Lecture 3: Dynamic Programming. Policy and Value Iterations

Anton Plaksin

Markov Decision Process

Markov Property

$$\mathbb{P}[S_{t+1}|S_t, A_t] = \mathbb{P}[S_{t+1}|S_1, A_1, S_2, A_2 \dots, S_t, A_t]$$

$$\mathbb{P}[R_t|S_t, A_t] = \mathbb{P}[R_t|S_1, A_1, S_2, A_2 \dots, S_t, A_t] = 1$$

Markov Decision Process $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite ($|\mathcal{S}| = n$) state space
- $\mathcal{A}$ is a finite ($|\mathcal{A}| = m$) action space
- $\mathcal{P}$ is a known transition probability function

$$\mathcal{P}(s'|s, a) = \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a]$$

- $\mathcal{P}_0$ is a known initial state probability function
- $\mathcal{R}$ is a known reward function

$$\mathcal{R}(s, a) = R_t \quad \Leftrightarrow \quad \mathbb{P}[R_t|S_t = s, A_t = a] = 1$$

- $\gamma \in [0, 1]$ is a discount coefficient

Stochastic Policy

$$\pi(a|s) \in [0,1], \quad a \in \mathcal{A}, \quad s \in \mathcal{S}$$

- Set π
- Agent starts from the initial state $S_0 \sim \mathcal{P}_0$
- acts $A_0 \sim \pi(\cdot|S_0)$
- gets the reward $R_0 = \mathcal{R}(S_0, A_0)$ and goes to the next state $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- acts $A_1 \sim \pi(\cdot|S_1)$
- gets the reward $R_1 = \mathcal{R}(S_1, A_1)$ and goes to the next state $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$
- $\ldots$
- $\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}, \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$

The Reinforcement Learning problem

$$\mathbb{E}_\pi[G] \quad \longrightarrow \quad \max_\pi$$

Value Function

- Set π and s
- Agent starts from the initial state $S_0 = s$
- acts $A_0 \sim \pi(\cdot|S_0)$
- gets the reward $R_0 = \mathcal{R}(S_0, A_0)$ and goes to the next state $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- acts $A_1 \sim \pi(\cdot|S_1)$
- gets the reward $R_1 = \mathcal{R}(S_1, A_1)$ and goes to the next state $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$
- $\ldots$
- $\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}, \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$

Value Function

$$v_\pi(s) = \mathbb{E}_\pi[G]$$

Deterministic Case

Remark

If Policy and Environment are deterministic (non-stochastic) then

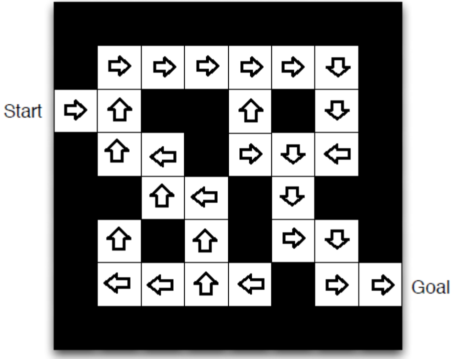
$$v_\pi(s) = G(\tau_\pi),$$

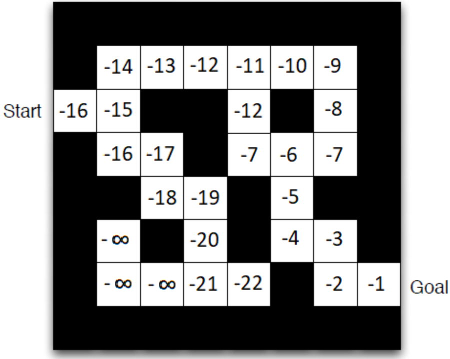
where $\tau_\pi : \mathbb{P}(\tau_\pi|\pi) = 1$.

Пример: Maze

v_π :

$R_t = -1, \quad \gamma = 1, \quad \pi$:





Bellman Expectation Equation

$$\tau = (S_0, A_0, S_1, A_1, S_2, A_2, \ldots), \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$$

$$\tilde{\tau} = (S_1, A_1, S_2, A_2, S_3, A_3 \ldots), \quad G(\tilde{\tau}) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_{t+1}, A_{t+1})$$

$$G(\tau) = \mathcal{R}(S_0, A_0) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} \mathcal{R}(S_t, A_t) = \mathcal{R}(S_0, A_0) + \gamma G(\tilde{\tau})$$

Bellman Expectation Equation for v_π

$$v_\pi(s) = \sum_a \pi(a|s) \Big(\mathcal{R}(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v_\pi(s') \Big)$$

How to solve Bellman Expectation Equation?

Bellman Expectation Equation for v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a) + \gamma \sum_{s'} \sum_a \pi(a|s)\mathcal{P}(s'|s,a) v_\pi(s')$$

$$\mathcal{R}_\pi(s) = \sum_a \pi(a|s)\mathcal{R}(s,a), \quad \mathcal{P}_\pi(s',s) = \sum_a \pi(a|s)\mathcal{P}(s'|s,a)$$

$$v_\pi(s) = \mathcal{R}_\pi(s) + \gamma \sum_{s'} \mathcal{P}_\pi(s',s) v_\pi(s')$$

$$v_\pi = \begin{pmatrix} v_\pi(s_1) \\ \cdots \\ v_\pi(s_n) \end{pmatrix}, \mathcal{R}_\pi = \begin{pmatrix} \mathcal{R}_\pi(s_1) \\ \cdots \\ \mathcal{R}_\pi(s_n) \end{pmatrix}, \mathcal{P}_\pi = \begin{pmatrix} \mathcal{P}_\pi(s_1,s_1) & \dots & \mathcal{P}_\pi(s_1,s_n) \\ \vdots & \ddots & \vdots \\ \mathcal{P}_\pi(s_n,s_1) & \dots & \mathcal{P}_\pi(s_n,s_n) \end{pmatrix}$$

How to solve Bellman Expectation Equation?

Bellman Expectation Equation for v_π

$$v_\pi(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')\Big)$$

$$v_\pi = \mathcal{R}_\pi + \gamma \mathcal{P}_\pi v_\pi$$

$$(E - \gamma \mathcal{P}_\pi) v_\pi = \mathcal{R}_\pi$$

$$v_\pi = (E - \gamma \mathcal{P}_\pi)^{-1} \mathcal{R}_\pi$$

Theorem

If $\gamma < 1$ then there exists a unique solution v_π of Bellman Expectation Equation.

Iterative Policy Evaluation (Fixed-Point Iteration)

Let π; $v^0(s)$, $s \in \mathcal{S}$, $K \in \mathbb{N}$.

For each $k \in \overline{0, K}$, do

$$v^{k+1}(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v^k(s')\Big), \quad s \in \mathcal{S}$$

or

$$v^{k+1} = \mathcal{R}_\pi + \gamma \mathcal{P}_\pi v^k$$

Theorem

$v^k \to v_\pi$, $k \to \infty$. Convergence rate $O(mn^2)$

Action-Value Function

Action-Value Function

$$q_\pi(s,a) = \mathbb{E}_\pi[G \mid S_0 = s,\ A_0 = a]$$

q_π and v_π

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a), \quad q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s')$$

Bellman Expectation Equation for q_π

$$q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a')$$

Policy Improvement

Partially Order for Policies

$$\pi' \geq \pi \quad \Leftrightarrow \quad v_{\pi'}(s) \geq v_\pi(s), \quad \forall s \in \mathcal{S}$$

Greedy Policy Improvement

$$\pi'(a|s) = \begin{cases} 1, \text{ if } a \in \text{argmax}_{a' \in \mathcal{A}}\, q_\pi(s, a') \\ 0, \text{ otherwise} \end{cases}$$

Policy Improvement Theorem

Let π. If π' is defined by Greedy Policy Improvement then

$$\pi' \geq \pi$$

Optimal Policy

(Optimal) Value Function and Action-Value Function

$$v_*(s) = \max_{\pi} v_\pi(s), \quad q_*(s,a) = \max_{\pi} q_\pi(s,a)$$

Optimal Policy Existence Theorem

There exists a (optimal) policy π_* such that

- $\pi_* \geq \pi, \ \forall \pi$
- $v_{\pi_*}(s) = v_*(s), \ \forall s \in \mathcal{S}$
- $q_{\pi_*}(s,a) = q_*(s,a), \ \forall s \in \mathcal{S}, \ \forall a \in \mathcal{A}$

Policy Iteration

Let π^0 and $L, K \in \mathbb{N}$.
For each $k \in \overline{0, K}$, do

- (Policy evaluation) Iterative Policy Evaluation

$$v^{l+1} = \mathcal{R}_{\pi^k} + \mathcal{P}_{\pi^k} v^l, \quad l \in \overline{0, L-1}.$$

  Define $q^L(s,a)$ by $v^L(s)$

- (Policy improvement) Greedy Policy Improvement

$$\pi^{k+1}(a|s) = \begin{cases} 1, & \text{if } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q^L(s, a') \\ 0, & \text{otherwise} \end{cases}$$

Theorem

$\pi^k \to \pi_*$, $k \to \infty$. Convergence rate $O(mn^2)$

Bellman Optimality Equations

Bellman Optimality Equations for v_*

$$v_*(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v_*(s') \right)$$

Bellman Optimality Equations for q_*

$$q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \max_{a' \in \mathcal{A}} q_*(s',a')$$

v_* and q_*

$$v_*(s) = \max_{a \in A} q_*(s,a), \quad q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v_*(s')$$

π_* and q_*

$$\pi_*(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q_*(s,a') \\ 0, & \text{иначе} \end{cases}$$

Value Iteration

Let $v^0(s)$, $s \in \mathcal{S}$ and $K \in \mathbb{N}$.

For each $k \in \overline{0, K}$, do

$$v^{k+1}(s) = \max_{a \in \mathcal{A}} \Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v^k(s') \Big), \quad s \in \mathcal{S}$$

Theorem

$v^k \to v_*$, $k \to \infty$. Convergence rate $O(mn^2)$

- Definitions of v_π, q_π, v_*, q_*, π_* will be used for the general MDP (when $\mathcal{S}$ and $\mathcal{A}$ are infinite, and $\mathcal{P}$ and $\mathcal{R}$ are unknown)
- Bellman Expectation Equation for v_π and q_π, and Bellman Optimality Equation for v_* and q_* as well as Policy Improvement Theorem and Optimal Policy Existence Theorem hold in the case of MDP in which $\mathcal{S}$ and $\mathcal{A}$ are finite, but $\mathcal{P}$ and $\mathcal{R}$ can be unknown.
- Policy Iteration and Value Iteration algorithms are only for the case of MDP in which $\mathcal{S}$ and $\mathcal{A}$ are finite, and $\mathcal{P}$ and $\mathcal{R}$ are known.