Value

CISC 7404 - Decision Making

Steven Morad

University of Macau

Review

Policy-Conditioned Returns

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Guaranteed optimal policy, given infinite compute

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Guaranteed optimal policy, given infinite compute

We must make approximations to implement trajectory optimization

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Guaranteed optimal policy, given infinite compute

We must make approximations to implement trajectory optimization

These approximations break optimality guarantees

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Guaranteed optimal policy, given infinite compute

We must make approximations to implement trajectory optimization

These approximations break optimality guarantees

Today, we will look at new algorithms based on the notion of **value**

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Guaranteed optimal policy, given infinite compute

We must make approximations to implement trajectory optimization

These approximations break optimality guarantees

Today, we will look at new algorithms based on the notion of **value**

Uses fewer approximations and can achieve optimal policy

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Guaranteed optimal policy, given infinite compute

We must make approximations to implement trajectory optimization

These approximations break optimality guarantees

Today, we will look at new algorithms based on the notion of **value**

Uses fewer approximations and can achieve optimal policy

Can model infinitely long returns

Policy-Conditioned Returns

Trajectory optimization is model-based algorithm

Guaranteed optimal policy, given infinite compute

We must make approximations to implement trajectory optimization

These approximations break optimality guarantees

Today, we will look at new algorithms based on the notion of **value**

Uses fewer approximations and can achieve optimal policy

Can model infinitely long returns

Expensive to train, but very cheap to use

Policy-Conditioned Returns

Recall the return from trajectory optimization

Policy-Conditioned Returns

Recall the return from trajectory optimization

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Policy-Conditioned Returns

Recall the return from trajectory optimization

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

This is an **action-conditioned** discounted return

Policy-Conditioned Returns

Recall the return from trajectory optimization

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

This is an **action-conditioned** discounted return

Conditioned/dependent on a sequence of actions

Policy-Conditioned Returns

Recall the return from trajectory optimization

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, \ldots] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, \ldots]$$

This is an **action-conditioned** discounted return

Conditioned/dependent on a sequence of actions

There is no structure to the actions

Policy-Conditioned Returns

Recall the return from trajectory optimization

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

This is an **action-conditioned** discounted return

Conditioned/dependent on a sequence of actions

There is no structure to the actions

- Random

Policy-Conditioned Returns

Recall the return from trajectory optimization

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

This is an **action-conditioned** discounted return

Conditioned/dependent on a sequence of actions

There is no structure to the actions

- Random
- Picked by humans

Policy-Conditioned Returns

Recall the return from trajectory optimization

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

This is an **action-conditioned** discounted return

Conditioned/dependent on a sequence of actions

There is no structure to the actions

- Random
- Picked by humans
- Maximize $\mathcal{G}$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Last time, we introduced the policy

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Last time, we introduced the policy

$$\pi : S \times \Theta \mapsto \Delta A$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, \ldots] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, \ldots]$$

Last time, we introduced the policy

$$\pi : S \times \Theta \mapsto \Delta A$$

Example policy, greedy policy

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Last time, we introduced the policy

$$\pi : S \times \Theta \mapsto \Delta A$$

Example policy, greedy policy

$$\pi(a_t \mid s_t; \theta_\pi) = \begin{cases} 1 \text{ if } a_t = \arg\max_{a_t \in A} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] \\ 0 \text{ otherwise} \end{cases}$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Last time, we introduced the policy

$$\pi : S \times \Theta \mapsto \Delta A$$

Example policy, greedy policy

$$\pi(a_t \mid s_t; \theta_\pi) = \begin{cases} 1 \text{ if } a_t = \arg\max_{a_t \in A} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] \\ 0 \text{ otherwise} \end{cases}$$

Must construct and evaluate decision tree at each timestep!

Policy-Conditioned Returns

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Policy-Conditioned Returns

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$\pi : S \times \Theta \mapsto \Delta A$$

Policy-Conditioned Returns

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$\pi : S \times \Theta \mapsto \Delta A$$

Conditioning the return on actions is annoying

Policy-Conditioned Returns

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$\pi : S \times \Theta \mapsto \Delta A$$

Conditioning the return on actions is annoying

Must compute infinitely many actions and outcomes for the return

Policy-Conditioned Returns

$$[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$\pi : S \times \Theta \mapsto \Delta A$$

Conditioning the return on actions is annoying

Must compute infinitely many actions and outcomes for the return

What if we condition on a policy, instead of specific actions?

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$a_0 \sim \pi(\cdot \mid s_0; \theta_\pi), \quad a_1 \sim \pi(\cdot \mid s_1; \theta_\pi), \quad a_2 \sim \pi(\cdot \mid s_2; \theta_\pi), \quad ...$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$a_0 \sim \pi(\cdot \mid s_0; \theta_\pi), \quad a_1 \sim \pi(\cdot \mid s_1; \theta_\pi), \quad a_2 \sim \pi(\cdot \mid s_2; \theta_\pi), \quad ...$$

Condition on distribution parameterized by θ_π instead of many actions

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$a_0 \sim \pi(\cdot \mid s_0; \theta_\pi), \quad a_1 \sim \pi(\cdot \mid s_1; \theta_\pi), \quad a_2 \sim \pi(\cdot \mid s_2; \theta_\pi), \quad ...$$

Condition on distribution parameterized by θ_π instead of many actions

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

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

$$a_0 \sim \pi(\cdot \mid s_0; \theta_\pi), \quad a_1 \sim \pi(\cdot \mid s_1; \theta_\pi), \quad a_2 \sim \pi(\cdot \mid s_2; \theta_\pi), \quad ...$$

Condition on distribution parameterized by θ_π instead of many actions

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

Remember, $\pi(a \mid s; \theta_\pi)$ provides a distribution over the action space

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

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

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

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

Now, return conditioned on the policy with θ_π

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

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

Now, return conditioned on the policy with θ_π

But remember, $\mathcal{R}(s_{t+1})$ hides the magic

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, a_0, a_1, ...]$$

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

Now, return conditioned on the policy with θ_π

But remember, $\mathcal{R}(s_{t+1})$ hides the magic

How does $\mathbb{E}[\mathcal{R}(s_{t+1})]$ change when we condition on θ_π?

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, \ldots] = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+1}) \sum_{s_{t+1} \in S} \Pr(s_{t+1} \mid s_0, a_0, \ldots, a_t)$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+1}) \sum_{s_{t+1} \in S} \Pr(s_{t+1} \mid s_0, a_0, ..., a_t)$$

Question: What changes when we condition on θ_π?

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+1}) \sum_{s_{t+1} \in S} \Pr(s_{t+1} \mid s_0, a_0, ..., a_t)$$

Question: What changes when we condition on θ_π?

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+1}) \sum_{s_{t+1} \in S} \Pr(s_{t+1} \mid s_0, a_0, ..., a_t)$$

Question: What changes when we condition on θ_π?

$$\Pr(s_{t+1} \mid s_0, a_0, ..., a_t) \Rightarrow \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+1}) \sum_{s_{t+1} \in S} \Pr(s_{t+1} \mid s_0, a_0, ..., a_t)$$

Question: What changes when we condition on θ_π?

$$\Pr(s_{t+1} \mid s_0, a_0, ..., a_t) \Rightarrow \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Maybe we can use $\Pr(s_{t+1} \mid s_t, a_t)$ to figure this out

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+1}) \sum_{s_{t+1} \in S} \Pr(s_{t+1} \mid s_0, a_0, ..., a_t)$$

Question: What changes when we condition on θ_π?

$$\Pr(s_{t+1} \mid s_0, a_0, ..., a_t) \Rightarrow \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Maybe we can use $\Pr(s_{t+1} \mid s_t, a_t)$ to figure this out

Question: What was $\Pr(s_{t+1} \mid s_t, a_t)$?

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...] = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_{t+1}) \sum_{s_{t+1} \in S} \Pr(s_{t+1} \mid s_0, a_0, ..., a_t)$$

Question: What changes when we condition on θ_π?

$$\Pr(s_{t+1} \mid s_0, a_0, ..., a_t) \Rightarrow \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Maybe we can use $\Pr(s_{t+1} \mid s_t, a_t)$ to figure this out

Question: What was $\Pr(s_{t+1} \mid s_t, a_t)$?

Answer: State transition function

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Policy-Conditioned Returns

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Policy-Conditioned Returns

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Issue: State transition function needs an action a_t

Policy-Conditioned Returns

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Issue: State transition function needs an action a_t

Policy π outputs a distribution over the action space

Policy-Conditioned Returns

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Issue: State transition function needs an action a_t

Policy π outputs a distribution over the action space

Question: What is $\Pr(s_{t+1} \mid s_t, \theta_\pi)$?

Policy-Conditioned Returns

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Issue: State transition function needs an action a_t

Policy π outputs a distribution over the action space

Question: What is $\Pr(s_{t+1} \mid s_t, \theta_\pi)$? Hint: Consider all possible actions

Policy-Conditioned Returns

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Issue: State transition function needs an action a_t

Policy π outputs a distribution over the action space

Question: What is $\Pr(s_{t+1} \mid s_t, \theta_\pi)$? Hint: Consider all possible actions

$$\Pr(s_{t+1} \mid s_t; \theta_\pi) = \sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

Policy-Conditioned Returns

$$\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$$

Issue: State transition function needs an action a_t

Policy π outputs a distribution over the action space

Question: What is $\Pr(s_{t+1} \mid s_t, \theta_\pi)$? Hint: Consider all possible actions

$$\Pr(s_{t+1} \mid s_t; \theta_\pi) = \sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

Combine the policy distribution with next state distribution

Policy-Conditioned Returns

$$\Pr(s_{t+1} \mid s_t; \theta_\pi) = \sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

Policy-Conditioned Returns

$$\Pr(s_{t+1} \mid s_t; \theta_\pi) = \sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

Write out the first few timesteps

Policy-Conditioned Returns

$$\Pr(s_{t+1} \mid s_t; \theta_\pi) = \sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

Write out the first few timesteps

$$\Pr(s_1 \mid s_0; \theta_\pi) = \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cdot \pi(a_0 \mid s_0; \theta_\pi)$$

Policy-Conditioned Returns

$$\Pr(s_{t+1} \mid s_t; \theta_\pi) = \sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

Write out the first few timesteps

$$\Pr(s_1 \mid s_0; \theta_\pi) = \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cdot \pi(a_0 \mid s_0; \theta_\pi)$$

$$\Pr(s_2 \mid s_0; \theta_\pi) = \sum_{s_1 \in S} \sum_{a_1 \in A} \mathrm{Tr}(s_2 \mid s_1, a_1) \cdot \pi(a_1 \mid s_1; \theta_\pi)$$

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

Policy-Conditioned Returns

$$\Pr(s_1 \mid s_0; \theta_\pi) = \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cdot \pi(a_0 \mid s_0; \theta_\pi)$$

$$\Pr(s_2 \mid s_0; \theta_\pi) = \sum_{s_1 \in S} \sum_{a_1 \in A} \mathrm{Tr}(s_2 \mid s_1, a_1) \cdot \pi(a_1 \mid s_1; \theta_\pi)$$
$$\cdot \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cdot \pi(a_0 \mid s_0; \theta_\pi)$$

Policy-Conditioned Returns

$$\Pr(s_1 \mid s_0; \theta_\pi) = \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cdot \pi(a_0 \mid s_0; \theta_\pi)$$

$$\Pr(s_2 \mid s_0; \theta_\pi) = \sum_{s_1 \in S} \sum_{a_1 \in A} \mathrm{Tr}(s_2 \mid s_1, a_1) \cdot \pi(a_1 \mid s_1; \theta_\pi) \cdot \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cdot \pi(a_0 \mid s_0; \theta_\pi)$$

Derive a general form for $\Pr(s_{n+1} \mid s_0; \theta_\pi)$

Policy-Conditioned Returns

$$\Pr(s_1 \mid s_0; \theta_\pi) = \sum_{a_0 \in A} \mathrm{Tr}(s_1 \mid s_0, a_0) \cdot \pi(a_0 \mid s_0; \theta_\pi)$$

$$\Pr(s_2 \mid s_0; \theta_\pi) = \sum_{s_1 \in S} \sum_{a_1 \in A} \mathrm{Tr}(s_2 \mid s_1, a_1) \cdot \pi(a_1 \mid s_1; \theta_\pi)$$

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

Derive a general form for $\Pr(s_{n+1} \mid s_0; \theta_\pi)$

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \dots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

Policy-Conditioned Returns

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

Policy-Conditioned Returns

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

Plug back into our expected reward

Policy-Conditioned Returns

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

Plug back into our expected reward

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Policy-Conditioned Returns

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

Plug back into our expected reward

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Need to plug expected reward back into expected discounted return

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Discounted return is discounted sum of rewards

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Discounted return is discounted sum of rewards

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

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Discounted return is discounted sum of rewards

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Discounted return is discounted sum of rewards

$$\begin{aligned}\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = & \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi) \\ & + \gamma \sum_{s_2 \in S} \mathcal{R}(s_2) \cdot \Pr(s_2 \mid s_0; \theta_\pi)\end{aligned}$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Discounted return is discounted sum of rewards

$$\begin{aligned} \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = & \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi) \\ & + \gamma \sum_{s_2 \in S} \mathcal{R}(s_2) \cdot \Pr(s_2 \mid s_0; \theta_\pi) \\ & + \gamma^2 \sum_{s_3 \in S} \mathcal{R}(s_3) \cdot \Pr(s_3 \mid s_0; \theta_\pi) \end{aligned}$$

Policy-Conditioned Returns

$$\mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0; \theta_\pi] = \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Discounted return is discounted sum of rewards

$$\begin{aligned}\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = & \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi) \\ & + \gamma \sum_{s_2 \in S} \mathcal{R}(s_2) \cdot \Pr(s_2 \mid s_0; \theta_\pi) \\ & + \gamma^2 \sum_{s_3 \in S} \mathcal{R}(s_3) \cdot \Pr(s_3 \mid s_0; \theta_\pi) \\ & \dots\end{aligned}$$

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

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

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

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

Where the state distribution is

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

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

Where the state distribution is

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \operatorname{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

Value Functions

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

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

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

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

These two equations form the basis of all reinforcement learning

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \dots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

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

These two equations form the basis of all reinforcement learning

- DQN
- DDPG/SAC
- A3C/PPO/GRPO

Goal: find the θ_π (policy parameters) to maximize the expected return

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

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

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

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

We have another name for $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

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

We have another name for $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

We call it the **value function** $V : S \times \Theta \mapsto \mathbb{R}$

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{s_1, \dots, s_n \in S} \prod_{t=0}^{n} \left(\sum_{a_t \in A} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta) \right)$$

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

We have another name for $\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi]$

We call it the **value function** $V : S \times \Theta \mapsto \mathbb{R}$

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240\ \mathrm{km/h}$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240 \text{ km/h}$ $\theta_\pi = \text{Race car driver}$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240 \text{ km/h}$ $\theta_\pi = \text{Race car driver}$ $V(s, \theta_\pi) = \text{good}$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240 \text{ km/h}$ $\theta_\pi = \text{Race car driver}$ $V(s, \theta_\pi) = \text{good}$

$s = 240 \text{ km/h}$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240 \text{ km/h}$ $\theta_\pi = \text{Race car driver}$ $V(s, \theta_\pi) = \text{good}$

$s = 240 \text{ km/h}$ $\theta_\pi = \text{Grandma}$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240 \text{ km/h}$ $\quad$ $\theta_\pi = \text{Race car driver}$ $\quad$ $V(s, \theta_\pi) = \text{good}$

$s = 240 \text{ km/h}$ $\quad$ $\theta_\pi = \text{Grandma}$ $\quad$ $V(s, \theta_\pi) = \text{not good}$

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240 \text{ km/h}$ $\quad$ $\theta_\pi = \text{Race car driver}$ $\quad$ $V(s, \theta_\pi) = \text{good}$

$s = 240 \text{ km/h}$ $\quad$ $\theta_\pi = \text{Grandma}$ $\quad$ $V(s, \theta_\pi) = \text{not good}$

We use the value function to direct the policy to good states

Value Functions

$$V(s_0, \theta_\pi) = \mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0; \theta_\pi] = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Value function takes any state s_0, and tells us how valuable s_0 is

Valuable states lead to good returns **under the current policy**

$s = 240 \text{ km/h}$ $\quad$ $\theta_\pi = \text{Race car driver}$ $\quad$ $V(s, \theta_\pi) = \text{good}$

$s = 240 \text{ km/h}$ $\quad$ $\theta_\pi = \text{Grandma}$ $\quad$ $V(s, \theta_\pi) = \text{not good}$

We use the value function to direct the policy to good states

It is a critical part of decision making

Value Functions

With the value function, we can use any state as a starting state

Value Functions

With the value function, we can use any state as a starting state

The state does not need to be the start of a trajectory

Value Functions

With the value function, we can use any state as a starting state

The state does not need to be the start of a trajectory

Example: Consider the sequence of states

$$s_0 = S_a, s_1 = S_b, s_2 = S_c, \ldots$$

Value Functions

With the value function, we can use any state as a starting state

The state does not need to be the start of a trajectory

Example: Consider the sequence of states

$$s_0 = S_a, s_1 = S_b, s_2 = S_c, \ldots$$

We can compute

$$V(s_0 = S_a, \theta_\pi), V(s_0 = S_b, \theta_\pi), V(s_0 = S_c, \theta_\pi)$$

To find the value of any state $S_a, S_b, S_c, \ldots$

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Answer: We can use the value of a state to make decisions

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Answer: We can use the value of a state to make decisions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Answer: We can use the value of a state to make decisions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

Given all your preferences ($\mathcal{R}$) and thoughts (θ_π), we can determine which life is better for you

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Answer: We can use the value of a state to make decisions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

Given all your preferences ($\mathcal{R}$) and thoughts (θ_π), we can determine which life is better for you

$V(s, \theta_\pi)$ considers your future friends, income, wife/husband, etc

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Answer: We can use the value of a state to make decisions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

Given all your preferences ($\mathcal{R}$) and thoughts (θ_π), we can determine which life is better for you

$V(s, \theta_\pi)$ considers your future friends, income, wife/husband, etc

Combines all this info into one value, a single number of "goodness"

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Answer: We can use the value of a state to make decisions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

Given all your preferences ($\mathcal{R}$) and thoughts (θ_π), we can determine which life is better for you

$V(s, \theta_\pi)$ considers your future friends, income, wife/husband, etc

Combines all this info into one value, a single number of "goodness"

$$V(S_a, \theta_\pi) = 1032$$

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? He writes the return 100 different ways. How is the value function useful?

Answer: We can use the value of a state to make decisions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

Given all your preferences ($\mathcal{R}$) and thoughts (θ_π), we can determine which life is better for you

$V(s, \theta_\pi)$ considers your future friends, income, wife/husband, etc

Combines all this info into one value, a single number of "goodness"

$$V(S_a, \theta_\pi) = 1032 \qquad V(S_b, \theta_\pi) = 945$$

Value Functions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

Value Functions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

$$V(S_a, \theta_\pi) = 1032 \qquad V(S_b, \theta_\pi) = 945$$

Value Functions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

$$V(S_a, \theta_\pi) = 1032 \qquad V(S_b, \theta_\pi) = 945$$

This value leads us to the right decisions

Value Functions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

$$V(S_a, \theta_\pi) = 1032 \qquad V(S_b, \theta_\pi) = 945$$

This value leads us to the right decisions

Some optimal decisions are hard for humans to make

Value Functions

$$S_a = \text{Live in Macau}, S_b = \text{Live in Beijing}$$

$$V(S_a, \theta_\pi) = 1032 \qquad V(S_b, \theta_\pi) = 945$$

This value leads us to the right decisions

Some optimal decisions are hard for humans to make

With value, we can be sure we make the right decision

Exercise

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you ($\mathcal{R}$)

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you ($\mathcal{R}$)
- 3 life goals as states $S_x, S_y, S_z \in G$ (e.g., friends, money, hobby, etc)

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you $(\mathcal{R})$
- 3 life goals as states $S_x, S_y, S_z \in G$ (e.g., friends, money, hobby, etc)
- Assign a reward $\mathcal{R}$ for each goal, and choose discount factor γ

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you ($\mathcal{R}$)
- 3 life goals as states $S_x, S_y, S_z \in G$ (e.g., friends, money, hobby, etc)
- Assign a reward $\mathcal{R}$ for each goal, and choose discount factor γ

For each location $s_0 \in \{S_a, S_b\}$:

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you $(\mathcal{R})$
- 3 life goals as states $S_x, S_y, S_z \in G$ (e.g., friends, money, hobby, etc)
- Assign a reward $\mathcal{R}$ for each goal, and choose discount factor γ

For each location $s_0 \in \{S_a, S_b\}$:

- Write probability of reaching goals $\Pr(s_g \mid s_0); s_g \in \{S_x, S_y, S_z\}$

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you ($\mathcal{R}$)
- 3 life goals as states $S_x, S_y, S_z \in G$ (e.g., friends, money, hobby, etc)
- Assign a reward $\mathcal{R}$ for each goal, and choose discount factor γ

For each location $s_0 \in \{S_a, S_b\}$:

- Write probability of reaching goals $\Pr(s_g \mid s_0); s_g \in \{S_x, S_y, S_z\}$
- Estimate time to accomplish each goal $t_g; g \in \{S_x, S_y, S_z\}$

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you ($\mathcal{R}$)
- 3 life goals as states $S_x, S_y, S_z \in G$ (e.g., friends, money, hobby, etc)
- Assign a reward $\mathcal{R}$ for each goal, and choose discount factor γ

For each location $s_0 \in \{S_a, S_b\}$:

- Write probability of reaching goals $\Pr(s_g \mid s_0); s_g \in \{S_x, S_y, S_z\}$
- Estimate time to accomplish each goal $t_g; g \in \{S_x, S_y, S_z\}$

$$V(s_0, \theta_\pi) = \sum_{s_g \in \{S_x, S_y, S_z\}} \gamma^{t_g} \mathcal{R}(s_g) \cdot \Pr(s_g \mid s_0; \theta_\pi)$$

Exercise

- Think of two places you want to live after graduation $s_0 \in \{S_a, S_b\}$
- Consider your behavior (θ_π) and what is important to you ($\mathcal{R}$)
- 3 life goals as states $S_x, S_y, S_z \in G$ (e.g., friends, money, hobby, etc)
- Assign a reward $\mathcal{R}$ for each goal, and choose discount factor γ

For each location $s_0 \in \{S_a, S_b\}$:

- Write probability of reaching goals $\Pr(s_g \mid s_0); s_g \in \{S_x, S_y, S_z\}$
- Estimate time to accomplish each goal $t_g; g \in \{S_x, S_y, S_z\}$

$$V(s_0, \theta_\pi) = \sum_{s_g \in \{S_x, S_y, S_z\}} \gamma^{t_g} \mathcal{R}(s_g) \cdot \Pr(s_g \mid s_0; \theta_\pi)$$

Where should you live?

TD Value Functions

TD Value Functions

Note: We can define the value function in different ways

TD Value Functions

Note: We can define the value function in different ways

Always approximates the expected discounted return starting from s_0

TD Value Functions

Note: We can define the value function in different ways

Always approximates the expected discounted return starting from s_0

We call the following equation the **Monte Carlo** value function

$$V(s_0, \theta_\pi) = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

TD Value Functions

Note: We can define the value function in different ways

Always approximates the expected discounted return starting from s_0

We call the following equation the **Monte Carlo** value function

$$V(s_0, \theta_\pi) = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Difficult to compute the Monte Carlo value function

TD Value Functions

Note: We can define the value function in different ways

Always approximates the expected discounted return starting from s_0

We call the following equation the **Monte Carlo** value function

$$V(s_0, \theta_\pi) = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Difficult to compute the Monte Carlo value function

Must have a terminal state, we cannot compute infinite sum

TD Value Functions

Note: We can define the value function in different ways

Always approximates the expected discounted return starting from s_0

We call the following equation the **Monte Carlo** value function

$$V(s_0, \theta_\pi) = \sum_{n=0}^{\infty} \gamma^n \sum_{s_{n+1} \in S} \mathcal{R}(s_{n+1}) \cdot \Pr(s_{n+1} \mid s_0; \theta_\pi)$$

Difficult to compute the Monte Carlo value function

Must have a terminal state, we cannot compute infinite sum

Let us try to delete the infinite sum

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Factor out initial timestep $t = 0$ out of the outer sum

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Factor out initial timestep $t = 0$ out of the outer sum

$$V(s_0, \theta_\pi) = \gamma^0 \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=1}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=1}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=1}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Rewrite sum starting from $t = 0$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=1}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Rewrite sum starting from $t = 0$

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

Factor out γ

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

Factor out γ

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \gamma \sum_{t=0}^{\infty} \gamma^{t} \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

Split Pr using Markov property

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_0; \theta_\pi)$$

Split Pr using Markov property

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_1; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_1; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_1; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

Move sum and Pr outside

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_1; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

Move sum and Pr outside

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

Question: What is this term?

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

Question: What is this term?

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

Question: What is this term?

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+1} \in S} \mathcal{R}(s_{t+1}) \cdot \Pr(s_{t+1} \mid s_0; \theta_\pi)$$

$$V(s_1, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \cdot \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi)$$

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_1; \theta_\pi)$$

Replace infinite sum with value function

$$V(s_0, \theta_\pi) = \left(\sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi) \right) + \gamma V(s_1, \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \left(\sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi) \right) + \gamma V(s_1, \theta_\pi)$$

TD Value Functions

$$V(s_0, \theta_\pi) = \left(\sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi) \right) + \gamma V(s_1, \theta_\pi)$$

First term is expected reward

TD Value Functions

$$V(s_0, \theta_\pi) = \left(\sum_{s_1 \in S} \mathcal{R}(s_1) \cdot \Pr(s_1 \mid s_0; \theta_\pi) \right) + \gamma V(s_1, \theta_\pi)$$

First term is expected reward

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

TD Value Functions

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

TD Value Functions

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

This is a huge finding!

TD Value Functions

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

This is a huge finding!

Value function has a recursive definition

TD Value Functions

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

This is a huge finding!

Value function has a recursive definition

Represent policy-conditioned discounted return without an infinite sum

TD Value Functions

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

This is a huge finding!

Value function has a recursive definition

Represent policy-conditioned discounted return without an infinite sum

We call this the **Temporal Difference** (TD) value function

TD Value Functions

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

This is a huge finding!

Value function has a recursive definition

Represent policy-conditioned discounted return without an infinite sum

We call this the **Temporal Difference** (TD) value function

Compute the return with a single transition $s_0 \rightarrow s_1$

TD Value Functions

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

This is a huge finding!

Value function has a recursive definition

Represent policy-conditioned discounted return without an infinite sum

We call this the **Temporal Difference** (TD) value function

Compute the return with a single transition $s_0 \rightarrow s_1$

Evaluate infinite-depth decision tree with one function

TD Value Functions

To summarize, we can represent the value function in two ways:

TD Value Functions

To summarize, we can represent the value function in two ways:

The Monte Carlo value function

TD Value Functions

To summarize, we can represent the value function in two ways:

The Monte Carlo value function

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, \theta_\pi]$$

TD Value Functions

To summarize, we can represent the value function in two ways:

The Monte Carlo value function

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, \theta_\pi]$$

The Temporal Difference value function

TD Value Functions

To summarize, we can represent the value function in two ways:

The Monte Carlo value function

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, \theta_\pi]$$

The Temporal Difference value function

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

TD Value Functions

To summarize, we can represent the value function in two ways:

The Monte Carlo value function

$$V(s_0, \theta_\pi) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_0, \theta_\pi]$$

The Temporal Difference value function

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

They produce the same result, but with different computation

Q Functions

Q Functions

We saw two forms of the value function

Q Functions

We saw two forms of the value function

The value function relies on a policy

Q Functions

We saw two forms of the value function

The value function relies on a policy

But our goal was to find a policy, so how does value help?

Q Functions

We saw two forms of the value function

The value function relies on a policy

But our goal was to find a policy, so how does value help?

Special connection between an optimal policy and the value function

Q Functions

We saw two forms of the value function

The value function relies on a policy

But our goal was to find a policy, so how does value help?

Special connection between an optimal policy and the value function

We can modify the value function to find an optimal policy

Q Functions

We saw two forms of the value function

The value function relies on a policy

But our goal was to find a policy, so how does value help?

Special connection between an optimal policy and the value function

We can modify the value function to find an optimal policy

We call the modified value function, a Q function

Q Functions

Consider the Temporal Difference value function

Q Functions

Consider the Temporal Difference value function

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

Q Functions

Consider the Temporal Difference value function

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

We conditioned the value function on policy parameters

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

Q Functions

Consider the Temporal Difference value function

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

We conditioned the value function on policy parameters

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

With trajectory optimization we conditioned on actions

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...]$$

Q Functions

Consider the Temporal Difference value function

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

We conditioned the value function on policy parameters

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

With trajectory optimization we conditioned on actions

$$\mathbb{E}[\mathcal{G}(\boldsymbol{\tau}) \mid s_0, a_0, a_1, ...]$$

What if we wanted a mix of both?

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

Q Functions

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

Q Functions

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

This expectation means:

Q Functions

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

This expectation means:

- Take a specific action a_0 (trajectory optimization)

Q Functions

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

This expectation means:

- Take a specific action a_0 (trajectory optimization)
- Follow $\pi(\cdot \mid s_t; \theta_\pi)$ for all future actions $a_1, a_2, \ldots$ (value function)

We call this the **Q function**

Q Functions

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

This expectation means:

- Take a specific action a_0 (trajectory optimization)
- Follow $\pi(\cdot \mid s_t; \theta_\pi)$ for all future actions $a_1, a_2, \ldots$ (value function)

We call this the **Q function**

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

Q Functions

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

This expectation means:

- Take a specific action a_0 (trajectory optimization)
- Follow $\pi(\cdot \mid s_t; \theta_\pi)$ for all future actions $a_1, a_2, \ldots$ (value function)

We call this the **Q function**

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

We can derive the Q function from the value function

Q Functions

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

Q Functions

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

First, introduce the action a_0

Q Functions

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

First, introduce the action a_0

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

Condition the initial reward on the action

Q Functions

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

First, introduce the action a_0

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

Condition the initial reward on the action

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

Q Functions

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

First, introduce the action a_0

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

Condition the initial reward on the action

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

Call it the Q function

Q Functions

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

First, introduce the action a_0

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

Condition the initial reward on the action

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

Call it the Q function

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

Q Functions

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

Q Functions

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

The Q function tells us:

Q Functions

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

The Q function tells us:

- The value of an action a_0

Q Functions

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

The Q function tells us:

- The value of an action a_0
- In state s_0

Q Functions

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

The Q function tells us:

- The value of an action a_0
- In state s_0
- If we follow $\pi(a_t \mid s_t; \theta_\pi)$ afterwards

Q Functions

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

The Q function tells us:

- The value of an action a_0
- In state s_0
- If we follow $\pi(a_t \mid s_t; \theta_\pi)$ afterwards

Question: How can we use the Q function for decision making?

Q Functions

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

The Q function tells us:

- The value of an action a_0
- In state s_0
- If we follow $\pi(a_t \mid s_t; \theta_\pi)$ afterwards

Question: How can we use the Q function for decision making?

Hint: We can evaluate Q for every possible action

Q Functions

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

The Q function tells us:

- The value of an action a_0
- In state s_0
- If we follow $\pi(a_t \mid s_t; \theta_\pi)$ afterwards

Question: How can we use the Q function for decision making?

Hint: We can evaluate Q for every possible action

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

Q Functions

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

Q Functions

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

This is a very powerful equation

Q Functions

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

This is a very powerful equation

- Compute $Q(s_0, a_0)$ for all a_0

Q Functions

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

This is a very powerful equation

- Compute $Q(s_0, a_0)$ for all a_0
- Pick the a_0 that maximizes $Q(s_0, a_0)$

Q Functions

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

This is a very powerful equation

- Compute $Q(s_0, a_0)$ for all a_0
- Pick the a_0 that maximizes $Q(s_0, a_0)$
- This a_0 is **guaranteed** to be the optimal action

Q Functions

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

This is a very powerful equation

- Compute $Q(s_0, a_0)$ for all a_0
- Pick the a_0 that maximizes $Q(s_0, a_0)$
- This a_0 is **guaranteed** to be the optimal action

This considers the effect of a_0 on the **infinite** future

Q Functions

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

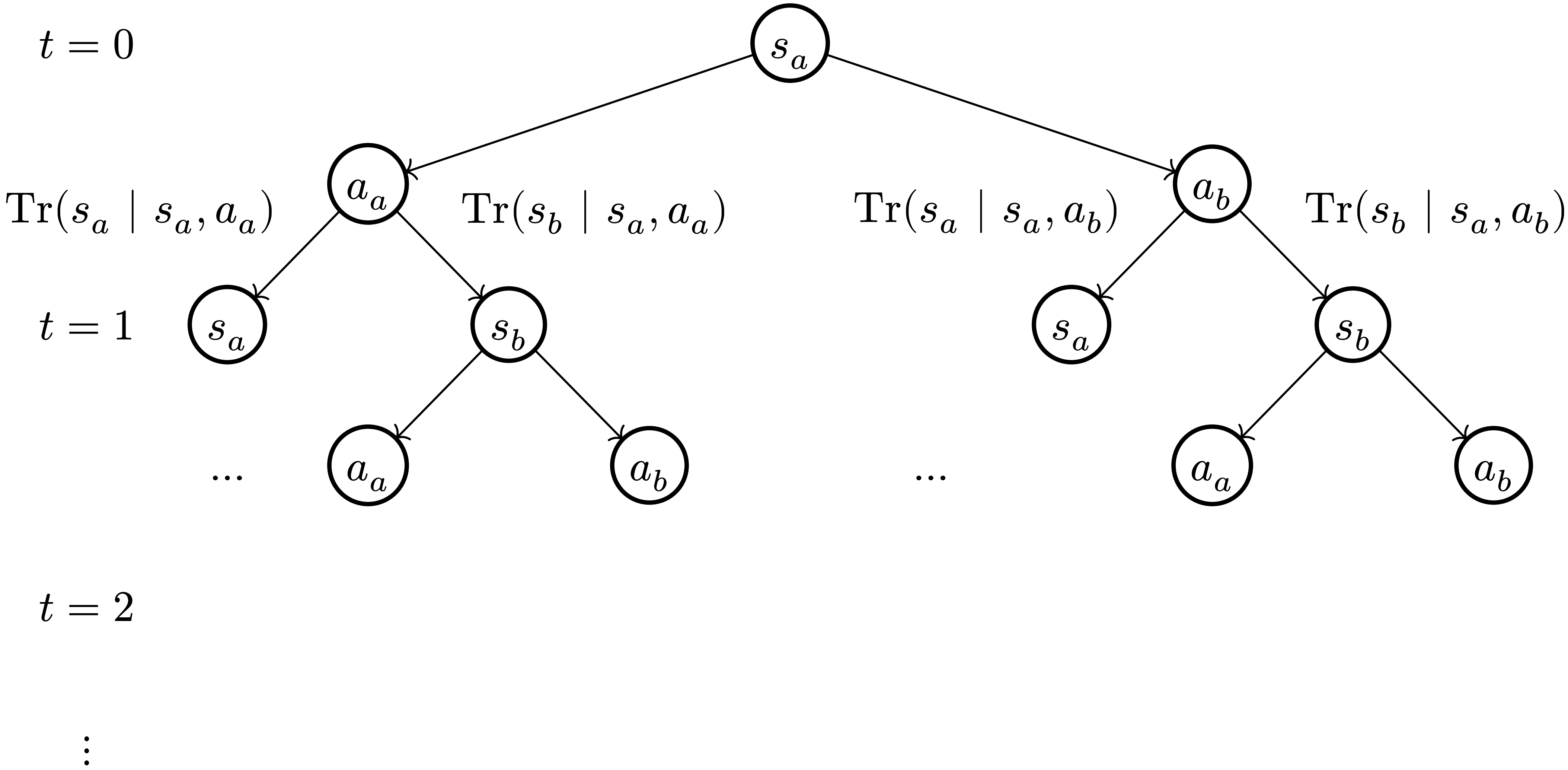
This is a very powerful equation

- Compute $Q(s_0, a_0)$ for all a_0
- Pick the a_0 that maximizes $Q(s_0, a_0)$
- This a_0 is **guaranteed** to be the optimal action

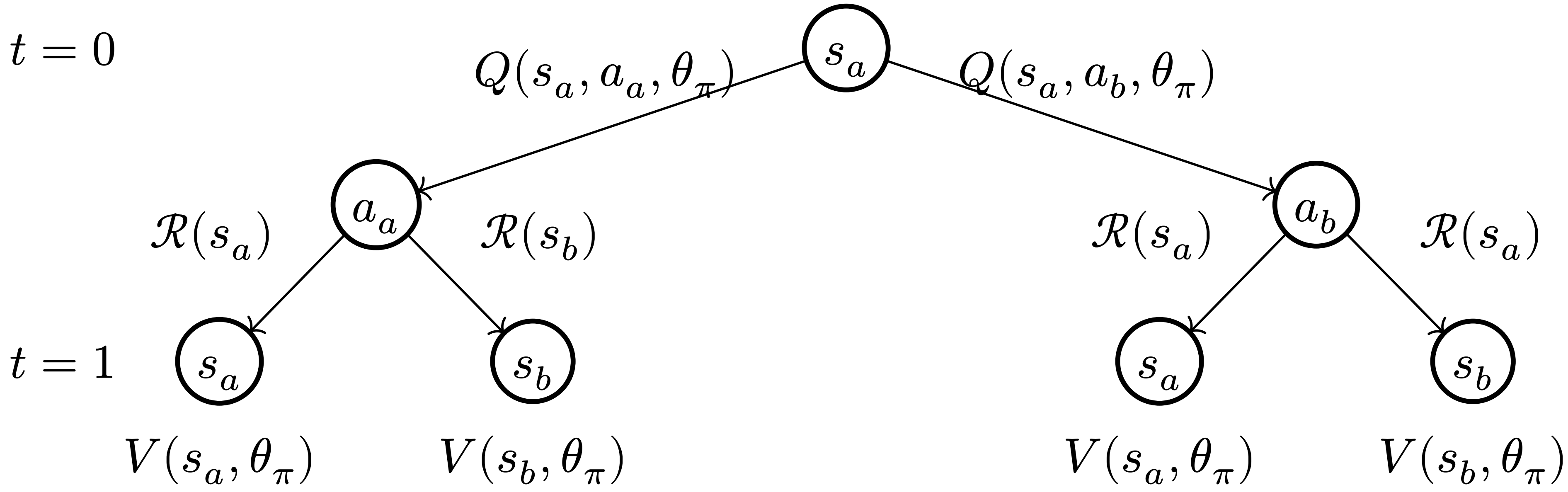
This considers the effect of a_0 on the **infinite** future

We collapsed the infinite decision tree into a single level

Q Functions


$t = 0$
s_a
a_a
a_b
$\mathrm{Tr}(s_a \mid s_a, a_a)$
$\mathrm{Tr}(s_b \mid s_a, a_a)$
$\mathrm{Tr}(s_a \mid s_a, a_b)$
$\mathrm{Tr}(s_b \mid s_a, a_b)$
$t = 1$
s_a
s_b
s_a
s_b
$\ldots$
a_a
a_b
$\ldots$
a_a
a_b
$t = 2$
$\vdots$

Q Functions


$t = 0$
s_a
$Q(s_a, a_a, \theta_\pi)$
$Q(s_a, a_b, \theta_\pi)$
a_a
a_b
$\mathcal{R}(s_a)$
$\mathcal{R}(s_b)$
$\mathcal{R}(s_a)$
$\mathcal{R}(s_a)$
$t = 1$
s_a
s_b
s_a
s_b
$V(s_a, \theta_\pi)$
$V(s_b, \theta_\pi)$
$V(s_a, \theta_\pi)$
$V(s_b, \theta_\pi)$

Q Learning

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Cheap to train, expensive to use

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Cheap to train, expensive to use

Closer to traditional control theory

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Cheap to train, expensive to use

Closer to traditional control theory

Model-free

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Cheap to train, expensive to use

Closer to traditional control theory

Model-free

We do not know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Cheap to train, expensive to use

Closer to traditional control theory

Model-free

We do not know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Expensive to train, cheap to use

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Cheap to train, expensive to use

Closer to traditional control theory

Model-free

We do not know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Expensive to train, cheap to use

Closer to deep learning

Q Learning

Q learning is a **model-free** algorithm first discovered in the 1980s

Model-based

We know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Cheap to train, expensive to use

Closer to traditional control theory

Model-free

We do not know $\mathrm{Tr}(s_{t+1} \mid s_t, a_t)$

Expensive to train, cheap to use

Closer to deep learning

Q Learning

Q learning is still popular today

[1] *Simplifying Deep Temporal Difference Learning*. ICLR. 2024.

[2] *Exclusively Penalized Q-Learning for Offline Reinforcement Learning*. NeurIPS. 2025.

Q Learning

Q learning is still popular today

Works well with deep neural networks

[1] *Simplifying Deep Temporal Difference Learning*. ICLR. 2024.

[2] *Exclusively Penalized Q-Learning for Offline Reinforcement Learning*. NeurIPS. 2025.

Q Learning

Q learning is still popular today

Works well with deep neural networks

Researchers are still improving it[1][2]

[1] *Simplifying Deep Temporal Difference Learning*. ICLR. 2024.

[2] *Exclusively Penalized Q-Learning for Offline Reinforcement Learning*. NeurIPS. 2025.

Q Learning

Q learning is still popular today

Works well with deep neural networks

Researchers are still improving it[1][2]

In fact, our lab is using it in our research right now

---

[1] *Simplifying Deep Temporal Difference Learning*. ICLR. 2024.

[2] *Exclusively Penalized Q-Learning for Offline Reinforcement Learning*. NeurIPS. 2025.

Q Learning

Q learning is still popular today

Works well with deep neural networks

Researchers are still improving it[1][2]

In fact, our lab is using it in our research right now

We now have all the information we need to implement Q learning

---

[1] *Simplifying Deep Temporal Difference Learning*. ICLR. 2024.

[2] *Exclusively Penalized Q-Learning for Offline Reinforcement Learning*. NeurIPS. 2025.

Q Learning

Our Q function relies on the value function for some θ_π

Q Learning

Our Q function relies on the value function for some θ_π

Right now, it is not clear what the policy is

Q Learning

Our Q function relies on the value function for some θ_π

Right now, it is not clear what the policy is

So how can we use the Q function without knowing the policy?

Q Learning

Our Q function relies on the value function for some θ_π

Right now, it is not clear what the policy is

So how can we use the Q function without knowing the policy?

Let us find out

Q Learning

Start with the Q function

Q Learning

Start with the Q function

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

Q Learning

Start with the Q function

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

We want to take the action that maximizes Q

Q Learning

Start with the Q function

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

We want to take the action that maximizes Q

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

Q Learning

Start with the Q function

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

We want to take the action that maximizes Q

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

Recall Monte Carlo value function

Q Learning

Start with the Q function

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

We want to take the action that maximizes Q

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

Recall Monte Carlo value function

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

Q Learning

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

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

Q Learning

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

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

Take optimal action a_0, now must find optimal a_1

Q Learning

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

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

Take optimal action a_0, now must find optimal a_1

$$\arg\max_{a_1 \in A} Q(s_1, a_1, \theta_\pi) = \arg\max_{a_1 \in A} (\mathbb{E}[\mathcal{R}(s_2) \mid s_1, a_1] + \gamma V(s_2, \theta_\pi))$$

Q Learning

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

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

Take optimal action a_0, now must find optimal a_1

$$\arg\max_{a_1 \in A} Q(s_1, a_1, \theta_\pi) = \arg\max_{a_1 \in A} (\mathbb{E}[\mathcal{R}(s_2) \mid s_1, a_1] + \gamma V(s_2, \theta_\pi))$$

Take optimal action a_1, now must find optimal a_2

Q Learning

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

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

Take optimal action a_0, now must find optimal a_1

$$\arg\max_{a_1 \in A} Q(s_1, a_1, \theta_\pi) = \arg\max_{a_1 \in A}(\mathbb{E}[\mathcal{R}(s_2) \mid s_1, a_1] + \gamma V(s_2, \theta_\pi))$$

Take optimal action a_1, now must find optimal a_2

$$\arg\max_{a_2 \in A} Q(s_2, a_2, \theta_\pi) = \arg\max_{a_2 \in A}(\mathbb{E}[\mathcal{R}(s_3) \mid s_2, a_2] + \gamma V(s_3, \theta_\pi))$$

Q Learning

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

$$\arg\max_{a_1 \in A} Q(s_1, a_1, \theta_\pi) = \arg\max_{a_1 \in A}(\mathbb{E}[\mathcal{R}(s_2) \mid s_1, a_1] + \gamma V(s_2, \theta_\pi))$$

$$\arg\max_{a_2 \in A} Q(s_2, a_2, \theta_\pi) = \arg\max_{a_2 \in A}(\mathbb{E}[\mathcal{R}(s_3) \mid s_2, a_2] + \gamma V(s_3, \theta_\pi))$$

There is a pattern. What policy causes this pattern?

Q Learning

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

$$\underset{a_1 \in A}{\arg\max}\, Q(s_1, a_1, \theta_\pi) = \underset{a_1 \in A}{\arg\max}(\mathbb{E}[\mathcal{R}(s_2) \mid s_1, a_1] + \gamma V(s_2, \theta_\pi))$$

$$\underset{a_2 \in A}{\arg\max}\, Q(s_2, a_2, \theta_\pi) = \underset{a_2 \in A}{\arg\max}(\mathbb{E}[\mathcal{R}(s_3) \mid s_2, a_2] + \gamma V(s_3, \theta_\pi))$$

There is a pattern. What policy causes this pattern?

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

The policy uses the Q function

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

The policy uses the Q function

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

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

The policy uses the Q function

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

The Q function uses the policy

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \text{arg max}_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

The policy uses the Q function

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

The Q function uses the policy

Question: Can we simplify the Q function using the policy?

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

The policy uses the Q function

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

The Q function uses the policy

Question: Can we simplify the Q function using the policy?

$$V(s_0, \theta_\pi) = \max_{a \in A} Q(s_0, a, \theta_\pi)$$

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

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

$$V(s_0, \theta_\pi) = \max_{a \in A} Q(s_0, a, \theta_\pi)$$

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

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

$$V(s_0, \theta_\pi) = \max_{a \in A} Q(s_0, a, \theta_\pi)$$

Replace V with Q

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

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

$$V(s_0, \theta_\pi) = \max_{a \in A} Q(s_0, a, \theta_\pi)$$

Replace V with Q

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

Q Learning

Definition: In Temporal Difference Q learning, we learn Q using

Q Learning

Definition: In Temporal Difference Q learning, we learn Q using

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

Q Learning

Definition: In Temporal Difference Q learning, we learn Q using

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

Definition: In Monte Carlo Q learning, we learn Q using

Q Learning

Definition: In Temporal Difference Q learning, we learn Q using

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

Definition: In Monte Carlo Q learning, we learn Q using

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Q Learning

Definition: In Temporal Difference Q learning, we learn Q using

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

Definition: In Monte Carlo Q learning, we learn Q using

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0] + \underbrace{\sum_{t=1}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]}_{\text{Return following } \pi}$$

Q Learning

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

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Q Learning

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

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

If we want to learn the left hand side, we must know the right hand side

Q Learning

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

$$Q(s_0, a_0, \theta_\pi) = \mathbb{E}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

If we want to learn the left hand side, we must know the right hand side

Question: How do we find these terms?

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Question: How do we find these terms?

Answer: Empirical expectation from episode data $(s_t, a_t, \mathcal{R}(s_{t+1}))$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Question: How do we find these terms?

Answer: Empirical expectation from episode data $(s_t, a_t, \mathcal{R}(s_{t+1}))$

$$\boldsymbol{E} = \begin{bmatrix} s_0 & s_1 & s_2 & \dots \\ a_0 & a_1 & a_2 & \dots \\ r_0 & r_1 & r_2 & \dots \end{bmatrix}^\top$$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Question: How to find these terms?

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Question: How to find these terms?

$$\boldsymbol{E} = \begin{bmatrix} s_0 & s_1 & s_2 & \ldots \\ a_0 & a_1 & a_2 & \ldots \\ r_0 & r_1 & r_2 & \ldots \end{bmatrix}^\top$$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Question: How to find these terms?

$$\boldsymbol{E} = \begin{bmatrix} s_0 & s_1 & s_2 & \ldots \\ a_0 & a_1 & a_2 & \ldots \\ r_0 & r_1 & r_2 & \ldots \end{bmatrix}^\top$$

TD: $\gamma \max_{a \in A} Q(s_{t+1}, a, \theta_\pi)$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Question: How to find these terms?

$$\boldsymbol{E} = \begin{bmatrix} s_0 & s_1 & s_2 & \dots \\ a_0 & a_1 & a_2 & \dots \\ r_0 & r_1 & r_2 & \dots \end{bmatrix}^\top$$

TD: $\gamma \max_{a \in A} Q(s_{t+1}, a, \theta_\pi)$

MC: $\gamma r_{t+1} + \gamma^2 r_{t+2} + \dots$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Question: How to find these terms?

$$\boldsymbol{E} = \begin{bmatrix} s_0 & s_1 & s_2 & \ldots \\ a_0 & a_1 & a_2 & \ldots \\ r_0 & r_1 & r_2 & \ldots \end{bmatrix}^\top$$

TD: $\gamma \max_{a \in A} Q(s_{t+1}, a, \theta_\pi)$

MC: $\gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$

We know the right hand side, use it to learn the left hand side

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Assume $Q(s, a, \theta_\pi)$ has error η with right hand side

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Assume $Q(s, a, \theta_\pi)$ has error η with right hand side

Use the error to update the Q function

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Assume $Q(s, a, \theta_\pi)$ has error η with right hand side

Use the error to update the Q function

$$Q_{i+1}(s, a, \theta_\pi) = Q_i(s, a, \theta_\pi) - \eta$$

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Assume $Q(s, a, \theta_\pi)$ has error η with right hand side

Use the error to update the Q function

$$Q_{i+1}(s, a, \theta_\pi) = Q_i(s, a, \theta_\pi) - \eta$$

Improve convergence with a learning rate α

Q Learning

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q(s_1, a, \theta_\pi)$$

$$Q(s_0, a_0, \theta_\pi) = \hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi]$$

Assume $Q(s, a, \theta_\pi)$ has error η with right hand side

Use the error to update the Q function

$$Q_{i+1}(s, a, \theta_\pi) = Q_i(s, a, \theta_\pi) - \eta$$

Improve convergence with a learning rate α

$$Q_{i+1}(s, a, \theta_\pi) = Q_i(s, a, \theta_\pi) - \alpha \cdot \eta$$

Q Learning

Monte Carlo update:

Q Learning

Monte Carlo update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

Q Learning

Monte Carlo update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

Q Learning

Monte Carlo update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = Q_i(s_0, a_0, \theta_\pi) - \left(\hat{\mathbb{E}} \; [\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi] \right)$$

Q Learning

Monte Carlo update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = \underbrace{Q_i(s_0, a_0, \theta_\pi)}_{\text{Predicted value}} - \left(\hat{\mathbb{E}} \; [\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi] \right)$$

Q Learning

Monte Carlo update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = \underbrace{Q_i(s_0, a_0, \theta_\pi)}_{\text{Predicted value}} - \underbrace{\left(\hat{\mathbb{E}} \left[\mathcal{R}(s_1) \mid s_0, a_0 \right] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}} \left[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi \right] \right)}_{\text{Empirical value}}$$

Q Learning

Monte Carlo update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = \underbrace{Q_i(s_0, a_0, \theta_\pi)}_{\text{Predicted value}} - \underbrace{\left(\hat{\mathbb{E}} \; [\mathcal{R}(s_1) \mid s_0, a_0] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}}[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi] \right)}_{\text{Empirical value}}$$

If we visit all $s, a \in S \times A$, guaranteed convergence to true Q function

Q Learning

Monte Carlo update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = \underbrace{Q_i(s_0, a_0, \theta_\pi)}_{\text{Predicted value}} - \underbrace{\left(\hat{\mathbb{E}} \left[\mathcal{R}(s_1) \mid s_0, a_0 \right] + \sum_{t=1}^{\infty} \gamma^t \hat{\mathbb{E}} \left[\mathcal{R}(s_{t+1}) \mid s_1; \theta_\pi \right] \right)}_{\text{Empirical value}}$$

If we visit all $s, a \in S \times A$, guaranteed convergence to true Q function

$\lim_{i \to \infty} \eta = 0$

Q Learning

Temporal Difference update:

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = Q_i(s_0, a_0, \theta_\pi) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q_i(s_1, a, \theta_\pi) \right)$$

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = Q_i(s_0, a_0, \theta_\pi) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q_i(s_1, a, \theta_\pi) \right)$$

小心! If s_1 is a terminal state, future value is 0 ($\neg d = \text{not terminated}$)

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = \underbrace{Q_i(s_0, a_0, \theta_\pi)}_{\text{Predicted value}} - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q_i(s_1, a, \theta_\pi) \right)$$

小心! If s_1 is a terminal state, future value is 0 ($\neg d =$ not terminated)

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = \underbrace{Q_i(s_0, a_0, \theta_\pi)}_{\text{Predicted value}} - \underbrace{\left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q_i(s_1, a, \theta_\pi) \right)}_{\text{Empirical value}}$$

小心! If s_1 is a terminal state, future value is 0 ($\neg d =$ not terminated)

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

$$\eta = \underbrace{Q_i(s_0, a_0, \theta_\pi)}_{\text{Predicted value}} - \underbrace{\left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q_i(s_1, a, \theta_\pi)\right)}_{\text{Empirical value}}$$

小心! If s_1 is a terminal state, future value is 0 ($\neg d = \text{not terminated}$)

If we visit all $s, a \in S \times A$, guaranteed convergence to true Q function

Q Learning

Temporal Difference update:

$$Q_{i+1}(s_0, a_0, \theta_\pi) = Q_i(s_0, a_0, \theta_\pi) - \alpha \cdot \eta$$

The error η is the difference between true and predicted value

Predicted value

$$\eta = Q_i(s_0, a_0, \theta_\pi) - \left(\hat{\mathbb{E}}[\mathcal{R}(s_1) \mid s_0, a_0] + \neg d \gamma \max_{a \in A} Q_i(s_1, a, \theta_\pi)\right)$$

Empirical value

小心! If s_1 is a terminal state, future value is 0 ($\neg d = \text{not terminated}$)

If we visit all $s, a \in S \times A$, guaranteed convergence to true Q function

$\lim_{i \to \infty} \eta = 0$

Q Learning

Last thing, we must collect episodes to train Q!

Q Learning

Last thing, we must collect episodes to train Q!

Can run policy in environment to create episodes

Q Learning

Last thing, we must collect episodes to train Q!

Can run policy in environment to create episodes

```
states, next_states, rewards, terminateds = [], [], [], []
state = environment.reset()
while not terminated:
    action = policy.sample(state)
    next_state, reward, terminated = environment.step(action)

    states.append(state), next_states.append(next_state), ...
    state = next_state

episode = (states, next_states, rewards, terminateds)
```

Q Learning

What policy do we sample actions from?

Q Learning

What policy do we sample actions from?

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Q Learning

What policy do we sample actions from?

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Question: Any issues?

Q Learning

What policy do we sample actions from?

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Question: Any issues?

Answer: Always sample the same action (exploit, no exploration)

Q Learning

What policy do we sample actions from?

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Question: Any issues?

Answer: Always sample the same action (exploit, no exploration)

If Q function is wrong, always sample bad actions

Q Learning

What policy do we sample actions from?

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Question: Any issues?

Answer: Always sample the same action (exploit, no exploration)

If Q function is wrong, always sample bad actions

Without correct actions, Q function will not improve!

Q Learning

What policy do we sample actions from?

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Question: Any issues?

Answer: Always sample the same action (exploit, no exploration)

If Q function is wrong, always sample bad actions

Without correct actions, Q function will not improve!

Question: What can we do?

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Epsilon greedy policy!

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Epsilon greedy policy!

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} (1 - \varepsilon) \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ \frac{\varepsilon}{|A|} \text{ for } a \in A \end{cases}$$

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Epsilon greedy policy!

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} (1 - \varepsilon) \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ \frac{\varepsilon}{|A|} \text{ for } a \in A \end{cases}$$

Sample random action with probability ε

Q Learning

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} 1 \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ 0 \text{ otherwise} \end{cases}$$

Epsilon greedy policy!

$$\pi(a_0 \mid s_0; \theta_\pi) = \begin{cases} (1 - \varepsilon) \text{ if } a_0 = \arg\max_{a \in A} Q(s_0, a, \theta_\pi) \\ \frac{\varepsilon}{|A|} \text{ for } a \in A \end{cases}$$

Sample random action with probability ε

In the limit, we sample all possible actions in all states

Q Learning

Can we visualize Q learning?

Q Learning

Can we visualize Q learning?

Navigation example, reward of 1 for reaching center tile

Q Learning

Can we visualize Q learning?

Navigation example, reward of 1 for reaching center tile

https://user-images.githubusercontent.com/1883779/113412338-97430100-93d5-11eb-856c-ef0f420d1acb.gif

Q Learning

Can we visualize Q learning?

Navigation example, reward of 1 for reaching center tile

https://user-images.githubusercontent.com/1883779/113412338-97430100-93d5-11eb-856c-ef0f420d1acb.gif

https://mohitmayank.com/interactive_q_learning/q_learning.html

Q Learning

So far:

Q Learning

So far:

- Defined training objective (TD and MC updates)

Q Learning

So far:

- Defined training objective (TD and MC updates)
- Defined dataset (episodes)

Q Learning

So far:

- Defined training objective (TD and MC updates)
- Defined dataset (episodes)
- Need to define model (Q function)!

Q Learning

So far:

- Defined training objective (TD and MC updates)
- Defined dataset (episodes)
- Need to define model (Q function)!

Next time, we will use deep neural networks

Q Learning

So far:

- Defined training objective (TD and MC updates)
- Defined dataset (episodes)
- Need to define model (Q function)!

Next time, we will use deep neural networks

Today and for homework, use a simple matrix

Q Learning

Model the Q function as a matrix

Q Learning

Model the Q function as a matrix

Each state is a row, each action is a column in a matrix

Q Learning

Model the Q function as a matrix

Each state is a row, each action is a column in a matrix

$$\begin{bmatrix} Q(S_1, A_1) & Q(S_1, A_2) & \dots \\ Q(S_2, A_1) & Q(S_2, A_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Q Learning

Model the Q function as a matrix

Each state is a row, each action is a column in a matrix

$$\begin{bmatrix} Q(S_1, A_1) & Q(S_1, A_2) & \dots \\ Q(S_2, A_1) & Q(S_2, A_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$Q_{i,j}$ gives Q value for state $s = S_i$ and action $a = A_j$

Homework

Homework

You have everything you need to solve homework

Homework

You have everything you need to solve homework

Due in 2 weeks (Weds 12 March, 23:59)

Homework

You have everything you need to solve homework

Due in 2 weeks (Weds 12 March, 23:59)

Download and submit `.py` and `.ipynb` files

Homework

You have everything you need to solve homework

Due in 2 weeks (Weds 12 March, 23:59)

Download and submit `.py` and `.ipynb` files

Uses turnitin for checking

Homework

You have everything you need to solve homework

Due in 2 weeks (Weds 12 March, 23:59)

Download and submit `.py` and `.ipynb` files

Uses turnitin for checking

https://colab.research.google.com/drive/1xtBxAaVc3ax6_j59RC3NLQQPFcIEoau-?usp=sharing