Study Notes for Dynamic Programming

Mike Mingcheng Wei

May 8, 2020

Contents

Chapter 1

Finite-Horizon MDP

In general, for the finite-horizon problem with Markov Decision Process, we can

- Focus on MD policy;
- The optimal value function satisfies OE;
- The optimal policy is the solution of OE;

Hence, we can first write down the OE for this finite-horizon problem, and then analyze OE and find the policy or decision rule to solve for the OE. In finding the optimal policy, we can find the structure of the decision rule by using the techniques at the last of this notes.

1.1 Primliminary

The focus of dynamic programming is the sequential control (also known as dynamic or real time control) of a discrete-time dynamic system with **additive reward / cost**[1].

In this chapter, we only focus on MDP: Markov Decision Processes: both the reward and transition probability depend only on current state-action pair but not on the history of state-action sequence.

Hence, base on this MDP assumption, there must exist one optimal MD policy, so we only need to consider MD policy within HR policy domain.[2]

Definition 1 *If time horizon, $N < \infty$, then it is **finite horizon** DP model*

Definition 2 *If time horizon, $N = \infty$, then it is **infinite horizon** DP model*

Definition 3 *We say the decision rule is **History Dependent, H,** if it is a decision rule depends on past history.*

Definition 4 *We say the decision rule is **Markovian, M,** if it is a decision rule depends on past only through the current state s and currente stage t;*

Definition 5 *A **deterministic decision rule, D,** is that a decision rule, $d_t(\cdot)$, choose an action with certainty;*

Definition 6 *A **randomized decision rule, R,** is that a decision rule, $d_t(\cdot)$, choose an action $a \in A_s$ with probability $q(a)$;*

[1] If the period cost is not additive or not separetable, then DP is not very efficient tool.

[2] This assumption, MDP, is critial to prove MD policy is optimal within the domain of HR policy. (The main proof is Thm 4.4.2, which rely on the assumption of reward and transition probability depending only on current state-action pair.)

Definition 7 *A **history-dependent-deterministic, HD, rule** is deterministic and depends on past history.*

Definition 8 *A **history-dependent-random, HR, rule** is random and depends on past history.*

Definition 9 *A **markovian-deterministic, MD, rule** is deterministic and depends on past only through the current state s and currente stage t;.*

Definition 10 *A **markovian-random, MR, rule** is random and depends on past only through the current state s and currente stage t;.*

Definition 11 *The **open-loop policy** is to select orders $\{a_1, ..., a_{N-1}\}$ at once at time $t = 1$, without observing demand.*

Definition 12 *The **close-loop policy** is based on obvserved demand to make decision at the begining of each time period and prescribed by a sequence of decision rules $\{d_1, ..., d_{N-1}\}$.*

1.1.1 Notation

Parameter	Explaination
N	Time horizon or the number of times decisions is made
t	Decision epoch, $t = 1, 2, ..., N$
s_t	System state; observed at the beginning of period t
S	State Space
a_t	Control; decision to be selected at time t
A_s	Action space of state s
$r_t(s_t, a_t)$	The reward / cost in time t (MDP)
$r_N(s_N)$	The terminal reward at the last period N
$P(s_{t+1}\|s_t, a_t)$	The transition probability (MDP)
$d_t(\cdot)$	Decision rule at time t
π	A policy: a sequence of decision rules to be used ($\pi = (d_1, ..., d_{N-1})$)
$v_N^\pi(s)$	Expected total reward over N periods, given the system starts in state s and policy π is used $v_N^\pi(s) = E_s^\pi \left[\sum_{t=1}^{N-1} r_t(s_t, a_t) + r_N(s_N)\right]$
$v_N^*(s)$	Optimal value function of expected total reward; $v_N^*(s) = \sup_{\pi \in \Pi^{HR}} v_N^\pi(s)$
$u_t^\pi(s)$	Expected total reward for using π from time t to the end of period N: $u_t^\pi(s) = E_s^\pi \left[\sum_{n=t}^{N-1} r_n(s_n, a_n) + r_N(s_N)\right]$
$u_t^*(s)$	Best expected total reward attainable from time t to the end of period N: $u_t^*(s) = \sup_{\pi \in \Pi^{HR}} u_t^\pi(s)$

1.2 Main result

1.2.1 General Result

This section's result build on the main assumption: reward and transition probability depending only on current state-action pair. Hence, it is markov process.

Definition 13 *We say the reward is **uniformly bounded** if*

$$|r_t(s, a)| \leq M < \infty, \text{ for } s \in S \text{ and } a \in A_s$$

Definition 14 *We say the policy, π^*, is an **optimal policy** if*

$$v_N^{\pi^*}(s) = v_N^*(s), \text{ for } s \in S$$

Theorem 15 *If rewards are uniformly bounded and in finite horizon decision process, then $v_N^*(s)$ is well defined and always exists.*

However, the optimal policy π^* may not exist.

Definition 16 *A policy π_ε^* is said to be $\varepsilon-$**optimal policy** if*

$$v_N^{\pi_\varepsilon^*}(s) + \varepsilon > v_N^*(s), \text{ for } s \in S$$

Theorem 17 *If rewards are uniformly bounded and in finite horizon decision process, then there always exists a $\varepsilon-$**optimal policy** for any $\varepsilon > 0$.*

Algorithm 18 *A finite Horizon Policy Evaluation Algorithm:*
1. *Set $t = N$ and $u_N^\pi(s) = r_N(s)$ for all $s \in S$*
2. *if $t = 1$, stop, otherwise go to step 3;*
3. *reduce t by 1 and compute $u_t^\pi(s)$ for each s by $u_t^\pi(s) = r_t(s, d_t(s)) + \sum_{j \in S} p(j|s, d_t(s)) u_{t+1}^\pi(j)$;*
4. *return to step 2.*

This algorithm reduce this multistage problems by evaluating a sequence of simpler, inductively defined single-period problems.

Theorem 19 *Let $\pi \in \Pi^{HD}$ and suppose that $u_t^\pi(s)$ is obtained by using the policy iteration algorithm. Then for all $t \leq N$,*

$$u_t^\pi(s) = E_s^\pi \left[\sum_{n=t}^{N-1} r_n(s_n, a_n) + r_N(s_N) \right]$$

(This theorem implies $u_1^\pi(s) = v_N^\pi(s)$)
(This theorem is still valid for $\pi \in \Pi^{HR}$: Theorem 4.2.2 in Puterman)

Definition 20 *The **optimality equations (AKA bellman equations)** are defined as:*

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

for each history realization $h_t \in H_t$. The boundary conditions satisfy

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

for each $h_N = (s_1, a_1, ..., s_{N-1}, a_{N-1}, s_N)$

Lemma 21 *(Lemma 4.3.1) Let w be a real-valued function on an arbitrary discrete set W and $q(\cdot)$ be a probability distribution on W. Then*

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

Theorem 22 *(Theorem 4.3.2) suppose u_t is a solution of Equation (4.3.2) for $t = 1, ..., N-1$, and u_N satisfies Equation (4.3.3), then*
a. $u_t(h_t) = u_t^*(h_t)$ for all $h_t \in H_t$, $t = 1, ..., N$;
b. $u_1(s_1) = u_1^*(s_1) = v_N^*(s_1)$ for all $s_1 \in S$.
(The part a means optimality equations has unique solution, which is $u_t^*(h_t)$ for all $h_t \in H_t$.)
(THe part b indicate that we can find the optimal value function $v_N^*(s_1)$ by solving optimality equation.)

Theorem 23 *(Theorem 4.3.3)(If the optimal policy exists: Optimal policy) suppose $u_t^*, t = 1, ..., N$, are the solutions of the optimality equations and $\pi^* = (d_1^*, ..., d_{N-1}^*) \in \Pi^{HD3}$ satisfies:*

$$r_t(s_t, d_t^*(h_t)) + \sum\nolimits_{j \in S} p_t(j|s, d_t^*(h_t)) u_{t+1}^*(h_t, d_t^*(h_t), j) \tag{4.3.10}$$
$$= \max_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum\nolimits_{j \in S} p_t(j|s, a) u_{t+1}^*(h_t, a, j) \right\}, for t = 1, ..., N-1$$

(we assume sup is attained, so the optimal policy exists) Then
a. For each $t = 1, ..., N$, $u_t^{\pi^*}(h_t) = u_t^*(h_t), h_t \in H_t$;
b. π^* is an optimal policy and $v_N^{\pi^*}(s) = v_N^*(s), \forall s \in S$.

Theorem 24 *(Principle of Optimality: Bellman 1957) An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

(This is an implication of Theorem 4.3.3: at each decision epoch, the remaining decision $(d_t^*, ..., d_{N-1}^*)$ must be optimal because of the construction of equation 4.3.10)

Theorem 25 *(Theorem 4.3.4)(If the optimal policy doesn't exist: $\varepsilon-$optimal policy) Let $\varepsilon > 0$. suppose $u_t^*, t = 1, ..., N$, are the solutions of the optimality equations and $\pi^\varepsilon = (d_1^\varepsilon, ..., d_{N-1}^\varepsilon) \in \Pi^{HD}$ satisfies:*

$$r_t(s_t, d_t^\varepsilon(h_t)) + \sum\nolimits_{j \in S} p_t(j|s, d_t^\varepsilon(h_t)) u_{t+1}^*(h_t, d_t^\varepsilon(h_t), j) + \frac{\varepsilon}{N-1} \tag{4.3.10}$$
$$\geq \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum\nolimits_{j \in S} p_t(j|s, a) u_{t+1}^*(h_t, a, j) \right\}, for t = 1, ..., N-1$$

Then
a. For each $t = 1, ..., N-1$, $u_t^{\pi^\varepsilon}(h_t) + \frac{(N-t)\varepsilon}{N-1} = u_t^*(h_t), h_t \in H_t$;
b. π^ε is an $\varepsilon-$optimal policy and $v_N^{\pi^\varepsilon}(s) + \varepsilon \geq v_N^*(s), \forall s \in S$.

Theorem 26 *(Theorem 4.4.2) Let $u_t^*, t = 1, ..., N$, are the solutions of the optimality equations, Then*
a. For each $t = 1, ..., N$, $u_t^*(h_t)$ depends on h_t only through s_t;
b. For any $\varepsilon > 0$, there exists an $\varepsilon-$optimal policy which is deterministic and Markov
c. If there exists an $a' \in A_{s_t}$ that satisfies

$$r_t(s_t, a') + \sum\nolimits_{j \in S} p_t(j|s, a') u_{t+1}^*(h_t, a', j)$$
$$= \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum\nolimits_{j \in S} p_t(j|s, a) u_{t+1}^*(h_t, a, j) \right\}, for\ t = 1, ..., N-1$$

Then there exists an MD policy that is optimal.

(Hence, this theorem tell us that we can focus only on MD policy among HR policy: within the domain of HR policy, we only need to consider MD policy (MD optimal policy or MD $\varepsilon-$optimal policy).)

Theorem 27 *(Thm 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 $j, s \in S$ and $t = 1, ..., N$, or
c. A_s is compact, $r_t(s, a)$ is upper semicontinuous 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 lower semi-continuous in a for each $j, s \in S$ and $t = 1, ..., N$, or
Then there exists a deterministic Markovian policy which is optimal.
(This theorem says when there exist a MD optimal policy, not only $\varepsilon-$optimal policy)

[3]The reason why we can only consider HD policy instead of more general HR policy is because of Lemma 4.3.1. If there existed a history-dependent randomized policy which satisfied the obvious generalization of equation 4.3.10, as a result of Lemma 4.3.1, we could find a deterministic policy which satisfied equation 4.3.10.
Hence, we can only consider HD policy with in the domain of HR policy: HD is the dominant set in HR.

Summary 28 *1. $u_t^*(s)$ is the unique solution of optimality equation (Thm 4.3.2 a);*
2. The optimality equations can be used to determine the optimal policy when exists (Thm 4.3.3);
3. If the expected total reward under policy π, $u_t^\pi(s)$, satisfies optimality equations for $t = 1, ..., N$ then π is optimal (Thm 4.3.3)
4. Backward induction provides an efficient method for computing optimal value functions and policies; (Thm 4.3.2 b)
5. We only need to consider MD policy within the domain of HR policy (Lemma 4.3.1 and Thm 4.4.2)

1.2.2 Deterministic systems (Perfect prediction of furture; non-stochastic):

Theorem 29 *For deterministic system, since future is perfectly predictable, the reward achieved by an optimal closed-loop policy can also be achieved by an optimal open loop policy.*

The finite state deterministic systems can be represented by a graph and equivalent to a shortest path problem. For example,

- shortest path problem,
- traveling salesman problem,
- four queues problem,
- sequential allocation problem,
- constrained maximum likelihood problem,
- Hidden markov Model;(In order to estimate the transition probability given the state transition observation) (Viterbi Algorithm)
- Convolutional coding and decoding

All above problem can be solved by backward DP algorithm (or equivalently the forward DP algorithm). (those two algorithm is very simple, refer to Lecture (5) of Dr. Susan Xu's handout.)

1.2.3 Non-deterministic System (Furture is uncertain; stochastic model)

If the optimal policy / best expected reward has no obvious structure, then we can use backward induction.
The backward induction assumes that maxima are obtained in

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

so sup can be replaced by max so that we are assured to be able to find an MD optimal policy instead of an $\varepsilon-$optimal policy.

Algorithm 30 *The Backward Induction Algorithm:*
1. Set $t = N$ and $u_N^\pi(s) = r_N(s)$ for all $s \in S$
2. reduce t by 1 and compute $u_t^*(s)$ for each $s \in S$ by

$$u_t^*(s) = \max_{a \in A_{s_t}} \left\{ r_t(s, a) + \sum\nolimits_{j \in S} p(j|s, a) u_{t+1}^\pi(j) \right\} \tag{4.5.1}$$

set

$$A_{s_t,t} = \arg\max_{a \in A_{s_t}} \left\{ r_t(s, a) + \sum\nolimits_{j \in S} p(j|s, a) u_{t+1}^\pi(j) \right\} \tag{4.5.2}$$

and let $d_t^*(s_t) \in A_{s_t,t}^*$;
3. if $t = 1$, stop. Otherwise return to step 2.

Remark 31 *Properties of backward induction algorithm:*

a. It computes the expected total reward for the entire horizon and from each period to the end of the horizon;

b. It determines the optimal policy π^* as follows: $A^*_{s_t,t}$ represents the set of all optimal actions in state s_t at time t; Let $d^*_t(s_t) \in A^*_{s_t,t}$, then the optimal policy is $\pi^* = (d^*_1, ..., d^*_{N-1})$;

c. Complexity: Assume K states in each period and L actions in each state: backward induction algorithm requires $(N-1)LK^2$. (While direct evaluation of all MD policies require $L^{K(N-1)}(N-1)K^2$)

If optimal policy / best expected reward has obvious specific structure (Such as Sequential allocation problem), then we can use conjecture and induction:

1. We can use backward DP to calculate the last few periods' optimal policy / best expected reward for those periods
2. then make conjecture of the structure of the optimal policy / best expected reward
3. use induction to proof the optimality of the conjecture.

Sample Problem:

- So who is counting problem
- A gambling model
- A card game (S. Ross, Introduction to DP)
- Optimal Stopping problem
  - Secretary Problem (Hlynka and Sheahan 1988: the seretary problem for a random walk)
- A quality control model with learning effects (C. Fine 1988: a quality control model with learning effects)
- Airline yield management / revenue managment (Lauterbacher and Stidham 1999: The underlying MDP in the single-leg airline Yield Management Problem)

Chapter 2

Infinite-Horizon MDP

The focus of dynamic programming is the sequential control (also known as dynamic or real time control) of a discrete-time dynamic system with additive reward / cost.

In this chapter, we only focus on MDP: Markov Decision Processes: both the reward and transition probability depend only on current state-action pair but not on the history of state-action sequence.

Also, In this chapter, we assume **Stationary data**: reward $r(s,a)$, transition probability $P(j|s,a)$, and decision sets A_s do not depend on decision epoch t.

Analogous to finite-horizon MDP, the optimal value function of expected total reward is defined as

$$v^*(s) = \sup_{\pi\in\Pi^{HR}} E_s^\pi\left[\sum_{t=1}^{\infty} r(s_t,a_t)\right] = \sup_{\pi\in\Pi^{HR}} \lim_{N\to\infty} E_s^\pi\left[\sum_{t=1}^{N} r(s_t,a_t)\right], \text{ for } s\in S$$

However, in real practice, it will be more efficient to consider the infinite-horizon MDP by considering the limit of the finite period problem:

$$v_\infty^*(s) = \lim_{N\to\infty} v_{N+1}^*(s) = \lim_{N\to\infty}\left\{\sup_{\pi\in\Pi^{HR}} E_s^\pi\left[\sum_{t=1}^{N} r(s_t,a_t)\right]\right\}, \text{ for } s\in S$$

But, there are two technique problem:

1. Does the limit, $v_\infty^*(s) = \lim_{N\to\infty} v_N^*(s)$, exist?
2. If the limit exists, is it always true $v^*(s) = v_\infty^*(s)$?

Hence, because of those two problems, the infinite MDP has pitfalls:

- It is possible that $v^*(s) \neq v_\infty^*(s)$;
- The solution of the optimality equations may not be unique; [1]
- the policy determined by the optimality equations may not be optimal ; [2]

Definition 32 *A MDP model is said to be* ***unstable (instability)*** *if $v^*(s) \neq v_\infty^*(s)$ for some s.*

Hence, if a MDP is unstable, using the limit of the finite period problem to determine the infinite-horizon MDP is wrong. However, if we have stable problem, then it is directly to consider the limit of finite period problem. There are some ways, such as discounting, we can remove the instability of a MDP problem.

[1] In finite-horizon problem, by Theorem 4.3.2, the optimality has unique solution

[2] In finite-horizon problem, by Theorem 4.3.3, the policy determined by the optimality equations is optimal.

Definition 33 *Similarlly, we define the* ***optimal value function*** *of a infinite-horizon MDP as*

$$\textit{For Expected Total reward: } v^*(s) = \sup_{\pi \in \Pi^{HR}} v^\pi(s)$$

$$\textit{For Expected Total discounted reward: } v^*_\lambda(s) = \sup_{\pi \in \Pi^{HR}} v^\pi_\lambda(s)$$

Definition 34 *We define the* ***optimal policy***, $\pi^* \in \Pi^{HR,HD,MR,MD}$ *of a infinite-horizon MDP as*

$$\textit{For Expected Total reward: } v^{\pi^*}(s) = v^*(s)$$

$$\textit{For Expected Total discounted reward: } v^{\pi^*}_\lambda(s) = v^*_\lambda(s)$$

Theorem 35 *(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. $v^\pi_N(s) = v^{\pi'}_N(s)$ *for* $1 \le N < \infty$*; if* $r_N(s) = 0$ *and* $v^\pi(s) = \lim_{N\to\infty} v^\pi_N(s)$ *exists,* $v^\pi(s) = v^{\pi'}(s)$*;*

b. $v^\pi_\lambda(s) = v^{\pi'}_\lambda(s)$ *for* $0 \le \lambda < 1$*;*[3]

2.1 Use Finite counterpart to solve for Infinite Period problem

(This is a self summary of how to solve for infinite problem from its finite counterpart. And should be verified by reference for correctness.)

For a infinite problem we want to find optimal value function, $v(s)$, and optimal decision rule ψ such that Optimality equation holds $v^\psi(s) = r(s,a) + \lambda \sum_{j\in S} p(j|s,a) v^\psi(j)$.

Hence, we first find its finite period counterpart: $u_t(h_t) = \sup_{a\in A_{s_t}} \left\{ r_t(s_t,a) + \lambda \sum_{j\in S} p_t(j|s,a) u_{t+1}(s_t,a,j) \right\}$ and $u_N(s_N) = u^*_N(s_N) = r_N(s_N)$. Then hypothesis the functional properties, such as convexity or monotone, of $r_N(s_N)$. (As long as persevation holds, the functional properties can be transfered to the infinite case.) Base on those perfered properties, derive the optimal decision rule, ψ, for $u_{N-1}(s_{N-1}) = \sup_{a\in A_{s_t}} \left\{ r_{N-1}(s_{N-1},a) + \lambda \sum_{j\in S} p_t(j|s_N,a) u_N(s_{N-1},a,s_N) \right\}$ such that $u^*_{N-1}(s_{N-1}) = r_{N-1}(s_{N-1},\psi) + \lambda \sum_{j\in S} p_t(j|s_N,\psi) u_N$ (Attainment holds). Then check under optimal decision rule, ψ, whether $u^*_{N-1}(s_{N-1})$ preserve the functional properties we assumed for $r_N(s_N)$. (Check for perservation. The reason why we need preservation is that the optimal decision rule, ψ, may need certain functional properties hold for $u_{t+1}(s_t,\psi,s_{t+1})$ to be optimal for $u_t(s_t) = \sup_{a\in A_{s_t}} \left\{ r_t(s_t,a) + \lambda \sum_{j\in S} p_t(j|s,a) u_{t+1}(s_t,a,j) \right\}$. So if persevation holds, the optimal decision rule, ψ, can be recursively used as optimal decision rule as long as those functional properties holds.). If perservation holds, then we can recursively induct this process for $u^*_{N-2}(\cdot)$, $u^*_{N-2}(\cdot)$, .., $u^*_1(\cdot)$, such that the optimal decision rule, ψ,holds for every period and all $u^*_i(\cdot)$ perserve our assumed functional properties.

Remark 36 *For each N, we have $v^*_N(s) = u^*_1(s)$, and the optimal decision rule, ψ, is optimal for $v^*_N(s)$ and $v^*_N(s)$ perserve our assumed functional properties of $r_N(s_N)$. Let $N \to \infty$, and let $v^*_\infty(s) = \lim_{N\to\infty} v^*_N(s)$, so the optimal decision rule, ψ, is optimal for $v^*_\infty(s)$ and the functional properties may be perserved for $v^*_\infty(s)$. (Some properties such as convexity and continuouty indeed preserved under limitation, but be alter other possibilities.)*

For stable problem, such as discounted problem, the optimal value function for infinite problem, $v^*(s)$, equals the limiting optimal value function, $v^*_\infty(s)$. Hence, the optimal decision rule, ψ, is optimal for $v^*(s)$ and the functional properties may be perserved for $v^*(s)$.

[3]This means we only need to consider MR policy within the domain of all HR policy for expected total discounted reward, expected total reward, and average reward models.

Remark 37 *For discounted problem, the above argument can be understand by using contraction mapping argument.*

For any terminal value function $u_N^*(s_N)$ with some functional properties, if we can prove that there is optimal decision rule, ψ, such that attainment holds: $u_{N-1}^*(s_{N-1}) = r_{N-1}(s_{N-1}, \psi) + \lambda \sum_{j \in S} p_t(j|s_N, \psi) u_N^*(s_{N-1}, \psi, s_N)$. (Use simple notation, given some functional propperties for $u_N(s_N)$, the optimal decision rule hold for $u_{N-1}^*(\cdot) = Au_N^*(\cdot)$). If we can prove perservation holds for the recursive relationship under optimal decision rule, ψ, such that $Au_N^*(\cdot)$ has the same desired functional properties as $u_N^*(\cdot)$. Then $A^n u_N^*(\cdot)$ for any n, will have the desired functional properties as $u_N^*(\cdot)$ and decision rule, ψ, will be optimal for $A^n u_N^*(\cdot)$ as well.

Under discount model, the optimal operator A is an $\lambda-$contraction mapping. So there exists one and only one fixed point such that $v^*(\cdot) = \lim_{n \to \infty} A^n u_N^*(\cdot)$ and $v^*(\cdot) = Av^*(\cdot)$. Hence, $v^*(\cdot)$ has the same functional properties as $u_N^*(\cdot)$ (If those functional properties are perserved under limitation). And the optimal decision rule, ψ, is optimal for $v^*(\cdot)$.

Hence, for stable problem, such as discounted problem, OE is satisfied, $v^*(\cdot) = Av^*(\cdot)$, so the stationary decision rule, $(\psi, \psi, ...)$, is optimal for the infinite horizon problem. And $v^*(\cdot)$ has the same functional properties as $u_N^*(\cdot)$. (Hence, when solving finite period counterpart, the terminal value function can be any functional form and has any functional properties, as long as perservation holds, all those form and properties will transfer to the optimal value function of the original infinite periods model.)

2.2 The expected total discounted reward criterion model

In general, for infinite-horizon discounted total reward problem, the major result is in theorem 43. In summary, we can first write the OE and then find the stationary policy that satisfies OE. Similarly, we can use the techniques at the end of this notes to find the structure of policy.

2.2.1 Main result

For the expected total discounted reward criterion, the MDP criteria fo a fixed policy is defined as

$$v_\lambda^\pi(s) = \lim_{N \to \infty} \left\{ E_s^\pi \left[\sum_{t=1}^{N} \lambda^{t-1} r(s_t, a_t) \right] \right\}, \text{ for } s \in S, \, 0 \leq \lambda < 1, \text{ and } \pi \in \Pi^{HR}$$

Two ways of interpretation of discounting:

- Time value, so the reward is discounted;
- The horizon length is random (system earning is undiscounted reward, but the total horizon lenght it alive is random.)

In thie section, the following assumption are assumed:

Conjecture 38 *(Assumption 6.0.1)1. Stationary rewards and transition probabilities: reward $r(s, a)$ and transition probability $P(j|s, a)$ do not vary from decision epoch to decision epoch;*

2. Bounded rewards: $|r(s, a)| \leq M < \infty$ for all $a \in A_s$ and $s \in S$; [4]

3. Discounting: future rewards are discounted according to a discounted factor λ, $0 \leq \lambda < 1$.

4. Discrete state spaces: S is finite or countable.[5]

Theorem 39 *(Proposition 6.2.1) For expected total discounted reward criterion model, under the assumption 6.0.1 and $0 \leq \lambda \leq 1$, there exists a Markov deterministic policy, MD, with the same optimal value function as MR policy.* [6]

[4] Can be generalized under some circulstance

[5] Can be generalized under some circulstance

[6] So, we only need to consider the MD policy within the domain of HR policy

So, we can only consider MD policy within the domain of HR policy because of theorem 5.5.3 and proposition 6.2.1 under assumption 6.0.1:

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

Notation 40 *The notation of this section is summarized as below:*

Parameter	*Explaination*
S	*State Space*
π	*A policy:* $\pi = (d_1, d_2, d_3..., d_N, ...) \in \Pi^{MD}$
$\pi^{'}$	*A reduced policy:* $\pi^{'} = (d_2, d_3..., d_N, ...) \in \Pi^{MD}$
d^∞	*A stationary policy:* $\pi = (d, d, ...) = d^\infty$
$v_\lambda^\pi(s)$	*Expected total discounted reward given the system starts in state* s *and policy* π *is used,* $v_\lambda^\pi(s) = r(s, d_1(s)) + \lambda \sum_{j \in S} p_t(j\|s, d_1(s)) v_\lambda^{\pi^{'}}(j)$
v_λ^π	*a* $\|S\|-$*vector, with the* s^{th} *componenet* $v_\lambda^\pi(s)$
r_{d_t}	*a* $\|S\|-$*vector, with the* s^{th} *componenet* $r_{d_t}(s) = r(s, d_t(s))$
P_{d_t}	*a* $\|S\| \times \|S\|$ *matrix, with the* $(s,j)^{th}$ *entry* $p(j\|s, d_t(s))$

Hence, under the vector notation, we have

$$v_\lambda^\pi = r_{d_t} + \lambda P_{d_t} v_\lambda^{\pi^{'}}(s)$$

Definition 41 *A policy* π *is called* ***stationary policy*** *if* $\pi = \pi^{'}$. *(equilvalently,* $\pi = (d, d, ...) = d^\infty$*)*

Theorem 42 *(Evaluation of a stationary policy)(Theorem 2 of Dr. Susan Xu's handout)* $v_\lambda^{d^\infty}$ *is the unique bounded solution of*

$$v = r_d + \lambda P_d v, \text{ or equivalently } v = (I - \lambda P_d)^{-1} r_d$$

[7]

(Hence, the policy evaluation of a stationary policy is equivalent to solving a system of linear equations.)

Theorem 43 *(Complete regularity of discounted MDP) (Optimality equations and optimal policy) (Theorem 4 of Dr. Susan Xu's handout)*

1. The optimal value function v_λ^* *is the unique bounded solution of the optimality equations:*

$$v(s) = \max_{a \in A_s} \left\{ r(s,a) + \lambda \sum\nolimits_{j \in S} p(j|s,a) v(j) \right\}$$

2. If for each $s \in S$, $d(s)$ *is an action* $a \in A_s$ *that achieves the above max, then the deterministic stationary policy* $\pi = d^\infty$ *is optimal. That is* $v_\lambda^*(s) = v_\lambda^{d^\infty}(s)$[8]

Sample Problem:

- Selling an asset;
- Interactive process quality (Marcellus and Dada 1991: management science 37, 1364-1376);

[7] The optimal value function is given by this equation, but we can not get the optimal value function from this equation directly becasue we do not know the optimal policy and hence we do not know P_d and r_d.

[8] This theorem means: 1. Optimal value function satisfy the optimality equations; 2. Optimal value function is the unique solution of optimality equation; 3. the deterministic stationary policy identified by the optimality equation is optimal (not all optimal policy is stationary, but there exist one optimal policy which is stationary);

2.2.2 Contraction Mapping and Discounted MDP

The result of Theorem 43 rest on a very well-known theorem: Banach Fixed-point Theorem.

Let V be the set of bounded and real valued functions on set S.

Definition 44 *For every function $v \in V$, the **supremum norm** or **sup norm** of v is defined by*

$$\|v\| = \sup_{s \in S} |v(s)|$$

Definition 45 *A mapping $T : V \to V$ is said to be a **contraction mapping** if there exists scalar $0 \leq \lambda < 1$ such that*

$$\|Tu - Tv\| \leq \lambda \|u - v\|, \text{ for all } u, v \in V$$

(roughly speaking, a mapping is a contraction mapping if it can decrease or shrink the maximal distance between its elements)

Definition 46 *An $n-$stage contraction mapping is defined as $T : V \to V$ such that there is scalar $0 \leq \lambda < 1$ and*

$$\|T^n u - T^n v\| \leq \lambda^n \|u - v\|, \text{ for all } u, v \in V$$

For any stationary policy $d^\infty = (d, d, ...)$ and $v \in V$, define two operators, L_d for expected discounted return under a special stationary policy and L for expected discounted return under optimal stationary policy, by:

$$L_d v = r_d + \lambda P_d v$$

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

Theorem 47 *(Proposition 2 of Dr. Susan Xu's handout) Both L_d and L are contraction mapping on V (V is the set of bounded functions) for $0 \leq \lambda < 1$.*

Theorem 48 *(Banach Fixed Point Theorem) (Theorem 6 of Dr. Susan Xu's handout) Suppose V is complete, normed linear space (called banach space) and T is a contraction mapping*[9]

a). There exists a unique function $v^* \in V$ (called the fixed point of T) such that $Tv^* = v^*$;

b). For any $v^0 \in V$, $\lim_{n \to \infty} \left\|T^n v^0 - v^*\right\| = 0$. In other words, $T^n v^0$ converges to v^* as $n \to \infty$.

Theorem 49 *(Theorem 5 of Dr. Susan Xu's handout) Let $v^0 \in V$, $\varepsilon > 0$, and $v^{n+1} = Lv^n$ for $n \geq 1$. Then*

a). v^n converges in norm to v_λ^*,

b). There exists a finite integer N such that $\left\|v^{n+1} - v^n\right\| < \frac{\varepsilon(1-\lambda)}{2\lambda}$, for $n \geq N$

c). $\left\|v^{n+1} - v_\lambda^*\right\| < \frac{\varepsilon}{2}$, for $n \geq N$

d). Let $d_\varepsilon^\infty = (d_\varepsilon, d_\varepsilon, ...)$ be the policy determined by the value iteration algorithm. Then d_ε^∞ is an $\varepsilon-$optimal policy: $\left\|v_\lambda^{d_\varepsilon^\infty} - v_\lambda^*\right\| < \varepsilon$

Some Methods to determine whether a mapping is contraction mapping:

Let's $\psi : R^n \to R^n$ be a mapping, which we want to show is contraction mapping. Define the matrix of derivatives of best response function as

$$A = \begin{bmatrix} 0 & \partial\psi_1/\partial x_2 & ... & \partial\psi_1/\partial x_n \\ \partial\psi_2/\partial x_1 & 0 & ... & \partial\psi_2/\partial x_n \\ ... & ... & ... & ... \\ \partial\psi_n/\partial x_1 & \partial\psi_n/\partial x_2 & ... & 0 \end{bmatrix}$$

[9] V is banach space, so it is normed linear space. However, T is not necessarily an linear mapping. E.g. L_d is linear mapping, but L is non-linear mapping.

and let $\rho(A) = \{\max|\lambda| : Ax = \lambda x, x \neq 0\}$, the largest absolute eigenvalues, states the spectral radius of matrix A. Then, from Horn and Johnson 1996's Matrix Analysis and Cachon and Netessine in chapter 2 of Handbook of quantitative supply Chain Analysis, edited by Simchi-Levi, Wu, and Shen:

Theorem 50 *The mapping $\psi(x) : R^n \to R^n$ is contraction iff $\rho(A) < 1$.*

Lemma 51 *Let A be a matirx, $\rho(A)$ is its spectral radius and $\|\|$ is a consistent matrix norm. Then*

1. *for each $k \in N$: $\rho(A) \leq \left\|A^k\right\|^{1/k}, \forall k \in N$.*
2. *$\lim_{k\to\infty} A^k = 0$ iff $\rho(A) < 1$.*
3. *$\rho(A) = \lim_{k\to\infty} \left\|A^k\right\|^{1/k}$.*

Hence, the most convience way to shown $\rho(A) < 1$ is by using the above lemma: $\rho(A) \leq \|A\|$ by letting $k = 1$ and consistent norm as the maximum column-sum and maximum row-sum norms[10]. Hence, to verify the contraction mapping, it is sufficient to verify that no column sum or no row sum of matirx A exceeds one:

$$\sum_{i=1}^{n} \left|\frac{\partial \psi_k}{\partial x_i}\right| < 1, \text{ or } \sum_{i=1}^{n} \left|\frac{\partial \psi_i}{\partial x_k}\right| < 1, \text{ for } \forall k$$

2.2.3 Computational Methods for Discounted MDP

Value Iteration (Successive approximation)

Algorithm 52 *(Value Iteration Algorithm) Base on contraction mapping:*

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

$$v^{n+1}(s) = \max_{a\in A_{s_t}} \left\{r(s,a) + \lambda \sum\nolimits_{j\in S} p(j|s,a) u^{\pi}_{t+1}(j)\right\} \quad (i.e.\ v^{n+1} = Lv^n)$$

3. *If*

$$\left\|v^{n+1} - v^n\right\| < \frac{\varepsilon(1-\lambda)}{2\lambda}$$

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

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

$$d_\varepsilon(s) \in \arg\max_{a\in A_{s_t}} \left\{r(s,a) + \lambda \sum\nolimits_{j\in S} p(j|s,a) u^{\pi}_{t+1}(j)\right\}$$

and stop.

By using value iteration algorithm, we can get an $\varepsilon-$optimal policy, whose convergence is guaranteed by Theorem 49.

Policy Iteration

Algorithm 53 *The Policy Iteration Algorithm (Howard 1960)*

1. *Set $n = 0$ and select an arbitrary decision rule d_0;*
2. *(Policy evaluation: solve the expected discounted reward for a particular stationary policy d^∞_n) Obtain $v^{d^\infty_n}$ by solving the system of linear equations*

$$v^{d^\infty_n}(s) = r(s, d_n(s)) + \lambda \sum\nolimits_{j\in S} p(j|s, d_n(s)) v^{d^\infty_n}(s), \text{ for } j \in S \quad (i.e.\ v = L_d v)$$

[10]This is equal to letting $k \to \infty$ and define the norm as Euclidean norm.

3. (Policy Improvement: find a new policy d_{n+1}^{∞} that improves d_n^{∞}) Obtain a new decision rule d_{n+1} to satisfy

$$d_{n+1}(s) \in \arg\max_{a \in A_{s_t}} \left\{ r(s,a) + \lambda \sum\nolimits_{j \in S} p(j|s,a) v^{d_n^{\infty}}(s) \right\}, \text{ for } j \in S$$

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

4. If $d_{n+1}(s) = d_n(s)$ for each $s \in S$, stop and $d^* = d_n$. Otherwise increment n by 1 and return to step 2.

Theorem 54 *1. Let v^n and v^{n+1} be the successive values generated by the policy iteration algorithm. Then $v^{n+1} \geq v^n$.*

2. If $v^{n+1} = v^n$, then $v^n = v^{n+1} = v_\lambda^*$;

3. If state space S and action space A are finite, then the algorithm terminates in a finite number of iterations.[11]

Also, there are some other modified policy iteration algorithm to take the advantage of value iteration and policy iteration, such as "A Modified policy iteration algorithm of order m" in Dr. Xu's lecture 14.

Linear Programming

This method bases on the monotonicity property of contraction mapping, which states if $v \geq u$, then $Lv \geq Lu$.[12]

Suppose $v \in V$ such that $v \geq Lv$. Then $v \geq Lv \geq L^2 v \geq ... \geq v_\lambda^*$ as $n \to \infty$. In other word, if $v \in V$ such that $v \geq Lv$, then v is an upper bound of v_λ^*. [13]

Also, $v_\lambda^* \in V$ and $v_\lambda^* \geq Lv_\lambda^*$, so it is the smallest solution for this monotonicity of contraction mapping. So, this serves as the basis for linear programming formulation.

Algorithm 55 *(Primal Linear Programming)*

$$\left\{ \begin{array}{c} \min \sum_{s \in S} \alpha_s v(s) \\ s.t.\ v(s) \geq r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v(j),\ a \in A_s \text{ and } s \in S \\ \sum_{s \in S} \alpha_s = 1 \text{ and } \alpha_s > 0 \end{array} \right\}$$

α_s is understood as the probability that the MDP start in state $s \in S$.

(Dual Linear Programming)

$$\left\{ \begin{array}{c} \min \sum_{s \in S} \sum_{a \in A_s} r(s,a) x(s,a) \\ s.t.\ \sum_{a \in A_s} x(j,a) - \lambda \sum_{s \in S} \sum_{a \in A_s} p(j|s,a) x(s,a) = \alpha_j,\ j \in S \\ \sum_{s \in S} \alpha_s = 1 \text{ and } x(s,a) > 0,\ a \in A_s \text{ and } s \in S \end{array} \right\}$$

Theorem 56 *(Primal LP)An optimal solution $\{v^*(s), s \in S\}$ to primal LP problem satisfies the optimality equations, hence $v^* = v_\lambda^*$.*

Theorem 57 *(Dual LP) 1. Any feasible dual solution x defines a stationary randomized policy $d_x^{\infty} = \{d_x, d_x, ...\}$ where*

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

2. If x^* is an optimal solution to the dual LP, then $d_{x^*}^{\infty}$ is an optimal policy;

3. There exists a bounded optimal basic feasible solution x^* to the dual LP and $d_{x^*}^{\infty}$ defined by x^* is a stationary deterministic optimal policy. (x^* is basic feasible solution of an LP if it cannot be expressed as a convex combination of any other solutions of the LP)

[11] Hence, the policy iteration has this major advantage over value iteration, which in general converges in an infinite number of iterations.

But in policy evaluation require solve $|S|$ number of LP problem in policy iteration algorithm.

[12] This monotonicity is not hold in general for contraction mapping, but it is hold for the contraction mapping defined by $Lv = \max_{d \in D^{MD}} \{r_d + \lambda P_d v\}$. (Prove similar to Theorem 6.2.2)

[13] The condition, $v \in V$ such that $v \geq Lv$, is an assumption. So, we can first find a very large v so that using contraction mapping operator L will lead to $v \geq Lv$.

2.3 The expected total reward criterion models

We can only consider MD policy within the domain of HR policy because of theorem 5.5.3 and proposition 6.2.1[14].

For the expected total reward criterion, the MDP criteria fo a fixed policy is defined as

$$v^{\pi}(s) = \lim_{N\to\infty} v^{\pi}_{N+1}(s) = \lim_{N\to\infty} \left\{ E^{\pi}_{s}\left[\sum\nolimits_{t=1}^{N} r(s_t, a_t)\right]\right\}, \text{ for } s \in S \text{ and } \pi \in \Pi^{HR} \tag{5.2.1}$$

where $v^{\pi}_{N+1}(s)$ is the extected total reward with N period and terminal reward 0.

Because the limit of equation 5.2.1 may diverge for some policies, we **restrict attention to models in which the limit exists for all polcies**. Henc we need to provide a condition that ensures the limit exists.

let $x^{+} = \max\{x, 0\}$ and $x^{-} = \max\{-x, 0\}$. Define

$$v^{\pi}_{+}(s) = E^{\pi}_{s}\left[\sum\nolimits_{t=1}^{\infty} r^{+}(s_t, a_t)\right]$$

$$v^{\pi}_{-}(s) = E^{\pi}_{s}\left[\sum\nolimits_{t=1}^{\infty} r^{-}(s_t, a_t)\right]$$

Hence, if, for all $s \in S$ and $\pi \in \Pi^{HR}$, either $v^{\pi}_{+}(s)$ or $v^{\pi}_{-}(s)$ is finite, then the limit in equation 5.2.1 exists and satisfies

$$v^{\pi}(s) = v^{\pi}_{+}(s) - v^{\pi}_{-}(s)$$

Conjecture 58 *(Assumption 7.1.1) For all $s \in S$ and $\pi \in \Pi^{HR}$, either $v^{\pi}_{+}(s)$ or $v^{\pi}_{-}(s)$ is finite.*

Definition 59 *For each $s \in S$, there exists an $a \in A_s$, for which $r(s_t, a_t) \geq 0$ and $v^{\pi}_{+}(s)$ is finite for all π. Then it is called **positive bounded models DP**, because all rewards are positive.*

Definition 60 *For each $s \in S$ and $a \in A_s$, $r(s_t, a_t) \leq 0$ and for some π, $v^{\pi}_{-}(s) > -\infty$ for all s. Then it is called **negative bounded models DP**, because all rewards are negative and maximizing negative reward is equivalent to minimizing positive cost.* [15]

Definition 61 *For each $s \in S$, both $v^{\pi}_{+}(s)$ and $v^{\pi}_{-}(s)$ are finite for all π. Then it is called **convergent model DP**.*

Theorem 62 *(Theorem 7.1.3) (Optimality equation for total reward model) Suppose assumption 7.1.1 holds (either $v^{\pi}_{+}(s)$ or $v^{\pi}_{-}(s)$ is finite). Then the optimal value function v^{*} satisfies optimality equation:*

$$v(s) = \sup_{a\in A_{s_t}} \left\{ r(s, a) + \sum\nolimits_{j\in S} p(j|s, a)v(j)\right\}, \text{ for } s \in S$$

(Notes: v^{*} is not the unique solution of OE, while in finite-horizon model and infinite-horizon discounted models v^{*} is the unique solution of OE. In fact, if $v(s)$ is a solution of OE, then $v(s) + c$ is also a solution of OE)

Theorem 63 *(Theorem 7.1.6) If $\pi^{*} \in \Pi^{HR}$ is optimal. Then $v^{\pi^{*}}$ satisfies the optimality equation.*

Definition 64 *A decision rule $d \in D^{MD}$ is **conserving decision rule**[16] if*

$$d(s) = \arg\max_{a\in A_{s_t}} \left\{ r(s, a) + \sum\nolimits_{j\in S} p(j|s, a)v^{*}(j)\right\}, \text{ for } s \in S$$

[14] In Proposition 6.2.1, the result hold for $\lambda = 1$, which is the expected total reward criterion model

[15] The positive model and negative model are not equivalent with signs reversed. In positive models, we want to always continuou to get positve reward to maximize total reward, while in negative models, we always want to find a policy to terminate to stop costing more in the furture.

[16] Conserving decision rule is stationary policy.

(In other words, d is conserving if it satisfies $r_d + P_d v^* = v^*$)[17]

(By using conserving decision rule d for another period, we conserve the optimal reward v^*)

(In the discounted case, a stationary policy d^∞ is optimal if and only if d is conserving)

(For expected total reward case, that d is conserving is a necessary condition, but not a sufficient condition.)

Theorem 65 *(Theorem 7.1.7) (Optimal decision rule for total reward model)*

a). (Necessary) If d^∞ is optimal, then d is conserving. In other word, if d^∞ is optimal, then

$$d(s) = \arg\max_{a \in A_{s_t}} \left\{ r(s,a) + \sum\nolimits_{j \in S} p(j|s,a) v^*(j) \right\}, \text{ for } s \in S \tag{7.1.9}$$

b). (Sufficient) if d is conserving and

$$\lim_{N \to \infty} \left\{ \sup_{s \in S} E_s^d \left[v^*(s_{N+1}) \right] \right\} \leq 0 \tag{7.1.10}$$

then d^∞ is optimal.

(Equation 7.1.10 means d drives the system to the states in which there is little opportunity for positve future reward. $v^*(s_{N+1})$ means after N steps, the system arrive at state s_{N+1} and have the optimal value function of $v^*(s_{N+1})$. $E_s^d [v^*(s_{N+1})]$ means if the system start at state s and using decision rule d, then its expected optimal value function after N steps is $E_s^d [v^*(s_{N+1})]$)

(In discounted model, all decision rules are equalizing: $\lim_{N \to \infty} \left\{ \sum_{t=N}^{\infty} \lambda^{t-1} E_s^d r(s_t, d) \right\} = 0$)

(In negative model, $v^* \leq 0$, so all the decision rules are equalizing; consequently, a conserving stationary policy is optimal for a negative model)

(In positive model, if d is conserving and equalizing, then d^∞ is optimal).

Definition 66 *A decision ruleis said to be **equalizing decision rule**, if it satisfies equation 7.1.10.*

2.3.1 Positive Bounded DP models (max reward)

In this section, we focus our attention on positve bounded DP models, which must satisfies the following assumption:

Conjecture 67 *1. (Assumption 7.1.1) For all $s \in S$ and $\pi \in \Pi^{HR}$, either $v_+^\pi(s)$ or $v_-^\pi(s)$ is finite.*

2. (Assumption 7.2.1) $v_+^\pi(s)$ is finite for all $\pi \in \Pi^{HR}$ and $s \in S$;

3. (Assumption 7.2.2) For each $s \in S$, there exists at least one $a \in A_s$ with $r(s,a) \geq 0$

Let V^+ be the set of nonnegative bounded functions.

Theorem 68 *(Theorem 7.2.2) Let $v \in V^+$ be any function satisfying $v > \max_{d \in D} \{r_d + P_d v\} = Lv$, then $v \geq v^*$*

Because v^* satisfies OE and $v^* \in V^+$, we have the following theorem

Theorem 69 *(Theorem 7.2.3)*

a). $v^* \in V^+$ is the smallest non-negative solution of the Optimality Equation.

b). $v^{d^\infty} \in V^+$ is the smallest nonnegative solution of $v = r_d + P_d v = Lv$

Theorem 70 *(Theorem 7.2.5)(Optimal policy)*[18]

a). (Thm 7.1.7) If d is conserving and equalizing, then the stationary policy d^∞ is optimal.

b). (Thm 7.1.6) A policy π^* is optimal if and only if its value function satisfies OE: $v^{\pi^*} = \max_{a \in A_{s_t}} \left\{ r_{d^*} + P_{d^*} v^{\pi^*} \right\}$[19]

[17] $v^*(s)$ is defined as $v^*(s) = \max_{a \in A_{s_t}} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a) v^*(j) \right\}$, so if d is conserving, then d solves this problem and, hence, $v^*(s) = r(s,d) + \sum_{j \in S} p(j|s,d) v^*(j)$

[18] There is no guarantee that there exists a stationary optimal policy, but if there is stationary policy which is conserving and equalizing, then it is optimal.

[19] This means there may be some policy, which is not stationary (hence, not conserving and equilizing), is optimal iff it satisfies OE.

Theorem 71 *(Theorem 7.2.12) (Primal theorem for Value iteration algorithm) In a positve bounded model, let $v^0 = 0$ and set $v^{n+1} = Lv^n$. Then v^n converges pointwise and monotonically to v^*.*

Computational methods

The computational methods are similar to Discounted model with some modification, so they are ignored here. Please refer to Dr. Xu's Notes. (Lecture 16)

2.3.2 Negative Bounded DP models (min cost)

In this section, we focus our attention on negative bounded DP models, which must satisfies the following assumption:

Conjecture 72 *1. (Assumption 7.1.1) For all $s \in S$ and $\pi \in \Pi^{HR}$, either $v_+^\pi(s)$ or $v_-^\pi(s)$ is finite.*
2. $v_+^\pi(s) = 0$ for all $s \in S$ and all $\pi \in \Pi^{HR}$ and $s \in S$;
3. there exists a $\pi \in \Pi^{HR}$ with $v^\pi(s) > -\infty$ for all $s \in S$,

Similarly, as discussed in the general case for expected total reward model, the negative models

- The optimal value function in a negative model satisfies the optimality equation
- the solution of OE for the negative model is not unique
- Stationary policy derived from OE is optimal.

Similar to Positve bounded DP models, Let V^- be the set of nonpositive functions (maybe unbounded). we have the following two theorem (Thm 7.3.2 and Thm 7.3.3)

Theorem 73 *(Theorem 7.3.2) Let $v \in V^-$ be any function satisfying $v \leq \max_{d \in D}\{r_d + P_d v\} = Lv$, then $v \leq v^*$*

Because v^* satisfies OE and $v^* \in V^-$, we have the following theorem

Theorem 74 *(Theorem 7.3.3)*
a). $v^* \in V^-$ is the maximal nonpositive solution of the Optimality Equation.
b). $v^{d^\infty} \in V^-$ is the maximal nonpositive solution of $v = r_d + P_d v = Lv$

Computational methods

The computational methods are similar to Discounted model with some modification, so they are ignored here. Please refer to Dr. Xu's Notes. (Lecture 17)

Result	Discounted Models	Positive Bounded Models	Negative Models
Optimality Equations	v^* is the unique solution of OE in V	v^* is minimal solution of OE in V^+	v^* is maximal solution of OE in V^-
Optimal policy	d is conserving	d is conserving and equalizing	d is conserving
Value Iteraton converges	If initial value function $v^0 \in V$	If initial value function $0 \leq v^0 \leq v^*$	If initial value function $0 \geq v^0 \geq v^*$ and either A or S finite
Policy Iteraton convergence	New policy always improves old policy. Find opt policy if termination occurs	New policy may not be an improvement of old policy. Find opt policy if termination occurs	New policy improves the old policy. But may terminate at a suboptimal policy
Solution by LP	Yes	Yes	No

Chapter 3

Techniques

3.1 How to get optimal policy structure propeties

3.1.1 Conjecture and Induction

If optimal policy / best expected reward has obvious specific structure (Such as Sequential allocation problem), then we can use conjecture and induction:

1. We can use backward DP to calculate the last few periods' optimal policy / best expected reward for those periods
2. then make conjecture of the structure of the optimal policy / best expected reward
3. use induction to proof the optimality of the conjecture.

3.1.2 Use finite-horizon model to get the structure of infinite-horizon model structure.

This method still holds for infinite-horizon model, in which we can use finite-horizon counterpart, and make conjecture on the finite-horizon counterpart. Then we can proof our conjecture for infinite-horizon model.

If the discounted MDP is stable, then $v_\lambda^n \to v_\lambda^*$ as $n \to \infty$. So the v_λ^* will inherit the same property of v_λ^n, and there exists an optimal stationary policy with the same special structure. This idea has been used extensively in inventory theory, equipment maintenance models, and queueing control. For example, if v_λ^n is increasing as $n \to \infty$, then v_λ^* is increasing.

3.1.3 Monotone optimal policy (base on supermodular function)

In MDP, the optimality equations are defined by

$$u_t^*(s) = \max_{a \in A_{s_t}} \underbrace{\left\{ r_t(s,a) + \sum\nolimits_{j \in S} p(j|s,a) u_{t+1}^\pi(j) \right\}}_{=w_t(s,a)}$$

And the optimal decision rule is define as

$$d_t^*(s) = \arg\max_{a \in A_{s_t}} \underbrace{\left\{ r_t(s,a) + \sum\nolimits_{j \in S} p(j|s,a) u_{t+1}^\pi(j) \right\}}_{=w_t(s,a)}$$

If $w_t(s,a)$ is supermodular function, then $d_t^*(s)$ is monotone increasing.

For proof $w_t(s,a)$ of supermodular, please refer to "Study notes for basic mathematics" section 1.2.1: supermodular function;

For proof $d_t^*(s)$ of the is monotone increasing and monotone increasing in other cases, please refer to "Study notes for basic mathematics" section 1.2.2: properties under optimization;

3.1.4 Myopic Policies (Base on Chapter 6 of Porteus)

Notation 75 *$s_{t+1} = z(a, X_t)$: s_{t+1} is next state given current state is X_t and action a is given;*

$A_* = \bigcup_{s \in S} A(s)$: The union of all feasible decision for some set of state S (This set of states can be subset of all states);

$S(a) = \{s \in S | a \in A(s)\}$: the set of states for which action a is admissible;

$r_S(s)$: separatable immediate reward / cost with respect to states;

$r_A(a)$: separatable immediate reward / cost with respect to action;

$\psi(a) = r_A(a) + \alpha E[r_S(z(a, X))]$: Truncated reward function;

a^*: a maximizer of $\psi(\cdot)$;

Theorem 76 *(Veinott, 1965; Sobel, 1981) If the following two conditions hold:*

1. Immediate reward / cost is additively separatable: $r(s, a) = r_S(s) + r_A(a)$;

2. Transition depends on current period state and action: $P(j|s, a) = P(j|s, a, h_{t-1})$;

*3. $z(a^*, X) \in S(a^*)$ for every realization of X; (This means if system starts in $S(a^*)$ and the admissible decision a^* is made, then the next state vistied will be in $S(a^*)$. Hence, $S(a^*)$ is called the set of **consistent states**.)*

4. $v_T(s) = r_S(s)$ for all s;

Then There exists an optimal policy that selects action a^* whenever a consistent state is visited.

In particular, action a^* is optimal at state s for all $s \in S(a^*)$. Futhermore, once a consistent state is visited, then only consistent states will be visted thereafter.