

Markov Decision Processes: Discrete Stochastic Dynamic Programming

Peng Lingwei

August 13, 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 | 14 |
| 6.4 | POLICY ITERATION | 15 |
| 6.5 | MODIFIED POLICY ITERATION | 17 |
| 6.5.1 | Convergence Rates | 19 |
| 6.6 | SPANS, BOUNDS, STOPPING CRITERIA, AND RELATIVE VALUE ITEARTION | 19 |
| 6.6.1 | The Span Seminorm | 19 |
| 6.6.2 | Bounds on the Value of a Discounted Markov Decision Process | 21 |
| 6.6.3 | Stopping Criteria | 21 |

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 variable 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

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

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

□

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 $\limsup_{n \rightarrow \infty} \left[\frac{\|y_n - y^*\|}{\|y_0 - y^*\|} \right]^{1/n}$

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

6.3.2 Value Iteration**Theorem 12.** $(d_\epsilon)^\infty$ is ϵ -optimal.*Proof.*

$$\|v^{n+1} - v^n\| = \|Lv^n - Lv^{n-1}\| \leq \lambda^{n-1} \|v^1 - v^0\|$$

so

$$\exists N, \forall n > N \geq 1 + \log \left(\frac{\epsilon(1 - \lambda)}{\lambda^2 \|v^1 - v^0\|} \right), \|v^{n+1} - v^n\| < \epsilon(1 - \lambda)/(2\lambda).$$

$$\begin{aligned} \|v^{d_\epsilon^\infty} - v^{n+1}\| &= \|L_{d_\epsilon} v^{d_\epsilon^\infty} - v^{n+1}\| \\ &\leq \|L_{d_\epsilon} v^{d_\epsilon^\infty} - L_{d_\epsilon} v^{n+1}\| + \|Lv^{n+1} - Lv^n\| \\ &\leq \lambda \|v^{d_\epsilon^\infty} - v^{n+1}\| + \lambda \|v^{n+1} - v^n\| \\ \|v^{d_\epsilon^\infty} - v^{n+1}\| &\leq \frac{\lambda}{1 - \lambda} \|v^{n+1} - v^n\|. \end{aligned}$$

$$\text{Analogously, } \|v^{n+1} - v^*\| \leq \frac{\lambda}{1 - \lambda} \|v^{n+1} - v^n\|.$$

$$\|v^{d_\epsilon^\infty} - v^*\| \leq \|v^{d_\epsilon^\infty} - v^{n+1}\| + \|v^{n+1} - v^*\| \leq \epsilon$$

□

Theorem 13. (monotone). If $u \succeq v$, then $Lu \succeq Lv$.*Proof.*

$$\begin{aligned} Lu - Lv &= \max_{d \in D^{MD}} (r_d + \lambda P_d u) - \max_{d \in D^{MD}} (r_d + \lambda P_d v) \\ &= \max_{d \in D^{MD}} (r_d + \lambda P_d u) - (r_{d_v} + \lambda P_{d_v} v) \\ &\succeq (r_{d_v} + \lambda P_{d_v} u) - (r_{d_v} + \lambda P_{d_v} v) \\ &= \lambda P_{d_v} (u - v) \succeq \vec{0} \end{aligned}$$

□

Therefore, if $Lv^0 \succeq (\preceq)v^0$, then value iteration converges monotonically to v^* .

Theorem 14. (Convergence of value iteration).

1. $\|v^{n+1} - v_\lambda^*\| = \|Lv^n - Lv_\lambda^*\| \leq \lambda\|v^n - v_\lambda^*\|$
2. $\frac{\|v^n - v_\lambda^*\|}{\|v^0 - v_\lambda^*\|} \leq \lambda^n \Rightarrow \limsup_{n \rightarrow \infty} \left[\frac{\|v^n - v_\lambda^*\|}{\|v^0 - v_\lambda^*\|} \right]^{1/n} \leq \lambda$
3. $\|v^n - v_\lambda^*\| \leq \frac{\lambda^n}{1-\lambda} \|\lambda^1 - \lambda^0\|$

If we want change inequality into equality, we need $v^0 \succeq (\preceq)v^*$ and $v^1 - v^* = \lambda(v^0 - v^*)$

6.4 POLICY ITERATION

Algorithm 2 Policy Iteration Algorithm

```

Select an arbitrary rule  $d_0 \in D^{MD}$ .
for  $n = 1, 2, \dots$  do
  Policy evaluation:  $v^n = (I - \lambda P_{d_n})^{-1} r_{d_n}$ 
  Policy improvement:  $d_{n+1} \in \arg \max_{d \in D^{MD}} \{r_d + \lambda P_d v^n\}$ 
  if  $d_{n+1} = d_n$  then
    break.
  end if.
end for.
return  $d_{n+1}$ 

```

Proposition 2. In policy iteration algorithm $v^{n+1} \geq v^n$.

Proof.

$$\begin{aligned}
 r_{d_{n+1}} + \lambda P_{d_{n+1}} v^n &\geq r_{d_n} + \lambda P_{d_n} v^n = v^n \\
 v^{n+1} &= (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} \geq v^n
 \end{aligned}$$

□

If states and actions are finite, the algorithm can terminate in finite number of iterations.

Definition 6. Operator $B : V \rightarrow V$,

$$Bv = \max_{d \in D^{MD}} \{r_d + (\lambda P_d - I)v\} = Lv - v.$$

Proposition 3. $\forall u, v \in V$ and $d_v \in D_v$.

$$Bu \geq Bv + (\lambda P_{d_v} - I)(u - v) \Rightarrow (\lambda P_{d_v} - I) \in \partial_v(Bv)$$

Proof.

$$\begin{aligned}
 Bu - Bv &= \max_{d \in D^{MD}} \{r_d + (\lambda P_d - I)u\} - \max_{d \in D^{MD}} \{r_d + (\lambda P_d - I)v\} \\
 &\succeq \{r_{d_v} + (\lambda P_{d_v} - I)u\} - \{r_{d_v} + (\lambda P_{d_v} - I)v\} \\
 &\succeq (\lambda P_{d_v} - I)(u - v)
 \end{aligned}$$

□

Proposition 4. Suppose the sequence $\{v^n\}$ is obtained from the policy iteration algorithm. Then, for any $d_{v^n} \in D_{v^n}$.

$$v^{n+1} = v^n - (\lambda P_{d_{v^n}} - I)^{-1} B v^n$$

Proof.

$$\begin{aligned} v^{n+1} &= (I - \lambda P_{d_{v^n}})^{-1} r_{d_{v^n}} - v^n + v^n \\ &= v^n - (\lambda P_{d_{v^n}} - I)^{-1} [r_{d_{v^n}} + (\lambda P_{d_{v^n}} - I)v^n] \\ &= v^n - (\lambda P_{d_{v^n}} - I)^{-1} B v^n \end{aligned}$$

□

Definition 7. $V_B = \{v \in V; Bv \geq 0\}$ ($v \in V_B \Rightarrow v \preceq v^*$).

Definition 8. $Zv = v - (\lambda P_{d_v} - I)^{-1} Bv$.

Lemma 5. Let $v \in V_B, d_v \in D_v, v \succeq u$. Then $Zv \succeq Lu, Zv \in V_B, Zv \succeq v$.

Proof.

$$\begin{aligned} Zv &= v - (\lambda P_{d_v} - I)^{-1} Bv \succeq v + Bv = Lv \succeq Lu \\ B(Zv) &\succeq Bv + (\lambda P_{d_v} - I)(Zv - v) = \vec{0} \\ Zv &= v + (I - \lambda P_{d_v})^{-1} Bv \succeq v \end{aligned}$$

□

Theorem 15. (Policy iteration converges monotonically).

Proof. Let $u^k = L^k v^0$ and $v^k = Z^k v_0$. We inductively show that $v^k \in V_B$ and $u^k \leq v^k \leq v_\lambda^*$.

First, if $k = 0$, then $u^0 = v^0$ and

$$Bv^0 \succeq r_{d_0} + (\lambda P_{d_0} - I)v^0 = \vec{0},$$

therefore, $v^0 \in V_B$ and $v^0 \preceq v_\lambda^*$. Above all, $k = 0, u^0 \preceq v^0 \preceq v_\lambda^*$.

Then, we assume $k \leq n, u^k \preceq v^k \preceq v_\lambda^*$ and $Bv^k \succeq \vec{0}$.

$$\begin{aligned} v^{n+1} &= Zv^n \in V_B \Rightarrow v^{n+1} \preceq v_\lambda^*. \\ v^k &\succeq u^k, v^{n+1} = Zv^n \succeq Lu^n = u^{n+1} \end{aligned}$$

□

Theorem 16. (Convergence Rate). If policy iteration's sequence $\{v^n\}$ satisfies $\|P_{d_{v^n}} - P_{d_{v_\lambda^*}}\| \leq K\|v^n - v_\lambda^*\|$ (for some K), then

$$\|v^{n+1} - v_\lambda^*\| \leq \frac{K\lambda}{1-\lambda} \|v^n - v_\lambda^*\|^2$$

Proof. Let $U_n = \lambda P_{d_{v^n}} - I$ and $U_* = \lambda P_{d_{v_\lambda^*}} - I$. then

$$\begin{aligned} Bv^n &\succeq Bv_\lambda^* + U_*(v^n - v_\lambda^*) = U_*(v^n - v_\lambda^*) \Rightarrow U_n^{-1} Bv^n \preceq U_n^{-1} U_*(v^n - v_\lambda^*) \\ 0 &\preceq v_\lambda^* - v^{n+1} = v_\lambda^* - v^n + U_n^{-1} Bv^n \preceq U_n^{-1} (U_n - U_*)(v_\lambda^* - v^n) \\ \|v_\lambda^* - v^{n+1}\| &\preceq \|U_n^{-1}\| \|U_n - U_*\| \|v_\lambda^* - v^n\| \preceq \frac{\lambda}{1-\lambda} \|P_{d_{v^n}} - P_{d_{v_\lambda^*}}\| \|v_\lambda^* - v^n\| \end{aligned}$$

□

Consider that $\|P_{d_v^n} - P_{d_{v_\lambda^*}}\| \leq K\|v^n - v_\lambda^*\|$ is unsatisfying, for the unknown v_λ^* , we can change into a general condition:

$$\forall u, v \in V, \|P_{d_v} - P_{d_u}\| \leq K\|v - u\|$$

$$\forall u, v \in V, \|P_{d_v} - P_{d_{v_\lambda^*}}\| \leq K\|v - v_\lambda^*\|$$

6.5 MODIFIED POLICY ITERATION

Algorithm 3 Modified Policy Iteration Algorithm (MPI)

Require: $\epsilon > 0, \{m_0, m_2, \dots\}$.

Ensure: $v^0 \in V_B$.

```

for  $n = 0, 1, \dots$  do
     $d_{n+1} \in \arg \max_{d \in D} \{r_d + \lambda P_d v^n\}$ 
     $u_n^0 = r_{d_{n+1}} + \lambda P_{d_{n+1}} v^n$ 
    if  $\|u_n^0 - v^n\| < \epsilon(1 - \lambda)/(2\lambda)$  then break
    end if.
    for  $k = 0, 1, \dots, m_n$  do
         $u_n^{k+1} = r_{d_{n+1}} + \lambda P_{d_{n+1}} u_n^k = L_{d_{n+1}} u_n^k$ 
    end for.
     $(v^{n+1} = L_{d_{n+1}}^{m_n+1} v^n)$ 
end for.
return  $d_{n+1}$ 

```

In policy iteration, we have

$$v^{n+1} = v^n - (\lambda P_{d_v^n} - I)^{-1} B v^n = v^n + \sum_{k=0}^{\infty} (\lambda P_{d_{n+1}}^k B v^n)$$

Proposition 5. *Modified policy iteration algorithm equals:*

$$v^{n+1} = v^n - (\lambda P_{d_v^n} - I)^{-1} B v^n = v^n + \sum_{k=0}^{m_n} (\lambda P_{d_{n+1}}^k) B v^n$$

Proof.

$$\begin{aligned}
 v^{n+1} &= v^n + \sum_{k=0}^{m_n} (\lambda P_{d_{n+1}})^k [r_{d_{n+1}} + \lambda P_{d_{n+1}} v^n - v^n] \\
 &= r_{d_{n+1}} + \lambda P_{d_{n+1}} r_{d_{n+1}} + \dots + (\lambda P_{d_{n+1}})^{m_n} r_{d_{n+1}} + (\lambda P_{d_{n+1}})^{m_n+1} v^n \\
 &= (L_{d_{n+1}})^{m_n+1} v^n
 \end{aligned}$$

□

The preceeding proposition shows that order 0 modified policy iteration equals to value iteration, and order ∞ modified policy iteration equals to policy iteration.

The graph of algorithm: Bv lines and 45-degree lines.

Denote the operator $U^m : V \rightarrow V$,

$$U^m v = \max_{d \in D} \sum_{k=0}^m (\lambda P_d)^k r_d + (\lambda P_d)^{m+1} v.$$

Proposition:

1. $\|U^m u - U^m v\| \leq \lambda^{m+1} \|u - v\|$;
2. The sequence $w^{n+1} = U^m w^n$ converges in norm to v_λ^* ;

Proof. Assume w^* is the fixed point of U^m , and let $d^* \in D^{MD}$ be the v_λ^* -improving decision rule.

$$\begin{aligned} v_\lambda^* &= L^m v_\lambda^* = \sum_{k=0}^m (\lambda P_{d^*})^k r_{d^*} + (\lambda P_{d^*})^{m+1} v_\lambda^* \preceq U^m v_\lambda^* \preceq (U^m)^n v_\lambda^* \rightarrow w^*, \\ w^* &= U^m w^* \preceq L^m w^* \rightarrow v_\lambda^* \end{aligned}$$

□

3. v_λ^* is the unique fixed point of U^m ;
4. $\|w^{n+1} - v_\lambda^*\| \preceq \lambda^{m+1} \|w^n - v_\lambda^*\|$

Denote the MPI operator $W^m : V \rightarrow V$,

$$W^m v = v + \sum_{k=0}^m (\lambda P_{d_v})^k Bv$$

Lemma 6. For $u \in V$ and $v \in V$ satisfying $u \succeq v \Rightarrow U^m u \succeq W^m v$. Furthermore, if $u \in V_B$, then $W^m u \succeq U^0 v = Lv$.

Proof. Let $d_v \in D$ is v -improving and $d_u \in D$ is u -improving. Then

$$\begin{aligned} U^m u - W^m v &\succeq \sum_{k=0}^m (\lambda P_{d_v})^k r_{d_v} + (\lambda P_{d_v})^{m+1} u - \sum_{k=0}^m (\lambda P_{d_v})^k r_{d_v} - (\lambda P_{d_v})^{m+1} v \\ &= (\lambda P_{d_v})^{m+1} (u - v) \succeq 0. \end{aligned}$$

For $u \in V_B$,

$$W^m u = u + \sum_{k=0}^m (\lambda P_{d_u})^k Bu \succeq u + Bu = Lu \succeq r_{d_v} + \lambda P_{d_v} u \succeq Lv$$

□

Lemma 7. $u \in V_B \Rightarrow w = W^m u \in V_B$.

Proof.

$$\begin{aligned} Bw &\succeq Bu + (\lambda P_{d_u} - I)(w - u) = Bu + (\lambda P_{d_u} - I) \sum_{k=0}^m (\lambda P_{d_u})^k Bu \\ &= (\lambda P_{d_u})^{m+1} Bu \succeq \vec{0} \end{aligned}$$

□

Theorem 17. (The monotonical convergence of MPI).

Proof. Define three sequence $\{v^n\}, \{y^n\}, \{w^n\}$ which corresponds to W^{m_n}, L , and U^{m_n} , and $v^0 = y^0 = w^0 \in V_B$. We will show that $v^n \in V_B, v^{n+1} \succeq v^n$, and $w^n \succeq v^n \succeq y^n$.

According preceeding lemma, $v^0 \in V_B \Rightarrow v^n \in V_B$.

We can get monotonous by $v^{n+1} = v^n + \sum_{m=0}^{m_n} (\lambda P_{d_n})^m Bv^n \succeq v^n$.

By conduction, we assum $w^n \succeq v^n \succeq y^n$ the preceeding lemma also proofs that $U^{m_n} w^n \succeq W^{m_n} v^n \succeq Ly^n$. \square

Noting: $W^{m_n+k} v^n$ can be small than $W^{m_n} v^n$

6.5.1 Convergence Rates

Theorem 18. Suppose $v^0 \in V_B$ and $\{v^n\}$ is generated by modified policy iteration, d_n is a v^n -improving decision rule, and d^* is a v_λ^* -improving decision rule.

$$\|v^{n+1} - v_\lambda^*\| \leq \left(\frac{\lambda(1 - \lambda^{m_n})}{1 - \lambda} \|P_{d_n} - P_{d^*}\| + \lambda^{m_n+1} \right) \|v^n - v_\lambda^*\|. \quad (11)$$

Proof.

$$\begin{aligned} 0 \leq v_\lambda^* - v^{n+1} &= v_\lambda^* - v^n - \sum_{k=0}^{m_n} (\lambda P_{d_n})^k Bv^n \\ &\leq v_\lambda^* - v^n + \sum_{k=0}^{m_n} (\lambda P_{d_n})^k (I - \lambda P_{d^*})(v^n - v_\lambda^*) \\ &= \lambda(P_{d_n} - P_{d^*}) \sum_{k=0}^{m_n-1} (\lambda P_{d_n})^k (v^n - v_\lambda^*) - \lambda^{m_n+1} P_{d_n}^{m_n} P_{d^*} (v^n - v_\lambda^*) \end{aligned}$$

Taking norms yields the result. \square

If $\lim_{n \rightarrow \infty} \|P_{d_n} - P_{d^*}\| = 0$, then $\|v^{n+1} - v_\lambda^*\| \leq (\lambda^{m_n+1} + \epsilon) \|v^n - v_\lambda^*\|$.

If $m_n \rightarrow \infty$, $\limsup_{n \rightarrow \infty} \frac{\|v^{n+1} - v_\lambda^*\|}{\|v^n - v_\lambda^*\|} = 0$.

6.6 SPANS, BOUNDS, STOPPING CRITERIA, AND RELATIVE VALUE ITEARTION

6.6.1 The Span Seminorm

1. $\Lambda(v) = \min_{s \in S} v(s), \Upsilon(v) = \max_{s \in S} v(s)$;
2. $sp(v) = \max_{s \in S} v(s) - \min_{s \in S} v(s) = \Upsilon(v) - \Lambda(v)$
 - $\forall v \in V, sp(v) \geq 0$;
 - $\forall v, u \in V, sp(u + v) \leq sp(u) + sp(v)$;
 - $\forall k \in \mathbb{R}, sp(kv) = |k|sp(v)$;
 - $\forall k \in \mathbb{R}, sp(v + ke) = sp(v)$;
 - $sp(v) = sp(-v)$;

$$\bullet \text{ } sp(v) \leq 2\|v\|_\infty \leq 2\|v\|_2 \leq 2\|v\|_1$$

Proposition 6. Let $v \in V, d \in D$. Then $sp(P_d v) \leq \gamma_d sp(v)$,
 $\gamma_d = \max_{s, s' \in S \times S} \sum_{j \in S} \max \{0, P_d(j|s) - P_d(j|s')\}$.

Proof. Let $b(s, s'; j) = \min \{P(j|s), P(j|s')\}$

$$\begin{aligned} sp(Pv) &= \max_{s, s' \in S \times S} \sum_{j \in S} [P(j|s) - b(s, s'; j)]v(j) - \sum_{j \in S} [P(j|s') - b(s, s'; j)]v(j) \\ &\leq \max_{s, s' \in S \times S} \sum_{j \in S} [P(j|s) - b(s, s'; j)]\Upsilon(v) - \sum_{j \in S} [P(j|s') - b(s, s'; j)]\Lambda(v) \\ &= \max_{s, s' \in S \times S} \left[1 - \sum_{j \in S} b(s, s'; j) \right] sp(v) = \max_{s, s' \in S \times S} \left[1 - \sum_{j \in S} \min \{P(j|s), P(j|s')\} \right] sp(v) \\ &= \max_{s, s' \in S \times S} \left[1 - \sum_{j \in S} (P(j|s) + P(j|s') - |P(j|s) - P(j|s')|)/2 \right] sp(v) \\ &= \max_{s, s' \in S \times S} \left[\frac{1}{2} \sum_{j \in S} |P(j|s) - P(j|s')| \right] sp(v) \\ &= \max_{s, s' \in S \times S} \sum_{j \in S} \max \{0, P(j|s) - P(j|s')\} sp(v) \end{aligned}$$

$$(|x - y| = x + y - 2 \min(x, y), \max(0, x - y) = x - \min(x, y), \max(0, y - x) = y - \min(x, y)) \quad \square$$

$\exists v' \in V$ such that $sp(Pv) = sp(v)$:

1. P's rows are equal $\Rightarrow \gamma_d = 0 \Rightarrow sp(Pv) = 0 = 0 \cdot sp(v)$;
2. Let s^*, s'^* be $\sum_{j \in S} \max \{0, P(j|s^*) - P(j|s'^*)\} = \max_{s, s' \in S \times S} \sum_{j \in S} \max \{0, P(j|s) - P(j|s')\}$,
then $v(j) = 1_{\{P(j|s^*) > P(j|s'^*)\}}$. $sp(v') = 1$ and $sp(Pv) \geq \sum_{j \in S} P(j|s^*)v(j) - \sum_{j \in S} P(j|s'^*)v(j) = \sum_{j \in S} \max \{0, P(j|s^*) - P(j|s'^*)\} = \gamma_d sp(v)$

γ_d is referred to as the Hajnal measure or delta coefficient of P_d , which upper bounds the subradius (modulus of the second largest eigenvalue) of P_d , $\sigma_s(P_d)$. γ_d equals to 0 if all rows of P_d are equal, and equals to 1 if at least two rows of P_d are orthogonal.

Theorem 19. Let **span contraction** $T : V \rightarrow T$ and suppose there exists an $\alpha, 0 \leq \alpha < 1$ for which

$$sp(Tv - Tu) \leq \alpha \cdot sp(v - u)$$

then

1. $\exists v^* \in V, sp(Tv^* - v^*) = 0$ which called **span fixed point**. Furthermore, $Tv^* = v^* = v^* + ke$.
2. For sequence $\{v^n\}$ by $v^n = T^n v^0$, then $\lim_{n \rightarrow \infty} sp(v^n - v^*) = 0$.
3. $sp(v^{n+1} - v^*) \leq \alpha^n sp(v^0 - v^*)$

6.6.2 Bounds on the Value of a Discounted Markov Decision Process

Theorem 20. For $v \in V, m \geq -1$, and any v -improving decision rule d_v ,

$$G_m(v) = v + \sum_{i=1}^m (\lambda P_{d_v})^i Bv + \lambda^{m+1} (1 - \lambda)^{-1} \Lambda(Bv) \vec{1}, \quad \text{nondecreasing in } m$$

$$G^m(v) = v + \sum_{k=0}^m (\lambda P_{d_{v_\lambda^*}})^k Bv + \lambda^{m+1} (1 - \lambda)^{-1} \Upsilon(Bv) \vec{1}, \quad \text{nonincreasing in } m$$

$$G_m(v) \leq v_\lambda^{(d_v)^\infty} \leq v_\lambda^* \leq G^m(v)$$

Proof. We have $0 = Bv_\lambda^* \succeq Bv + (\lambda P_{d_v} - I)(v_\lambda^* - v)$. Since that $(I - \lambda P_{d_v})^{-1} \succeq 0$, then, $0 \succeq v - v_\lambda^* + (I - \lambda P_{d_v})^{-1} Bv$.

$$\begin{aligned} v_\lambda^* &\succeq v + \sum_{k=0}^m (\lambda P_{d_v})^k Bv + \sum_{k=m+1}^{\infty} (\lambda P_{d_v})^k [\Lambda(Bv)] \vec{1} \\ &= v + \sum_{k=0}^m (\lambda P_{d_v})^k Bv + \frac{\lambda^{m+1}}{1 - \lambda} [\Lambda(Bv)] \vec{1} \end{aligned}$$

Analogously, $Bv \succeq Bv_\lambda^* + (\lambda P_{d_{v_\lambda^*}} - I)(v - v_\lambda^*) \Rightarrow v_\lambda^* \preceq v + (I - \lambda P_{d_{v_\lambda^*}})^{-1} Bv \preceq v + \sum_{k=0}^m (\lambda P_{d_{v_\lambda^*}})^k Bv + \frac{\lambda^{m+1}}{1 - \lambda} [\Upsilon(Bv)] \vec{1}$. \square

Corollary 2.

$$\begin{aligned} v + (1 - \lambda)^{-1} \Lambda(Bv) \vec{1} &\preceq v + Bv + \lambda(1 - \lambda)^{-1} \Lambda(Bv) \vec{1} \preceq v_\lambda^{d_v^\infty} \\ &\preceq v_\lambda^* \preceq v + Bv + \frac{\lambda}{1 - \lambda} \Upsilon(Bv) \vec{1} \\ &\preceq v + (1 - \lambda)^{-1} \Upsilon(Bv) \vec{1} \end{aligned}$$

6.6.3 Stopping Criteria

Proposition 7. For $v \in V$ and $\epsilon > 0$ that

$$sp(Lv - v) = sp(Bv) < \frac{(1 - \lambda)}{\lambda} \epsilon$$

then,

$$\|Lv + \frac{\lambda}{1 - \lambda} \Lambda(Bv) \vec{e} - v_\lambda^*\| < \epsilon$$

and

$$\|v_\lambda^{d_v^\infty} - v_\lambda^*\| < \epsilon$$

Proof. ($w \leq x \leq y \leq z \Rightarrow 0 \leq y - x \leq z - w$).

$$0 \preceq v_\lambda^* - v - Bv - \frac{\lambda}{1 - \lambda} \Lambda(Bv) \vec{1} \preceq \frac{\lambda}{1 - \lambda} sp(Bv) \vec{1}$$

Because $Lv = Bv + v$, therefore we can get the first inequation by taking norms on both side. Analogously,

$$0 \preceq v_\lambda^* - v_\lambda^{d_v^\infty} \preceq \frac{\lambda}{1 - \lambda} sp(Bv) \vec{1}$$

\square

Here is something we need to know

$$\forall k, \arg \max_{d \in D} \{r_d + \lambda P_d(v + k\vec{1})\} = \arg \max_{d \in D} \{r_d + \lambda P_d v + \lambda k\vec{1}\} = \arg \max_{d \in D} \{r_d + \lambda P_d v\}$$

Theorem 21. $\gamma = \max_{s \in S, a \in A_s, s' \in S, a' \in A_{s'}} \left[1 - \sum_{j \in S} \min[p(j|s, a), p(j|s', a')] \right]$.
Then $\forall u, v \in V, sp(Lv - Lu) \leq \lambda \gamma sp(v - u)$.

Proof.

$$\begin{aligned} sp(Lv - Lu) &\leq \max_{s \in S} (Lv(s) - Lu(s)) - \min_{s \in S} (Lv(s) - Lu(s)) \\ &\leq \max_{s \in S} (L_{d_v} v(s) - L_{d_v} u(s)) - \min_{s \in S} (L_{d_u} v(s) - L_{d_u} u(s)) \\ &= \max_{s \in S} (P_{d_v}(v - u)(s)) - \min_{s \in S} (\lambda P_{d_u}(v - u)(s)) \\ &\leq sp \left(\lambda \begin{bmatrix} P_{d_v} \\ P_{d_u} \end{bmatrix} (v - u) \right) \leq \lambda \gamma_{d_v, d_u} (v - u) \leq \lambda \gamma (v - u) \end{aligned}$$

□

If $u = Lv$ then $\forall v \in V, sp(B^2 v) \leq \lambda \gamma sp(Bv)$. For value iteration,

$$\|v^{n+2} - v^{n+1}\| = \|Bv^{n+1}\| = \|B^2 v^n\| \leq \lambda \|Bv^n\| = \lambda \|v^{n+1} - v^n\|$$

$$sp(v^{n+2} - v^{n+1}) = sp(B^2 v^n) \leq \lambda \gamma sp(Bv^n) = \lambda \gamma sp(v^{n+1} - v^n)$$

We can use γ' instead of γ : $\gamma \leq 1 - \sum_{j \in S} \min_{s \in S, a \in A_s} p(j|s, a) = \gamma'$.

Corollary 3. Let $v^0 \in V$, $\{v^n\}$ has been generated using value iteration. Then

1. $\lim_{n \rightarrow \infty} sp(v^n - V_\lambda^*) = 0$;
2. $\forall n, sp(v^{n+1} - v_\lambda^*) \leq (\lambda \gamma)^n sp(v^0 - v_\lambda^*)$;
3. $sp(v^{n+1} - v^n) \leq (\lambda \gamma)^n sp(v^1 - v^0)$.

In chapter8, the following algorithm is useful.

Algorithm 4 Relative Value Iteration Algorithm

Require: $\epsilon > 0$

Ensure: $u^0 \in V$, choose s_0 set $w^0 = u^0 - u^0(s_0)\vec{1}$

for $n = 0, 1, \dots$ **do**

$$u^{n+1} = Lw^n$$

$$w^{n+1} = u^{n+1} - u^{n+1}(s_0)\vec{1}$$

if $sp(u^{n+1} - u^n) < (1 - \lambda)\epsilon/\lambda$ **then** break

end if.

end for.

return $d_\epsilon \in \arg \max_{d \in D} \{r_d + \lambda P_d u^n\}$
