

Markov Decision Processes: Discrete Stochastic Dynamic Programming

Peng Lingwei

August 7, 2019

Contents

4	Chapter4: Finite-Horizon Markov Decision Processes	2
4.1	OPTIMALITY CRITERIA	2
4.1.1	Some Preliminaries	2
4.1.2	The Expected Total Reward Criterion	2
4.1.3	Optimal Policies	3
4.2	FINITE-HORIZON POLICY EVALUATION	3
4.3	OPTIMALITY EQUATIONS AND THE PRINCIPLE OF OPTIMALITY	4
4.4	OPTIMALITY OF DETERMINISTIC MARKOV POLICIES	6
4.5	BACKWARD INDUCTION	7
4.6	OPTIMALITY OF MONOTONE POLICIES	7
4.6.1	Structured Policies	7
4.6.2	Superadditive Functions	7
4.7	Optimality of Monotone Policies	7
5	Infinite-Horizon Models: Foundations	8
5.1	THE VALUE OF A POLICY	8
5.2	MARKOV POLICIES	8
6	Discounted Markov Decision Problems	10
6.1	POLECY EVALUATION (Stationary Policy)	10
6.2	OPTIMALITY EQUATIONS	10
6.3	VALUE ITERATION AND ITS VARIANTS	13
6.3.1	Rates of Convergence	13
6.3.2	Value Iteration	13

4 Chapter4: Finite-Horizon Markov Decision Processes

4.1 OPTIMALITY CRITERIA

4.1.1 Some Preliminaries

About MDP:

1. $\pi = (d_1, d_2, \dots, d_{N-1}) \in \Pi^{HR}$;
2. $h_N = (s_1, a_1, s_2, \dots, s_N)$
3. Rewards sequence: $\{r_1(s_1, a_1), r_2(s_2, a_2), \dots, r_{N-1}(s_{N-1}, a_{N-1}), r_N(s_N)\}$
 - $\pi \in \Pi^{HD}, \{r_1(X_1, d_1(H_1)), \dots, r_{N-1}(X_{N-1}, d_{N-1}(H_{N-1})), r_N(X_N)\}$
 - $\pi \in \Pi^{MD}, \{r_1(X_1, d_1(X_1)), \dots, r_{N-1}(X_{N-1}, d_{N-1}(X_{N-1})), r_N(X_N)\}$
4. $R = (R_1, R_2, \dots, R_N)$, where $R_t = r_t(X_t, Y_t)$, and $|R_t| \leq M < \infty$.
5. $\mathbb{P}_R^\pi(r_1, r_2, \dots, r_N) = \mathbb{P}^\pi[\{(s_1, a_1, \dots, s_N) : (r(s_1, a_1), \dots, r_N(s_N)) = (r_1, \dots, r_N)\}]$

Definition:

1. The random vairable U is stochastically greater than V:

$$\forall t \in \mathbb{R}, \quad P(V > t) \leq P(U > t).$$

2. Probability distribution P_2 is stochastically greater than P_1 if:

$$\forall t \in \mathbb{R}, \quad \int_t^\infty p_1(t)dt \leq \int_t^\infty p_2(t)dt.$$

3. The random vector $\vec{U} = (U_1, \dots, U_n)$ is stochastically greater than the random vector $\vec{V} = (V_1, \dots, V_n)$:

$$\forall f \in \{f : \mathbb{R}^n \rightarrow \mathbb{R} | \vec{v} \preceq \vec{u} \Rightarrow f(\vec{v}) \leq f(\vec{u})\}, \quad \mathbb{E}[f(\vec{V})] \leq \mathbb{E}[f(\vec{U})]$$

4.1.2 The Expected Total Reward Criterion

The expected total reward criterion:

1. $\pi \in \Pi^{HR}$: $v_N^\pi(s) = \mathbb{E}_S^\pi \left\{ \sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right\}$.
2. $\pi \in \Pi^{HD}$: $v_N^\pi(s) = \mathbb{E}_S^\pi \left\{ \sum_{t=1}^{N-1} r_t(X_t, d_t(H_t)) + r_N(X_N) \right\}$.
3. Discounted reward: $\pi \in \Pi^{HR}$,
 $v_{N,\lambda}^\pi(s) = \mathbb{E}_S^\pi \left\{ \sum_{t=1}^{N-1} \lambda^{t-1} r_t(X_t, d_t(H_t)) + \lambda^{N-1} r_N(X_N) \right\}$.

Taking the discount factor into account does not effect any theoretical results or algorithms in the finite-horizon case but might effect the decision maker's preference for policies.

4.1.3 Optimal Policies

Definition:

1. Optimal policy $\pi^* : \forall \pi \in \Pi^{HR}, v_N^{\pi^*} \succeq v_N^\pi$.
2. ϵ -optimal policy, $\pi^* : \forall \pi \in \Pi^{HR}, v_N^{\pi^*} + \epsilon \succeq v_N^\pi$.
3. Optimal value: $v_N^* = \sup_{\pi \in \Pi^{HR}} v_N^\pi$.
4. We can get $v_N^{\pi^*} = v_N^*$ and $v_N^{\pi^*} + \epsilon > v_N^*$.
5. Considering initial state distribution P_1 : $v_N^{\pi, P_1} = \sum_{s \in S} v_N^\pi(s) P_1\{X_1 = s\}$.

Markov decision problem = Markov decision process + Optimality criteria

4.2 FINITE-HORIZON POLICY EVALUATION

1. $\pi = (d_1, d_2, \dots, d_{N-1}) \in \Pi^{HR}$
2. Define: $u_t^\pi(h_t) = \mathbb{E}_{h_t}^\pi \left\{ \sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\}$, $(u_t^\pi : H_t \rightarrow \mathbb{R})$.
And we define $u_N^\pi(h_N) = r_N(s_N)$.
3. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{HD}$):

$$\begin{aligned} \hat{u}_t^\pi(h_t) &= r_t(s_t, d_t(h_t)) + \sum_{s' \in S} p_t(s' | s_t, d_t(h_t)) \hat{u}_{t+1}^\pi(h_t, d_t(h_t), s'). \quad ((h_t, d_t(h_t), s') \in H_{t+1}) \\ &= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^\pi \left\{ \hat{u}_{t+1}^\pi(h_t, d_t(h_t), X_{t+1}) \right\} \end{aligned}$$

Proof. Part proof with backward induction hypothesis ($u_{h_{t+1}}^\pi = \hat{u}_{h_{t+1}}^\pi$):

$$\begin{aligned} \hat{u}_t^\pi(h_t) &= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^\pi \left\{ u_{t+1}^\pi(h_t, d_t(h_t), X_{t+1}) \right\} \\ &= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^\pi \left\{ \mathbb{E}_{h_{t+1}}^\pi \left\{ \sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\} \right\} \\ &= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^\pi \left\{ \sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\} \\ &= \mathbb{E}_{h_t}^\pi \left\{ r_t(s_t, d_t(h_t)) + \sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\} = u_t^\pi(h_t) \end{aligned}$$

□

4. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{HR}$):

$$\hat{u}_t^\pi(h_t) = \sum_{a \in A_{s_t}} q_{d_t(h_t)}(a) \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s' | s_t, a) \hat{u}_{t+1}^\pi(h_t, a, s') \right\}$$

5. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{MD}$):

$$\hat{u}_t^\pi(s_t) = r_t(s_t, d_t(s_t)) + \sum_{s' \in S} p_t(s' | s_t, d_t(s_t)) \hat{u}_{t+1}^\pi(s').$$

6. The computation complexity. There are K states and L actions, then:

- If $\pi \in \Pi^{HD}$, then requiring $K \sum_{i=0}^{N-1} (KL)^i$ multiplications.
- If $\pi \in \Pi^{MD}$, then requiring $(N-1)K^2L$ multiplications.

4.3 OPTIMALITY EQUATIONS AND THE PRINCIPLE OF OPTIMALITY

Optimality equations (Bellman equations or functional equations).

We start study this equation:

$$u_t^*(h_t) = \sup_{\pi \in \Pi^{HR}} u_t^\pi(h_t)$$

When minimizing costs instead of maximizing rewards, we sometimes refer to u_t^* as a **cost-to-go** function.

Definition 1. (Optimality equations).

$$\hat{u}_t(h_t) = \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) \hat{u}_{t+1}(h_t, a, s') \right\}, \quad s.t. \hat{u}_N(h_N) = r_N(s_N). \quad (1)$$

If A_{s_t} is finite, it can be replaced by max. Then, $\forall h_t, \hat{u}_t(h_t) = u_t^*(h_t)$.

Proof. The proof is in two parts.

Let arbitrary $\pi' = (d'_1, d'_2, \dots, d'_{N-1}) \in \Pi^{HR}$.

Step1:

First, we have $u_N^{\pi'}(h_N) = \hat{u}_N(h_N) = u_N^*(h_N)$.

Then, because we take the operation sup, we reasonably have $\hat{u}_{N-1}(h_{N-1}) \geq u_{N-1}^*(h_{N-1})$.

Assuming that $\forall h_t \in H_t$, and $t = n+1, \dots, N$, we have $\hat{u}_t(h_t) \geq u_t^*(h_t)$.

$$\begin{aligned} \hat{u}_n(h_n) &= \sup_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) \hat{u}_{n+1}(h_n, a, s') \right\} \\ &\geq \sup_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) u_{n+1}^*(s_n, a, s') \right\} \\ &\geq \sum_{a \in A_{s_n}} q_{d'_n(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) u_{n+1}^{\pi'}(s_n, a, s') \right\} \\ &\geq u_n^{\pi'}(h_n) \end{aligned}$$

Which means that, $\forall \pi \in \Pi^{HR}, \hat{u}_n(h_n) \geq u_n^\pi(h_n)$.

Step2:

$\forall \epsilon$, we can construct $\pi' \in \Pi^{HR}$ for which: $u_n^{\pi'}(h_n) + (N-n)\epsilon \geq \hat{u}_n(h_n)$.

To do this, construct a policy $\pi' = (d'_1, d'_2, \dots, d'_{N-1}) \in \Pi^{HR}$ by choosing $d_n(h_n)$ to satisfy

$$\sum_{a \in A_{s_t}} q_{d'_n(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) \hat{u}_{n+1}(s_n, a, s') \right\} + \epsilon \geq \hat{u}_n(h_n).$$

First, we have $u_N^{\pi'}(h_N) = u_N(h_N)$.

Then, we assume that $u_t^{\pi'}(h_t) + (N - t)\epsilon \geq u_t(h_t)$ for $t = n + 1, \dots, N$.

$$\begin{aligned} u_n^{\pi'}(h_n) &= \sum_a q_{d'_n(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) u_{n+1}^{\pi'}(s_n, a, s') \right\} \\ &\geq \sum_a q_{d'_n(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) \hat{u}_{n+1}(s_n, a, s') \right\} - (N - n - 1)\epsilon \\ &\geq \hat{u}_n(h_n) - (N - n)\epsilon \end{aligned}$$

Step3: $u_n^*(h_n) + (N - n)\epsilon \geq u_n^{\pi'}(h_n) + (N - n)\epsilon \geq u_n(h_n) \geq u_n^*(h_n)$.

The lefting question is

$$\int_{a \in A_{s_n}} q_{d'_n(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) u_{n+1}^{\pi'}(s_n, a, s') \right\} da$$

□

Theorem 1. Suppose $u_t^*, t = 1, \dots, N$ are solutions of the optimality equation (max version). Then we can construct a corresponding policy $\pi^* = (d_1^*, d_2^*, \dots, d_{N-1}^*) \in \Pi^{HD}$ satisfies

$$d_t^*(h_t) \in \arg \max_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) u_{t+1}^*(h_t, a, s') \right\}$$

for $t = 1, \dots, N - 1$. Then

1. $u_t^{\pi^*}(h_t) = u_t^*(h_t), \quad h_t \in H_t.$
2. $v_N^{\pi^*}(s) = v_N^*(s), \quad s \in S.$

Proof. Clearly, $u_N^{\pi^*}(h_N) = u_N^*(h_N), h_N \in H_N$.

We assume that $u_{n+1}^{\pi^*}(h_{n+1}) = u_{n+1}^*(h_{n+1})$,

$$\begin{aligned} u_n^*(h_n) &= \max_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) u_{n+1}^*(h_n, a, s') \right\} \\ &= r_n(s_n, d_n^*(h_n)) + \sum_{s' \in S} p_n(s'|s_n, d_n^*(h_n)) u_{n+1}^*(h_n, d_n^*(h_n), s') \\ &= r_n(s_n, d_n^*(h_n)) + \sum_{s' \in S} p_n(s'|s_n, d_n^*(h_n)) u_{n+1}^{\pi^*}(h_n, d_n^*(h_n), s') \\ &= u_n^{\pi^*}(h_n) \end{aligned}$$

□

Theorem 2. Let $\epsilon > 0$ be arbitrary and suppose $u_t^*, t = 1, \dots, N$ are solutions of the optimality equation (sup version, a is continuous). Then we can construct a corresponding policy $\pi^\epsilon = (d_1^\epsilon, d_2^\epsilon, \dots, d_{N-1}^\epsilon) \in \Pi^{HD}$ satisfies

$$\left\{ r_t(s_t, d_t^\epsilon) + \sum_{s' \in S} p_t(s'|s_t, d_t^\epsilon) u_{t+1}^*(h_t, d_t^\epsilon, s') \right\} + \frac{\epsilon}{N - 1}$$

$$\geq \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) u_{t+1}^*(h_t, a, s') \right\}$$

for $t = 1, \dots, N-1$. Then

$$1. \quad u_t^{\pi^\epsilon}(h_t) + (N-t) \frac{\epsilon}{N-1} \geq u_t^*(h_t), \quad h_t \in H_t.$$

$$2. \quad v_N^{\pi^\epsilon}(s) + \epsilon = v_N^*(s), \quad s \in S.$$

The proof is analogous.

4.4 OPTIMALITY OF DETERMINISTIC MARKOV POLICIES

Theorem 3. Let $u_t^*(h_t)$ is the solution of the optimality equations, then:

1. $\forall t = 1, \dots, N$, $u_t^*(h_t)$ depends on h_t only through s_t .
2. $\forall \epsilon > 0$, there exists an ϵ -optimal policy which is deterministic and Markov.
3. if a is reachable, then there exists an optimal policy which is deterministic Markov.

Proof. First, we have $\forall h_{N-1} \in H_{N-1}, a_{N-1} \in A_{S_{N-1}}, u_N^*(h_N) = u_N^*(s_N) = r_N(s_N)$. Then, we assume that $\forall n = t+1, \dots, N, u_n^*(h_n) = u_n^*(s_n)$.

$$\begin{aligned} u_t^*(h_t) &= \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) u_{t+1}^*(h_t, a, s') \right\} \\ &= \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) u_{t+1}^*(s') \right\} \\ &= u_t^*(s_t) \end{aligned}$$

□

We have established that

$$v_N^*(s) = \sup_{\pi \in \Pi^{HR}} v_N^\pi(s) = \sup_{\pi \in \Pi^{MD}} v_N^\pi(s), \quad s \in S$$

Proposition 1. Assume S is finite or countable, and that

1. A_s is finite for each $s \in S$, or
2. A_s is compact; $p_t(s'|s, a), r_t(s, a)$ is continuous in a , and $|r_t(s, a)| \leq M < \infty$
3. A_s is compact; $r_t(s, a)$ is upper semicontinuous in a ; and $|r_t(s, a)| \leq M < \infty$; $p_t(s'|s, a)$ is lower semi-continuous in a .

Then there exists a deterministic Markovian policy which is optimal. (Which means that sup is reachable.)

4.5 BACKWARD INDUCTION

The terms “backward induction” and “dynamic programming” are synonymous. Key assumption: optimal action is obtainable.

Definition 2. (*The backward induction algorithm*).

1. $\forall s \in S$, let $\hat{u}_N(s) = r_N(s)$.
2. $t = N - 1 : 1$, we calculate that

$$\forall s \in S, \hat{u}_t(s) = \max_{a \in A_s} \left\{ r_t(s, a) + \sum_{s' \in S} p_t(s' | s, a) \hat{u}_{t+1}(s') \right\}$$

4.6 OPTIMALITY OF MONOTONE POLICIES

4.6.1 Structured Policies

4.6.2 Superadditive Functions

Definition 3. Let X and Y be partially ordered sets and $g : X \times Y \rightarrow \mathbb{R}$. We say g is **superadditive** if for $x^+ \geq x^-$ and $y^+ \geq y^-$, we have

$$g(x^+, y^+) + g(x^-, y^-) \geq g(x^+, y^-) + g(x^-, y^+)$$

If the reverse inequality above holds, $g(x, y)$ is said to be **subadditive**. If superadditive function g is twice differentiable, we have $\frac{\partial^2 g(x, y)}{\partial x \partial y} \geq 0$.

Lemma 1. Let

$$f(x) = \max_y \left\{ y \in \arg \max_{y' \in Y} g(x, y') \right\}$$

If g is a superadditive function, then $f(x)$ is monotone nondecreasing in x .

Proof. Let corresponding numbers: (x^+, y^+) and (x^-, y^-) , where $y^+ = f(x^+)$ and $y^- = f(x^-)$. We assume that $x^+ > x^-$, but $y^+ \leq y^-$, then:

1. By the definition of $f(x)$, we have $g(x^-, y^-) \geq g(x^-, y^+)$.
2. By the definition of superadditive, we have $g(x^+, y^-) + g(x^-, y^+) \geq g(x^-, y^-) + g(x^+, y^+)$.
3. Then we have $g(x^+, y^-) \geq g(x^+, y^+)$, which contradicts with the definition of f .

□

4.7 Optimality of Monotone Policies

Leaving...

5 Infinite-Horizon Models: Foundations

- S is finite or countable.
- stationary policy: $d^\infty = (d, d, \dots)$

5.1 THE VALUE OF A POLICY

1. **Expected total reward** of policy $\pi \in \Pi^{HR}$:

$$v^\pi(s) = \lim_{n \rightarrow \infty} \mathbb{E}_S^\pi \left\{ \sum_{t=1}^n r(X_t, Y_t) \right\} = \lim_{n \rightarrow \infty} v_{n+1}^\pi(s) \quad (2)$$

If the limit exists and when interchanging the limits and expectation is valid, we have

$$v^\pi(s) = \mathbb{E}_S^\pi \left\{ \sum_{t=1}^{\infty} r(X_t, Y_t) \right\} \quad (3)$$

2. **Expected total discounted reward** of policy $\pi \in \Pi^{HR}$:

$$v_\lambda^\pi(s) = \lim_{n \rightarrow \infty} \mathbb{E}_S^\pi \left\{ \sum_{t=1}^n \lambda^{t-1} r(X_t, Y_t) \right\} \quad (4)$$

For $0 \leq \lambda \leq 1$, the limits exists when $\sup_{s \in S} \sup_{a \in A_s} |r(s, a)| = M < \infty$. When the limit exists and interchainging the limit and expectation are valid, we have

$$v_\lambda^\pi(s) = \mathbb{E}_S^\pi \left\{ \sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t) \right\} \quad (5)$$

3. **Average reward or gain** of policy $\pi \in \Pi^{HR}$:

$$g^\pi(s) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}_S^\pi \left\{ \sum_{t=1}^n r(X_t, Y_t) \right\} = \lim_{n \rightarrow \infty} \frac{1}{n} v_{n+1}^\pi(s) \quad (6)$$

If the limit doesn't exist, we define:

$$g_-^\pi(s) = \liminf_{n \rightarrow \infty} \frac{1}{n} v_{n+1}^\pi(s), \quad g_+^\pi(s) = \limsup_{n \rightarrow \infty} \frac{1}{n} v_{n+1}^\pi(s).$$

5.2 MARKOV POLICIES

Theorem 4. $\forall \pi = (d_1, d_2, \dots) \in \Pi^{HR}$. Then, for each $s_1 \in S_1$, $\exists \pi' = (d'_1, d'_2, \dots) \in \Pi^{MR}$, satisfying

$$\forall t, \quad P^{\pi'} \{X_t = s', Y_t = a | X_1 = s_1\} = P^\pi \{X_t = s', Y_t = a | X_1 = s_1\} \quad (7)$$

Proof. We construct the randomized Markov decision rule $d'_t \in \pi'$ by

$$q_{d'_t(s')}(a) = P^\pi \{Y_t = a | X_t = s', X_1 = s_1\}$$

Then,

$$P^{\pi'} \{Y_t = a | X_t = s'\} = P^{\pi'} \{Y_t = a | X_t = s', X_1 = s_1\} = P^\pi \{Y_t = a | X_t = s', X_1 = s_1\}$$

We use induction method. Clearly the theorem holds with $t = 1$. We assume that the theorem holds for $t = 1, 2, \dots, n-1$. Then,

$$\begin{aligned} P^\pi \{X_n = s' | X_1 = s_1\} &= \sum_{s \in S} \sum_{a \in A_s} P^\pi \{X_{n-1} = s, Y_{n-1} = a | X_1 = s_1\} p(s' | s, a) \\ &= \sum_{s \in S} \sum_{a \in A_s} P^{\pi'} \{X_{n-1} = s, Y_{n-1} = a | X_1 = s_1\} p(s' | s, a) \\ &= P^{\pi'} \{X_n = s' | X_1 = s_1\} \\ P^{\pi'} \{X_n = s', Y_n = a | X_1 = s_1\} &= P^{\pi'} \{Y_n = a | X_n = s'\} P^{\pi'} \{X_n = s' | X_1 = s_1\} \\ &= P^\pi \{Y_n = a | X_n = s', X_1 = s_1\} P^\pi \{X_n = s' | X_1 = s_1\} \\ &= P^\pi \{X_n = s', Y_n = a | X_1 = s_1\} \end{aligned}$$

□

Note that, in the above theorem, π' depends on the initial state X_1 . When the state at decision epoch 1 is chosen according to a probability distribution, then π' is depended on the distribution instead of $X_1 = s_1$.

Corollary 1. $\forall \mathcal{D}_1 \sim X_1, \pi \in \Pi^{HR}, \exists \pi' \in \Pi^{MR}$ for which

$$P^{\pi'} \{X_t = s', Y_t = a\} = P^\pi \{X_t = s', Y_t = a\}$$

Noting that

$$\begin{aligned} v_N^\pi(s) &= \sum_{t=1}^{N-1} \sum_{s' \in S} \sum_{a \in A_{s'}} r(s', a) P^\pi \{X_t = s', Y_t = a | X_1 = s_1\} \\ &\quad + \sum_{s' \in S} \sum_{a \in A_{s'}} r_N(s') P^\pi \{X_N = s', Y_N = a | X_1 = s_1\} \\ v_\lambda^\pi(s) &= \sum_{t=1}^{\infty} \sum_{s' \in S} \sum_{a \in A_{s'}} \lambda^{t-1} r(s', a) P^\pi \{X_t = s', Y_t = a | X_1 = s_1\} \end{aligned} \tag{8}$$

6 Discounted Markov Decision Problems

Assumptions in this chapter:

1. Stationary rewards and transition probabilities; $r(s, a)$ and $p(s'|s, a)$ do not vary from decision epoch to decision epoch.
2. Bounded rewards; $|r(s, a)| \leq M < \infty$.
3. Discount factor λ .
4. Discrete state spaces.

6.1 POLECY EVALUATION (Stationary Policy)

$$v_\lambda^*(s) = \sup_{\pi \in \Pi^{HR}} v_\lambda^\pi(s) = \sup_{\pi \in \Pi^{MR}} v_\lambda^\pi(s)$$

Let $\pi = (d_1, d_2, \dots) \in \Pi^{MR}$, then

$$v_\lambda^\pi(s_1) = \mathbb{E}_{s_1}^\pi \left\{ \sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t) \right\} = r_{d_1} + \lambda P_{d_1} v_\lambda^{\pi'=\{d_2, d_3, \dots\}}$$

Let $d^\infty = (d, d, \dots)$, then $v_\lambda^{d^\infty}(s_1) = r_d(s_1) + \lambda P_d v_\lambda^{d^\infty}$.
Let $\forall v \in V, L_d v = r_d + \lambda P_d v$, then $v_\lambda^{d^\infty} = L_d v_\lambda^{d^\infty}$, which means $v_\lambda^{d^\infty}$ is a fixed point of L_d in V .

Theorem 5. Suppose $0 \leq \lambda < 1$. Then $\forall d^\infty$ with $d \in D^{MR}$, $\vec{v}_\lambda^{d^\infty}$ is the unique solution in V of $\vec{v} = r_d + \lambda P_d \vec{v}$, and $\vec{v}_\lambda^{d^\infty} = (I - \lambda P_d)^{-1} r_d$.

Proof. Key theorem: $\|P_d\| = 1$ and $\sigma(\lambda P_d) \leq \|\lambda P_d\| = \lambda \leq 1$, then $(I - \lambda P_d)^{-1}$ exists.

$$\vec{v} = (I - \lambda P_d)^{-1} r_d = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} r_d = \vec{v}_\lambda^{d^\infty}$$

□

Lemma 2. 1. $\vec{u} \succeq \vec{0} \Rightarrow (I - \lambda P_d)^{-1} \vec{u} \succeq \vec{u} \succeq \vec{0}$

2. $\vec{u} \succeq \vec{v} \Rightarrow (I - \lambda P_d)^{-1} \vec{u} \succeq (I - \lambda P_d)^{-1} \vec{v}$

3. $\vec{u} \succeq \vec{0} \Rightarrow \vec{u}^T (I - \lambda P_d)^{-1} \succeq \vec{u}^T$

6.2 OPTIMALITY EQUATIONS

Optimality equations or Bellman equations (in discounted MDP):

$$v(s) = \sup_{a \in A_s} \left\{ r(s, a) + \sum_{s' \in S} \lambda p(s'|s, a) v(s') \right\}$$

Lemma 3. $\forall v \in V, 0 \leq \lambda < 1, \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\} = \sup_{d \in D^{MR}} \{r_d + \lambda P_d v\}$

Proof. First, $D^{MD} \subset D^{MR}$, so $\sup_{d \in D^{MD}} \{r_d + \lambda P_d v\} \preceq \sup_{d \in D^{MR}} \{r_d + \lambda P_d v\}$.
Second, $\forall d^{MR} \in D^{MR}$,

$$\sum_{a \in A_s} q_{d^{MR}}(a) \left[r(s, a) + \sum_{s' \in S} \lambda p(s'|s, a) v(s') \right] \leq \sup_{a \in A_s} \left\{ r(s, a) + \sum_{s' \in S} \lambda p(s'|s, a) v(s') \right\}$$

which means,

$$r_{d^{MR}} + \lambda P_{d^{MR}} v \preceq \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\} \Rightarrow \sup_{d \in D^{MR}} \{r_d + \lambda P_d v\} \preceq \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\}$$

□

Definition 4. (Bellman operator).

$$\forall v \in V, \mathcal{L}v = \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\} \quad (9)$$

If the supremum is attained for all $v \in V$, we define L by

$$\forall v \in V, Lv = \max_{d \in D^{MD}} \{r_d + \lambda P_d v\} \quad (10)$$

Theorem 6. Suppose there exists a $v \in V$ for which

1. $v \succeq \mathcal{L}v \Rightarrow v \succeq v_\lambda^*$;
2. $v \preceq \mathcal{L}v \Rightarrow v \preceq v_\lambda^*$;
3. $v = \mathcal{L}v \Rightarrow v$ is unique and $v = v_\lambda^*$.

Proof. First, we proof 1.

$\forall \pi = (d_1, d_2, \dots) \in \Pi^{MR}$,

$$\begin{aligned} v &\succeq \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\} = \sup_{d \in D^{MR}} \{r_d + \lambda P_d v\} \\ &\succeq r_{d_1} + \lambda P_{d_1} v = \sum_{t=1}^n (\lambda P^\pi)^{t-1} r_{d_t} + (\lambda P^\pi)^n v \\ v - v_\lambda^\pi &\succeq (\lambda P^\pi)^n v - \sum_{t=n+1}^{\infty} (\lambda P^\pi)^{t-1} r_{d_t} \\ &\succeq -\lambda^n \|v\|_\infty \cdot \vec{e} - \lambda^n \cdot \frac{M}{1-\lambda} \cdot \vec{e} \end{aligned}$$

Because r is bounded, so $\forall \epsilon, \exists N$, when $n \geq N$, we have

$$v \succeq v_\lambda^\pi - \epsilon \cdot \vec{e}$$

$$v \succeq \sup_{\pi \in \Pi^{MR}} v_\lambda^\pi = \sup_{\pi \in \Pi^{HR}} v_\lambda^\pi = v_\lambda^*$$

Second, we proof 2.

If $v \preceq \mathcal{L}v$, by definition of sup, we have

$$\forall \epsilon, \exists d \in D^{MD}, v \preceq r_d + \lambda P_d v + \epsilon \cdot \vec{e}$$

$$\Rightarrow v \preceq (I - \lambda P_d)^{-1} (r_d + \epsilon \cdot \vec{e}) = v_\lambda^{d^\infty} + (1 - \lambda)^{-1} \epsilon \cdot \vec{e} \preceq \sup_{\pi \in \Pi^{HR}} v_\lambda^\pi + (1 - \lambda)^{-1} \epsilon \cdot \vec{e}$$

□

The following norm is supremum norm.

Theorem 7. (Banach Fixed-Point Theorem). Suppose U is a Banach space and $T : U \rightarrow U$ is a contraction mapping with contraction parameter λ . Then

1. there exists a unique v^* in U such that $Tv^* = v^*$;
2. $\forall v^0 \in U, \lim_{n \rightarrow \infty} v^n = \lim_{n \rightarrow \infty} T^n v^0 = v^*$.

Proof.

$$\begin{aligned} \forall m \geq 1, \quad \|v^{n+m} - v^n\| &\leq \sum_{k=0}^{m-1} \|v^{n+k+1} - v^{n+k}\| = \sum_{k=0}^{m-1} \|T^{n+k}v^1 - T^{n+k}v^0\| \\ &\leq \sum_{k=0}^{m-1} \lambda^{n+k} \|v^1 - v^0\| = \frac{\lambda^n(1 - \lambda^m)}{(1 - \lambda)} \|v^1 - v^0\| \end{aligned}$$

It follows that $\{v^n\}$ is a Cauchy sequence. From the completeness of U , it follows that $\{v^n\}$ has a limit $v^\infty \in U$.

$$\begin{aligned} 0 \leq \|Tv^\infty - v^\infty\| &\leq \|Tv^\infty - v^n\| + \|v^n - v^\infty\| \\ &= \|Tv^\infty - Tv^{n-1}\| + \|v^n - v^\infty\| \leq \lambda \|v^\infty - v^{n-1}\| + \|v^n - v^\infty\| \rightarrow 0 \end{aligned}$$

which means that v^∞ is a fixed point of T . Let u^* and v^* are fixed points of T , then

$$\|u^* - v^*\| = \|Tu^* - Tv^*\| \leq \lambda \|u^* - v^*\| \Rightarrow u^* = v^*$$

□

Lemma 4. Suppose that $0 \leq \lambda < 1$; then L and \mathcal{L} are contraction mappings on V .

Proof. Let $u, v \in V$, corresponding optimal actions are a_u, a_v , fix $s \in S$, without loss of generality, let $Lu(s) \geq Lv(s)$.

$$\begin{aligned} 0 \leq Lu(s) - Lv(s) &= r(s, a_u) + \sum_{s' \in S} \lambda p(s'|s, a_u) u(s') - Lv(s) \\ &\leq \sum_{s' \in S} \lambda p(s'|s, a_u) (u(s') - v(s')) \leq \lambda \|u - v\|_\infty \end{aligned}$$

$\forall s \in S$, we have $|Lu(s) - Lv(s)| \leq \lambda \|u - v\|_\infty$

The proof of \mathcal{L} is analogue.

□

Theorem 8. Suppose $0 \leq \lambda < 1$, S is finite or countable, and $r(s, a)$ is bounded. If V is a complete normed linear space, there exists a unique $v^* \in V$ satisfying $Lv^* = v^*$, and $v^* = v_\lambda^*$.

Definition 5. For $v \in V$, call a decision rule $d_v \in D^{MD}$ v -improving if

$$d_v \in \arg \max_{d \in D^{MD}} \{r_d + \lambda P_d v\} \Leftrightarrow L_{d_v} v = Lv$$

Clarify:

1. $v_\lambda^{d_v^\infty}$ needs not be greater than or equal to v .
2. Even if $r_{d_v} + \lambda P_{d_v} v \succeq v$, $v_\lambda^{d_v^\infty}$ exceeds v in some component only if $r_{d_v}(s') + \lambda P_{d_v} v(s') > v(s')$.
3. d^* , v_λ^* -improving, is called conserving decision rule.

Theorem 9. If supremum is attained, then $\exists d \in D^{MD}, d^\infty \in \Pi^{MD}$, satisfies $v_\lambda^{d^\infty} = v_\lambda^*$. So we can calculate that $v_\lambda^* = \sup_{d \in D^{MD}} v_\lambda^{d^\infty}$.

Proof.

$$v_\lambda^* = L v_\lambda^* = L_{d_{v_\lambda^*}} v_\lambda^* \Rightarrow v_\lambda^* = v_\lambda^{d_{v_\lambda^*}^\infty}$$

□

Theorem 10. Assume S is discrete, and either

1. A_s is finite for each $s \in S$, or
2. A_s is compact, $r(s, a)$ is continuous in a for each $s \in S$, and for each $s' \in S$ and $s \in S$, $p(s'|s, a)$ is continuous in a , or
3. A_s is compact, $r(s, a)$ is upper semicontinuous in a for each $s \in S$, and for each $s' \in S$ and $s \in S$, $p(s'|s, a)$ is lower semicontinuous in a .

Then there exists an optimal deterministic stationary policy.

If the supremum is not attained in $\mathcal{L}v$, then optimal policies need not exist.

Theorem 11. Support S is finite or countable, then for all $\epsilon > 0$ there exists an ϵ -optimal deterministic stationary policy.

Proof. Take d_ϵ satisfying

$$r_{d_\epsilon} + \lambda P_{d_\epsilon} v_\lambda^* \succeq \sup_{d \in D^{MD}} \{r_d + \lambda P_d v_\lambda^*\} - (1 - \lambda)\epsilon \vec{1} = v_\lambda^* - (1 - \lambda)\epsilon \vec{1}$$

$$v_\lambda^{d_\epsilon^\infty} = (I - \lambda P_{d_\epsilon})^{-1} r_{d_\epsilon} \succeq v_\lambda^* - (1 - \lambda)\epsilon (I - \lambda P_{d_\epsilon})^{-1} \vec{1} = v_\lambda^* - \epsilon \vec{1}$$

□

6.3 VALUE ITERATION AND ITS VARIANTS

6.3.1 Rates of Convergence

Rate of Convergence

1. linear convergence or quadratic convergence: $\|y_{n+1} - y^*\| \leq K \|y_n - y^*\|^\alpha$;
2. superlinearly convergence: $\limsup_{n \rightarrow \infty} \frac{\|y_{n+1} - y^*\|}{\|y_n - y^*\|} = 0$;
3. asymptotic average rate of convergence $\limsup_{n \rightarrow \infty} \left[\frac{\|y_n - y^*\|}{\|y_0 - y^*\|} \right]^{1/n}$

6.3.2 Value Iteration

Algorithm 1 Value Iteration Algorithm

Require: $\epsilon > 0$

Ensure: $v^0 \in V$

for $n = 1, 2, \dots$ **do**

$\forall s \in S, v^{n+1}(s) = \max_{a \in A_s} \{r(s, a) + \sum_{s' \in S} \lambda p(s'|s, a) v^n(s')\}$

if $\|v^{n+1} - v^n\| < \epsilon(1 - \lambda)/(2\lambda)$ **then**

break.

end if.

end for.

return $d_\epsilon(s) \in \arg \max_{a \in A_s} \{r(s, a) + \sum_{s' \in S} \lambda p(s'|s, a) v^{n+1}(s')\}$
