MDP

Blitz Course

Fabien Pesquerel and Odalric-Ambrym Maillard

January 6, 2021

MDP

4-tuple (S, A, p, r):

1. S: State space
2. A: Action space
3. $p : (s', s, a) \in S \times S \times A \mapsto p(s'|s, a)$: transition model
4. $r : (s, a) \in S \times A \mapsto r(s, a)$: reward model

Policy

Mapping

$$\pi : S \to A$$

(State) Value function

Value function of a policy:

$$V_\gamma^\pi(s) = \mathbb{E}_{\pi,MDP}\left(\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right)$$

State-action Value function

State-action value function of a policy:

$$Q^{\pi}_{\gamma}(s,a) = \mathbb{E}_{\pi,MDP}\left(\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s,\ a_0 = a\right)$$

Optimal policies

An optimal policy π^* is associated to **the** value function that is uniformly dominant:

$$\forall \pi, \forall s, \;\; V^{\pi^*}(s) \geq V^{\pi}(s)$$

Bellman operator - Value function

V^π is the unique fixed point: $V = T^\pi V$.

$$V(s) = r\left(s, \pi(s)\right) + \gamma \sum_{s' \in S} p\left(s' | s, \pi(s)\right) V(s')$$

- T^π is called the Bellman operator of the policy π.
- $V = T^\pi V$ is a linear system of equations.
- T^π is a γ-contraction.

Bellman operator - State-action value function

Q^π is the unique fixed point: $Q = T^\pi Q$.

$$Q(s,a) = r\left(s,a\right) + \gamma \sum_{s' \in S} p\left(s' | s, a\right) Q(s', \pi(s'))$$

- T^π is called the Bellman operator of the policy π.
- $Q = T^\pi Q$ is a linear system of equations.
- T^π is a γ-contraction.

Computing value functions

- Invert the system
- Compositions of the contraction map
- Monte-Carlo estimation

Optimal Bellman operator - Value function

V^* is the unique fixed point: $V = T^*V$.

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

- T^* is called the optimal Bellman operator.
- $V = T^*V$ is a **not** a linear system of equations.
- T^* is a γ-contraction.

Optimal Bellman operator - State-action Value function

Q^* is the unique fixed point: $Q = T^*Q$.

$$Q(s,a) = r\left(s,a\right) + \gamma \sum_{s' \in S} p\left(s'|s,a\right) \max_{a' \in A} Q(s',a')$$

- T^* is called the optimal Bellman operator.
- $Q = T^*Q$ is **not** a linear system of equations.
- T^* is a γ-contraction.

Computing optimal value functions

- Invert the system (hard)
- Compositions of the contraction map (Value Iteration)
- Monte-Carlo estimation followed by $\max$ non-linearities (Policy Iteration)

Algorithms

- **Value Iteration** : Iterate the optimal Bellman operator of your choice. Once convergence to V^* or Q^* is assumed, compute π^* as the greedy policy with respect to Q^*.
- **Policy Iteration** : Alternate between *policy evaluation*, *i.e.* evaluate the current policy, and *policy improvement*, *i.e.* greedy selection of actions according to the current policy.