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 but 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 but 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 but 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

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

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

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

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

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

Conditioned/dependent on a sequence of actions

There is no structure to the actions

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

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

Last time, we introduced the policy

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

Example policy, greedy policy

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

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

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

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

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

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 a function 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 a function 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 a function parameterized by θ_π instead of many actions

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

The function outputs a distribution over the action space $\pi(a \mid s; \theta_\pi)$

Policy-Conditioned Returns

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

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

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

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

Now conditioned on the policy

Policy-Conditioned Returns

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

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

Now conditioned on the policy

But remember, $\mathcal{R}(s_{t+1})$ hides much of 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 conditioned on the policy

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

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

Policy-Conditioned Returns

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

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?

$$\Pr(s_{t+1} \mid s_0; \theta_\pi)$$

Policy-Conditioned Returns

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

Question: What changes when we condition on θ_π?

$$\Pr(s_{t+1} \mid s_0; \theta_\pi)$$

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?

$$\Pr(s_{t+1} \mid s_0; \theta_\pi)$$

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?

$$\Pr(s_{t+1} \mid s_0; \theta_\pi)$$

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

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

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

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

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

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

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

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

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

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

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

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

Where the state distribution is

Policy-Conditioned Returns

Definition: General form of policy-conditioned discounted return

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

Where the state distribution is

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

Value Functions

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \ldots, a_n \in A} \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \dots, a_n \in A} \sum_{s_1, \dots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

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

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \ldots, a_n \in A} \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

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

These two equations form the basis of all reinforcement learning

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \ldots, a_n \in A} \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

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

These two equations form the basis of all reinforcement learning

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

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \dots, a_n \in A} \sum_{s_1, \dots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

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

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \ldots, a_n \in A} \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

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

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

Value Functions

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \ldots, a_n \in A} \sum_{s_1, \ldots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

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

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

$$\Pr(s_{n+1} \mid s_0; \theta_\pi) = \sum_{a_0, \dots, a_n \in A} \sum_{s_1, \dots, s_n \in S} \prod_{t=0}^{n} \mathrm{Tr}(s_{t+1} \mid s_t, a_t) \cdot \pi(a_t \mid s_t; \theta_\pi)$$

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

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**

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

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**

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

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**

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

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

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**

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

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

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**

$s = 240 \text{ km/h}$	$\theta_\pi = \text{Race car driver}$	$V(s, \theta_\pi) = \text{good}$
$s = 240 \text{ km/h}$	$\theta_\pi = \text{Grandpa}$	$V(s, \theta_\pi) = \text{not good}$

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**

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

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

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

$$s_a, s_b, s_c$$

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

$$s_a, s_b, s_c$$

We can compute

$$V(s_a, \theta_\pi), V(s_b, \theta_\pi), V(s_c, \theta_\pi)$$

To find the value of any state

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? How is the value function useful?

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? How is the value function useful?

We can use the value of a state to make decisions

Value Functions

Question: Why does Prof. Steven keep showing stupid equations? How is the value function useful?

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? How is the value function useful?

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? How is the value function useful?

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? How is the value function useful?

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? How is the value function useful?

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? How is the value function useful?

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

Question: Why does Prof. Steven keep showing stupid equations? How is the value function useful?

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_a, s_b

Exercise

- Think of two places you want to live after graduation 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_a, s_b
- Consider your behavior (θ_π) and what is important to you ($\mathcal{R}$)
- Top 3 life goals as states $s_x, s_y, s_z \in G$ (e.g., friends, money, hobby, etc)

Exercise

Exercise

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

- Estimate probability of reaching each goal $\Pr(s_g \mid s_0)$; $s_g \in \{s_x, s_y, s_z\}$
- Estimate time to accomplish each goal $t = \ldots$

$$V(s_0, \theta_\pi) = \sum_{s_g \in \{s_x, s_y, s_z\}} \gamma^t \mathcal{R}(s_g) \cdot \Pr(s_g \mid s_0; \theta_\pi)$$

TD Value Functions

TD Value Functions

Note: We define the value function in many different ways

TD Value Functions

Note: We define the value function in many different ways

It always approximates the expected discounted return from s_0

TD Value Functions

Note: We define the value function in many different ways

It always approximates the expected discounted return 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 define the value function in many different ways

It always approximates the expected discounted return 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 define the value function in many different ways

It always approximates the expected discounted return 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 either have a terminal states

TD Value Functions

Note: We define the value function in many different ways

It always approximates the expected discounted return 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 either have a terminal states
- Or build the infinite decision tree

TD Value Functions

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

TD Value Functions

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

Infinite sums are icky

TD Value Functions

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

Infinite sums are icky

They make everything difficult and intractable

TD Value Functions

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

Infinite sums are icky

They make everything difficult and intractable

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

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

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

TD Value Functions

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

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

TD Value Functions

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

$$+\gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

Move sum and Pr outside

TD Value Functions

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

$$+ \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \sum_{s_1} \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi) \Pr(s_1 \mid s_0; \theta_\pi)$$

Move sum and Pr outside

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

TD Value Functions

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

TD Value Functions

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

Question: What is this term?

TD Value Functions

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

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

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

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

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

TD Value Functions

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

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

TD Value Functions

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

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

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

TD Value Functions

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

$$+ \sum_{s_1} \Pr(s_1 \mid s_0; \theta_\pi) \gamma \sum_{t=0}^{\infty} \gamma^t \sum_{s_{t+2} \in S} \mathcal{R}(s_{t+2}) \Pr(s_{t+2} \mid s_{t+1}; \theta_\pi)$$

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

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

This is a huge finding!

TD Value Functions

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

TD Value Functions

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

Value function has a recursive definition

TD Value Functions

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

Value function has a recursive definition

Represent policy-conditioned discounted return without an infinite sum

TD Value Functions

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

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

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

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

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

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 $s_0 \to s_1$

Evaluate infinite-depth decision tree with a single function call

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

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

TD Value Functions

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

The Monte Carlo value function

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

The Temporal Difference value function

TD Value Functions

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

The Monte Carlo value function

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

The Temporal Difference 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 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 it does not tell us the policy

Q Functions

We saw two forms of the value function

The value function relies on a policy

But it does not tell us the policy

How can we use the value function to find an optimal policy?

Q Functions

Consider the value function

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

With trajectory optimization we conditioned on actions

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

We conditioned the value function on policy parameters

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

What if we wanted a mix of both?

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

Q Functions

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

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

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

First, introduce the action a_0

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

Q Functions

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

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

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

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

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

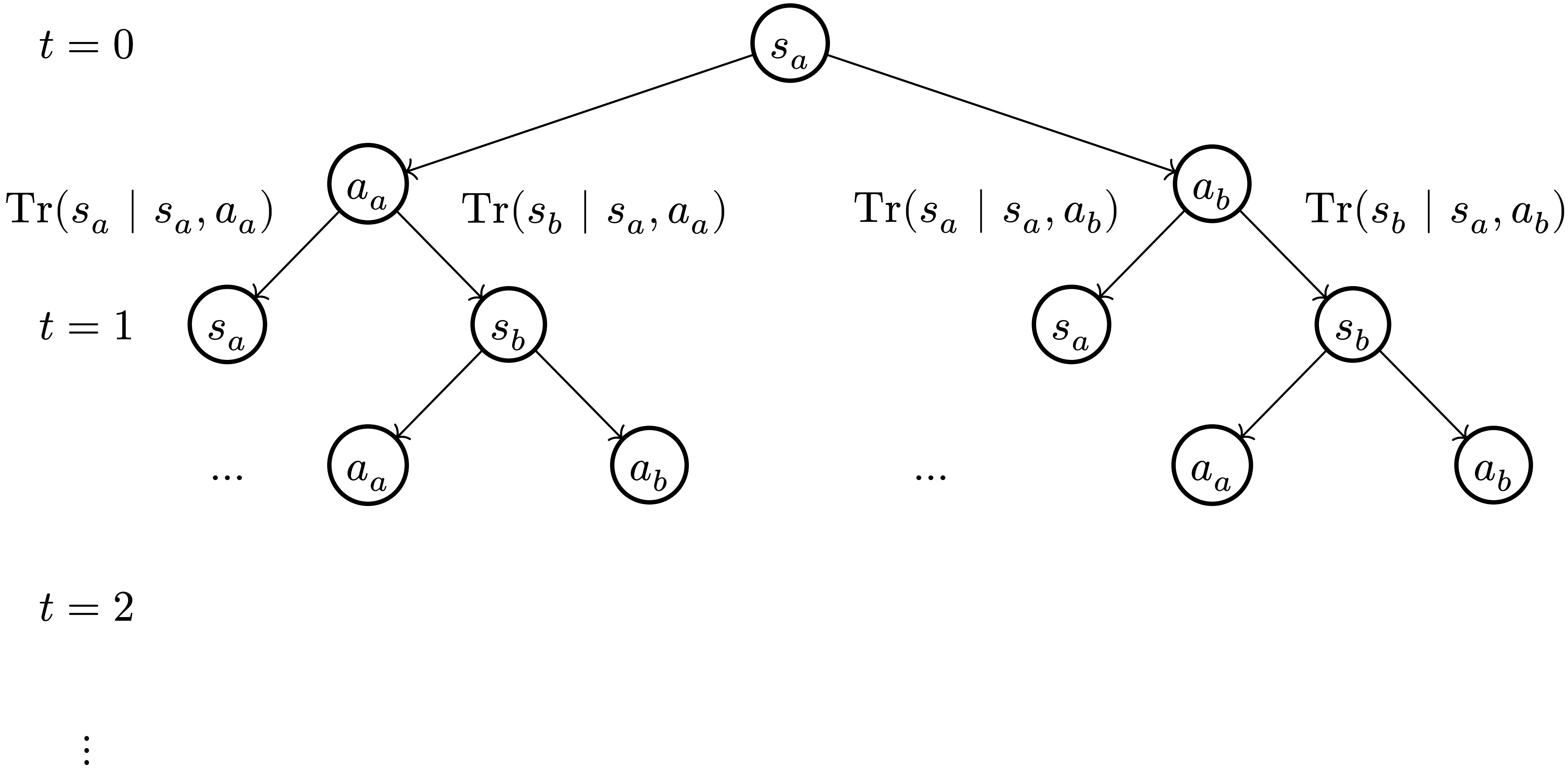
This is a very powerful equation

- 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 for the **infinite** future

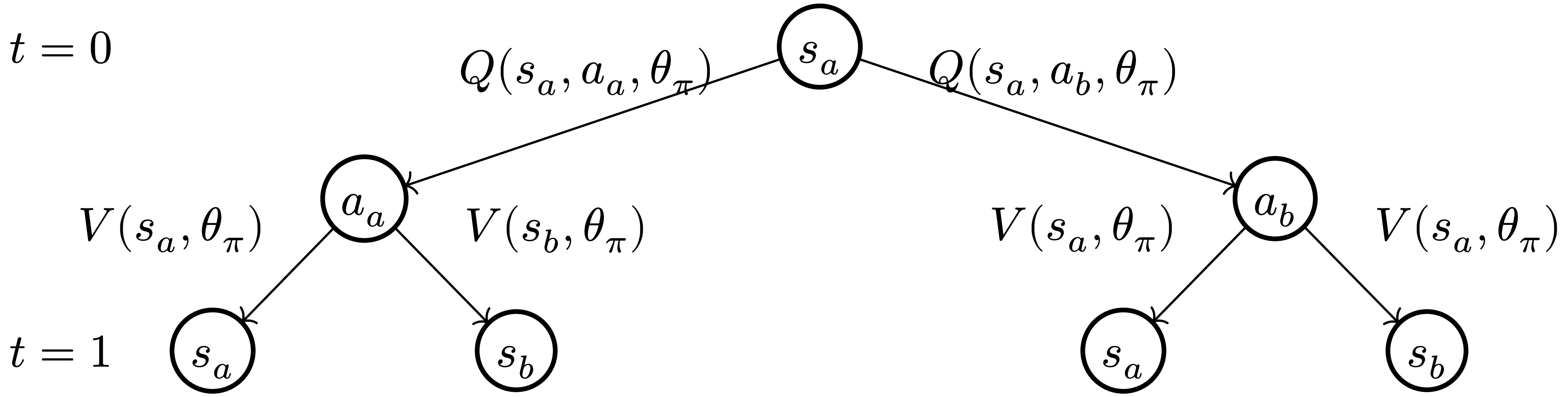
We collapsed the infinite decision tree into a single level

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
...
a_a
a_b
...
a_a
a_b
$t = 2$
$\vdots$

Q Functions



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

Q Functions

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

Q Functions

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

We will use either Q or V in every other algorithm in the course!

It is the core of decision making

Q Learning

Q Learning

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

It is still used heavily today

In fact, I am using it in our research right now

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

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

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 the definition of 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

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

What should our policy be?

Well we know the following is optimal

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

What if we say we follow a policy that maximizes Q?

Q Learning

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

What is the value function for this policy?

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

So we can rewrite the Q function without V

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

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

Homework

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