

L09

Introduction to Markov Decision Processes and RL

50.579 Optimization for Machine Learning

Ioannis Panageas

ISTD, SUTD

The framework

A finite Markov Decision Process (MDP) is defined as follows:

- A finite state space S .
- A finite action space A .
- A transition model P where $P(s'|s, a)$ is the probability of transitioning into state s' upon taking action a in state s . P is a matrix of size $(S \cdot A) \times S$.
- Reward function $r : S \times A \rightarrow [0, 1]$.
- A discounted factor $\gamma \in [0, 1)$.

The framework

A finite Markov Decision Process (MDP) is defined as follows:

- A finite state space S .
- A finite action space A .
- A transition model P where $P(s'|s, a)$ is the probability of transitioning into state s' upon taking action a in state s . P is a matrix of size $(S \cdot A) \times S$.
- Reward function $r : S \times A \rightarrow [0, 1]$.
- A discounted factor $\gamma \in [0, 1)$.

The **goal** is to find a **stationary** policy $\pi : S \rightarrow A$ such that the function

$$V^\pi(s) = (1 - \gamma) \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s \right]$$

is **maximized**. This is the **Infinite Time Horizon** case. $V^\pi(s) \in [0, 1]$

Example

Example (Navigation). Suppose you are given a *grid map*. The state of the agent is their *current location*. The four *actions* might be moving 1 step along each of east, west, north or south. The transitions in the simplest setting are deterministic. There is a goal g that is trying to reach. *Reward* is one if the agent reaches the goal and zero otherwise.

Example

Example (Navigation). Suppose you are given a *grid map*. The state of the agent is their *current location*. The four *actions* might be moving 1 step along each of east, west, north or south. The transitions in the simplest setting are deterministic. There is a goal g that is trying to reach. *Reward* is one if the agent reaches the goal and zero otherwise.

The *optimal behavior* π in this setting corresponds to finding the shortest path from the initial to the goal state.

Example

Example (Navigation). Suppose you are given a *grid map*. The state of the agent is their *current location*. The four *actions* might be moving 1 step along each of east, west, north or south. The transitions in the simplest setting are deterministic. There is a goal g that is trying to reach. *Reward* is one if the agent reaches the goal and zero otherwise.

The *optimal behavior* π in this setting corresponds to finding the shortest path from the initial to the goal state.

The value function of a state s , given the aforementioned policy is

$$V^\pi(s) = (1 - \gamma)\gamma^d,$$

where d is the number of steps to reach the goal from s .

State-action value function

Definition (Q-function). *In discounted infinite horizon problems, for any policy π , the state-action value function $Q : S \times A \rightarrow \mathbb{R}$ is given by*

$$Q^\pi(s, a) = (1 - \gamma) \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \text{ and } \pi(s_\tau) = a_\tau \forall \tau \right]$$

State-action value function

Definition (Q-function). In discounted infinite horizon problems, for any policy π , the state-action value function $Q : S \times A \rightarrow \mathbb{R}$ is given by

$$Q^\pi(s, a) = (1 - \gamma) \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \text{ and } \pi(s_\tau) = a_\tau \forall \tau \right]$$

By definition of Q function (for fixed policy) the following **equations** must be satisfied:

$$V^\pi(s) = Q^\pi(s, \pi(s)) \text{ and } Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} [V^\pi(s')]$$

namely $Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s' | s, a) V^\pi(s')$.

State-action value function

$$Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s').$$

$$Q^\pi(s, a) = (1 - \gamma)r(s, a) + \sum_{s'} P(s'|s_0 = s, a_0 = a) \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) | s_1 = s', \pi].$$

State-action value function

$$Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s').$$

$$\begin{aligned} Q^\pi(s, a) &= (1 - \gamma)r(s, a) + \sum_{s'} P(s'|s_0 = s, a_0 = a) \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) | s_1 = s', \pi]. \\ &= (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s_0 = s, a_0 = a) \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_1 = s', \pi]. \end{aligned}$$

State-action value function

$$Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s').$$

$$\begin{aligned} Q^\pi(s, a) &= (1 - \gamma)r(s, a) + \sum_{s'} P(s'|s_0 = s, a_0 = a) \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) | s_1 = s', \pi]. \\ &= (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s_0 = s, a_0 = a) \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_1 = s', \pi]. \\ &= (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s_0 = s, a_0 = a) V^\pi(s'). \end{aligned}$$

State-action value function

$$Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s').$$

$$\begin{aligned} Q^\pi(s, a) &= (1 - \gamma)r(s, a) + \sum_{s'} P(s'|s_0 = s, a_0 = a) \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) | s_1 = s', \pi]. \\ &= (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s_0 = s, a_0 = a) \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_1 = s', \pi]. \\ &= (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s_0 = s, a_0 = a) V^\pi(s'). \end{aligned}$$

Similarly, one can show that

$$V^\pi(s) = (1 - \gamma)r(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^\pi(s').$$

State-action value function

$$V^\pi(s) = Q^\pi(s, \pi(s)) \text{ and } Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} [V^\pi(s')]$$

namely $Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s')$.

That gives

$$Q^\pi = (1 - \gamma)r + \gamma P V^\pi.$$

State-action value function

$$V^\pi(s) = Q^\pi(s, \pi(s)) \text{ and } Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V^\pi(s')]$$

namely $Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s')$.

That gives

$$Q^\pi = (1 - \gamma)r + \gamma P V^\pi.$$

By substitution

$$Q^\pi = (1 - \gamma)r + \gamma P^\pi Q^\pi,$$

where P^π is a $(S \cdot A) \times (S \cdot A)$ that is induced by the P and the policy π .

State-action value function

$$V^\pi(s) = Q^\pi(s, \pi(s)) \text{ and } Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V^\pi(s')]$$

namely $Q^\pi(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s')$.

That gives

$$Q^\pi = (1 - \gamma)r + \gamma P V^\pi.$$

By substitution

$$Q^\pi = (1 - \gamma)r + \gamma P^\pi Q^\pi,$$

where P^π is a $(S \cdot A) \times (S \cdot A)$ that is induced by the P and the policy π .

We conclude that

$$Q^\pi = (1 - \gamma)(I - \gamma P^\pi)^{-1}r,$$

where $(I - \gamma P^\pi)^{-1}$ is invertible (why?).

Bellman Equations

We would like to find the optimal stationary policy, that is we want to

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

Lemma (Bellman Equations). *The following must hold:*

$$V^*(s) = \max_{a \in A} \left\{ (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^*(s') \right\}.$$

Equivalently for Q-function

$$Q^*(s, a) = (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, \pi(s)) \max_{b \in A} Q^*(s', b)$$

Bellman Equations

Proof.

$$\begin{aligned} V^*(s) &= \max_{\pi} V^{\pi}(s). \\ &= \max_{\pi} (1 - \gamma) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi, s, a \right]. \end{aligned}$$

Bellman Equations

Proof.

$$\begin{aligned} V^*(s) &= \max_{\pi} V^{\pi}(s). \\ &= \max_{\pi} (1 - \gamma) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi, s, a \right]. \\ &= \max_{a, \pi'} (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi', s' \right]. \end{aligned}$$

Bellman Equations

Proof.

$$\begin{aligned} V^*(s) &= \max_{\pi} V^{\pi}(s). \\ &= \max_{\pi} (1 - \gamma) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi, s, a \right]. \\ &= \max_{a, \pi'} (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi', s' \right]. \\ &= \max_{a, \pi'} (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^{\pi'}(s'). \end{aligned}$$

Bellman Equations

Proof.

$$\begin{aligned} V^*(s) &= \max_{\pi} V^{\pi}(s). \\ &= \max_{\pi} (1 - \gamma) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi, s, a \right]. \\ &= \max_{a, \pi'} (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi', s' \right]. \\ &= \max_{a, \pi'} (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^{\pi'}(s'). \\ &= \max_a \left\{ (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \max_{\pi'} V^{\pi'}(s') \right\}. \end{aligned}$$

Bellman Equations

Proof.

$$\begin{aligned} V^*(s) &= \max_{\pi} V^{\pi}(s). \\ &= \max_{\pi} (1 - \gamma) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi, s, a \right]. \\ &= \max_{a, \pi'} (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \mathbb{E} \left[\sum_t \gamma^t r(s_t, a_t) \mid \pi', s' \right]. \\ &= \max_{a, \pi'} (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^{\pi'}(s'). \\ &= \max_a \left\{ (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \max_{\pi'} V^{\pi'}(s') \right\}. \\ &= \max_a \left\{ (1 - \gamma) r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^*(s') \right\}. \end{aligned}$$

Bellman Operator

Definition (Bellman Operator). *Let's define the following operator T :*

$$TW(x) = \max_{a \in A} (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s' | s, a)W(s')$$

Bellman Operator

Definition (Bellman Operator). *Let's define the following operator T :*

$$TW(x) = \max_{a \in A} (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s' | s, a)W(s')$$

Claim (Bellman Operator). *V^* is the unique fixed point of the operator.*

Bellman Operator

Definition (Bellman Operator). *Let's define the following operator T :*

$$TW(x) = \max_{a \in A} (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s' | s, a)W(s')$$

Claim (Bellman Operator). *V^* is the unique fixed point of the operator.*

Proof. We have already shown it is a fixed point. We will show that T is **contracting!** (Banach Fixed point Theorem).

Bellman Operator

Definition (Bellman Operator). Let's define the following operator T :

$$TW(x) = \max_{a \in A} (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a)W(s')$$

Claim (Bellman Operator). V^* is the unique fixed point of the operator.

Proof. We have already shown it is a fixed point. We will show that T is **contracting!** (Banach Fixed point Theorem).

Consider f, f' and observe that

$$|\max_a f(a) - \max_{a'} f'(a')| \leq \max_a |f(a) - f'(a)|$$

Bellman Operator

Proof cont. Assume a maximizes $f(a)$ and moreover $f(a) \geq \max_{a'} f'(a')$ (w.l.o.g due to symmetry). Then we get

$$f(a) - \max_{a'} f'(a') \leq f(a) - f'(a) \leq \max_b f(b) - f'(b).$$

Bellman Operator

Proof cont. Assume a maximizes $f(a)$ and moreover $f(a) \geq \max_{a'} f'(a')$ (w.l.o.g due to symmetry). Then we get

$$f(a) - \max_{a'} f'(a') \leq f(a) - f'(a) \leq \max_b f(b) - f'(b).$$

Therefore

$$\|TV - TV'\|_\infty = (1 - \gamma) \left[\max_a r(s, a) + \sum_{s'} P(s'|a, s) V(s') - \max_{a'} r(s, a') - \sum_{s'} P(s'|a', s) V'(s') \right]$$

Bellman Operator

Proof cont. Assume a maximizes $f(a)$ and moreover $f(a) \geq \max_{a'} f'(a')$ (w.l.o.g due to symmetry). Then we get

$$f(a) - \max_{a'} f'(a') \leq f(a) - f'(a) \leq \max_b f(b) - f'(b).$$

Therefore

$$\begin{aligned} \|TV - TV'\|_\infty &= (1 - \gamma) \left[\max_a r(s, a) + \sum_{s'} P(s'|a, s) V(s') - \max_{a'} r(s, a') - \sum_{s'} P(s'|a', s) V'(s') \right]_\infty \\ &\leq (1 - \gamma) \max_a \left[r(s, a) + \sum_{s'} P(s'|a, s) V - r(s, a) - \sum_{s'} P(s'|a, s) V' \right]_\infty \end{aligned}$$

Bellman Operator

Proof cont. Assume a maximizes $f(a)$ and moreover $f(a) \geq \max_{a'} f'(a')$ (w.l.o.g due to symmetry). Then we get

$$f(a) - \max_{a'} f'(a') \leq f(a) - f'(a) \leq \max_b f(b) - f'(b).$$

Therefore

$$\begin{aligned} \|TV - TV'\|_\infty &= (1 - \gamma) [\max_a r(s, a) + \sum_{s'} P(s'|a, s)V(s') - \max_{a'} r(s, a') - \sum_{s'} P(s'|a', s)V'(s')]_\infty \\ &\leq (1 - \gamma) \max_a [r(s, a) + \sum_{s'} P(s'|a, s)V - r(s, a) - \sum_{s'} P(s'|a, s)V']_\infty \\ &= (1 - \gamma) \max_a [\sum_{s'} P(s'|a, s)(V - V')]_\infty \end{aligned}$$

Bellman Operator

Proof cont. Assume a maximizes $f(a)$ and moreover $f(a) \geq \max_{a'} f'(a')$ (w.l.o.g due to symmetry). Then we get

$$f(a) - \max_{a'} f'(a') \leq f(a) - f'(a) \leq \max_b f(b) - f'(b).$$

Therefore

$$\begin{aligned} \|TV - TV'\|_\infty &= (1 - \gamma) [\max_a r(s, a) + \sum_{s'} P(s'|a, s)V(s') - \max_{a'} r(s, a') - \sum_{s'} P(s'|a', s)V'(s')]_\infty \\ &\leq (1 - \gamma) \max_a [r(s, a) + \sum_{s'} P(s'|a, s)V - r(s, a) - \sum_{s'} P(s'|a, s)V']_\infty \\ &= (1 - \gamma) \max_a [\sum_{s'} P(s'|a, s)(V - V')]_\infty \\ &\leq (1 - \gamma) \|V - V'\|_\infty \max_a [\sum_{s'} P(s'|a, s)] = (1 - \gamma) \|V - V'\|_\infty. \end{aligned}$$

Value Iteration

Idea: We build a sequence of value functions. Let V_0 be any vector, then iterate the application of the optimal Bellman operator so that given V_k at iteration k we compute

$$V_{k+1} = TV_k.$$

Value Iteration

Idea: We build a sequence of value functions. Let V_0 be any vector, then iterate the application of the optimal Bellman operator so that given V_k at iteration k we compute

$$V_{k+1} = TV_k.$$

The policy will be given at every iteration as

$$\pi_k = \arg \max_a (1 - \gamma)r(s, a) + \gamma \sum_{s'} P(s'|s, a) V_k(s')$$

After $k = \frac{\log(1/\epsilon)}{\log(1/\gamma)}$ we have error ϵ .

Policy Iteration

Idea: We build a sequence of policies. Let π_0 be any stationary policy. At each iteration k we perform the two following steps:

1. **Policy evaluation** given π_k , compute V^{π_k} .
2. **Policy improvement**: we compute the *greedy* policy π_{k+1} from V^{π_k} as:

$$\pi_{k+1}(x) \in \arg \max_{a \in A} \left[r(x, a) + \gamma \sum_y p(y|x, a) V^{\pi_k}(y) \right].$$

The iterations continue until $V^{\pi_k} = V^{\pi_{k+1}}$.

Conclusion

- Introduction to Markov Decision Processes.
 - Policy Iteration
 - Value Iteration
 - Bellman Equations
- I hope you enjoyed the lectures! Next week we have the **presentations**.