Exam Proofs

Markov Decision Theory and Algorithmic Methods (191531920)

16 December 2022
8:45-10:45h

You may use the Accompanying Material.

Instructions:

- This exam contains one proposition and two theorems. You have to construct a proof for one of the three and may ignore the other two.
- Construct the proof for the proposition or theorem of your choice according to Polya's four steps for problem solving (see the Accompanying Material).
- Your work will be graded according to the rubric in the Accompanying Material.
- Note: the grading will be based on your proof for *only one* proposition or theorem. So if, during this exam, you work on multiple proofs, hand in only the proof that you want to be considered for grading.

Prove Proposition 4.7.3:

Suppose that the maximum is attained in

$$\max_{a \in A'} \left\{ r_t(s,a) + \sum_{j=0}^{\infty} p_t(j|s,a) u(j) \right\}$$

and that

1. $r_t(s,a)$ is nondecreasing in s for all $a \in A'$ and $t = 1, \ldots, N-1$,
2. $q_t(k|s,a)$ is nondecreasing in s for al $k \in S$, $a \in A'$, and $t = 1, \ldots, N-1$, and
3. $r_N(s)$ is nondecreasing in s.

Then $u_t^*(s)$ is nondecreasing in s for $t = 1, \ldots, N$.

For this proof, you may not use the Accompanying Material beyond Proposition 4.7.3.

Prove Theorem 6.3.3 parts d and e:

Let $v^0 \in V$ and let $\{v^n\}$ denote the iterates of value iteration. Then the following global convergence rate properties hold for the value iteration algorithm:

(d) for all n,

$$||v^n - v_\lambda^*|| \leq \frac{\lambda^n}{1-\lambda} ||v^1 - v^0||,$$

(e) for any $d_n \in \arg\max_{d \in D} \{r_d + \lambda P_d v^n\}$,

$$||v_\lambda^{(d_n)^\infty} - v_\lambda^*|| \leq \frac{2\lambda^n}{1-\lambda} ||v^1 - v^0||.$$

For this proof, you may not use the Accompanying Material beyond Theorem 6.3.3.

Prove Theorem 6.9.4 parts a, d and e:

Suppose $|r(s,a)| \leq M < \infty$ for all $a \in A_s$ and $s \in S$. Then:

(a) There exists a bounded optimal basic feasible solution x^* to the dual LP.

(d) Suppose $\{d^*\}^\infty$ is an optimal policy for the discounted Markov decision problem. Then x_{d^*} is an optimal solution for the dual linear program.

(e) Suppose $\{d^*\}^\infty$ is a deterministic optimal policy for the discounted Markov decision problem. Then x_{d^*} is an optimal basic solution for the dual linear program.

For this proof, you may not use the Accompanying Material beyond Theorem 6.9.4.

Accompanying Material

for

Exam Proofs

December 16, 8:45-10:45h

Part of the course

Markov Decision Theory and Algorithmic Methods

(191531920)

This material consists of:

- Polya's four steps for problem solving
  according to which the proof should be constructed

- Rubric MDT&AM – Proofs
  according to which the proof will be graded

- Theorems, lemma's, propositions and corollary's
  that are part of the course's study material, from the book 'Markov Decision Processes – Discrete Stochastic Dynamic Programming' by Martin L. Puterman

Students are allowed to use this material while working on their Exam Proofs.

Polya's four steps for problem solving

Step 1. Getting Started. State the important information and summarize the problem. If possible, include a diagram. Note any assumptions you're making. You may complete this step by answering the following questions:

- What is given? Can you single out each condition?
- Can you explain what each condition means?
- Can you anticipate which role each condition will play in the final result?
- What needs to be proved?
- What is the intuitive interpretation of this result/problem?
- How can this result/problem be formulated mathematically?

Step 2. Devise Plan. Devise a plan of attack before diving into the solution. Break down the problem into smaller, manageable segments. Identify which mathematical relationship you can apply. You may complete this step by answering the following questions:

- What results and mathematical relations are relevant?
- Are the assumptions of these mathematical relations met?
- How do I need to apply these results and relations in this particular case?
- Which executable steps should I make?

You should be able to finish the plan with the phrase: **`And this will solve the problem'.**

Step 3. Execute Plan. Carry out your plan, explaining each step. The argument should be easy to follow. Articulate your thought process at each step (including roadblocks). Any variables should be clearly defined, and your diagrams should be labeled.

Step 4. Evaluate Proof. Check your proof for reasonableness. For example:

- Can you interpret the solution and/or the result intuitively?
- Look back at your solution. Did you use all conditions?
- Which role did each condition play in the solution?
- In your solution, could you relax some assumptions of the problem? What will change in your solution?
- If you got stuck and could not solve the problem, can you specify why? For example, which condition couldn't you use? Or, which step couldn't you make?
- Is the problem related to other problems or results?
- Does the solution/result make sense in extreme or special cases?
- If relevant, what are the implications of all above in practice?

Rubric MDT&AM - Proofs. Total points: 10

		0 points	1 point	2 points
Step 1 (2 pt)	*Problem statement*	Problem statement copied from the exercise, not explained in own words.	Problem statement is partially written down in own words.	Problem statement is written down in own words, it is clear that the student understands the problem.
	Intuitive interpretation	No intuitive explanation/interpretation of the conditions and the result.	Interpretation of the conditions and the result is not specific.	Clear interpretation of the condition and the result, for instance, by providing an example.
Step 2 (2 pt)	*Plan (writing)*	There is no concrete plan of attack.	There is a rough plan of attack, but no details are given.	The plan is written down clearly, it is clear how to execute this plan.
	Plan (correctness)	Execution of the plan will not solve the problem.	Execution of the plan will partially solve the problem.	Execution of the plan will solve the problem.
	Relevant theory	The relevant theory and results are not written down.	Some of the relevant theory and results are written down.	All relevant theory and results are written down.
	Assumptions (if applicable)	Assumptions are not stated	Some of the assumptions are stated.	All assumptions are stated.
Step 3 (2 x 2pt)	*Execution (writing)*	The solution is written down poorly: variables are not defined, steps are not justified.	The solution is written completely, but some variables are not defined and/or some steps/explanation are missing.	The solution is written down completely, all variables are defined, all steps are explained.
	Execution (content)	The problem is not solved.	The problem is partially solved.	The problem is solved correctly.
Step 4 (2 pt)	*Checking the solution and the answer*	The student does not check whether the result is correct.	The student checks whether the result is correct.	The student checks whether the result is correct and provides an explanation why the result makes sense (or why the result seems to be incorrect).
	Interpretation	No interpretation of the result is given.	There is only a shallow interpretation of the result that does not go beyond this specific problem.	There is a comprehensive interpretation of the result, for example: analysis what happens when conditions change; discussion of practical implications; connection to other topics in the course. If the problem is not solved, there is an analysis of what went wrong.

$$P^* = C - \lim_{N\to\infty} P^N. \tag{A.3}$$

$$PP^* = P^*P = P^*P^* = P^*. \tag{A.4}$$

Proposition A.3. When S is finite

a. the spectral radius of W, $\sigma(W) < 1$, and

b. $(I - W)^{-1}$ exists and satisfies

$$(I - W)^{-1} = \sum_{n=0}^{\infty} W^n.$$

$$(I - P)^{\#} = (I - P + P^*)^{-1}(I - P^*) \equiv H_P. \tag{A.14}$$

$$(I - P)H_P = H_P(I - P) = I - P^*, \tag{A.17}$$

$$H_P P^* = P^* H_P = 0, \tag{A.18}$$

$$H_P = Z_P - P^*, \tag{A.19}$$

$$Z_P P^* = P^*, \tag{A.20}$$

Proposition B.1. Let X be a complete separable metric space.

a. If f and g are u.s.c. on X, then $f + g$ is u.s.c. on X.

Theorem B.2. Suppose C is a compact subset of X, and f is u.s.c. on X. Then there exists an x^* in C such that $f(x^*) \geq f(x)$ for all $x \in C$.

Proposition B.3. Let X be a countable set, Y a complete separable metric space, and $q(x, y)$ a bounded non-negative real-valued function that is l.s.c. in y for each $x \in X$. Let $f(x)$ be a bounded nonpositive real-valued function on X for which $\sum_{x\in X} f(x)$ is finite. Then

$$h(y) = \sum_{x\in X} f(x)q(x, y)$$

is u.s.c. on Y.

Corollary C.4. Let Q be a bounded linear transformation on a Banach space V, and suppose that $\sigma(Q) < 1$. Then $(I - Q)^{-1}$ exists and satisfies

$$(I - Q)^{-1} = \lim_{N\to\infty} \sum_{n=0}^{N} Q^n. \tag{C.10}$$

Proposition C.5. Let $\{A_n\}$ be a sequence of matrices converging in norm to an invertible matrix A. Then for n sufficiently large A_n^{-1} exists and

$$\lim_{n\to\infty} \|A_n^{-1} - A^{-1}\| = 0.$$

Lemma C.6. Let $\{A_n\}$ be a sequence of matrices which converges in norm to A with $\|A\| < \infty$, and $\{x_n\}$ a sequence of vectors which converges in norm to x with $\|x\| < \infty$. Then $\{A_n x_n\}$ converges in norm to Ax.

Lemma C.7. Let $\{A_n\}$ and $\{B_n\}$ denote sequences of square matrices of the same dimension. Suppose that $\{A_n\}$ converges in norm to A with $\|A\| < \infty$, and that $\{B_n\}$ converges in norm to B with $\|B\| < \infty$. Then $\{A_n B_n\}$ converges in norm to AB.

Theorem D.1.

a. If either the primal or dual problem has a finite optimal solution, then it has an optimal solution which is a basic feasible solution.

b. (Weak duality) If x is feasible for the primal, and y is feasible for the dual, then $c^T x \geq b^T y$.

c. (Strong duality) If the primal problem has a bounded optimal solution x^*, then the dual has a bounded optimal solution y^*, and

$$c^T x^* = b^T y^*.$$

d. (Complementary slackness) Necessary and sufficient conditions for (x^*, u^*) to be an optimal solution for the augmented primal, and (y^*, v^*) to be an optimal solution for the augmented dual, are that they are non-negative and satisfy

$$(y^*)^T u^* = (y^*)^T(Ax^* - b) = 0 \tag{D.5}$$

and

$$(v^*)^T x^* = (c^T - (y^*)^T A)x^* = 0 \tag{D.6}$$

Proposition 4.1.1. Suppose $U = (U_1, \ldots, U_n)$ and $V = (V_1, \ldots, V_n)$ are random vectors such that $P[V_1 > t] \leq P[U_1 > t]$ for all $t \in R^1$, and for $j = 2, \ldots, n$, $v_i \leq u_i$ for $i = 1, \ldots, j - 1$ implies

$$P\left[V_j > t | V_1 = v_1, \ldots, V_{j-1} = v_{j-1}\right] \leq P\left[U_j > t | U_1 = u_1, \ldots, U_{j-1} = u_{j-1}\right].$$

Then U is stochastically greater than V. Further, for any nondecreasing function g: $R^1 \to R^1$,

$$E\left[g(V_j) | V_1 = v_1, \ldots, V_{j-1} = v_{j-1}\right] \leq E\left[g(U_j) | U_1 = u_1, \ldots, U_{j-1} = u_{j-1}\right].$$

$$u_t^\pi(h_t) = E_{h_t}^\pi\left\{\sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N)\right\}, \tag{4.2.1}$$

The Finite Horizon-Policy Evaluation Algorithm (for fixed $\pi \in \Pi^{HD}$)

1. Set $t = N$ and $u_N^\pi(h_N) = r_N(s_N)$ for all $h_N = (h_{N-1}, a_{N-1}, s_N) \in H_N$.
2. If $t = 1$, stop, otherwise go to step 3.
3. Substitute $t - 1$ for t and compute $u_t^\pi(h_t)$ for each $h_t = (h_{t-1}, a_{t-1}, s_t) \in H_t$ by

$$u_t^\pi(h_t) = r_t(s_t, d_t(h_t)) + \sum_{j \in S} p_t(j|s_t, d_t(h_t)) u_{t+1}^\pi(h_t, d_t(h_t), j), \tag{4.2.2}$$

noting that $(h_t, d_t(h_t), j) \in H_{t+1}$.

4. Return to 2.

$$u_t^\pi(h_t) = r_t(s_t, d_t(h_t)) + E_{h_t}^\pi\{u_{t+1}^\pi(h_t, d_t(h_t), X_{t+1})\}. \tag{4.2.3}$$

Theorem 4.2.1. Let $\pi \in \Pi^{HD}$ and suppose u_t^π, $t \le N$, has been generated by the policy evaluation algorithm. Then, for all $t \le N$, (4.2.1) holds, and $v_N^\pi(s) = u_1^\pi(s)$ for all $s \in S$.

$$u_t^\pi(h_t) = \sum_{a \in A_{s_t}} q_{d_t(h_t)}(a)\left\{r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j)\right\}. \tag{4.2.6}$$

Theorem 4.2.2. Let $\pi \in \Pi^{HR}$ and suppose u_t^π, $t \le N$ has been generated by the policy evaluation algorithm with (4.2.6) replacing (4.2.2). Then, for all $t \le N$, (4.2.1) holds and $v_N^\pi(s) = u_1^\pi(s)$ for all $s \in S$.

$$u_t^*(h_t) = \sup_{\pi \in \Pi^{HR}} u_t^\pi(h_t). \tag{4.3.1}$$

Lemma 4.3.1. Let w be a real-valued function on an arbitrary discrete set W and let $q(\cdot)$ be a probability distribution on W. Then

$$\sup_{u \in W} w(u) \ge \sum_{u \in W} q(u) w(u)$$

$$u_t(h_t) = \sup_{a \in A_{s_t}}\left\{r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j)\right\} \tag{4.3.2}$$

$$u_N(h_N) = r_N(s_N) \tag{4.3.3}$$

$$u_t(h_t) = \max_{a \in A_{s_t}}\left\{r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j)\right\}. \tag{4.3.4}$$

Theorem 4.3.2. Suppose u_t is a solution of (4.3.2) for $t = 1, \ldots, N-1$, and u_N satisfies (4.3.3). Then

a. $u_t(h_t) = u_t^*(h_t)$ for all $h_t \in H_t$, $t = 1, \ldots, N$, and

b. $u_1(s_1) = v_N^*(s_1)$ for all $s_1 \in S$.

Theorem 4.3.3. Suppose u_t^*, $t = 1, \ldots, N$ are solutions of the optimality equations (4.3.4) subject to boundary condition (4.3.3), and that policy $\pi^* = (d_1^*, d_2^*, \ldots, d_{N-1}^*) \in \Pi^{HD}$ satisfies

$$r_t(s_t, d_t^*(h_t)) + \sum_{j \in S} p_t(j|s_t, d_t^*(h_t)) u_{t+1}^*(h_t, d_t^*(h_t), j)$$

$$= \max_{a \in A_{s_t}}\left\{r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(h_t, a, j)\right\} \tag{4.3.10}$$

for $t = 1, \ldots, N-1$.

Then

a. For each $t = 1, 2, \ldots, N$,

$$u_t^{\pi^*}(h_t) = u_t^*(h_t), \quad h_t \in H_t. \tag{4.3.11}$$

b. π^* is an optimal policy, and

$$v_N^{\pi^*}(s) = v_N^*(s), \quad s \in S. \tag{4.3.12}$$

Theorem 4.3.4. Let $\varepsilon > 0$ be arbitrary and suppose u_t^*, $t = 1, \ldots, N$ are solutions of the optimality equations (4.3.2) and (4.3.3). Let $\pi^\varepsilon = (d_1^\varepsilon, d_2^\varepsilon, \ldots, d_{N-1}^\varepsilon) \in \Pi^{HD}$ satisfy

$$r_t(s_t, d_t^\varepsilon(h_t)) + \sum_{j \in S} p_t(j|s_t, d_t^\varepsilon(h_t)) u_{t+1}^*(h_t, d_t^\varepsilon(h_t), j) + \frac{\varepsilon}{N-1}$$

$$\ge \sup_{a \in A_{s_t}}\left\{r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(h_t, a, j)\right\} \tag{4.3.13}$$

for $t = 1, 2, \ldots, N-1$. Then,

a. For each $t = 1, 2, \ldots, N-1$,

$$u_t^{\pi^\varepsilon}(h_t) + (N-t)\frac{\varepsilon}{N-1} \geq u_t^*(h_t), \quad h_t \in H_t. \tag{4.3.14}$$

b. π^ε is an ε-optimal policy, that is

$$v_N^{\pi^\varepsilon}(s) + \varepsilon \geq v_N^*(s), \quad s \in S. \tag{4.3.15}$$

Theorem 4.4.1.

a. For any $\varepsilon > 0$, there exists an ε-optimal policy which is deterministic history dependent. Any policy in Π^{HD} which satisfies (4.3.13) is ε-optimal.

b. Let u_t^* be a solution of (4.3.2) and (4.3.3) and suppose that for each t and $s_t \in S$, there exists an $a' \in A_{s_t}$ for which

$$r_t(s_t, a') + \sum_{j \in S} p_t(j|s_t, a') u_{t+1}^*(h_t, a', j)$$
$$= \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(h_t, a, j) \right\} \tag{4.4.1}$$

for all $h_t = (s_{t-1}, a_{t-1}, s_t) \in H_t$. Then there exists a deterministic history-dependent policy which is optimal.

Theorem 4.4.2. Let u_t^*, $t = 1, \ldots, N$ be solutions of (4.3.2) and (4.3.3). Then

a. For each $t = 1, \ldots, N$, $u_t^*(h_t)$ depends on h_t only through s_t.

b. For any $\varepsilon > 0$, there exists an ε-optimal policy which is deterministic and Markov.

c. If there exists an $a' \in A_{s_t}$ such that (4.4.1) holds for each $s_t \in S$ and $t = 1, 2, \ldots, N-1$, there exists an optimal policy which is deterministic and Markov.

Proposition 4.4.3. Assume S is finite or countable, and that

a. A_s is finite for each $s \in S$, or

b. A_s is compact, $r_t(s, a)$ is continuous in a for each $s \in S$, there exists an $M < \infty$ for which $|r_t(s, a)| \leq M$ for all $a \in A_s$, $s \in S$, and $p_t(j|s, a)$ is continuous in a for each $j \in S$ and $s \in S$ and $t = 1, 2, \ldots, N$, or

c. A_s is a compact, $r_t(s, a)$ is upper semicontinuous (u.s.c.) in a for each $s \in S$, there exists an $M < \infty$ for which $|r_t(s, a)| \leq M$ for all $a \in A_s$, $s \in S$, and for each $j \in S$ and $s \in S$, $p_t(j|s, a)$ is lower semi-continuous (l.s.c.) in a and $t = 1, 2, \ldots, N$.

Then there exists a deterministic Markovian policy which is optimal.

$$u_t^*(s_t) = \max_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(j) \right\}. \tag{4.5.1}$$

$$A_{s_t,t}^* = \arg\max_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(j) \right\}. \tag{4.5.2}$$

Theorem 4.5.1. Suppose u_t^*, $t = 1, \ldots, N$ and $A_{s_t,t}^*$ $t = 1, \ldots, N-1$ satisfy (4.5.1) and (4.5.2); then,

a. for $t = 1, \ldots, N$ and $h_t = (h_{t-1}, a_{t-1}, s_t)$

$$u_t^*(s_t) = \sup_{\pi \in \Pi^{\mathrm{HR}}} u_t^\pi(h_t), \quad s_t \in S.$$

b. Let $d_t^*(s_t) \in A_{s_t,t}^*$ for all $s_t \in S$, $t = 1, \ldots, N-1$, and let $\pi^* = (d_1^*, \ldots, d_{N-1}^*)$. Then $\pi^* \in \Pi^{\mathrm{MD}}$ is optimal and satisfies

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

and

$$u_t^{\pi^*}(s_t) = u_t^*(s_t), \quad s_t \in S$$

for $t = 1, \ldots, N$.

$$g(x^+, y^+) + g(x^-, y^-) \geq g(x^+, y^-) + g(x^-, y^+). \tag{4.7.1}$$

Lemma 4.7.1. Suppose g is a superadditive function on $X \times Y$ and for each $x \in X$, $\max_{y \in Y} g(x, y)$ exists. Then

$$f(x) = \max\left\{ y' \in \arg\max_{y \in Y} g(x, y) \right\} \tag{4.7.3}$$

is monotone nondecreasing in x.

$$q_t(k|s, a) = \sum_{j=k}^{\infty} p_t(j|s, a)$$

$$w_t(s, a) = r_t(s, a) + \sum_{j=0}^{\infty} p_t(j|s, a) u_t^*(j) \tag{4.7.4}$$

Lemma 4.7.2. Let $\{x_j\}, \{x'_j\}$ be real-valued non-negative sequences satisfying

$$\sum_{j=k}^{\infty} x_j \geq \sum_{j=k}^{\infty} x'_j \tag{4.7.5}$$

for all k, with equality holding in (4.7.5) for $k = 0$.

Suppose $v_{j+1} \geq v_j$ for $j = 0, 1, \ldots$, then

$$\sum_{j=0}^{\infty} v_j x_j \geq \sum_{j=0}^{\infty} v_j x'_j, \tag{4.7.6}$$

where limits in (4.7.6) exist but may be infinite.

$$\max_{a \in A'} \left\{ r_t(s,a) + \sum_{j=0}^{\infty} p_t(j|s,a) u(j) \right\} \tag{4.7.8}$$

Proposition 4.7.3. Suppose the maximum in (4.7.8) is attained and that

1. $r_t(s,a)$ is nondecreasing (nonincreasing) in s for all $a \in A'$ and $t = 1, \ldots, N-1$,
2. $q_t(k|s,a)$ is nondecreasing in s for all $k \in S$, $a \in A'$, and $t = 1, \ldots, N-1$, and
3. $r_N(s)$ is nondecreasing (nonincreasing) in s.

Then $u_t^*(s)$ is nondecreasing (nonincreasing) in s for $t = 1, \ldots, N$.

Theorem 4.7.4. Suppose for $t = 1, \ldots, N-1$ that

1. $r_t(s,a)$ is nondecreasing in s for all $a \in A'$,
2. $q_t(k|s,a)$ is nondecreasing in s for all $k \in S$ and $a \in A'$,
3. $r_t(s,a)$ is a superadditive (subadditive) function on $S \times A'$,
4. $q_t(k|s,a)$ is a superadditive (subadditive) function on $S \times A'$ for all $k \in S$, and
5. $r_N(s)$ is nondecreasing in s.

Then there exist optimal decision rules $d_t^*(s)$ which are nondecreasing (nonincreasing) in s for $t = 1, \ldots, N-1$.

Theorem 4.7.5. Suppose for $t = 1, \ldots, N-1$ that

1. $r_t(s,a)$ is nonincreasing in s for all $a \in A'$,
2. $q_t(k|s,a)$ is nondecreasing in s for all $k \in S$ and $a \in A'$,
3. $r_t(s,a)$ is a superadditive function on $S \times A'$,
4. $\sum_{j=0}^{\infty} p_t(j|s,a)u(j)$ is a superadditive function on $S \times A'$ for nonincreasing u, and
5. $r_N(s)$ is nonincreasing in s.

Then there exist optimal decision rules $d_t^*(s)$ which are nondecreasing in s for $t = 1, \ldots, N$.

Lemma 4.7.6. Let $g(s,a)$ be a real-valued function on $S \times A$, with $A = \{0,1\}$ and $S = \{0, 1, \ldots\}$. If $g(s,a)$ satisfies

$$[g(s+1,1) - g(s+1,0)] - [g(s,1) - g(s,0)] \geq 0 \tag{4.7.9}$$

for all s, it is superadditive.

$$\sup_{s \in S} \sup_{a \in A_s} |r(s,a)| = M < \infty, \tag{5.1.4}$$

$$v_\nu^\pi(s) = E_s^\pi \left[E_\nu \left\{ \sum_{t=1}^{\nu} r(X_t, Y_t) \right\} \right], \tag{5.3.1}$$

Proposition 5.3.1. Suppose that (5.1.4) holds and ν has a geometric distribution with parameter λ. Then $v_\nu^\pi(s) = v_\lambda^\pi(s)$ for all $s \in S$.

Theorem 5.5.1. Let $\pi = (d_1, d_2, \ldots) \in \Pi^{HR}$. Then, for each $s \in S$, there exists a policy $\pi' = (d'_1, d'_2, \ldots) \in \Pi^{MR}$, satisfying

$$P^{\pi'}\{X_t = j, Y_t = a | X_1 = s\} = P^{\pi}\{X_t = j, Y_t = a | X_1 = s\} \tag{5.5.1}$$

for $t = 1, 2, \ldots$.

Corollary 5.5.2. For each distribution of X_1, and any history-dependent policy π, there exists a randomized Markov policy π' for which

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

Theorem 5.5.3. Suppose $\pi \in \Pi^{HR}$, then for each $s \in S$ there exists a $\pi' \in \Pi^{MR}$ (which possibly varies with s) for which

a. $v_N^{\pi'}(s) = v_N^{\pi}(s)$ for $1 \leq N < \infty$;

b. $v_\lambda^{\pi'}(s) = v_\lambda^{\pi}(s)$ for $0 \leq \lambda < 1$;

c. $g_+^{\pi'}(s) = g_+^{\pi}(s)$, $g_-^{\pi'}(s) = g_-^{\pi}(s)$, and $g^{\pi}(s) = g^{\pi'}(s)$, whenever $g_+^{\pi'}(s) = g_-^{\pi'}(s)$ and $g_+^{\pi}(s) = g_-^{\pi}(s)$, and

d. if $r_N(s) = 0$ and $v^{\pi}(s) \equiv \lim_{N \to \infty} v_N^{\pi}(s)$ exists, $v^{\pi'}(s) = v^{\pi}(s)$.

Lemma 5.6.1. Suppose S is discrete, $|r(s,a)| \leq M$ for all $a \in A_s$ and $s \in S$, and $0 \leq \lambda \leq 1$. Then, for all $v \in V$ and $d \in D^{MR}$, $r_d + \lambda P_d v \in V$.

$$v_\lambda^\pi = \sum_{t=1}^{\infty} \lambda^{t-1} P_\pi^{t-1} r_{d_t}, \tag{5.6.5}$$

Assumption 6.0.2. *Bounded rewards*; $|r(s,a)| \le M < \infty$ for all $a \in A_s$ and $s \in S$

Theorem 6.1.1. Suppose $0 \le \lambda < 1$. Then for any stationary policy d^∞ with $d \in D^{MR}$, $v_\lambda^{d^\infty}$ is the unique solution in V of

$$v = r_d + \lambda P_d v. \tag{6.1.9}$$

Further, $v_\lambda^{d^\infty}$ may be written as

$$v_\lambda^{d^\infty} = (I - \lambda P_d)^{-1} r_d. \tag{6.1.10}$$

Lemma 6.1.2. Suppose $0 \le \lambda < 1$ and $u \in V$ and $v \in V$. Then, for any $d \in D^{MR}$,

a. if $u \ge 0$, then $(I - \lambda P_d)^{-1} u \ge 0$ and $(I - \lambda P_d)^{-1} u \ge u$;
b. if $u \ge v$, then $(I - \lambda P_d)^{-1} u \ge (I - \lambda P_d)^{-1} v$; and
c. if $u \ge 0$, then $u^T (I - \lambda P_d)^{-1} \ge 0$ and $u^T (I - \lambda P_d)^{-1} \ge u^T$.

$$\mathscr{L} v \equiv \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\}, \tag{6.2.3}$$

Proposition 6.2.1. For all $v \in V$ and $0 \le \lambda \le 1$,

$$\sup_{d \in D^{MD}} \{r_d + \lambda P_d v\} = \sup_{d \in D^{MR}} \{r_d + \lambda P_d v\}. \tag{6.2.5}$$

Theorem 6.2.2. Suppose there exists a $v \in V$ for which

a. $v \ge \mathscr{L} v$, then $v \ge v_\lambda^*$;
b. $v \le \mathscr{L} v$, then $v \le v_\lambda^*$;
c. $v = \mathscr{L} v$, then v is the only element of V with this property and $v = v_\lambda^*$.

Theorem 6.2.3. (Banach Fixed-Point Theorem) Suppose U is a Banach space and $T: U \to U$ is a contraction mapping. Then

a. there exists a unique v^* in U such that $Tv^* = v^*$; and
b. for arbitrary v^0 in U, the sequence $\{v^n\}$ defined by

$$v^{n+1} = Tv^n = T^{n+1} v^0 \tag{6.2.10}$$

converges to v^*.

Proposition 6.2.4. Suppose that $0 \le \lambda < 1$; then L and $\mathscr{L}$ are contraction mappings on V.

Theorem 6.2.5. Suppose $0 \le \lambda < 1$, S is finite or countable, and $r(s,a)$ is bounded.

a. Then there exists a $v^* \in V$ satisfying $Lv^* = v^*$ ($\mathscr{L} v^* = v^*$). Further, v^* is the only element of V with this property and equals v_λ^*.
b. For each $d \in D^{MR}$, there exists a unique $v \in V$ satisfying $L_d v = v$. Further, v is the unique solution and equals $v_\lambda^{d^\infty}$.

Theorem 6.2.6. A policy $\pi^* \in \Pi^{HR}$ is optimal if and only if $v_\lambda^{\pi^*}$ is a solution of the optimality equation.

Theorem 6.2.7. Let S be discrete, and suppose that the supremum is attained in (6.2.3) for all $v \in V$. Then

a. there exists a conserving decision rule $d^* \in D^{MD}$;
b. if d^* is conserving, the deterministic stationary policy $(d^*)^\infty$ is optimal; and
c. $v_\lambda^* = \sup_{d \in D^{MD}} v_\lambda^{d^\infty}$.

Corollary 6.2.8. Suppose for each $v \in V$ and $s \in S$, there exists an $a_s^v \in A_s$, such that

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

Then there exists a deterministic stationary optimal policy $(d^*)^\infty$. Further, if $d^*(s) = a_s^*$ where

$$a_s^* \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v_\lambda^*(j) \right\},$$

then $(d^*)^\infty$ is optimal.

Theorem 6.2.9. Suppose there exists

a. a conserving decision rule, or
b. an optimal policy.

Then there exists a deterministic stationary policy which is optimal.

Theorem 6.2.10. Assume S is discrete, and either

a. A_s is finite for each $s \in S$, or
b. A_s is compact, $r(s,a)$ is continuous in a for each $s \in S$, and, for each $j \in S$ and $s \in S$, $p(j|s,a)$ is continuous in a, or
c. A_s is compact, $r(s,a)$ is upper semicontinuous (u.s.c.) in a for each $s \in S$, and for each $j \in S$ and $s \in S$, $p(j|s,a)$ is lower semicontinuous (l.s.c.) in (a).

Then there exists an optimal deterministic stationary policy.

Theorem 6.2.11. Suppose S is finite or countable, then for all $\varepsilon > 0$ there exists an ε-optimal deterministic stationary policy.

Value Iteration Algorithm

1. Select $v^0 \in V$, specify $\varepsilon > 0$, and set $n = 0$.
2. For each $s \in S$, compute $v^{n+1}(s)$ by

$$v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v^n(j) \right\}. \tag{6.3.2}$$

3. If

$$\|v^{n+1} - v^n\| < \varepsilon(1-\lambda)/2\lambda, \tag{6.3.3}$$

go to step 4. Otherwise increment n by 1 and return to step 2.

4. For each $s \in S$, choose

$$d_\varepsilon(s) \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v^{n+1}(j) \right\} \tag{6.3.4}$$

and stop.

$$v^{n+1} = Lv^n, \tag{6.3.5}$$

Theorem 6.3.1. Let $v^0 \in V$, $\varepsilon > 0$, and let $\{v^n\}$ satisfy (6.3.5) for $n \geq 1$. Then

a. v^n converges in norm to v_λ^*,
b. finite N for which (6.3.3) holds for all $n \geq N$,
c. the stationary policy $(d_\varepsilon)^\infty$ defined in (6.3.4) is ε-optimal, and
d. $\|v^{n+1} - v_\lambda^*\| < \varepsilon/2$ whenever (6.3.3) holds.

Proposition 6.3.2.

a. Let $u \in V$ and $v \in V$ with $v \geq u$. Then $Lv \geq Lu$.
b. Suppose for some N that $Lv^N \leq (\geq) v^N$, then $v^{N+m+1} \leq (\geq) v^{N+m}$ for all $m \geq 0$.

Theorem 6.3.3. Let $v^0 \in V$ and let $\{v^n\}$ denote the iterates of value iteration. Then the following global convergence rate properties hold for the value iteration algorithm:

a. convergence is linear at rate λ,
b. its asymptotic average rate of convergence equals λ,
c. it converges $O(\lambda^n)$,
d. for all n,

$$\|v^n - v_\lambda^*\| \leq \frac{\lambda^n}{1-\lambda} \|v^1 - v^0\|, \tag{6.3.7}$$

e. for any $d_n \in \arg\max_{d \in D} \{r_d + \lambda P_d v^n\}$,

$$\|v_\lambda^{(d_n)^\infty} - v_\lambda^*\| \leq \frac{2\lambda^n}{1-\lambda} \|v^1 - v^0\|. \tag{6.3.8}$$

The Policy Iteration Algorithm

1. Set $n = 0$, and select an arbitrary decision rule $d_0 \in D$.
2. (Policy evaluation) Obtain v^n by solving

$$(I - \lambda P_{d_n}) v = r_{d_n}. \tag{6.4.1}$$

3. (Policy improvement) Choose d_{n+1} to satisfy

$$d_{n+1} \in \arg\max_{d \in D} \{r_d + \lambda P_d v^n\}, \tag{6.4.2}$$

setting $d_{n+1} = d_n$ if possible.

4. If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise increment n by 1 and return to step 2.

Proposition 6.4.1. Let v^n and v^{n+1} be successive values generated by the policy iteration algorithm. Then $v^{n+1} \geq v^n$.

Theorem 6.4.2. Suppose S is finite and, for each $s \in S$, A_s is finite. Then the policy iteration algorithm terminates in a finite number of iterations, with a solution of the optimality equation and a discount optimal policy $(d^*)^\infty$.

Primal Linear Program

$$\text{Minimize} \sum_{j \in S} \alpha(j) v(j)$$

subject to

$$v(s) - \sum_{j \in S} \lambda p(j|s,a) v(j) \geq r(s,a)$$

for $a \in A_s$ and $s \in S$, and $v(s)$ unconstrained for all $s \in S$.

Dual Linear Program

$$\text{Maximize} \sum_{s \in S} \sum_{a \in A_s} r(s,a) x(s,a)$$

subject to

$$\sum_{a \in A_j} x(j,a) - \sum_{s \in S} \sum_{a \in A_s} \lambda p(j|s,a) x(s,a) = \alpha(j) \tag{6.9.2}$$

and $x(s,a) \geq 0$ for $a \in A_s$ and $s \in S$.

Theorem 6.9.1

a. For each $d \in D^{MR}$, $s \in S$ and $a \in A_s$ define $x_d(s,a)$ by

$$x_d(s,a) = \sum_{j \in S} \alpha(j) \sum_{n=1}^{\infty} \lambda^{n-1} P^{d^\infty}(X_n = s, Y_n = a | X_1 = j). \tag{6.9.3}$$

Then $x_d(s,a)$ is a feasible solution to the dual problem.

b. Suppose $x(s,a)$ is a feasible solution to the dual problem, then, for each $s \in S$, $\sum_{a \in A_s} x(s,a) > 0$. Define the randomized stationary policy d_x^∞ by

$$P\{d_x(s) = a\} = \frac{x(s,a)}{\sum_{a' \in A_s} x(s,a')}. \tag{6.9.4}$$

Then $x_{d_x}(s,a)$ as defined in (6.9.3) is a feasible solution to the dual LP and $x_{d_x}(s,a) = x(s,a)$ for all $a \in A_s$ and $s \in S$.

$$\sum_{s \in S} \alpha(s) v_\lambda^{d_x^\infty}(s) = \sum_{s \in S} \sum_{a \in A_s} x(s,a) r(s,a). \tag{6.9.8}$$

$$\sum_{s \in S} \alpha(s) v_\lambda^{d^\infty}(s) = \sum_{s \in S} \sum_{a \in A_s} x_d(s,a) r(s,a) \tag{6.9.9}$$

Let X denote the set of $x(s,a)$ satisfying (6.9.2) and $x(s,a) \geq 0$ for $a \in A_s$ and $s \in S$; $\mathscr{X}$ denote the mapping from x into D^{MR} defined by (6.9.4) and $\mathscr{D}$ denote the mapping from D^{MR} into X defined by (6.9.3).

Corollary 6.9.2. $\mathscr{X}$ and $\mathscr{D}$ are 1-1 and onto so that

$$\mathscr{X} = \mathscr{D}^{-1} \quad \text{and} \quad \mathscr{D} = \mathscr{X}^{-1}.$$

Proposition 6.9.3.

a. Let x be a basic feasible solution to the dual LP. Then $d_x \in D^{MD}$.

b. Suppose that $d \in D^{MD}$, then x_d is a basic feasible solution to the dual LP.

Theorem 6.9.4. Suppose Assumption 6.0.2 holds. Then:

a. There exists a bounded optimal basic feasible solution x^* to the dual LP.

b. Suppose x^* is an optimal solution to the dual linear program, then $(d_{x^*})^\infty$ is an optimal policy.

c. Suppose x^* is an optimal basic solution to the dual linear program, then $(d_{x^*})^\infty$ is an deterministic optimal policy.

d. Suppose $(d^*)^\infty$ is an optimal policy for the discounted Markov decision problem. Then x_{d^*} is an optimal solution for the dual linear program.

e. Suppose $(d^*)^\infty$ is a deterministic optimal policy for the discounted Markov decision problem. Then x_{d^*} is an optimal basic solution for the dual linear program.

Proposition 6.9.5. For any positive vector α, the dual linear program has the same optimal basis. Hence, $(d_{x^*})^\infty$ does not depend on α.

Assumption 8.0.2. *Bounded rewards*; $|r(s,a)| \le M < \infty$ for all $a \in A_s$ and $s \in S$

$$g^\pi(s) = \lim_{N\to\infty} \frac{1}{N} v_{N+1}^\pi(s) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P_\pi^{n-1} r_{d_n}(s). \tag{8.1.3}$$

Proposition 8.1.1.

a. Let S be countable. Let $d^\infty \in \Pi^{SR}$ and suppose that the limiting matrix of P_d, P_d^* is stochastic. Then the limit in (8.1.3) exists and

$$g^{d^\infty}(s) = \lim_{N\to\infty} \frac{1}{N} v_{N+1}^{d^\infty}(s) = P_d^* r_d(s). \tag{8.1.4}$$

b. If S is finite, (8.1.4) holds.

Theorem 8.1.2. For each $\pi \in \Pi^{HR}$ and $s \in S$, there exists a $\pi' \in \Pi^{MR}$ for which

a. $g_+^{\pi'}(s) = g_+^\pi(s)$,
b. $g_-^{\pi'}(s) = g_-^\pi(s)$, and
c. $g^{\pi'}(s) = g^\pi(s)$ whenever $g_-^\pi(s) = g_+^\pi(s)$.

Proposition 8.2.1. Suppose P^* is stochastic. Then if j and k are in the same closed irreducible class, $g(j) = g(k)$. Further, if the chain is irreducible, or has a single recurrent class and possibly some transient states, $g(s)$ is a constant function.

Proposition 8.2.2. Suppose $g(s) = 0$ for all $s \in S$.

a. Then

$$h(s) = C\text{-}\lim_{N\to\infty} E_s\left\{\sum_{t=1}^{N} r(X_t)\right\} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} v_k(s)$$

and

b.

$$h(s) = \lim_{N\to\infty} E_s\left\{\sum_{t=1}^{N} r(X_t)\right\} = \lim_{N\to\infty} v_{N+1}(s) \tag{8.2.6}$$

whenever the limit exists.

Theorem 8.2.3. Assume finite S. Let ν denote the nonzero eigenvalue of $I - P$ with the smallest modulus. Then, for $0 < \rho < |\nu|$,

$$v_\lambda = (1+\rho)\left[\rho^{-1} y_{-1} + \sum_{n=0}^{\infty} \rho^n y_n\right] \tag{8.2.8}$$

where $y_{-1} = P^* r = g$, $y_0 = H_P r = h$, and $y_n = (-1)^n H_P^{n+1} r$ for $n = 1, 2, \ldots$.

Theorem 8.2.6. Let S be finite and let g and h denote the gain and bias of a MRP with transition matrix P and reward r.

a. Then

$$(I - P)g = 0 \tag{8.2.11}$$

and

$$g + (I - P)h = r \tag{8.2.12}$$

b. Suppose g and h satisfy (8.2.11) and (8.2.12), then $g = P^* r$ and $h = H_P r + u$ where $(I - P)u = 0$.
c. Suppose g and h satisfy (8.2.11), (8.2.12) and $P^* h = 0$, then $h = H_P r$.

Corollary 8.2.7. Suppose P is unichain or irreducible. Then the average reward $P^* r = ge$ and it is uniquely determined by solving

$$ge + (I - P)h = r. \tag{8.2.14}$$

Suppose g and h satisfy (8.2.14), then $g = P^* r$ and $h = H_P r + ke$ for arbitrary scalar k. Furthermore, if g and h satisfy (8.2.14) and $P^* h = 0$, then $h = H_P r$.

Theorem 8.2.8. In a Markov reward process with transition matrix P and reward r, let y_n, $n = -1, 0, \ldots$ denote the coefficients of the Laurent series expansion of v_λ.

a. Then

$$(I - P)y_{-1} = 0, \tag{8.2.19}$$

$$y_{-1} + (I - P)y_0 = r, \tag{8.2.20}$$

and for $n = 1, 2, \ldots$

$$y_{n-1} + (I - P)y_n = 0. \tag{8.2.21}$$

b. Suppose for some $M \geq 0$, that $w_{-1}, w_0, \ldots, w_M$ satisfy (8.2.19), (8.2.20), and if $M \geq 1$, (8.2.21) for $n = 1, 2, \ldots, M$. Then $w_{-1} = y_{-1}$, $w_0 = y_0, \ldots, w_{M-1} = y_{M-1}$ and $w_M = y_M + u$ where $(I - P)u = 0$.

c. Suppose the hypotheses of part (b) hold and in addition $P^* w_M = 0$. Then $w_M = y_M$.

Corollary 8.2.9. Suppose u, v, and w satisfy

$$(I - P)u = 0, \quad u + (I - P)v = r \quad \text{and} \quad v + (I - P)w = 0 \tag{8.2.22}$$

then $u = g$, $v = h$, and $w = y_1 + z$ where $(I - P)z = 0$.

Theorem 8.4.1. Suppose S is countable.

a. If there exists a scalar g and an $h \in V$ which satisfy $B(g, h) \leq 0$, then

$$ge \geq g_+^*. \tag{8.4.4}$$

b. If there exists a scalar g and $h \in V$ which satisfy $B(g, h) \geq 0$, then

$$ge \leq \sup_{d \in D^{MD}} g_-^{d^\infty} \leq g_-^*. \tag{8.4.5}$$

c. If there exists a scalar g and an $h \in V$ for which $B(g, h) = 0$, then

$$ge = g^* = g_+^* = g_-^*. \tag{8.4.6}$$

Corollary 8.4.2. Suppose that

$$\lim_{N \to \infty} N^{-1} E_s^\pi \{h(X_N)\} = 0 \tag{8.4.9}$$

for all $\pi \in \Pi^{HR}$ and $s \in S$. Then results (a) – (c) in Theorem 8.4.1 hold.

Theorem 8.4.3. Suppose S and A_s are finite Assumption 8.0.2 holds and the model is unichain.

a. Then there exists a $g \in R^1$ and an $h \in V$ for which

$$0 = \max_{d \in D} \{r_d - ge + (P_d - I)h\}.$$

b. If (g', h') is any other solution of the average reward optimality equation, then $g = g'$.

Theorem 8.4.4. Suppose there exists a scalar g^*, and an $h^* \in V$ for which $B(g^*, h^*) = 0$. Then, if d^* is h^*-improving, $(d^*)^\infty$ is average optimal.

Theorem 8.4.5. Suppose S is finite and A_s is finite for each $s \in S$, $r(s, a)$ is bounded and the model is unichain. Then

a. there exists a stationary average optimal policy,

b. there exists a scalar g^* and an $h^* \in V$ for which $B(g^*, h^*) = 0$,

c. any stationary policy derived from an h^*-improving decision rule is average optimal, and

d. $g^* e = g_+^* = g_-^*$.

Assumption 8.4.1. For each $s \in S$, $r(s, a)$ is a bounded, continuous function of a.

Assumption 8.4.2. For each $s \in S$ and $j \in S$, $p(j|s, a)$ is a continuous function of a.

Proposition 8.4.6. Let $\{P_n\}$ denote a sequence of unichain transition probability matrices and suppose

$$\lim_{n \to \infty} \|P_n - P\| = 0, \tag{8.4.16}$$

then

a. $\lim_{n \to \infty} \|P_n^* - P^*\| = 0$, and

b. $\lim_{n \to \infty} \|H_{P_n} - H_P\| = 0$.

Theorem 8.4.7. Suppose S is finite, A_s is compact, the model is unichain, and Assumptions 8.4.1 and 8.4.2 hold.

a. Then there exists a $g \in R^1$ and an $h \in V$ for which

$$0 = \max_{d \in D} \{r_d - ge + (P_d - I)h\}.$$

b. If (g', h') is any other solution of the average reward optimality equation, then $g = g'$.

c. There exists a $d^* \in D^{MD}$ for which $(d^*)^\infty$ is average optimal; further, if d is h-improving, then d^∞ is average optimal.

$$Lv \equiv \max_{d \in D} \{r_d + P_d v\} \tag{8.5.2}$$

Value Iteration Algorithm

1. Select $v^0 \in V$, specify $\varepsilon > 0$ and set $n = 0$.
2. For each $s \in S$, compute $v^{n+1}(s)$ by

$$v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j) \right\}. \tag{8.5.3}$$

3. If

$$sp(v^{n+1} - v^n) < \varepsilon, \tag{8.5.4}$$

go to step 4. Otherwise increment n by 1 and return to step 2.

4. For each $s \in S$, choose

$$d_\varepsilon(s) \in \arg\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j) \right\} \tag{8.5.5}$$

and stop.

Theorem 6.6.6 shows that

$$sp(v^{n+2} - v^{n+1}) \le \gamma sp(v^{n+1} - v^n), \tag{8.5.6}$$

where

$$\gamma = \max_{s \in S, a \in A_s, s' \in S, a' \in A_{s'}} \left[1 - \sum_{j \in S} \min\{ p(j|s,a), p(j|s',a') \} \right]. \tag{8.5.7}$$

We say that an operator $T: V \to V$ is a *J-stage span contraction* if there exists an α, $0 \le \alpha < 1$ and a non-negative integer J for which

$$sp(T^J u - T^J v) \le \alpha\, sp(u - v) \tag{8.5.8}$$

for all u and v in V.

Proposition 8.5.1. Suppose T is a J-stage span contraction. Then for any $v^0 \in V$, the sequence $v^n = T^n v^0$ satisfies

$$sp(v^{nJ+1} - v^{nJ}) \le \alpha^n sp(v^1 - v^0) \tag{8.5.9}$$

for all non-negative integers n.

Theorem 8.5.2. Suppose there exists an integer $J \ge 1$ such that, for every pair of deterministic Markov policies π_1 and π_2,

$$\eta(\pi_1, \pi_2) \equiv \min_{(s,u) \in S \times S} \sum_{j \in S} \min\left\{ P^J_{\pi_1}(j|s), P^J_{\pi_2}(j|u) \right\} > 0. \tag{8.5.10}$$

a. Then L defined in (8.5.2) is J-step contraction operator on V with contraction coefficient

$$\gamma' = 1 - \min_{\pi_1, \pi_2 \in \Pi^{MD}} \eta(\pi_1, \pi_2). \tag{8.5.11}$$

b. For any $v^0 \in V$, let $v^n = L^n v^0$. Then, given $\varepsilon > 0$, there exists an N such that

$$sp(v^{nJ+1} - v^{nJ}) \le \varepsilon$$

for all $n \ge N$.

Theorem 8.5.3. Suppose either

a. $0 \le \gamma < 1$, where γ is given in (8.5.7),
b. there exists a state $s' \in S$ and an integer K such that, for any deterministic Markov policy π, $P^K_\pi(s'|s) > 0$ for all $s \in S$, or
c. all policies are unichain and $p(s|s,a) > 0$ for all $s \in S$ and $a \in A_s$.

Then (8.5.10) holds for all π_1 and π_2 in Π^{MD} and the conclusions of Theorem 8.5.2 follow.

Theorem 8.5.4. Suppose that all stationary policies are unichain and that every optimal policy has an aperiodic transition matrix. Then, for all $v^0 \in V$ and any $\varepsilon > 0$, the sequence of $\{v^n\}$ generated by the value iteration algorithm satisfies (8.5.4) for some finite N.

Theorem 8.5.5. Suppose the hypotheses of Theorem 8.4.5 hold, then for $v \in V$,

$$\min_{s\in S}[Lv(s) - v(s)] \le g^{d^\infty} \le g^* \le \max_{s\in S}[Lv(s) - v(s)], \qquad (8.5.12)$$

where d is any v-improving decision rule.

Theorem 8.5.6. Suppose (8.5.4) holds and d_ε satisfies (8.5.5).

a. Then $(d_\varepsilon)^\infty$ is an ε-optimal policy.

b. Define

$$g' = \tfrac{1}{2}\Big[\max_{s\in S}\big(v^{n+1}(s) - v^n(s)\big) + \min_{s\in S}\big(v^{n+1}(s) - v^n(s)\big)\Big]. \qquad (8.5.13)$$

Then $|g' - g^*| < \varepsilon/2$ and $|g' - g^{(d_\varepsilon)^\infty}| < \varepsilon/2$.

Theorem 8.5.7. Suppose (8.5.10) holds for all π_1 and π_2 in Π^{MD}, then, for any $\varepsilon > 0$, value iteration satisfies (8.5.4) in a finite number of iterations, identifies an optimal policy through (8.5.5), and obtains an approximation to g^* through (8.5.13). Further, for $n = 1, 2, \ldots$,

$$sp(v^{nJ+1} - v^{nJ}) \le (\gamma')^n sp(v^1 - v^0). \qquad (8.5.14)$$

The Unichain Policy Iteration Algorithm

1. Set $n = 0$ and select an arbitrary decision rule $d_n \in D$.
2. (Policy evaluation) Obtain a scalar g_n and an $h_n \in V$ by solving

$$0 = r_{d_n} - ge + (P_{d_n} - I)h. \qquad (8.6.1)$$

3. (Policy Improvement) Choose d_{n+1} to satisfy

$$d_{n+1} \in \arg\max_{d\in D}\{r_d + P_d h_n\}, \qquad (8.6.2)$$

setting $d_{n+1} = d_n$ if possible.

4. If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise increment n by 1 and return to step 2.

Proposition 8.6.3. Suppose d_n is determined in the improvement step of the policy iteration algorithm, and h_n is any solution of (8.6.1). Then

$$h^{d_{n+1}^\infty} = h^{d_n^\infty} - P^*_{d_{n+1}} h^{d_n^\infty} + H_{d_{n+1}} B(g_n, h_n). \qquad (8.6.10)$$

Lemma 8.6.4. Let $d \in D$. Suppose P_d is unichain and, in canonical form, R_d denotes recurrent states of P_d, T_d denotes its transient states, and H_d is expressed as

$$H_d = \begin{bmatrix} H_d^{RR} & H_d^{RT} \\ H_d^{TR} & H_d^{TT} \end{bmatrix}, \qquad (8.6.11)$$

where the partition corresponds to states in R_d and T_d. Then

a. $H_d^{RT} = 0$;

b. $H_d^{TT} = (I - P_d^{TT})^{-1}$, where P_d^{TT} denotes the restriction of P_d to its transient states, and

c. if $u(s) = 0$ for $s \in R_d$ and $u(s) \ge 0$ for $s \in T_d$, then $H_d u \ge u \ge 0$.

Proposition 8.6.5. Suppose d_{n+1} is determined in step 3 of the policy iteration algorithm. If $B(g_n, h_n)(s) = 0$ for all s that are recurrent under d_{n+1}, and $B(g_n, h_n)(s_0) > 0$ for some s_0 which is transient under d_{n+1} then

$$h^{d_{n+1}^\infty}(s) > h^{d_n^\infty}(s),$$

for some s which is transient under d_{n+1}.

Theorem 8.6.6. Suppose all stationary policies are unichain, and the set of states and actions are finite, then policy iteration converges in a finite number of iterations to a solution (g^*, h) of the optimality equation $B(g, h) = 0$ and an average optimal stationary policy $(d^*)^\infty$.

Primal Linear Program.

Minimize g

subject to

$$g + h(s) - \sum_{j\in S} p(j|s, a)h(j) \ge r(s, a), \qquad a \in A_s \text{ and } s \in S,$$

with g and $h(s)$ unconstrained.

Dual Linear Program.

$$\text{Maximize} \sum_{s\in S}\sum_{a\in A_s} r(s,a)x(s,a) \tag{8.8.2}$$

subject to

$$\sum_{a\in A_j} x(j,a) - \sum_{s\in S}\sum_{a\in A_s} p(j|s,a)x(s,a) = 0, \qquad j\in S \tag{8.8.3}$$

$$\sum_{s\in S}\sum_{a\in A_s} x(s,a) = 1, \tag{8.8.4}$$

and $x(s,a) \geq 0$ for $a \in A_s$, $s \in S$.

$$\sum_{s\in S} p_d(j|s)\pi(s) = \pi(j) \qquad j\in S \tag{8.8.5}$$

$$\sum_{j\in S} \pi(j) = 1 \tag{8.8.6}$$

$$q_{d_x(s)}(a) = \frac{x(s,a)}{\sum_{a'\in A_s} x(s,a')}, \qquad a\in A_s, \quad s\in S. \tag{8.8.8}$$

Proposition 8.8.5. In a unichain model, for any $d \in D^{MR}$, the system of equations (8.8.5) subject to (8.8.6) has the unique solution π_d in which $\pi_d(s) > 0$ for $s \in R_d$ and $\pi_d(s) = 0$ for $s \in T_d$.

Theorem 8.8.6. Suppose that the transition probability matrix of every stationary policy is unichain.

a. Then for $d \in D^{MR}$,

$$x_d(s,a) = \begin{cases} q_{d(s)}(a)\pi_d(s) & s\in R_d \\ 0 & s\in T_d \end{cases}$$

is a feasible solution to the dual LP.

b. Let $x(s,a)$ be a feasible solution to the dual,

$$S_x \equiv \left\{ s\in S: \sum_{a\in A_s} x(s,a) > 0 \right\} \tag{8.8.11}$$

and define d_x by (8.8.8) for $s \in S_x$ and arbitrary for $s \in S/S_x$. Then $S_x = R_{d_x}$ and $x_{d_x}(s,a) = x(s,a)$ for $a \in A_s$ and $s \in R_{d_x}$.

Corollary 8.8.7.

a. Let x be a basic feasible solution to the dual LP and suppose that d_x is defined as in Theorem 8.8.6(b). Then, for $s \in S_x$, $d_x(s)$ is deterministic and satisfies

$$d_x(s) = \begin{cases} a & \text{if } x(s,a) > 0 \quad \text{for } s\in S_x \\ \text{arbitrary} & \text{for } s\in S/S_x. \end{cases}$$

b. Suppose that $d(s)$ is deterministic; then $x_d = \pi_d$ is a basic feasible solution to the dual LP.

Corollary 8.8.8. There exists a bounded optimal basic feasible solution x^* to the dual and the policy $(d_{x^*})^\infty$ defined by

$$d_{x^*}(s) = \begin{cases} a & \text{if } x^*(s,a) > 0 \quad \text{for } s\in S_{x^*} \\ \text{arbitrary} & \text{for } s\in S/S_{x^*} \end{cases}$$

is an optimal policy.