Markov Decision Processes & Stochastic Games

Stelios Stavroulakis & Fivos Kalogiannis

Algorithmic Game Theory, Fall 2022

Motivation

- Real world problems are often **sequential**
- Going through states requires taking actions. Taking action now affects the future

The Markov Decision Process (MPD) captures the above aspects and provides a **general framework** for sequential decision-making.

Formalism

The Markov Decision Process is represented as a **discrete-time dynamicsl system** reactive to the actions taken by the agent. Formally, MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$

- A finite state space $\mathcal{S}$
- A finite action space $\mathcal{A}$
- A transition model $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
- A reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$
- A discount factor $\gamma \in [0, 1)$
- An initial state distribution $\mu \in \Delta(\mathcal{S})$

Policies

A decision-making protocol, a strategy in which the agent chooses actions.

Below is a deterministic policy:

➡️	➡️	➡️	🍇
⬆️	➡️	➡️	⬆️

Policies can also be stochastic, here is a stochastic one:

➡️ 100%	➡️ 100%	➡️ 100%	🍇
⬆️ 50%, ➡️ 50%	⬆️ 50%, ➡️ 50%	⬆️ 50%, ➡️ 50%	⬆️ 100%

Policies

- Policies can use history: $\pi : \mathcal{H} \to \Delta(\mathcal{A})$
- Or be **Markovian**: $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
- Policies can be stationary: $\pi_t = \pi, \forall t$
- Or be non-stationary: $\exists\, t, t' : \pi_t \neq \pi_{t'}$

Values

Pick a policy, how good is that policy at every state?

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

- The value is the expected discounted sum of rewards collected under policy π.
- Values allow to query the quality of the current situation instead of waiting to observe the long-run outcome.

Our Goal ⚽

Given a state s, the goal of the agent is to find a **Markovian** policy π that maximizes the value:

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

- The $\max()$ operator is over all (possibly non-stationary and randomized) policies.
- Access to $V^{\star}$ yields optimal behavior if:

$$\pi^{\star}(s) = \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V(s') \right\}$$

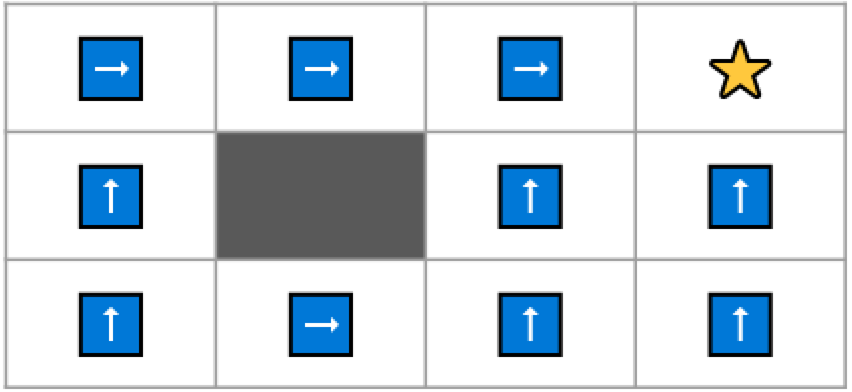
Examples

Navigation

- State: Current location
- Actions: 4 cardinal directions
- Transitions: Deterministic
- Rewards: 1 if goal reached, else 0

Optimal policy: Shortest path from initial to goal state

Optimal value γ^d



0.729	0.81	0.9	☆
0.656		0.81	0.9
0.590	0.656	0.729	0.81

Optimal Policies

> **Definition**: Recall that $\pi_1 \geq \pi_2$ if and only if $v_{\pi_1}(s) \geq v_{\pi_2}(s)\ \forall s \in \mathcal{S}$

An optimal policy $\pi^\star$ is one which is as good as or better than any other policy π'. The value function associated with that policy achieves maximum value in every state s:

$$V^{\pi^\star}(s) = \mathbb{E}_{\pi^\star}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_t = s] = \max_{\pi} V^{\pi}(s)\ \forall s \in \mathcal{S}$$

All optimal policies have the same optimal value function which we denote by $V^\star$

Bellman Equations

bike

The Bellman equations allow us to relate the value of the **current state** with the value of **future states** without waiting to observe rewards.

$$V(s) = r(s, \pi) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi)V(s')$$

Bellman Evaluation Operator

The Bellman evaluation operator $T^\pi : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ defined by its action on $\mathcal{S}$ via any $V : S \to \mathbb{R}$ in the following way:

$$(T^\pi V)(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi(s))V(s')$$

Notice the fixed point of this operator V^π

$$T^\pi V^\pi = V^\pi$$

T^π is an affine linear operator yielding a linear system of equations.

Contraction and Monotonicity of $\mathcal{T}_\pi$

Basic definitions:

Distance function: For value functions V, V' we define their distance as the maximum absolute value of the differences between values:

$$d(V, V') = \max_{s \in \mathcal{S}} |V(s) - V(s')|$$

Contraction Mapping: A function f is a contraction mapping if:

$$\exists\, k \in [0, 1) : d(f(x), f(y)) \leq k d(x, y)\ \forall x, y$$

Claim: The contraction property holds for Bellman evaluation operator $T_\pi \forall\ \pi$.

Proof: For any value V, V' and any policy π we have:

$$\begin{aligned} d(\mathcal{T}_\pi V, \mathcal{T}_\pi V') &= \max_{s \in \mathcal{S}} |\gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s))(V(s') - V'(s'))| \\ &\leq \max_{s \in \mathcal{S}} \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s))|(V(s') - V'(s'))| \\ &\leq \max_{s \in \mathcal{S}} \gamma \max_{s' \in \mathcal{S}} |(V(s') - V'(s'))| \\ &= \gamma d(V, V') \end{aligned}$$

Therefore, $\mathcal{T}_\pi$ is a contraction mapping.

Bellman Optimality Equations

The optimal value is given by the Bellman Optimality Equation defined below:

By substituting $\pi^\star$ into the Bellman equation and leveraging the fact that an optimal deterministic policy always exists, we replace the policy distribution over actions with best action:

$$V^\star(s) = \max_a \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^\star(s') \right), \ \forall s \in \mathcal{S}$$

Bellman Optimality Operator

Similarly to the Bellman evaluation operator, the Bellman optimality operator $\mathcal{T}$ is defined as:

$$(\mathcal{T}V)(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V(s') \right\}$$

The optimal value $V^\star$ is a fixed point of the operator $\mathcal{T}$.

Contraction property of $\mathcal{T}$

Consider an arbitrary $V, V', \forall\pi$ and we can write:

$$\text{Case 1: } (\mathcal{T}_\pi V)(s) \leq (\mathcal{T}_\pi V')(s) + \gamma d(V, V')$$

$$\text{Case 2: } (\mathcal{T}_\pi V')(s) \leq (\mathcal{T}_\pi V)(s) + \gamma d(V, V')$$

For any fixed s, we take the $\max$ on both sides in Case 1 (same for Case 2):

$$\max_{\pi\in\Pi}\{\mathcal{T}_\pi V(s)\} = \max_{\pi(s)\in\mathcal{A}}\{\mathcal{T}_\pi V(s)\} \leq \max_{\pi(s)\in\mathcal{A}}\{\mathcal{T}_\pi V'(s)\} + \gamma d(V, V')$$

$$\Rightarrow \mathcal{T}V(s) \leq \mathcal{T}V'(s) + \gamma d(V, V')$$

Similarly, Case 2 yields $\mathcal{T}V'(s) \leq \mathcal{T}V(s) + \gamma d(V, V')$.

Therefore: $|\mathcal{T}V(s) - \mathcal{T}V'(s)| \leq \gamma d(V, V') \; \forall s \in \mathcal{S}$

The optimal value is unique!

- When $\gamma \in (0, 1)$, $\mathcal{T}^{\pi}$ is a max-norm **contraction**
- The **fixed-point equation** $\mathcal{T}^{\pi} V = V$ has a unique solution by the **Banach Fixed Point Theorem**.
- Unique solution is exactly V^{π}!

Why bother?

- The uniqueness of the optimal value $V^\star$ provides a guarantee that no matter out initialization, given that the Bellman operator is a contraction mapping, converges to the (unique) optimal value!
- An algorithm that iteratively applies the Bellman operator, will always converge, and the values in each state will be simultaneously optimal at every state s.

How to solve MDPs

There are many ways to solve MDPs, each with their own benefits and drawbacks.

- **Dynamic Programming (DP)**.
  (+) Well developed mathematically
  (-) It requires the **full description** of the model of the environment (functions $P, r, \forall s, a \in \mathcal{S} \times \mathcal{A}$)
- **Monte Carlo methods (MC)**
  (+) Do not require full model and are conceptually simple (just sample trajectories)
  (-) Noisy
  (-) Update are always done at the
- **Temporal Difference Methods** (a combination of DP and MC), and more...

Value Iteration (DP)

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

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

which means, $\forall s \in \mathcal{S}$:

$$\begin{aligned} V_{k+1}(s) &= \max_{a \in \mathcal{A}} \mathbb{E}[r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s, a_t = a] \\ &= \max_{a \in \mathcal{A}} \sum_{s'} P(s'|s,a)[r(s,a) + \gamma V_k(s')] \end{aligned}$$

$\{V_k\}$ will converge to $V^\star$ and the value at the fixed point $V^\star$ is optimal.

Value Iteration (DP)

- We know that the Bellman optimality operator $T^\star$ has a unique fixed point. We found one above and the uniqueness of it is settles from the contraction property of the Bellman operator.
- $V^\star$ is a fixed point of $T^\star$ by the Bellman optimality equation
- By the Banach fixed point theorem, value iteration converges to $V^\star$ at a geometric rate.

The policy will be given at every iteration as:

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

After $k = \frac{\log \frac{1}{\epsilon}}{\log \frac{1}{\gamma}}$ steps, we have error ϵ.