \xrightarrow{a_{t+1} \sim \pi(a_t|s_t)} \boxed{s_{t+1} \sim p(s_{t+1}|s_t, a_t)} \rightarrow$$

Trajectory

Inifinite horizon trajectory $\tau \sim \pi$

$\tau = (s_0, a_0, s_1, a_1, ...)$, where:

$s_0 \sim \rho_0$

$a_t \sim \pi(s_t)$

$s_{t+1} \sim p(\cdot|s_t, a_t)$

Trajectory

$$\boxed{s_0 \sim \rho_0} \rightarrow \cdots \xrightarrow{a_{t+1} \sim \pi(a_t|s_t)} \boxed{s_{t+1} \sim p(s_{t+1}|s_t, a_t)} \rightarrow$$

Trajectory

Inifinite horizon trajectory $\quad \tau \sim \pi$

$\tau = (s_0, a_0, s_1, a_1, ...)$, where:

$s_0 \sim \rho_0$

$a_t \sim \pi(s_t)$

$s_{t+1} \sim p(\cdot|s_t, a_t)$

From the probability perspective: τ is a random variable.

Trajectory

$$\boxed{s_0 \sim \rho_0} \rightarrow \cdots \xrightarrow{a_{t+1} \sim \pi(a_t|s_t)} \boxed{s_{t+1} \sim p(s_{t+1}|s_t, a_t)} \rightarrow$$

Trajectory

Inifinite horizon trajectory $\quad \tau \sim \pi$

$\tau = (s_0, a_0, s_1, a_1, ...)$, where:

$s_0 \sim \rho_0$

$a_t \sim \pi(s_t)$

$s_{t+1} \sim p(\cdot|s_t, a_t)$

From the probability perspective: τ is a random variable.

Overloading notations: We will use $s_t(\tau)$ and $a_t(\tau)$ to denote *the* t^{th} *state of trajectory* τ and *the* t^{th} *action of trajectory* τ, respectively. In that case, $s_t(\tau)$ and $a_t(\tau)$ serve as random variables, just like τ

Probability and Likelihood of Trajectory

Carnegie Mellon

Probability and Likelihood of a Trajectory

Given *environment* and *policy* π, a trajectory $\tau = (s_0, a_0, s_1, a_1, ...)$ has probability

$$p_\pi(\tau) = p(s_0) \cdot \prod_{t=0}^{\infty} \Big(\pi(a_t|s_t) \cdot p(s_{t+1}|s_t, a_t)\Big) \tag{1}$$

Equivalently, τ has the likelihood

$$\log p_\pi(\tau) = \log p(s_0) + \sum_{t=0}^{\infty} \Big(\log \pi(a_t|s_t) + \log p(s_{t+1}|s_t, a_t)\Big) \tag{2}$$

Pseudo-Finite Horizon Trajectory

- In theory, a trajectory $\tau = (s_0, a_0, s_1, a_1, ...)$ can last forever
- In practice, we can only similuate finitely many steps of τ

$$\tau_T(s_0, a_0, s_1, a_1, ...) = (s_0, a_0, s_1, a_1, ..., s_T, a_T) \tag{3}$$

- Probability

$$p_\pi(\tau_T) = p(s_0) \cdot \prod_{t=0}^{T} \Big(\pi(a_t|s_t) \cdot p(s_{t+1}|s_t, a_t)\Big) \tag{4}$$

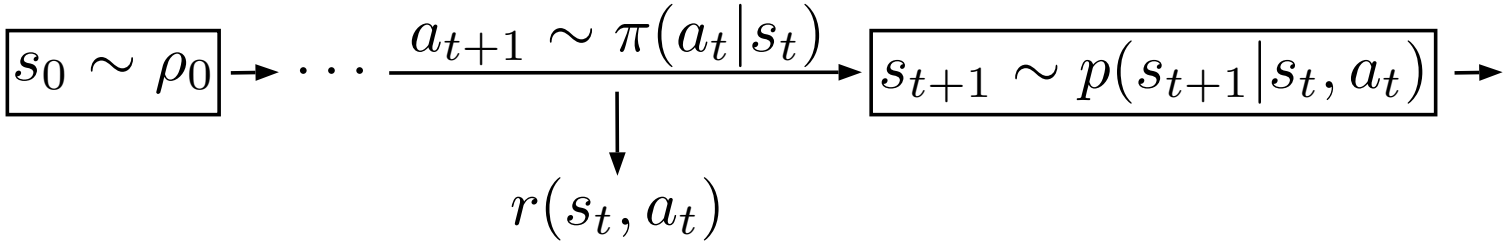
- Likelihood

$$\log p_\pi(\tau_T) = \log p(s_0) + \sum_{t=0}^{T} \Big(\log \pi(a_t|s_t) + \log p(s_{t+1}|s_t, a_t)\Big) \tag{5}$$

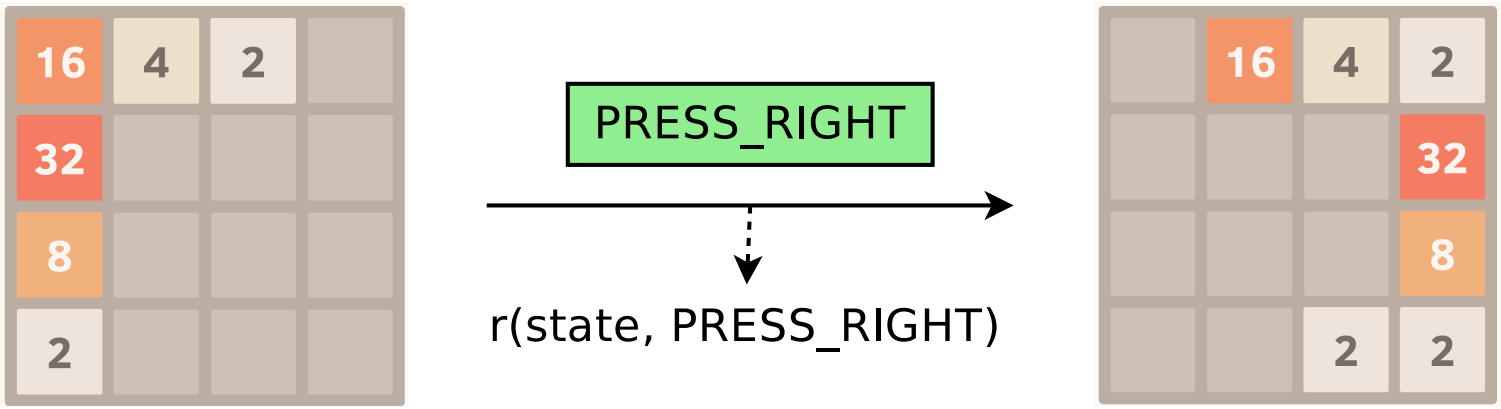
Reward

Reward

Instant reward	$r(s, a)$	Received upon taking action a at state s
Discount factor	γ	Further rewards count less
Discounted reward at T	$R_T(\tau)$	$\sum_{t=0}^{\infty} \gamma^t r(s_{T+t}, a_{T+t})$



Reward: An Example



RL Objective

Objective

Find a policy π^* that maximizes the expected discounted reward of all trajectories

$$\pi^* = \underset{\pi}{\text{argmax}}\, \mathbb{E}_{\tau \sim \pi} \left[R_0(\tau) \right] \tag{6}$$

where

$$R_0(\tau = (s_0, a_0, s_1, a_1, ...)) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \tag{7}$$

RL Objective so far

Three forms of RL expected discounted reward

Trajectory form	Discounted reward form
$\mathbb{E}_{\tau\sim\pi}\left[R_0(\tau)\right]$	$\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$

Outline

Carnegie Mellon

Episodic Policy Gradient

Carnegie Mellon

REINFORCE equation; episodic version [Williams, 1992]

Let $\pi(a|s;\theta)$ be a parameterized policy and let $J(\theta)$ be the RL objective

$$J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{8}$$

Then one can compute the gradient $\nabla_\theta J$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\cdot\nabla_\theta \log p_\pi(\tau)\Big] \tag{9}$$

Episodic Policy Gradient

REINFORCE equation; episodic version [Williams, 1992]

Let $\pi(a|s;\theta)$ be a parameterized policy and let $J(\theta)$ be the RL objective

$$J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{8}$$

Then one can compute the gradient $\nabla_\theta J$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\cdot\nabla_\theta \log p_\pi(\tau)\Big] \tag{9}$$

Why *episodic?* Because it has to see the whole *episode*, aka, *trajectory* τ

$$\nabla_\theta \log p_\pi(\tau) = \nabla_\theta\Big[\log\rho_0(s_0) + \sum_{t=0}^{\infty}\Big(\log p(s_{t+1}|s_t,a_t) + \log\pi(a_t|s_t;\theta)\Big)\Big] \tag{10}$$

$$= \sum_{t=0}^{\infty}\nabla_\theta \log\pi(a_t|s_t;\theta) \tag{11}$$

Derivation of REINFORCE

$$\nabla_\theta J(\theta) \tag{12}$$

$$\tag{13}$$

$$\tag{14}$$

$$\tag{15}$$

$$\tag{16}$$

$$\tag{17}$$

Derivation of REINFORCE

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{12}$$

(13)

(14)

(15)

(16)

(17)

Derivation of REINFORCE

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{12}$$

$$= \nabla_\theta \sum_\tau R_0(\tau) p_\pi(\tau) \tag{13}$$

(14)

(15)

(16)

(17)

Derivation of REINFORCE

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{12}$$

$$= \nabla_\theta \sum_\tau R_0(\tau) p_\pi(\tau) \tag{13}$$

$$= \sum_\tau R_0(\tau) \nabla_\theta p_\pi(\tau) \tag{14}$$

$$\tag{15}$$

$$\tag{16}$$

$$\tag{17}$$

Derivation of REINFORCE

Carnegie Mellon

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{12}$$

$$= \nabla_\theta \sum_\tau R_0(\tau) p_\pi(\tau) \tag{13}$$

$$= \sum_\tau R_0(\tau) \nabla_\theta p_\pi(\tau) \tag{14}$$

$$= \sum_\tau R_0(\tau) \cdot \frac{\nabla_\theta p_\pi(\tau)}{p_\pi(\tau)} \cdot p_\pi(\tau) \tag{15}$$

$$\tag{16}$$

$$\tag{17}$$

Derivation of REINFORCE

Carnegie Mellon

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{12}$$

$$= \nabla_\theta \sum_\tau R_0(\tau) p_\pi(\tau) \tag{13}$$

$$= \sum_\tau R_0(\tau) \nabla_\theta p_\pi(\tau) \tag{14}$$

$$= \sum_\tau R_0(\tau) \cdot \frac{\nabla_\theta p_\pi(\tau)}{p_\pi(\tau)} \cdot p_\pi(\tau) \tag{15}$$

$$= \sum_\tau R_0(\tau) \cdot \nabla_\theta \log p_\pi(\tau) \cdot p_\pi(\tau) \tag{16}$$

$$\tag{17}$$

Derivation of REINFORCE

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{12}$$

$$= \nabla_\theta \sum_\tau R_0(\tau) p_\pi(\tau) \tag{13}$$

$$= \sum_\tau R_0(\tau) \nabla_\theta p_\pi(\tau) \tag{14}$$

$$= \sum_\tau R_0(\tau) \cdot \frac{\nabla_\theta p_\pi(\tau)}{p_\pi(\tau)} \cdot p_\pi(\tau) \tag{15}$$

$$= \sum_\tau R_0(\tau) \cdot \nabla_\theta \log p_\pi(\tau) \cdot p_\pi(\tau) \tag{16}$$

$$= \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau) \cdot \nabla_\theta \log p_\pi(\tau)\Big] \tag{17}$$

Derivation of REINFORCE

Carnegie Mellon

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{12}$$

$$= \nabla_\theta \sum_\tau R_0(\tau) p_\pi(\tau) \tag{13}$$

$$= \sum_\tau R_0(\tau) \nabla_\theta p_\pi(\tau) \tag{14}$$

$$= \sum_\tau R_0(\tau) \cdot \frac{\nabla_\theta p_\pi(\tau)}{p_\pi(\tau)} \cdot p_\pi(\tau) \tag{15}$$

$$= \sum_\tau R_0(\tau) \cdot \nabla_\theta \log p_\pi(\tau) \cdot p_\pi(\tau) \tag{16}$$

$$= \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau) \cdot \nabla_\theta \log p_\pi(\tau)\Big] \tag{17}$$

REINFORCE doesn't care about **the reward function** $R_0(\tau)$**!**

A Deeper Look at $\nabla_\theta \log p_\pi(\tau)$

$$\nabla_\theta \log p_\pi(\tau) \tag{18}$$

$$\tag{19}$$

$$\tag{20}$$

$$\tag{21}$$

$$\tag{22}$$

A Deeper Look at $\nabla_\theta \log p_\pi(\tau)$

$$\nabla_\theta \log p_\pi(\tau) = \nabla_\theta \log \left(p(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t;\theta) \cdot p(s_{t+1}|s_t, a_t) \right) \tag{18}$$

(19)

(20)

(21)

(22)

A Deeper Look at $\nabla_\theta \log p_\pi(\tau)$

$$\nabla_\theta \log p_\pi(\tau) = \nabla_\theta \log \left(p(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t;\theta) \cdot p(s_{t+1}|s_t, a_t) \right) \tag{18}$$

$$= \nabla_\theta \left[\log p(s_0) + \sum_{t=0}^{\infty} \left(\log \pi(a_t|s_t;\theta) + \log p(s_{t+1}|s_t, a_t) \right) \right] \tag{19}$$

(20)

(21)

(22)

A Deeper Look at $\nabla_\theta \log p_\pi(\tau)$

$$\nabla_\theta \log p_\pi(\tau) = \nabla_\theta \log \left(p(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t;\theta) \cdot p(s_{t+1}|s_t,a_t) \right) \tag{18}$$

$$= \nabla_\theta \left[\log p(s_0) + \sum_{t=0}^{\infty} \left(\log \pi(a_t|s_t;\theta) + \log p(s_{t+1}|s_t,a_t) \right) \right] \tag{19}$$

$$= \nabla_\theta \log p(s_0) + \sum_{t=0}^{\infty} \left(\nabla_\theta \log \pi(a_t|s_t;\theta) + \nabla_\theta \log p(s_{t+1}|s_t,a_t) \right) \tag{20}$$

$$\tag{21}$$

$$\tag{22}$$

A Deeper Look at $\nabla_\theta \log p_\pi(\tau)$

$$\nabla_\theta \log p_\pi(\tau) = \nabla_\theta \log \left(p(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t;\theta) \cdot p(s_{t+1}|s_t,a_t) \right) \tag{18}$$

$$= \nabla_\theta \left[\log p(s_0) + \sum_{t=0}^{\infty} \left(\log \pi(a_t|s_t;\theta) + \log p(s_{t+1}|s_t,a_t) \right) \right] \tag{19}$$

$$= \nabla_\theta \log p(s_0) + \sum_{t=0}^{\infty} \left(\nabla_\theta \log \pi(a_t|s_t;\theta) + \nabla_\theta \log p(s_{t+1}|s_t,a_t) \right) \tag{20}$$

$$= \sum_{t=0}^{\infty} \nabla_\theta \log \pi(a_t|s_t;\theta) \tag{21}$$

$$\tag{22}$$

A Deeper Look at $\nabla_\theta \log p_\pi(\tau)$

$$\nabla_\theta \log p_\pi(\tau) = \nabla_\theta \log \left(p(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t;\theta) \cdot p(s_{t+1}|s_t, a_t) \right) \tag{18}$$

$$= \nabla_\theta \left[\log p(s_0) + \sum_{t=0}^{\infty} \left(\log \pi(a_t|s_t;\theta) + \log p(s_{t+1}|s_t, a_t) \right) \right] \tag{19}$$

$$= \nabla_\theta \log p(s_0) + \sum_{t=0}^{\infty} \left(\nabla_\theta \log \pi(a_t|s_t;\theta) + \nabla_\theta \log p(s_{t+1}|s_t, a_t) \right) \tag{20}$$

$$= \sum_{t=0}^{\infty} \nabla_\theta \log \pi(a_t|s_t;\theta) \tag{21}$$

$$\tag{22}$$

REINFORCE doesn't care about **the environment!**

REINFORCE for Temporal Structures

Carnegie Mellon

Temporal Structure: Stuffs that involves multiple steps (t in equations)

REINFORCE for Temporal Structures

Temporal Structure: Stuffs that involves multiple steps (t in equations)

We have the following equation (proof skipped for now)

$$\nabla_\theta \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R_t(\tau)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] \tag{23}$$

REINFORCE for Temporal Structures

Carnegie Mellon

Temporal Structure: Stuffs that involves multiple steps (t in equations)

We have the following equation (proof skipped for now)

$$\nabla_\theta \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R_t(\tau)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] \tag{23}$$

REINFORCE for Temporal Structures

Carnegie Mellon

Put together so far

$$\nabla_\theta \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\Big] \tag{24}$$

$$= \mathbb{E}_{\tau\sim\pi}\Big[R_0(\tau)\sum_{t=0}^{\infty}\nabla_\theta \log \pi(a_t|s_t;\theta)\Big] \tag{25}$$

$$= \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R_t(\tau)\nabla_\theta \log \pi(a_t|s_t;\theta)\Big] \tag{26}$$

Monte Carlo Episodic Policy Gradient

Monte Carlo Episodic Policy Gradient

REINFORCE (Eqn 26) induces a method to estimate policy gradient

$$\nabla_\theta \hat{J}(\theta) = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t|s_t;\theta)\Big] \tag{27}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty} \gamma^t R_t(\tau_i) \nabla_\theta \log \pi(a_t(\tau_i)|s_t(\tau_i);\theta) \tag{28}$$

Eqn 28, albeit useful, has high variance and high sample complexity.

Baseline

To reduce the high variance in the previous slide.

(29)

Baseline

To reduce the high variance in the previous slide.

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R_t(\tau)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t \big(R_t(\tau) - b(s_t)\big)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] \tag{29}$$

Baseline

To reduce the high variance in the previous slide.

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R_t(\tau)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t \Big(R_t(\tau) - b(s_t)\Big)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] \tag{29}$$

Proof. skip for now :(

Baseline

To reduce the high variance in the previous slide.

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R_t(\tau)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t \Big(R_t(\tau) - b(s_t)\Big)\nabla_\theta \log\pi(a_t|s_t;\theta)\Big] \tag{29}$$

Proof. skip for now :(

Therefore

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}\gamma^t \Big(R_t(\tau_i) - b(s_t(\tau_i))\Big)\nabla_\theta \log\pi(a_t(\tau_i)|s_t(\tau_i);\theta) \tag{30}$$

Summary of REINFORCE

Carnegie Mellon

REINFORCE algorithm

1. Initialize $\pi(a|s;\theta)$
2. Repeat until converge
   - Sample N trajectories $\tau_i \sim \pi$
   - Compute $\nabla_\theta \hat{J}(\theta)$

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{\infty} \gamma^t \big(R_t(\tau_i) - b(s_t(\tau_i))\big) \nabla_\theta \log \pi(a_t(\tau_i)|s_t(\tau_i);\theta)$$

   - Update $b\big(s_t(\tau_i)\big)$ using ℓ_2 loss $\|b(s_t(\tau_i)) - R_t(\tau_i)\|^2$ or moving averages
   - Update $\pi(a|s;\theta)$

Notations

Carnegie Mellon

Name	Notation	Meaning
Value	$V_\pi(s)$	Expected discounted reward from state s $V_\pi(s) = \mathbb{E}_{a\sim\pi(a\|s)}\Big[r(s,a) + \gamma\mathbb{E}_{s'\sim p(s'\|s,a)}\Big[V_\pi(s')\Big]\Big]$
Q-value	$Q_\pi(s,a)$	Expected discounted reward if take action a at state s $Q_\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim p(s'\|s,a)}\Big[V_\pi(s')\Big]$
Advantage	$A_\pi(s,a)$	How better is the action than average $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$

Problems of REINFORCE

Recall from REINFORCE. For each trajectory τ, we do this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{31}$$

Problems of REINFORCE

Recall from REINFORCE. For each trajectory τ, we do this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{31}$$

Problems:

Problems of REINFORCE

Recall from REINFORCE. For each trajectory τ, we do this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau); \theta) \tag{31}$$

Problems:

- The ∞. No update until τ finishes.

Problems of REINFORCE

Recall from REINFORCE. For each trajectory τ, we do this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{31}$$

Problems:

- The ∞. No update until τ finishes.
- The rumors have it: high variance.

Problems of REINFORCE

Recall from REINFORCE. For each trajectory τ, we do this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{31}$$

Problems:

- The ∞. No update until τ finishes.
- The rumors have it: high variance.
- The rumors have it: high sample complexity.

Actor-Critic [Sutton et al., 1999]

A solution: fix this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{32}$$

(33)

Actor-Critic [Sutton et al., 1999]

A solution: fix this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{32}$$

$$\approx \sum_{t=0}^{T} \gamma^t Q_\pi(s_t, a_t) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{33}$$

Actor-Critic [Sutton et al., 1999]

A solution: fix this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{32}$$

$$\approx \sum_{t=0}^{T} \gamma^t Q_\pi(s_t, a_t) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{33}$$

How does this help? $Q_\pi(s_t, a_t)$ does **not** need τ to finish.

Actor-Critic [Sutton et al., 1999]

A solution: fix this

$$\sum_{t=0}^{\infty} \gamma^t R_t(\tau) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{32}$$

$$\approx \sum_{t=0}^{T} \gamma^t Q_\pi(s_t, a_t) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau);\theta) \tag{33}$$

How does this help? $Q_\pi(s_t, a_t)$ does **not** need τ to finish.

$\rightarrow$ Can update θ after simulating τ for T steps.

Actor-Critic [Sutton et al., 1999]

Something's wrong! How do we know $Q_\pi(s_t, a_t)$?

Actor-Critic [Sutton et al., 1999]

Something's wrong! How do we know $Q_\pi(s_t, a_t)$?

We can approximiate $Q_\pi(s, a)$ with

$$\hat{Q}(s, a; \theta_Q) \approx Q_\pi(s, a) \tag{34}$$

Actor-Critic [Sutton et al., 1999]

Something's wrong! How do we know $Q_\pi(s_t, a_t)$?
We can approximiate $Q_\pi(s, a)$ with

$$\hat{Q}(s, a; \theta_Q) \approx Q_\pi(s, a) \tag{34}$$

Plug it in

$$\sum_{t=0}^{T} \gamma^t \hat{Q}(s_t, a_t; \theta_Q) \nabla_\theta \log \pi(a_t(\tau)|s_t(\tau); \theta) \tag{35}$$

Actor-Critic [Sutton et al., 1999]

Something's wrong! How do we know $Q_\pi(s_t, a_t)$?
We can approximiate $Q_\pi(s, a)$ with

$$\hat{Q}(s, a; \theta_Q) \approx Q_\pi(s, a) \tag{34}$$

Plug it in

$$\sum_{t=0}^{T} \gamma^t \hat{Q}(s_t, a_t; \theta_Q) \nabla_\theta \log \pi(a_t(\tau) | s_t(\tau); \theta) \tag{35}$$

How do we estimate θ_Q? Simulate τ and use ℓ_2 error; Q-learning, etc.

Actor-Critic [Sutton et al., 1999]

Carnegie Mellon

Actor-Critic Algorithm

1. Initialize $\pi(a, s; \theta)$ and $Q(s, a; \theta_Q)$
2. Repeat until convergence
   - Sample N trajectories τ_i, each for T steps
   - Compute

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \hat{Q}(s_t, a_t; \theta_Q) \nabla_\theta \log \pi(a_t | s_t; \theta) \tag{36}$$

   - Update θ using $\nabla_\theta \hat{J}(\theta)$
   - Update θ_Q using $\sum_{i=1}^{N} \sum_{t=0}^{T} \left\| Q(s_t, a_t; \theta_Q) - R_t(\tau_i) \right\|^2$

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Actor-Critic estimates ∇_θ for each trajectory τ as follows

$$\sum_{t=0}^{T} \gamma^t \hat{Q}(s_t, a_t; \theta_Q) \nabla_\theta \log \pi(a_t|s_t; \theta) \tag{37}$$

Problems:

- ~~The ∞. No update until τ finishes.~~
- The rumors have it: high variance.
- The rumors have it: high sample complexity.

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Solution: back to the baseline form of REINFORCE

(38)

(39)

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Solution: back to the baseline form of REINFORCE

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\big(R_t(\tau)-b(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{38}$$

(39)

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Solution: back to the baseline form of REINFORCE

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\big(R_t(\tau)-b(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{38}$$

$$\approx\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{T}\gamma^t\big(\hat{R}_t-V_\pi(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{39}$$

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Solution: back to the baseline form of REINFORCE

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\big(R_t(\tau)-b(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{38}$$

$$\approx \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{T}\gamma^t\big(\hat{R}_t-V_\pi(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{39}$$

Same as Actor-Critic, we will parameterize something. This time

$$\hat{V}(s;\theta_V)\approx V_\pi(s) \tag{40}$$

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Solution: back to the baseline form of REINFORCE

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\big(R_t(\tau)-b(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{38}$$

$$\approx \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{T}\gamma^t\big(\hat{R}_t-V_\pi(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{39}$$

Same as Actor-Critic, we will parameterize something. This time

$$\hat{V}(s;\theta_V)\approx V_\pi(s) \tag{40}$$

With $V_\pi(s)$ known, one can estimate $\hat{R}_t$ the following quantity with n-steps sampling from t

$$\hat{R}_t = r(s_t,a_t)+\gamma r(s_{t+1},a_{t+1})+\cdots+\gamma^{n-1}r(s_{t+n-1},a_{t+n-1})+\gamma^{t+n}V(s_{t+n};\theta_V) \tag{41}$$

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Solution: back to the baseline form of REINFORCE

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t\big(R_t(\tau)-b(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{38}$$

$$\approx \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{T}\gamma^t\big(\hat{R}_t-V_\pi(s_t)\big)\nabla_\theta\log\pi(a_t|s_t;\theta)\Big] \tag{39}$$

Same as Actor-Critic, we will parameterize something. This time

$$\hat{V}(s;\theta_V)\approx V_\pi(s) \tag{40}$$

With $V_\pi(s)$ known, one can estimate $\hat{R}_t$ the following quantity with n-steps sampling from t

$$\hat{R}_t = r(s_t,a_t)+\gamma r(s_{t+1},a_{t+1})+\cdots+\gamma^{n-1}r(s_{t+n-1},a_{t+n-1})+\gamma^{t+n}V(s_{t+n};\theta_V) \tag{41}$$

How do we train θ_V? Small ℓ_2 estimate error in Eqn 41 to $V(s_t;\theta_V)$.

Advantage Actor-Critic [Mnih et al., 2016]

Carnegie Mellon

Advantage Actor-Critic Algorithm

1. Initialize $\pi(a, s; \theta)$ and $V(s, a; \theta_V)$
2. Repeat until convergence
   - Sample N trajectories τ_i
   - For each state s_t in any trajectory τ_i, estimate using n-step sampling

$$\hat{R}_t(\tau_i) = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \cdots + \gamma^{n-1} r(s_{t+n-1}, a_{t+n-1}) + \gamma^{t+n} V(s_{t+n}; \theta_V) \tag{42}$$

   - Compute

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \big(\hat{R}_t - V(s; \theta_V)\big) \nabla_\theta \log \pi(a_t | s_t; \theta) \tag{43}$$

   - Update θ using $\nabla_\theta \hat{J}(\theta)$
   - Update θ_V using $\sum_{i=1}^{N} \sum_{t=0}^{T} \left\| \hat{R}_t(\tau_i) - V(s_t(\tau_i); \theta_V) \right\|^2$

Summary of Actor-Critic Methods

Actor-Critic [Sutton et al., 1999]

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \hat{Q}(s_t, a_t; \theta_Q) \nabla_\theta \log \pi(a_t|s_t; \theta) \tag{44}$$

$$\mathcal{L}(\theta_Q) = \sum_{i=1}^{N} \sum_{t=0}^{T} \|Q(s_t, a_t; \theta_Q) - R_t(\tau_i)\|^2 \tag{45}$$

Summary of Actor-Critic Methods

Carnegie Mellon

Actor-Critic [Sutton et al., 1999]

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \hat{Q}(s_t, a_t; \theta_Q) \nabla_\theta \log \pi(a_t|s_t; \theta) \tag{44}$$

$$\mathcal{L}(\theta_Q) = \sum_{i=1}^{N} \sum_{t=0}^{T} \|Q(s_t, a_t; \theta_Q) - R_t(\tau_i)\|^2 \tag{45}$$

Advantage Actor-Critic [Mnih et al., 2016]

$$\hat{R}_t(\tau_i) = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \cdots + \gamma^{n-1} r(s_{t+n-1}, a_{t+n-1}) + \gamma^{t+n} V(s_{t+n}; \theta_V) \tag{46}$$

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \left(\hat{R}_t(\tau_i) - V(s; \theta_V) \right) \nabla_\theta \log \pi(a_t|s_t; \theta) \tag{47}$$

$$\mathcal{L}(\theta_V) = \sum_{i=1}^{N} \sum_{t=0}^{T} \left\| \hat{R}_t(\tau_i) - V(s_t(\tau_i); \theta_V) \right\|^2 \tag{48}$$

Summary of Actor-Critic Methods

Carnegie Mellon

Actor-Critic [Sutton et al., 1999]

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \hat{Q}(s_t, a_t; \theta_Q) \nabla_\theta \log \pi(a_t|s_t; \theta) \tag{44}$$

$$\mathcal{L}(\theta_Q) = \sum_{i=1}^{N} \sum_{t=0}^{T} \left\| Q(s_t, a_t; \theta_Q) - R_t(\tau_i) \right\|^2 \tag{45}$$

Advantage Actor-Critic [Mnih et al., 2016]

$$\hat{R}_t(\tau_i) = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \cdots + \gamma^{n-1} r(s_{t+n-1}, a_{t+n-1}) + \gamma^{t+n} V(s_{t+n}; \theta_V) \tag{46}$$

$$\nabla_\theta \hat{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \left(\hat{R}_t(\tau_i) - V(s; \theta_V) \right) \nabla_\theta \log \pi(a_t|s_t; \theta) \tag{47}$$

$$\mathcal{L}(\theta_V) = \sum_{i=1}^{N} \sum_{t=0}^{T} \left\| \hat{R}_t(\tau_i) - V(s_t(\tau_i); \theta_V) \right\|^2 \tag{48}$$

Etymology:

Name	Meaning	Role
Actor	The policy π	Samples trajectories
Critic	$Q(s, a; \theta_Q)$ or $V(s; \theta_V)$	Scales the temporal gradient $\nabla_\theta \log \pi(a_t\|s_t; \theta)$
Advantage	$\hat{R}_t(\tau) - V(s; \theta_V)$	Estimate $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$

Volodymyr Mnih, Adri Puigdomnech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 1999.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992.