ISyE 6664: Stochastic Optimization

Rui Gong

November 16, 2024

Acknowledgements

These notes are based on the ISyE 6664 lectures given by Professor *Hayriye Ayhan* in Fall 2024 at Georgia Institute of Technology and the course notes of *Weiwei (William) Kong*.

Contents

The notes shall mentioned most of the chapters 4, 6, 8, probably 9 and 11 of *Markov Decision Process: Discrete Stochastic Dynamic Programming* by *Martin L. Puterman*.

1 Markov Decision Processes (MDPs)

We study *sequential decision making process:* a Markov Process, where the set of available actions, the rewards and transition probabilities depend on the current state and the action taken that state. It has the following ingredients:

- Decision epoch
- State space
- Actions space
- Rewards
- Transition probabilities

Example 1.1.

- Inventory Model: A warehouse manager observes his on hand inventory at the end of each month. Based on how many units he has, he decides to purchase new items or not to order anything at all.
  - the demand is random.
  - purchase cost
  - holding cost
  - reveneu from sales
  - pending cost for shortage
- Machine Replacement: A machine deteriorates over time. The decision maker checks the condition of the machine at the end of everyday and decides to keep or replace the machine.
  - state dependent income
  - state dependent cost
  - replacement cost
- Admission Control: Consider a system with k servers, i.e. the capacity is k, with service times following $\exp(\mu)$. One type of calls enters at a Poisson rate with parameter λ_1 and reward r_1 and another type of calls enters at a Poisson rate with parameter λ_2 and reward r_2 with $r_1 > r_2$.
  You should always accept the higher reward customers, and only reject the other set when as a number of servers greater M has filled up, where M is to be determined.

1.1 Modeling MDPs

Definition 1.1: Ingredients of a MDP

- Decision Epochs: T: *set of decision epochs,* $T = \{1, \ldots, N\}$ *where* $N - 1$ *is the time of last decision, and* N *is the time with a determined reward.* $T = \{1, 2, \ldots\}$ *if there are infinitely many epochs.*
- State Space (of the Markov Chain): S
- Action Space: A_s: *the set of possible actions in state* $s \in S$*, and the total action space is*

$$A = \cup_{s \in S} A_s.$$

We can choose actions deterministically or randomly. Let us define

$$P(A_s) \text{ : collection of probability distributions on subsets of } A_s$$

and $q(\cdot) \in P(A_s)$*. Basically, when you are in state* s*, you choose a particular action* a *with probability* $q(a)$*.*

- Rewards: $r_t(s, a)$ *is the reward received when action* a *is chosen in state* s *at time* t*.*
  $r_t(s, a, j)$ *is the reward earned when action* a *is chosen in state* s *at epoch* t *and the state is* j *at epoch* $t + 1$*, then*

$$r_t(s, a) = \sum_{j \in S} P_t(j \mid s, a) r_t(s, a, j).$$

For T *being finite, the terminal reward* $r_N(s)$ *is the reward earned at decision epoch* N *if the state is* s *at time* N*.*

- Transition Probability:

$$p_t(j \mid s, a) \text{ : probability of being in state } j \text{ at decision epoch } t + 1$$
$$\text{given that } a \text{ is chosen in state } s \text{ at decision epoch } t.$$

The five-tuple

$$\{T, S, A, p(\cdot \mid \cdot, \cdot), r(\cdot, \cdot)\}$$

forms a Markov decision process (MDP).

Definition 1.2: Decision Rules and Policy

A decision rule *prescribes a procedure for action selection at a specified decision epoch.*
Markovian Deterministic Decision Rule:

$$d_t : S \mapsto A \text{ where } d_t(s) \in A_s,$$

where d_t represents the decision rules at decision epoch t.
Markovian Randomized Decision Rule

$$d_t : S \mapsto P(A) \text{ where } q_{d_t(s)}(\cdot) \in P\left(A_s\right),$$

where the decision rule in state s_t tells you a probability of possible actions.
History Dependent Deterministic Decision Rule H_t : *the set of all histories at decision epoch t, where $h_t \in H_t$ is a specific instance of history such that*

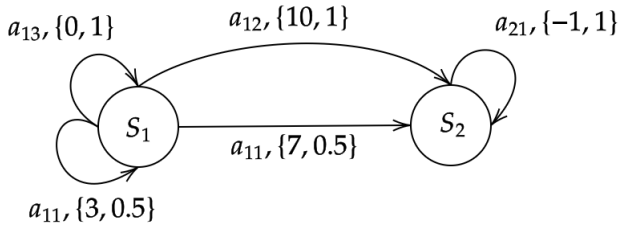
$$h_t = (s_1, a_1, s_2, a_2, \ldots, s_{t-1}, a_{t-1}, s_t) = (h_{t-1}, a_{t-1}, s_t),$$

then the rule can be represented as

$$d_t : H_t \to A, \text{ where } d_t(h_t) \in A_{s_t}.$$

History Dependent Randomized Decision Rule: $d_t : H_t \mapsto P(A)$.
Policy: *A policy π is a sequence of decision of rules. For finite epochs $T = \{1, \ldots, N\}$, $\pi = (d_1, d_2, \ldots, d_N)$; $T = \{1, 2, \ldots\}$, $pi = (d_1, d_2 \ldots)$. If $d_t = d$ for all $t \in T$, then $\pi = (d, d, \ldots) := d^\infty$ is called a stationary policy.*

Example 1.2. Consider the following plot of an MDP:



where for action a_{11}, it goes to S_2 with reward 7 and probability 0.5 OR go to S_1 with reward 3 with probability 0.5; similar interpretations for a_{12}, a_{13}, a_{21}. Specifically, $S = \{S_1, S_2\}, A_{S_1} = \{a_{11}, a_{12}, a_{13}\}, A_{S_2} = \{a_{21}\}, T = \{1, 2, 3\}$. For example, $P(S_1 \mid S_1, a_{11}) = 0.5, P(S_2 \mid S_1, a_{11}) = 0.5, r_{S_1, a_{11}, S_1)} = 3, r(S_1, a_{12}) = 10$.

Example 1.3 (Continued). A Markovian deterministic decision rule: $d_1(S_1) = a_{11}, d_1(S_2) = a_{21}, d_2(S_1) = a_{12}, d_2(S_2) = a_{21}$.
A history dependent deterministic decision rule: $d_1(S_1) = a_{11}, d_1(S_2) = a_{21}, d_2((S_1, a_{11}, S_1)) = a_{13}, d_2((S_1, a_{11}, S_2)) = a_{21}, d_2((S_1, a_{21}, S_2)) = a_{21}$
A Markovian randomized decision rule: $P(d_1(S_1) = a_{11}) = 0.6, P(d_1(S_1) = a_{13}) = 0.4, P(d_1(S_2) = a_{21}) = 1, P(d_2(S_1) = a_{11}) = 0.4, P(d_2(S_1) = a_{12}) = 0.6, P(d_2(S_2) = a_{21}) = 1$.

Example 1.4. An inventory manager checks his on-hand inventory at the end of each month. Depending on how many units he has on hand, he decides whether or not order new units from a supplier. Assume that newly purchased units arrive before the start of next month. Demand arrives during the month but orders are filled at the end of the month. Assume no backlogs are allowed, i.e., orders are lost if not enough inventories, and the warehouse has a capacity of M units. Let D_t be the monthly demand during month t and

$$P\left(D_t = j\right) = p_j \text{ for } j = 0, 1, 2, \ldots$$

Assume that if j units are purchased, the purchase cost is $C(j)$. The holding cost for u units is $h(u)$ and the revenue obtained from j units is $\rho(j)$. Finally, let $O(u)$ denote the wholesale purchase cost when u units are purchased and

$$O(u) = \begin{cases} k + C(u), & \text{if } u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The inventory manager would like to maximize his expected profit for the next N months. Let $g_N(s)$ be the terminal reward if there are s units left at time N.
Modeling this as a MDP, we have

$$\begin{aligned} T &= \{1, 2, \ldots, N\} \\ S &= \{0, 1, \ldots, M\} \\ A_s &= \{0, 1, \ldots, M - s\},\ \forall s \in S \\ p_t(j \mid s, a) &= \begin{cases} 0, & \text{if } j > s + a \\ p_{s+a-j}, & \text{if } 0 < j \leq s + a \\ \sum_{j=s+a}^{\infty} p_j & \text{if } j = 0 \end{cases} \\ r_t(s, a) &= -O(a) - h(s + a) + \sum_{j=0}^{s+a} \rho(j) p_j + \sum_{k=s+a+1}^{\infty} \rho(s + 1) p_j, \text{ for } t = 1, \ldots, N - 1 \\ r_N(s) &= g_N(s). \end{aligned}$$

Example 1.5. The condition of a machine used in a manufacturing process deteriorates over time. The condition of the machine is checked at predetermined discrete decision epochs. Let $S = \{0, 1, \ldots\}$ denote the state of the machine at each decision epoch. The higher the value of s is, the worse the condition of the machine. At each decision epoch, you can choose either to replace or keep as it is. Suppose replacements happen instantaneously. We assume in each period, the machine deteriorates by i states with probability $p(i)$. There is a fixed income of R units per period, a state dependent operating cost $h(s)$ where s is the state at the beginning of the period, and a replacement cost of K units. Suppose the objective is to maximize the long-run average profit. Modeling this as a MDP, we have

$$\begin{aligned} T &= \{1, 2, \ldots\} \\ S &= \{0, 1, \ldots\} \\ A_s &= \{0, 1\}, \text{ where 1 indicates a replacement action} \\ p_t(j \mid s, 0) &= \begin{cases} 0, & \text{if } j < s \\ p(j - s), & \text{if } j \geq s \end{cases} \\ p_t(j \mid s, 1) &= p(j) \\ r_t(s, 0) &= R - h(s) \\ r_t(s, 1) &= R - K - h(0) \end{aligned}$$

1.2 Finite Horizon MDPs

Throughout this subsection, let $T = \{1, \ldots, N\}$ and $\pi = (d_1, d_2, \ldots, d_{N-1})$. Let $V_N^{\pi}(s)$ be the total expected reward for an N period problem under policy π when the system state at the first decision epoch is s.

Suppose π is a randomized history dependent policy and

$$\begin{aligned} X_t &: \text{ state at time } t \\ Y_t &: \text{ action chosen at time } t, \end{aligned}$$

where $\{X_t\}$ is the Markov Chain representing state under policy π. Then,

$$V_N^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{N-1} r_t\left(X_t, Y_t\right) + r_N\left(X_N\right) \mid X_1 = s\right]$$

If, $\pi = (d_1, \ldots, d_{N-1})$ is a deterministic Markovian policy, then

$$V_N^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{N-1} r_t \left(X_t, d_t \left(X_t \right) \right) + r_N \left(X_N \right) \mid X_1 = s \right]$$

If instead, $\pi = (d_1, \ldots, d_{N-1})$ is a history dependent deterministic policy, then

$$V_N^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{N-1} r_t \left(X_t, d_t \left(h_t \right) \right) + r_N \left(X_N \right) \mid X_1 = s \right] \text{ with } h_t = \left(h_{t-1}, a_{t-1}, X_t \right),$$

where $|r(s,a) < M$ for all $a \in A_s, s \in S$.
If there exists $0 < \lambda < 1$ as a discount factor, then

$$V_N^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=1}^{N-1} \lambda^{t-1} r_t \left(X_t, d_t \left(h_t \right) \right) + \lambda^{N-1} r_N \left(X_N \right) \mid X_1 = s \right] \text{ with } h_t = \left(h_{t-1}, a_{t-1}, X_t \right).$$

Define Π : the set of all possible history dependent randomized policies. Our objective is to find π^* (among all history dependent randomized policies) such that

$$V_N^{\pi^*}(s) \geq V_N^\pi(s), \text{ for all } \pi \in \Pi$$

and we would also like to compute

$$V_N^*(s) = \sup_{\pi \in \Pi} V_N^\pi(s),$$

where $V_N^*(s) = \max_{\pi \in \Pi} V_N^\pi(s)$ if the supremum is attained.

Now for a policy $\pi = (d_1, d_2, \ldots, d_{N-1})$, let us define the total expected reward from t to $N-1$, given h_t, as

$$u_t^\pi \left(h_t \right) = \mathbb{E} \left[\sum_{n=t}^{N-1} r_n \left(X_n, d_n \left(h_n \right) \right) + r_n \left(X_N \right) \mid H_t = h_t \right]$$

for $t = 1, \ldots, N-1$ and $u_N \left(h_N \right) = r_N \left(s_N \right)$ for all $h_N = \left(h_{N-1}, a_{N-1}, s_N \right)$. If π is Markovian deterministic, then

$$u_t^\pi \left(s_t \right) = \mathbb{E} \left[\sum_{n=t}^{N-1} r_n \left(X_n, d_n \left(X_n \right) \right) + r_N \left(X_N \right) \mid X_t = s_t \right]$$

If $h_1 = s$, then

$$u_1^\pi(s) = V_N^\pi(s) = \text{ total expected reward}$$

Note that $V_N^\pi(s)$ is not dependent on t. From recursively figuring out $V_N^\pi(s)$ by calculating $u_t^\pi \left(h_t \right)$, we can compute $V_N^\pi(s)$.

1.2.1 Backward Dynamic Programming for Computing the Expected Reward for a Finite Horizon Problem

1. Set $t = N$ and $u_N^\pi \left(h_N \right) = r_N \left(s_N \right)$, the terminal reward, for all $h_N = \left(h_{N-1}, a_{N-1}, s_N \right)$. Go to Step 2.
2. If $t = 1$, stop; otherwise go to Step 3 .
3. Substitute $t-1$ for t and compute $u_t^\pi \left(h_t \right)$ as

$$u_t^\pi \left(h_t \right) = r_t \left(s_t, d_t \left(h_t \right) \right) + \sum_{j \in S} p_t \left(j \mid s_t, d_t \left(h_t \right) \right) u_{t+1}^\pi (\underbrace{h_t, d_t \left(h_t \right), j}_{h_{t+1}})$$

4. Return to Step 2.

For Markovian deterministic π, we have

$$u_t^\pi(h_t) = \underbrace{r_t(s_t, d_t(h_t))}_{\text{immediate reward}} + \underbrace{\sum_{j\in S} p(j \mid s_t, d_t(h_t))\, u_{t+1}^\pi(j)}_{\mathbb{E}^\pi_{h_t}[u_{t+1}]}$$

Theorem 1.3

Suppose that $\pi = (d_1, \ldots, d_{N-1})$ is a history dependent deterministic policy and u_t^π is obtained by the backward dynamic programming. Then for all $t \le N$,

$$u_t^\pi(h_t) = \mathbb{E}_{h_t}\left[\sum_{n=t}^{N-1} r_n(X_n, d_n(h_n)) + r_N(X_N)\right]$$

and $V_N^\pi(s) = u_1^\pi(h_1)$ for $h_1 = s$.

Proof. Let $t = N$, $u_N^\pi(h_N) = r_N(s_N)$ for all $h_N = (h_{N-1}, a_{N-1}, s_N)$. Suppose the result holds for $n = t+1, \ldots, N$ and we will prove that it holds for $n = t$.

$$\begin{aligned}
u_t^\pi(h_t) &= r_t(s_t, d_t(h_t)) + \sum_{j\in S} p(j \mid s_t, d_t(h_t))\, u_{t+1}^\pi(h_t, d_t(h_t), j) \\
&= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}\left[\mathbb{E}_{h_{t+1}}\left[\sum_{n=t+1}^{N-1} r_n(X_n, d_n(h_n)) + r_N(X_N)\right]\right] \\
&= r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}\left[\sum_{n=t+1}^{N-1} r_n(X_n, d_n(h_n)) + r_N(X_N)\right] \\
&= \mathbb{E}_{h_t}\left[\sum_{n=t}^{N-1} r_n(X_n, d_n(h_n)) + r_N(X_N)\right]
\end{aligned}$$

□

Suppose π were a randomized history dependent policy, then

$$u_t^\pi(h_t) = \sum_{a\in A_t} p(d_t(h_t) = a)\left(r_t(s_t, a) + \sum_{j\in S} p(j \mid s_t, a)\, u_{t+1}^\pi(h_t, a, j)\right)$$

1.2.2 Optimality Equations

We have

$$u_t^*(h_t) = \sup_{\pi\in\Pi} u_t(h_t),\ h_1 = s_1$$

where π belongs to the set of history dependent deterministic policies.

Lemma 1.4

Let w be a real valued function on an arbitrary discrete set W and let $q(\cdot)$ be a probability distribution on W. Then $\sup_{u\in W} w(u) \ge \sum_{u\in W} q(u)w(u)$

Proof. Let $w^* = \sup_{u\in W} w(u)$. Then

$$w^* = \sum_{u\in W} q(u)w^* \ge \sum_{u\in W} q(u)w(u)$$

□

That is, there is always a deterministic rule that performs as well/better than all randomized ones.

Optimality Equations for the N Period Problem Define

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

for $t = 1, \ldots, N-1$ and for $u_N(h_N) = r_N(s_N)$ for $h_N = (h_{N-1}, a_{N-1}, s_N)$.
If the supremum is obtained,

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

Recall that

$$u_t^*(h_t) = \sup_{\pi} u_t^{\pi}(h_t) \quad \text{and} \quad u_1^{\pi}(s) = V_N^{\pi}(s)$$

so by computing $u_1^*(s)$ like this, we will compute $V_N^*(s)$. In fact, we will show that, if we compute $u_t(h_t)$ as above, then it is actually $u_t^*(h_t)$ and hence we have $u_1(s_1) = V_N^*(s_1)$.

Theorem 1.5

Suppose that u_t is a solution to the optimality equations for $t = 1, \ldots, N-1$ with $u_N(s_N) = r_N(s_N)$. Then,
(a) $u_t(h_t) = u_t^*(h_t)$ for $t = 1, \ldots, N-1$
(b) $u_1(s_1) = V_N^*(s_1)$

Proof. We will first try to show taht $u_n(h_n) \geq u_n^*(h_n)$ for all $n = 1, \ldots, N$.
For $n = N$, $u_n(h_n) = r_N(s_N) = u_N^{\pi}(h_N)$ for all $\pi \in \Pi$ and $h_N = (h_{N_1}, a_{N-1}, s_N)$. Thus, the result holds for $n = N$. Assume it holds for $t = n+1, \ldots, N$, we will show that it holds for $t = n$ as well. Let $\pi = (d_1, \ldots, d_{N-1})$ be an arbitrary policy.

$$\begin{aligned}
u_n(h_n) &= \sup_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{j \in S} p_j(j \mid s_n, a)\, u_{n+1}(h_n, a, j) \right\} \\
&\geq \sup_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{j \in S} p_j(j \mid s_n, a)\, u_{n+1}^*(h_n, a, j) \right\} \\
&\geq r_n(s_n, d_n(h_n)) + \sum_{j \in S} P(j \mid s_n, d_n(h_n)) u_{n+1}^{\pi}(h_n, d_n(h_n), j) \\
&= u_n^{\pi}(h_n)
\end{aligned}$$

Since π is arbitrary,

$$u_n(h_n) \geq \sup_{\pi \in \Pi} u_n^{\pi}(h_n).$$

We will next show that for each $\epsilon > 0$, there exists π' such that

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

We will construct such a policy $\pi' = (d_1', d_2', \ldots, d_{N-1}')$ by choosing $d_n'(h_n)$ such that

$$r_n(s_n, d_n'(h_n)) + \sum_{j \in S} P_n(j \mid s_n, d_n'(h_n)) u_{n+1}^{\pi'}(h_n, d_n'(h_n), j) + \epsilon \geq u_n(h_n).$$

This π' exists by the definition of $u_n(h_n)$. Note $u_N^{\pi'} = r_N(s_N) = u_N(s_N)$ for $h_N = (h_{N-1}, a_{N-1}, s_N)$. Suppose the result holds for $t = n+1, \ldots, N$, then

$$\begin{aligned}
u_n^{\pi'}(h_n) &= r_n(s_n, d_n'(h_n)) + \sum_{j \in S} P_n(j \mid s_n, d_n'(h_n)) u_{n+1}^{\pi'}(h_n, d_n'(h_n), j) \\
&\geq r_n(s_n, d_n'(h_n)) + \sum_{j \in S} P_n(j \mid s_n, d_n'(h_n)) \left(u_{n+1}(h_n, d_n'(h_n), j) - (N-n-1)\epsilon\right) \\
&\geq r_n(s_n, d_n'(h_n)) + \left(\sum_{j \in S} P_n(j \mid s_n, d_n'(h_n)) u_{n+1}(h_n, d_n'(h_n), j)\right) + \epsilon - (N-n)\epsilon \\
&\geq u_n(h_n) - (N-n)\epsilon
\end{aligned}$$

But then for each n, we have

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

which impllies $u_n(h_n) = u_n^*(h_n)$. □

Now the above theorem shows us a way to iteratively compute $u_n^*(h_n)$ and hence $V_N^*(s_1)$ in the end. Now, the following theorem will show that in fact, we are able to compute an optimal policy based on the iterations.

Theorem 1.6

Suppose that u_t^* for $t = 1, \ldots, N$ are solutions to the optimality equations subject to the boundary condition and the policy $\pi^* = \left(d_1^*, \ldots, d_{N-1}^*\right)$ satisfies

$$\begin{aligned}
&r_t\left(s_t, d_t^*\left(h_t\right)\right) + \sum_{j \in S} p_t\left(j \mid s_t, d_t^*\left(h_t\right)\right) u_{t+1}^*\left(h_t, d_t^*\left(h_t\right), j\right) \\
&= \max_{a \in A_{s_t}} \left\{ r_t\left(s_t, a\right) + \sum_{j \in S} p_t\left(j \mid s_t, a\right) u_{t+1}^*\left(h_t, a, j\right) \right\} \text{ for } t = 1, \ldots, N-1
\end{aligned}$$

(a) $u_t^*\left(h_t\right) = u_t^{\pi^*}\left(h_t\right)$

(b) π^* is an optimal policy and $V_N^{\pi^*}(s) = V_N^*(s)$.

Proof.

(a) Trivially

$$u_N^*\left(s_N\right) = r_N\left(s_N\right) = u_N^{\pi^*}\left(s_n\right)$$

Suppose that this holds for $n = t+1, \ldots, N$. We will show that it also holds for $n = t$. We have

$$\begin{aligned}
u_t^*\left(h_t\right) &= \max_{a \in A_{s_t}} \left\{ r_t\left(s_t, a\right) + \sum_{j \in S} p_t\left(j \mid s_t, a\right) u_{t+1}^*\left(h_t, a, j\right) \right\} \\
&= r_t\left(s_t, d_t^*\left(h_t\right)\right) + \sum_{j \in S} p_t\left(j \mid s_t, a\right) u_{t+1}^{\pi^*}\left(h_t, d_t^*\left(h_t\right), j\right) \\
&= u_t^{\pi^*}\left(h_t\right)
\end{aligned}$$

(b) We have

$$V_N^{\pi^*}(s) = u_1^{\pi^*}(s) = u_1^*(s) = V_N^*(s)$$

Hence, the optimal policy $\pi^* = \left(d_1^*, \ldots, d_{N-1}^*\right)$ is defined as

$$d_t\left(h_t\right) \in \arg\max_{a \in A_{s_t}} \left\{ r_t\left(s_t, a\right) + \sum_{j \in S} p_t\left(j \mid s_t, a\right) u_{t+1}^*\left(h_t, a, j\right) \right\}$$

□

That is, when we do the iteration, if we always pick the action maximizing $u_t(h_t)$, we get an optimal policy. Now we show that we actually only need s_t rather than h_t and there exists a deterministic policy if all $u_t(h_t)$ are attained by a deterministic action.

Theorem 1.7

Let u_t^* for $t = 1, \ldots, N$ be the solution to the optimality equations together with the boundary conditions.

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

(b) If there exists $a^1 \in A_{s_t}$ such that

$$r_t\left(s_t, a^1\right) + \sum_{j \in S} p\left(j \mid s_t, a^1\right) u_{t+1}^*\left(h_t, a^1, j\right)$$

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

for all $t = 1, \ldots, N-1$ then there exists an optimal policy that is Markovian deterministic.

Proof.

(a) We have

$$u_N^*\left(h_N\right) = u_N^*\left(h_{N-1}, a_{N-1}, s_N\right) = r_N\left(s_N\right).$$

Thus, u_N^* depends on h_N only though s_N. The result holds for $n = N$. Let us assume it holds for $n = t+1, \ldots, N$ and we proceed to show that it also holds for $n = t$. We have

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

$$= \sup_{a \in A_{s_t}} \left\{ r_t\left(s_t, a\right) + \sum_{j \in S} p_t\left(j \mid s_t, a\right) u_{t+1}^*(j) \right\}$$

and the result holds for $n = t$.

(b) Given policy $\pi^* = \left(d_1^*, \ldots, d_{N-1}^*\right)$ we have, from a previous result,

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

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

□

Corollary 1.8

Let

$$\pi^{HR} : \text{ set of history dependent randomized policies}$$
$$\pi^{MD} : \text{ set of Markovian deterministic policies.}$$

Then,

$$V_N^*(s) = \sup_{\pi \in \pi^{HR}} V_N^\pi(s) = \sup_{\pi \in \pi^{MD}} V_N^\pi(s)$$

Proposition 1.9

Assume that S is finite or countable and if either one of the following conditions hold:

(a) A_s *is finite for each* $s \in S$.

(b) A_s *is compact for each* $s \in S$ *and*

$r_t(s,a)$ *is continuous in* a *for all* $s \in S$
$|r_t(s,a)| \leq M$ *for all* $a \in A_s, s \in S$
$p_t(j \mid s,a)$ *is continuous in* a *for each* $j \in S, s \in S$

(c) A_s *is compact for each* $s \in S$ *and*
$r_t(s,a)$ *is upper semicontinuous in* a *for all* $s \in S$,
$|r_t(s,a)| \leq M$ *for all* $a \in A_s, s \in S$,
$p_t(j \mid s,a)$ *is lower semicontinuous in* a *for each* $j \in S, s \in S$

then there exists a deterministic Markovian policy which is optimal.

Backward Induction Algorithm for the optimal policy and optimal total expected reward

(1) Set $t = N$ and $u_N^*(s_N) = r_N(s_N)$.

(2) Substitute $t-1$ for t and compute $u_t^*(s_t)$ for each $s_t \in S$ by

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

and set

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

(3) If $t = 1$ then stop. Otherwise go to step 2 .

Theorem 1.10

Suppose $\pi_t^*,\ t = 1, \ldots, N$ *and* $A_{s_t}^*$ *are obtained using backward dynamic programming.*

(i) For $t = 1, \ldots, N$ *and* $h_t = (h_{t-1}, a_t, s_t)$,

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

where Π *is the set of all history dependent randomized policies.*

(ii) Let $d_t^*(s_t) \in A_{s_t}^*$, *for all* $s_t \in S,\ t = 1, \ldots, N-1$ *and* $\pi^* = (d_1^*, d_2^*, \ldots, d_{N-1}^*)$. *The* π^* *is optimal,*

$$u_1^*(s) = V_N^*(s) = V_N^{\pi^*}(s).$$

Example 1.6 (Inventory problem revisited). Consider the setup

$$M = 3, h(u) = u, \rho(u) = 8u, N = 4, T = \{1, 2, 3, 4\}$$
$$A_s = \{0, \ldots, 3 - s\}$$

and

$$O(u) = \begin{cases} 4 + 2u, & u > 0 \\ 0, & u = 0 \end{cases}$$

with

$$P(D = 0) = \frac{1}{4}, P(D = 1) = \frac{1}{2}, P(D = 2) = \frac{1}{4}$$
$$r_N(0) = r_N(1) = r_N(2) = r_N(3) = 0$$

Now,

$$u_4^*(0) = u_4^*(1) = u_4^*(2) = u_4^*(3) = 0$$

and since

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

then

$$r(0,1) = -O(1) - h(1) + \rho(1)P(D = 1 \cup D = 2) = -6 - 1 + 8 \cdot \frac{3}{4} = -1$$
$$r(0,2) = -12 - 2 + 16 \cdot \frac{1}{4} + 8 \cdot \frac{1}{2} = -2$$
$$r(0,3) = -10 - 3 + 16 \cdot \frac{1}{4} + 8 \cdot \frac{1}{2} = -5$$
$$u_3^*(0) = \max\{0 + 1 \cdot 0, \underbrace{-1}_{=r(0,1)} + 0, \underbrace{-2}_{=r(0,2)}, \underbrace{-5}_{=r(0,3)}\} = 0, d_3^*(0) = 0$$

and continuing in this fashion, we will get

$$u_3^*(1) = 5, u_3^*(2) = 6, u_3^*(3) = 5$$
$$d_3^*(1) = 0, d_3^*(2) = 0, d_3^*(3) = 0$$

Next,

$$
\begin{aligned}
u_2^*(0) &= \max\left\{0, \underbrace{-6-1+8*\frac{3}{4}}_{\text{reward}} + \underbrace{\frac{3}{4}*0}_{\text{demand}\geq 1, u_3^*(0)} + \underbrace{\frac{1}{4}*5}_{\text{demand}=0, u_3^*(1)}, 2, \frac{1}{2}\right\} \\
&= \max\left\{0, \frac{1}{4}, 2, \frac{1}{2}\right\} \\
&= 2
\end{aligned}
$$

and $d_2^*(0) = 2$. Continuing, we will get

$$
d_1^*(s) = \begin{cases} 3, & s = 0 \\ 0, & \text{otherwise} \end{cases}, d_2^*(s) = \begin{cases} 2, & s = 0 \\ 0, & \text{otherwise} \end{cases}
$$

and $d_3^*(s) = 0$ for all $s \in \{1, 2, 3\}$. Finishing, we will get

$$
v_4^*(0) = \frac{67}{16}, v_4^*(1) = \frac{129}{16}, v_4^*(2) = \frac{97}{8}, v_4^*(3) = \frac{227}{16}
$$

1.3 Optimality of Monotone Policies

Consider

$$
u_t^*(s) = \max_{a \in A_s}\left\{r_t(s,a) + \sum_{j \in S} p_t(j \mid s, a) u_{t+1}^*(j)\right\}
$$

Definition 1.11

We say that $g(\cdot,\cdot)$ for $x^+ \geq x^-$ in X and $y^+ \geq y^-$ in Y is superadditive if

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

If $-g(\cdot,\cdot)$ is superadditive then $g(\cdot,\cdot)$ is subadditive.

Example 1.7. $g(x, y) = h(x) + f(y)$ is both superadditive and subadditive.

Lemma 1.12

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

$$
\rho(x) = \max\{y \in \arg\max_{y \in Y} g(x, y)\}
$$

is monotone non-decreasing in X.

Proof. Let $x^+ \geq x^-$ and choose $y \leq \rho\left(x^-\right)$. Then,

$$
g\left(x^-, \rho\left(x^-\right)\right) - g\left(x^-, y\right) \geq 0
$$

Since g is superadditive,

$$
\begin{aligned}
& g\left(x^+, \rho(x^-)\right) + g\left(x^-, y\right) \geq g\left(x^+, y\right) + g\left(x^-, \rho\left(x^-\right)\right) \\
\Longrightarrow\ & g\left(x^+, \rho\left(x^-\right)\right) \geq \underbrace{\left[g\left(x^-, \rho(x^-)\right) - g\left(x^-, y\right)\right]}_{\geq 0} + g\left(x^+, y\right) \\
\Longrightarrow\ & g\left(x^+, \rho\left(x^-\right)\right) \geq g\left(x^+, y\right), \forall y \leq \rho(x^-).
\end{aligned}
$$

then by definition, $\rho\left(x^{+}\right) \geq \rho\left(x^{-}\right)$ since

$$g\left(x^{+}, \rho\left(x^{+}\right)\right) \geq g\left(x^{+}, \rho\left(x^{-}\right)\right) \text{ and } g\left(x^{+}, y\right) \leq g\left(x^{+}, \rho\left(x^{-}\right)\right), \ \forall y \leq \rho\left(x^{-}\right).$$

if $\rho(x^+) < \rho(x^-)$, then $g\left(x^{+}, \rho\left(x^{+}\right)\right) = g\left(x^{+}, \rho\left(x^{-}\right)\right)$, but then by the definition of $\rho(x^+)$, we have $\rho(x^+) \geq \rho(x^-)$. □

Lemma 1.13

Let $g(s,a)$ be a function on $S \times A$, where $S = A = 0, 1, \cdots$ and suppose $g(s+1, a+1) + g(s,a) \geq g(s, a+1) + g(s+1, a)$ for all $a \in A$ and $s \in S$. Then g is superadditive.

Proof. Let $s^+ \geq s^-, a^+ \geq a^-$. We have

$$\begin{aligned}
& g(s^+, a^-) \\
\geq & g(s^+ - 1, a^+) + g(s^+, a^+ - 1) - g(s^+ - 1, a^+ - 1) \\
\geq & g(s^+ - 2, a^+) + g(s^+ - 1, a^+ - 1) - g(s^+ - 2, a^+ - 1) \\
& + g(s^+, a^+ - 2) + g(s^+ - 1, a^+ - 1) - g(s^+ - 1, a^+ - 2) - g(s^+ - 1, a^+ - 1) \\
= & g(s^+ - 2, a^+) + g(s^+, a^+ - 2) + g(s^+ - 1, a^+ - 1) \\
& - g(s^+ - 2, a^+ - 1) - g(s^+ - 1, a^+ - 2) \\
\geq & g(s^+, a^+ - 2) + g(s^+ - 2, a^+) - g(s^+ - 2, a^+ - 2) \\
& \vdots \\
\geq & g(s^+, a^-) + g(s^-, a^+) - g(s^-, a^-).
\end{aligned}$$

where from the second to the third line, we apply the assumption to both $g(s^+ - 1, a^+), g(s^+, a^+ - 1)$; from the fourth to the fifth line, we apply the assumption to $g(s^+ - 1, a^+ - 1) - g(s^+ - 2, a^+ - 1) - g(s^+ - 1, a^+ - 2)$.

Then, by adding $g(s^-, a^-)$ to both sides, we get

$$\begin{aligned}
& g(s^+, a^+) + g(s^-, a^-) \\
\geq & g(s^+, a^-) + g(s^-, a^+)
\end{aligned}$$

□

Lemma 1.14

Let $\{x_j\}, \{x'_j\}$ be real-valued sequences satisfying

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

for all $k \geq 0$ with equality holding for $k = 0$. Suppose $v_{j+1} \geq v_j$ for all $j = 0, 1, \ldots$. Then,

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

Proof. Set $v_{-1} = 0$. Then,

$$\begin{aligned}
\sum_{j=0}^{\infty} v_j x_j &= \sum_{j=0}^{\infty} x_j \sum_{i=0}^{j} (v_i - v_{i-1}) \\
&= \sum_{j=0}^{\infty} (v_j - v_{j-1}) \sum_{i=j}^{\infty} x_j \\
&= \sum_{j=1}^{\infty} (v_j - v_{j-1}) \sum_{i=j}^{\infty} x_j + v_0 \sum_{i=0}^{\infty} x_i \\
&\geq \sum_{j=1}^{\infty} (v_j - v_{j-1}) \sum_{i=j}^{\infty} x'_j + v_0 \sum_{i=0}^{\infty} x'_i \\
&= \sum_{j=0}^{\infty} v_j x'_j.
\end{aligned}$$

□

Note. A classical way to apply this lemma is that given two variables X, Y such that $P(X \geq a) \geq P(Y \geq a), \forall a$, then $\mathbb{E}[f(X)] \geq \mathbb{E}[f(y)]$ for every nondecreasing f.

Theorem 1.15

Assume that

1. $S = \{0, 1, \ldots\}$
2. $A_s = A$ *for all* $s \in S$

Suppose that

1. $r_t(s, a)$ *is non-decreasing (non-increasing) in* s *for all* $a \in A$ *and* $t = 1, \ldots, N-1$.
2. $\sum_{j=k}^{\infty} p_t(j \mid s, a)$ *is non-decreasing in* s *for all* $k \in S, a \in A$ *and* $t = 1, \ldots, N-1$.
3. $r_N(s)$ *is non-decreasing (non-increasing) in* s.

Then $u_t^*(s)$ *is non-decreasing (non-increasing) in* s *for all* $t = 1, \ldots, N$.

Proof. We know $u_N^*(s) = r_N(s)$ and thus the result holds for $t = N$. Now assume it holds for $n = t+1, \ldots, N$ and note that for $n = t$ we have

$$\begin{aligned}
u_t^*(s) &= \max_{a \in A_s} \left\{ r_t(s, a) + \sum_{j \in S} p_t(j \mid s, a) u_{t+1}^*(j) \right\} \\
&= r_t\left(s, a_s^*\right) + \sum_{j \in S} p_t\left(j \mid s, a_s^*\right) u_{t+1}^*(j)
\end{aligned}$$

Suppose that $s' \geq s$. We need to show $u_t^*\left(s'\right) \geq u_t^*(s)$. Now

$$\begin{aligned}
u_t^*(s) &= r_t\left(s, a_s^*\right) + \sum_{j \in S} p_t\left(j \mid s, a_s^*\right) u_{t+1}^*(j) \\
&\leq r_t\left(s', a_s^*\right) + \sum_{j \in S} p_t\left(j \mid s', a_s^*\right) u_{t+1}^*(j) \\
&\leq \max_{a \in A} \left\{ r_t\left(s', a\right) + \sum_{j \in S} p_t\left(j \mid s', a\right) u_{t+1}^*(j) \right\} \\
&= u_t^*\left(s'\right)
\end{aligned}$$

which follows from the assumptions of the theorem, induction hypothesis and the earlier lemma. □

Theorem 1.16

Assume that

1. $S = \{0, 1, \ldots\}$
2. $A_s = A$ *for all* $s \in S$.

Suppose that

(1) $r_t(s, a)$ *is non-decreasing in* s *for all* $a \in A$ *and* $t = 1, \ldots, N-1$.

(2) $\sum_{j=k}^{\infty} p_t(j \mid s, a)$ *is non-decreasing in* s *for all* $k \in S, a \in A$ *and* $t = 1, \ldots, N-1$.

(3) $r_t(s, a)$ *is a superadditive function on* $S \times A$.

(4) $\sum_{j=k}^{\infty} p_t(j \mid s, a)$ *is a superadditive function on* $S \times A$ *for every* $k \in S$.

(5) $r_N(s)$ *is non-decreasing in* s.

Then there exists an decision rules $d_t^*(s)$ *which are non-decreasing in* s *for all* $t = 1, \ldots, N-1$.

Proof. From 1, 2, and 5, we know that $u_t^*(s)$ is non-decreasing in s for all $t = 1, \ldots, N$ and so

$$\sum_{j=k}^{\infty} \left[p_t\left(j \mid s^+, a^+\right) + p_t\left(j \mid s^-, a^-\right)\right] \geq \sum_{j=k}^{\infty} \left[p_t\left(j \mid s^+, a^-\right) + p_t\left(j \mid s^-, a^+\right)\right]$$

for $s^+ \geq s^-, a^+ \geq a^-$, which implies, from the previous theorem, that

$$\sum_{j=0}^{\infty} \left[p_t\left(j \mid s^+, a^+\right) + p_t\left(j \mid s^-, a^-\right)\right] u_{t+1}^*(j) \geq \sum_{j=0}^{\infty} \left[p_t\left(j \mid s^+, a^-\right) + p_t\left(j \mid s^-, a^+\right)\right] u_{t+1}^*(j),$$

so $\sum_{j=0}^{\infty} p_t(j \mid s, a) u_{t+1}^*(j)$ is superadditive on $S \times A$. Since the sum of two superadditive functions is superadditive, then

$$r_t(s, a) + \sum_{j=0}^{\infty} p_t(j \mid s, a) u_{t+1}^*(j)$$

is superadditive and the result follows from Lemma 1.12. □

Theorem 1.17

Suppose for $t = 1, \ldots, N-1$ *that*

(1) $r_t(s, a)$ *is non-increasing in* s *for all* $a \in A$ *and* $t = 1, \ldots, N-1$.

(2) $\sum_{j=k}^{\infty} p_t(j \mid s, a)$ *is non-decreasing in* s *for all* $k \in S, a \in A$ *and* $t = 1, \ldots, N-1$.

(3) $r_t(s, a)$ *is a superadditive function on* $S \times A$.

(4) $\sum_{j=0}^{\infty} p_t(j \mid s, a)$ *is a superadditive function on* $S \times A$.

(5) $r_N(s)$ *is non-increasing in* s.

Then there exists an optimal decision rules $d_t^*(s)$ *which are non-decreasing in* s *for all* $t = 1, \ldots, N-1$.

Proof. From (1), (2), and (5) we have $u_t^*(s)$ non-increasing in s. Then from (3) and (4), we have

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

superadditive on $S \times A$. □

Backward Dynamic Programming for finding Monotone Optimal Policies Assume that for each t there is a monotone optimal decision rule. Suppose that $S = \{0, 1, \ldots, M\}$ and $A_s = A$ for all $s \in S$.

1. Set $t = N$ and $u_N^*(s) = r_N(s)$ for all $s \in S$.
2. Substitute $t-1$ for t, set $s = 0$ and $A_0 = A$.
   (a) Set
   $$u_t^*(s) = \max_{a \in A_s} \left\{ r_t(s,a) + \sum_{j \in S} p_t(j \mid s,a) u_{t+1}^*(j) \right\}$$
   (b) Set
   $$A_{s,t}^* = \arg\max_{a \in A_s} \left\{ r_t(s,a) + \sum_{j \in S} p_t(j \mid s,a) u_{t+1}^*(j) \right\}$$
   (c) If $s = M$ go to step 3, otherwise set
   $$A_{s+1} = \left\{ a \in A : a \geq \max \left\{ a' \in A_{s,t}^* \right\} \right\}$$
   (d) Substitute $s+1$ for s and return to (a).
3. If $t = 1$, stop; otherwise go to Step 2.

Example 1.8. Given $S = \{0, 1, \ldots\}$, the higher the worse the equipment is. From one decision epoch to the next, the equipment deteriorates i states with probability $p(i)$. We are also given, $A_s = \{0, 1\}$ where 0 is "do nothing" and 1 is replacing, R is the fixed income per period, $h(s)$ is the operating cost if the equipment is in state s, K is the replacement cost, $r_N(s)$ is the salvage of the equipment if it is in state s at time N. Assume $h(s)$ is non-decreasing in s and $r_N(s)$ is non-increasing in s. Let $T = \{1, \ldots, N\}$.

We have:

$$p(j \mid s, 0) = \begin{cases} 0, & \text{if } j < s \\ p(j-s), & \text{if } j \geq s \end{cases} \text{ and } p(j \mid s, 1) = p(j), i = 0, 1, 2, \ldots$$

and

$$r(s,0) = R - h(s) \text{ and } r(s,1) = R - K - h(0)$$

1. $r(s,a)$ is non-increasing in s. Clearly this holds for the rewards.
2. $r_N(s)$ is non-increasing in s.
3. $\sum_{j=k}^{\infty} p_t(j \mid s,a)$ is non-decreasing in s for all $k \in S$ and $a \in A$ since when we replace,
   $$\sum_{j=k+1}^{\infty} p(j \mid s+1, 1) - \sum_{j=k}^{\infty} p(j \mid s, 1) = \sum_{j=k}^{\infty} p(j) - \sum_{j=k}^{\infty} p(j) = 0$$
   Now when we do not replace, for $k > s$,
   $$\sum_{j=k}^{\infty} p(j \mid s+1, 0) - \sum_{j=k}^{\infty} p(j \mid s, 0) = \sum_{j=k}^{\infty} p(j-s-1) - \sum_{j=k}^{\infty} p(j-s) = p(k-s-1) \geq 0$$
   and for $k \leq s$, we have
   $$\sum_{j=k}^{\infty} p(j \mid s+1, 0) - \sum_{j=k}^{\infty} p(j \mid s, 0) = \sum_{j=s+1}^{\infty} p(j-s-1) - \sum_{j=s}^{\infty} p(j-s) = 0$$

4. $r(s,a)$ is superadditive on $S \times A$:

$$\begin{aligned} & r(s+1,1)+r(s,0) \geq r(s,1)+r(s+1,0) \\ \Longleftrightarrow & R-K-h(0)+R-h(s) \geq R-K-h(0)+R-h(s+1) \\ \Longleftrightarrow & h(s+1)-h(s) \geq 0 \end{aligned}$$

5. $\sum_{j=0}^{\infty} p(j \mid s,a)u(j)$ is superadditive on $S \times A$ for any non-increasing function u:

$$\begin{aligned} & \sum_{j=0}^{\infty} p(j \mid s+1,1)u(j)+\sum_{j=0}^{\infty} p(j \mid s,0)u(j) \geq \sum_{j=0}^{\infty} p(j \mid s,1)u(j)+\sum_{j=0}^{\infty} p(j \mid s+1,0)u(j) \\ \Longleftrightarrow & \sum_{j=0}^{\infty} p(j)u(j)+\sum_{j=s}^{\infty} p(j-s)u(j) \geq \sum_{j=0}^{\infty} p(j)u(j)+\sum_{j=s+1}^{\infty} p(j-s-1)u(j) \\ \Longleftrightarrow & \sum_{j=s}^{\infty} p(j-s)u(j) \geq \sum_{j=s+1}^{\infty} p(j-s-1)u(j) \\ \Longleftrightarrow & \sum_{j=s}^{\infty} p(j-s)u(j)-\sum_{j=s}^{\infty} p(j-s)u(j+1) \geq 0 \end{aligned}$$

since u is non-creasing.

6. $d_t^*(s)=\begin{cases} 0, \text{ if } & s \leq s_t^* \\ 1, \text{ if } & s>s_t^*, \forall t=1, \ldots, N-1 \end{cases}$

2 Infinite Horizon MDPs

We assume:

- Transition probabilities and rewards are stationary and $|r(s,a)| \leq M$
- We are given a discount factor $0 < \lambda < 1$.
- $\pi = (d_1, d_2, \ldots)$ is Markovian deterministic.
- $T = \{1, 2, 3, \ldots\}$.
- $v_\lambda^\pi(s)$: total expected discounted reward under policy π when the initial state is s and the discount factor is λ. Let $\{X_t : t \geq 1\}$ be the Markov Chain under policy π,

$$v_\lambda^\pi(s) = \mathbb{E}_s \left[\sum_{t=1}^{\infty} \lambda^{t-1} r\left(X_t, d_t\left(X_t\right)\right) \right]$$

- r_d : vector of rewards under decision rule d

$$r_{d_1} = \begin{bmatrix} r(s_1, d_1(s_1)) \\ r(s_2, d_1(s_2)) \\ \vdots \end{bmatrix}$$

- P_d : probability transition matrix under decision rule d

Let us denote $v_\lambda^*(s) = \sup_\pi v_\lambda^\pi(s)$. If it is attained, we would also like to find π^* where

$$v_\lambda^*(s) = v_\lambda^{\pi^*}(s).$$

If v_λ^π is the vector of total expected rewards, then

$$\begin{aligned} v_\lambda^\pi &= r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} r_{d_3} + \ldots \\ &= \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} r_d \\ &= r_{d_1} + \lambda P_{d_1} \left(r_{d_2} + \lambda P_{d_2} r_{d_3} + \ldots\right) \\ &= r_{d_1} + \lambda P_{d_1} v_\lambda^{\pi'} \end{aligned}$$

where $\pi' = (d_2, d_3, \ldots)$. Now if π is stationary, then

$$v_\lambda^\pi = r_d + \lambda P_d v_\lambda^\pi \implies v_\lambda^\pi = \left(I - \lambda P_d\right)^{-1} r_d$$

Theorem 2.1

For any stationary policy $\pi = d^\infty$, $v_\lambda^{d^\infty}$ is the unique solution of

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

and furthermore, $v_\lambda^{d^\infty}$ can be written as

$$v_\lambda^{d^\infty} = \left(I - \lambda P_d\right)^{-1} r_d = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} r_d = L_d v_\lambda^{d^\infty}$$

where $L_d(v) := r_d + \lambda P_d v$. Note the inverse exists because $\lambda < 1$.

Example 2.1. Consider a simple system with $S = \{s_1, s_2\}$ and $A_{s_1} = \{a_{11}, a_{12}\}$ and $A_{s_2} = \{a_{21}\}$. We have $p(s_1 \mid s_1, a_{11}) = 0.5$, $p(s_2 \mid s_1, a_{11}) = 0.5, p(s_2 \mid s_1, a_{12}) = 1$, and $p(s_2 \mid s_2, a_{21})$=1. Finally, $r(s_1, a_{11}, s_1) = 5, r(s_1, a_{11}, s_2) = 5, r(s_1, a_{12}) = 10$ and $r(s_2, a_{21}) = -1$. Consider the stationary policy that uses the decision rule $d(s_1) = a_{11}$ and $d(s_2) = a_{21}$. Compute $v_\lambda^{d^\infty}(s_1)$ and $v_\lambda^{d^\infty}(s_2)$.

We have $r_d = \begin{bmatrix} 5 \\ -1 \end{bmatrix}$ and

$$v_\lambda^{d^\infty}(s_1) = 5 + \lambda\left(0.5 v_\lambda^{d^\infty}(s_1) + 0.5 v_\lambda^{d^\infty}(s_2)\right)$$

$$v_\lambda^{d^\infty}(s_2) = -1 + \lambda v_\lambda^{d^\infty}(s_2) \implies v_\lambda^{d^\infty}(s_2) = \frac{-1}{1-\lambda}$$

and so after a substitution,

$$v_\lambda^{d^\infty}(s_1) = \frac{5 - 5.5\lambda}{(1-\lambda)(1-0.5\lambda)}.$$

Lemma 2.2

Suppose $0 \le \lambda < 1$. *Then for any Markovian deterministic decision rule d,*

(i) If $u \ge 0$ *then* $(I - \lambda P_d)^{-1} u \ge 0$ *and* $(I - \lambda P_d)^{-1} u \ge u$.

(ii) If $u \ge v$ *then* $(I - \lambda P_d)^{-1} u \ge (I - \lambda P_d)^{-1} v$.

(iii) If $u \ge 0$ *then* $u^T (I - \lambda P_d)^{-1} \ge 0$.

Proof. (i) and (iii): directly by

$$(I - \lambda P_d)^{-1} u = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} u \ge u \ge 0$$

(ii): follows from (i) by replacing u with $u - v$ □

2.1 Optimality Equations

Recall the optimality equation for the finite-horizon case:

$$v_n(s) = \sup_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j \mid s, a) v_{n+1}(j). \right\}$$

By taking the limit as $n \to \infty$ on both sides, we the optimality equation for the infinite-horizon case:

$$v_\lambda(s) = \underbrace{\sup_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j \mid s, a) v(j) \right\}}_{\mathcal{L}}$$

Let v be the vector of $v(s)$ for $s \in S$, then we write the above equation as $v = \mathcal{L}v$.
If the supremum is attained,

$$v_\lambda(s) = \underbrace{\max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda p(j \mid s, a) v^*(j) \right\}}_{L}$$

Theorem 2.3

Suppose that there exists a v such that

(i) $v \geq \mathcal{L}v$ *then* $v \geq v_\lambda^*$

(ii) $v \leq \mathcal{L}v$ *then* $v \leq v_\lambda^*$

(iii) $v = \mathcal{L}v$ *then* $v = v_\lambda^*$

Proof. (i) Let $\pi = (d_1, d_2, \ldots)$ and let us use the notation

$$\mathcal{L}v = \sup_\alpha \{r_\alpha + \lambda P_\alpha v\}$$
$$Lv = \max_\alpha \{r_\alpha + \lambda P_\alpha v\}$$

Then

$$\begin{aligned} v \geq & \sup_\alpha \{r_\alpha + \lambda P_\alpha v\} = \mathcal{L}v = r_{d_1} + \lambda P_{d_1} v \\ \geq & r_{d_1} + \lambda P_{d_1} (r_{d_2} + \lambda P_{d_2} v) \\ & \vdots \\ \geq & r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} r_{d_3} + \ldots + \lambda^{n-1} P_{d_1} \ldots P_{d_{n-1}} r_{d_n} + \lambda^n \underbrace{P_{d_1} \ldots P_{d_n}}_{P_\pi^n} v \end{aligned}$$

and also since

$$v_\lambda^\pi = r_{d_1} + \lambda P_{d_1} r_{d_2} + \ldots + \sum_{k=2}^{\infty} \lambda^k P_{d_1} \ldots P_{d_k} r_{d_{k+1}}$$

then

$$v - v_\lambda^\pi \geq \lambda^n P_{d_1} \ldots P_{d_n} v - \sum_{k=n}^{\infty} \lambda^k P_{d_1} \ldots P_{d_k} r_{d_{k+1}}$$

Next, if we define $\|v\| = \sup_{s \in S} |v(s)|$ then $\|\lambda^n P^n v\| \leq \lambda^n \|v\|$ then we can choose $\epsilon > 0$ such that there exists n sufficiently larges such that

$$-\frac{\epsilon}{2} e \leq \lambda^n P_{d_1} \ldots P_{d_n} v \leq \frac{\epsilon}{2} e$$

where e is a vector of ones. Also,

$$-\frac{\lambda^n M e}{(1-\lambda)} \leq \sum_{k=n}^{\infty} \lambda^k P_{d_1} \ldots P_{d_k} r_{d_{k+1}} \leq \frac{\lambda^n M e}{(1-\lambda)}$$

by $|r_{d_{k+1}}| \leq Me$, and so with can find n sufficiently large so that

$$v - v_\lambda^\pi \geq -\varepsilon e \implies v \geq \sup_\pi v_\lambda^\pi = v_\lambda^*$$

(ii) From the definition of $\mathcal{L}$, we know that for all $\epsilon > 0$ there exists α such that

$$v \leq r_\alpha + \lambda P_\alpha v + \epsilon e$$

which, by the previous lemma, implies

$$\begin{aligned}
&(I - \lambda P_\alpha)\, v \le r_\alpha + \epsilon e \\
\implies & v \le (I - \lambda P_\alpha)^{-1} (r_\alpha + \epsilon e) \\
\implies & v \le (I - \lambda P_\alpha)^{-1} r_\alpha + (I - \lambda P_\alpha)^{-1} \epsilon e
\end{aligned}$$

and hence

$$\begin{aligned}
v &\le v_\lambda^{\alpha^\infty} + \epsilon \sum_{k=1}^{\infty} \lambda^{k-1} P_\alpha^{k-1} e \\
&= v_\lambda^{\alpha^\infty} + \frac{\epsilon e}{1-\lambda} \\
&\le \sup_\pi v_\lambda^\pi = v_\lambda^*
\end{aligned}$$

where the last inequality is by pushing ϵ to 0.

(iii) Trivial.

□

Definition 2.4

Let U be a Banach space (complete normed linear space, e.g. $\mathbb{R}^n$). The operator $T : U \to U$ is a contraction mapping if $\exists \lambda$ with $0 \le \lambda < 1$ such that

$$\|Tv - Tu\| \le \lambda \|v - u\|$$

Theorem 2.5: Fixed Point Theorem

Suppose U is Banach space and $T : U \mapsto U$ is a contraction mapping. Then,

1. *$\exists v^* \in U$ unique such that $Tv^* = v^*$*
2. *for arbitrary $v^0 \in U$, the sequence $\{v^n\}$ defined by $v^{n+1} = Tv^n$ converges to v^*.*

Proof. (a) Directly

$$\begin{aligned}
\left\|v^{n+m} - v^n\right\| &= \left\| \sum_{k=0}^{m-1} v^{n+k+1} - \sum_{k=0}^{m-1} v^{n+k} \right\| \\
&\le \sum_{k=0}^{m-1} \left\|v^{n+k+1} - v^{n+k}\right\| \\
&= \sum_{k=0}^{m-1} \left\|T^{n+k} v^1 - T^{n+k} v^0\right\| \\
&\le \sum_{k=0}^{m-1} \lambda^{n+k} \left\|v^1 - v^0\right\| \\
&= \left\|v^1 - v^0\right\| \cdot \frac{\lambda^n (1 - \lambda^m)}{1 - \lambda}
\end{aligned}$$

and so $\{v^n\}$ is a Cauchy sequence and $\exists v^*$ such that $v^n \to v^*$. It remains to be seen that $Tv^* = v^*$. We have

$$\begin{aligned}
0 \le \|Tv^* - v^*\| &\le \|Tv^* - v^n\| - \|v^n - v^*\| \\
&\le \left\|Tv^* - Tv^{n-1}\right\| - \|v^n - v^*\| \\
&\le \lambda \left\|v^* - v^{n-1}\right\| - \|v^n - v^*\|.
\end{aligned}$$

Since $v^n \to v^*$ the the right hand side can be made arbitrarily small by picking large enough n. Hence $\|Tv^* - v^*\| = 0$ and $Tv^* = v^*$.
Suppose there exists v' such that $Tv' = v'$. Then,

$$\|v^* - v'\| = \|Tv^* - Tv'\| \leq \lambda \|v^* - v'\|$$

which is only possible if $\|v^* - v'\| = 0 \implies v^* = v'$.

□

Proposition 2.6

For $0 \leq \lambda < 1$, L and $\mathcal{L}$ are contraction mappings.

Proof. Let u and v be such that $Lv(s) \geq Lu(s)$ for $s \in S$ and

$$\max_{a \in A_s} \left\{ r(s,a) + \lambda \sum p(j \mid s,a) v(j) \right\} \geq \max_{a \in A_s} \left\{ r(s,a) + \lambda \sum p(j \mid s,a) u(j) \right\}$$

and suppose that

$$a_s^* \in \underset{a \in A_s}{argmax} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j \mid s,a) v(j) \right\}$$

Then

$$\begin{aligned} 0 \leq Lv(s) - Lu(s) &\leq r\left(s, a_s^*\right) + \lambda \sum_{j \in S} p\left(j \mid s, a_s^*\right) v(j) - r\left(s, a_s^*\right) - \lambda \sum_{j \in S} p\left(j \mid s, a_s^*\right) u(j) \\ &= \lambda \sum_{j \in S} p\left(j \mid s, a_s^*\right) [v(j) - u(j)] \\ &\leq \lambda \sum_{j \in S} p\left(j \mid s, a_s^*\right) \|v - u\| \\ &= \lambda \|v - u\| \end{aligned}$$

and we can have the similar result for $Lu(s) \geq Lv(s)$. Therefore,

$$|Lv(s) - Lu(s)| \leq \lambda \|v - u\| \implies \|Lv - Lu\| \leq \lambda \|v - u\|$$

and a similar argument can be made for $\mathcal{L}$. Note that L_d, through the same arguments, is also a contraction mapping. □

Theorem 2.7

1. *There exists a unique v^* satisfying $Lv^* = v^*$ ($\mathcal{L}v^* = v^*$) and $v^* = v_\lambda^*$.*
2. *For each d satisfying $L_d v = v$, there exists a unique solution $v = v_\lambda^\pi$ where $\pi = (d, d, \ldots)$. [$L_d v = r_d + \lambda P_d v$]*

Proof. By $L, \mathcal{L}$ being contraction mappings, we know there exists a unique solution v^* such that $Lv^* = v^*$. Then from Theorem 2.3, we know $v^* = v_\lambda^*$. Part (2) can be consider a special case of (1) where the only available policy is d. □

Theorem 2.8

A policy π^* is optimal if and only if $v_\lambda^{\pi^*}$ is a solution to the optimality equations.

Proof. If π^* is optimal then $v_\lambda^* = v_\lambda^{\pi^*}$ and hence $Lv_\lambda^{\pi^*} = v_\lambda^{\pi^*}$ by the above theorem. If $Lv_\lambda^{\pi^*} = v_\lambda^{\pi^*}$ then $v_\lambda^{\pi^*} = v_\lambda^*$ by Theorem 2.3 and hence π^* is optimal. □

Theorem 2.9

Suppose d is such that

$$L_{d^*} v_\lambda^* = r_{d^*} + \lambda P_{d^*} v_\lambda^* = v_\lambda^*$$

or $d^* \in \operatorname{argmax}\{r_d + \lambda P_d v_\lambda^*\}$ where we say that d^* is a conserving decision rule. Then, $(d^*)^\infty$ is an optimal decision policy and $v_\lambda^{(d^*)^\infty} = v_\lambda^*$.

Proof.

$$v_\lambda^* = Lv_\lambda^* = r_{d^*} + \lambda P_{d^*} v_\lambda^* = L_{d^*} v_\lambda^*,$$

then $v_\lambda^* = v_\lambda^{(d^*)^\infty}$ because $v_\lambda^{(d^*)^\infty}$ is the unique solution to $v = L_{d^*} v$. □

Theorem 2.10

Suppose there exists an optimal policy, then there exists an optimal stationary policy.

Proof. Given $\pi^* = (d_1, d_2, \ldots)$ and $\pi^* = (d_1, \pi')$. Then,

$$\begin{aligned} v_\lambda^* = v_\lambda^{\pi^*} &= r_{d_1} + \lambda P_{d_1} v_\lambda^{\pi'} \\ &\leq r_{d_1} + \lambda P_{d_1} v_\lambda^{\pi^*} \\ &\leq \sup_d \left\{ r_d + \lambda P_d v_\lambda^{\pi^*} \right\} \\ &= \mathcal{L} v_\lambda^{\pi^*} = v_\lambda^{\pi^*} \end{aligned}$$

and d_1 is a conserving decision rule which means it is an optimal decision rule. □

2.2 Algorithms

We will be considering:

1. Value Iteration
2. Policy Iteration
3. Linear Programming

Theorem 2.11

Suppose that S is countable. Then there exists a stationary optimal policy if

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

(b) A_s is compact for each $s \in S$, $r(s,a)$ is continuous in a for each s, and $p(j \mid s, a)$ is continuous in a for each $j \in S$ and $s \in S$, or

(c) A_s is compact for each $s \in S$, $r(s,a)$ is upper semicontinuous in a for each s, and $p(j \mid s, a)$ is lower semicontinuous in a for each $j \in S$ and $s \in S$.

2.2.1 Value Iteration

We wish to find a policy π_ϵ such that $v_\lambda^{\pi_\epsilon} \geq v_\lambda^*(s) - \epsilon$.

(1) Select $v^0 \in V, \epsilon > 0$ and set $n = 0$

(2) For each $s \in S$, compute $v^{n+1}(s)$ as

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

(3) If $\left\| v^{n+1} - v^n \right\| \leq \frac{\epsilon(1-\lambda)}{2\lambda}$ then go to step 4. Otherwise, increment n by 1 and go to step (2).

(4) For each $s \in S$, choose

$$d_\epsilon(s) \in \underset{a \in A_s}{arg\,max} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j \mid s, a) v^{n+1}(j) \right\}$$

Theorem 2.12

For value iteration, we have

(1) v^n converges to v_λ^*

(2) Stationary policy $(d_\epsilon)^\infty$ is an ϵ-optimal policy

Proof.

(1) Trivial, from fixed point theorem.

(2) We need to show that $\left\| v_\lambda^{(d_\epsilon)^\infty} - v_\lambda^* \right\| \leq \epsilon$, where $v_\lambda^{(d_\epsilon)^\infty}$ is the expected reward under the stationary policy $(d_\epsilon)^\infty$ satisfying $L_{(d_\epsilon)^\infty} v_\lambda^{(d_\epsilon)^\infty} = v_\lambda^{(d_\epsilon)^\infty}$. Note that

$$\left\| v_\lambda^{(d_\epsilon)^\infty} - v_\lambda^* \right\| \leq \left\| v_\lambda^{(d_\epsilon)^\infty} - v^{n+1} \right\| + \left\| v^{n+1} - v_\lambda^* \right\|.$$

First, we have

$$\begin{aligned}
\left\| v^{n+1} - v_\lambda^* \right\| &= \left\| \sum_{k=n+1}^{\infty} v^k - v^{k+1} \right\| \\
&\leq \sum_{k=n+1}^{\infty} \left\| v^k - v^{k+1} \right\| \\
&= \sum_{k=0}^{\infty} \left\| v^{k+n+1} - v^{k+n+2} \right\| \\
&= \sum_{k=0}^{\infty} \left\| L^{k+1} v^n - L^{k+1} v^{n+1} \right\| \\
&\leq \sum_{k=0}^{\infty} \lambda^{k+1} \left\| v^n - v^{n+1} \right\| \\
&\leq \sum_{k=0}^{\infty} \lambda^{k+1} \frac{\epsilon(1-\lambda)}{2\lambda} \\
&= \frac{\lambda}{1-\lambda} \cdot \frac{\epsilon(1-\lambda)}{2\lambda} \\
&= \frac{\epsilon}{2}
\end{aligned}$$

and

$$\begin{aligned}
\left\|v_\lambda^{(d_\epsilon)^\infty} - v^{n+1}\right\| &= \left\|L_{d_\epsilon} v_\lambda^{(d_\epsilon)^\infty} - v^{n+1}\right\| \\
&\leq \left\|L_{d_\epsilon} v_\lambda^{(d_\epsilon)^\infty} - Lv^{n+1}\right\| + \left\|Lv^{n+1} - v^{n+1}\right\| \\
&= \left\|L_{d_\epsilon} v_\lambda^{(d_\epsilon)^\infty} - L_{d_\epsilon} v^{n+1}\right\| + \left\|Lv^{n+1} - Lv^{n}\right\| \\
&\leq \lambda \left\|v_\lambda^{(d_\epsilon)^\infty} - v^{n+1}\right\| + \lambda \left\|v^{n+1} - v^{n}\right\| \\
\implies (1-\lambda)\left\|v_\lambda^{(d_\epsilon)^\infty} - v^{n+1}\right\| &\leq \lambda \left\|v^{n+1} - v^{n}\right\| \leq \frac{\epsilon(1-\lambda)}{2} \\
\implies \left\|v_\lambda^{(d_\epsilon)^\infty} - v^{n+1}\right\| &\leq \frac{\epsilon}{2}
\end{aligned}$$

where from the second to the third line, we first use the fact that

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

and then the definition $d_\epsilon(s) = \underset{a \in A_s}{arg\,max} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j \mid s, a) v^{n+1}(j) \right\}$, which implies that $Lv^{n+1}(s)$ is obtained under policy $d_\epsilon(s)$, that is, $L_{d_\epsilon} v^{n+1}(s) = Lv^{n+1}(s)$.
Combine the two results together, we get what we want.

□

Proposition 2.13

(1) Suppose $v \geq u$. Then $Lv \geq Lu$.

(2) Suppose that for some N, $Lv^N \leq (\geq) v^N$. Then $v^{N+m+1} \leq (\geq) v^{N+m}$ for all $m \geq 0$.

Proof.

1. Let $d' \in \operatorname{argmax} \{r_d + \lambda P_d u\}$. Then,

$$Lu = r_{d'} + \lambda P_{d'} u \leq r_{d'} + \lambda P_{d'} v \leq \max \{r_d + \lambda P_d v\} = Lv,$$

where the first inequality is by the fact that $P_{d'}$ is a nonnegative matrix.

2. Directly,

$$v^{N+m+1} = L^m L v^N \geq L^m v^N = v^{N+m}$$

and likewise for the $(\leq)$ case.

□

For the second property of the proposition above, it says that if such N exists, then from N, such property holds for all iterations after that. So if $v^1 \geq v^0$ in value iteration, then $\{v^n\} \to v_\lambda^*$ is monotone decreasing. For example, if $Lv^0 \geq v^0$, then $v^{n+1} \geq v^n$ for all n, similar for $\leq$, but for some problems, v^1, v^0 might not be comparable.

Definition 2.14

Let $y_n \to y^*$, so $\lim \|y_n - y^*\| = 0$. We say y_n converges of order α *if there exists a $k > 0$ such that*

$$\|y_{n+1} - y^*\| \leq k \|y_n - y^*\|^\alpha.$$

Theorem 2.15

(i) Convergence rate of value iteration is linear in λ.

(ii)

$$\limsup_{n\to\infty}\left[\frac{\|v^n - v_\lambda^*\|}{\|v^0 - v_\lambda^*\|}\right]^{\frac{1}{n}} \leq \lambda$$

(iii) For every n,

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

(iv) For every $d_\epsilon = \arg\max\{r_d + \lambda P_d v^n\}$,

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

Proof. (i) $\|v^{n+1} - v_\lambda^*\| = \|Lv^n - Lv_\lambda^*\| \leq \lambda\|v^n - v_\lambda^*\|$

(ii) Directly from (i)

(iii) Similar to the first part of the proof of Theorem 2.12.

(iv) Similar to the proof of Theorem 2.12.

□

2.2.2 Policy Iteration

(a) Set $n = 0$ and select arbitrary decision rule d_0

(b) (Policy Evaluation) Obtain v^n by solving

$$(I - \lambda P_{d_n})\, v^n = r_{d_n}$$

(c) (Policy Increment) Choose d_{n+1} such that

$$d_{n+1} \in \underset{d}{argmax}\,\{r_d + \lambda P_d v^n\}$$

and setting $d_{n+1} = d_n$ if possible. That is, if d_n is in the $\arg\max$ above, always pick $d_{n+1} = d_n$.

(d) If $d_{n+1} = d_n$ then stop and return $d^* = d_n$, otherwise increment n by 1 and return to (b)

- Advantages: Works well for solving d^* and even 1 iteration is a good heuristic.
- Disadvantages: Computing step (b)

Proposition 2.16

Let v^n, v^{n+1} be successive values generated by policy iteration. Then $v^{n+1} \geq v^n$.

Proof. Directly

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

□

Theorem 2.17

For a finite state and action space, policy iteration terminates after a finite number of step with a stationary (discounted) optimal policy $(d^*)^\infty$

That is, when we stop, our v^n solves the optimality equations and d_n is a conserving decision rule. It is finite because we have a finite number of actions and states.

Example 2.2. Recall example with

$$S = \{s_1, s_2\}, A_{s_1} = \{a_{11}, a_{12}\}, A_{s_2} = \{a_{21}\}$$

and

$$\begin{aligned} p(s_1 \mid s_1, a_{11}) &= \frac{1}{2} \\ p(s_2 \mid s_1, a_{11}) &= \frac{1}{2} \\ p(s_2 \mid s_1, a_{12}) &= 1 \\ p(s_2 \mid s_2, a_{21}) &= 1 \end{aligned}$$

and general $\lambda \in [0, 1)$. We also have

$$r(s_1, a_{11}) = 5, r(s_1, a_{12}) = 10, r(s_2, a_{21}) = -1$$

The policy iteration is:

(1) Let $d_0(s_1) = a_{11}$ and $d_0(s_2) = a_{21}$

(2) $\equiv$ (b) Get

$$v_\lambda^{(d_0)^\infty}(s_1) = \frac{5 - 5.5\lambda}{(1 - 0.5\lambda)(1 - \lambda)} \text{ and } v_\lambda^{(d_0)^\infty}(s_2) = \frac{-1}{1 - \lambda}$$

(3) $\equiv$ (c) Get

$$\begin{aligned} d_1(s_1) &\in \operatorname{argmax}\left\{5 + \frac{1}{2}v_\lambda^{(d_0)^\infty}(s_1) + \frac{1}{2}v_\lambda^{(d_0)^\infty}(s_2), 10 + v_\lambda^{(d_0)^\infty}(s_2)\right\} \\ \implies d_1(s_1) &\in \arg\max\left\{\frac{(5 - 5.5\lambda)}{(1 - 0.5\lambda)(1 - \lambda)}, \frac{2(5 - 5.5\lambda)}{1 - \lambda}\right\} \end{aligned}$$

Now if $\lambda > \frac{10}{11}$, we have $d_1(s_1) = a_{11}$, otherwise we have $d_1(s_1) = a_{12}$.

For example, let $\lambda = 0.95$ and $d_0(s_1) = a_{12}, d_0(s_2) = a_{21}$. Then

$$\begin{aligned} v_0 &= r_{d_0} + \lambda P_{d_0} v_0 \\ v_0(s_1) &= 10 + 0.95 v_0(s_2) \implies v_0(s_1) = -9 \\ v_0(s_2) &= -1 + 0.95 v_0(s_2) \implies v_0(s_2) = -20 \end{aligned}$$

And hence

$$d_s(1) = \arg\max\{\underbrace{5 + 0.95(0.5(-9) + 0.5 * 20)}_{a_{11}}, \underbrace{10 + 0.95(-20)}_{a_{12}}\} = \arg\max\{-8.775, -9\} = a_{11}.$$

Hence, $d_1(s_1) = a_{11}, d_1(s_2) = a_{21}$ which is different from d_0, we need to run the iteration again and we go back to the analysis above.

2.2.3 Modified Policy Iteration

Let $\{m_n\}$ be a sequence of non-negative integers.

(1) Select v^0, specify $\epsilon > 0$, and set $n = 0$.

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

$$d_{n+1} \in \underset{d}{argmax} \{r_d + \lambda P_d v^n\}$$

and setting $d_{n+1} = d_n$ if possible (when $n > 0$).

(3) (Partial Policy Evaluation)

a. Set $k = 0$ and

$$u_n^0 = \max_{d \in D} \{r_d + \lambda P_d v^n\}$$

or equivalently,

$$u_n^0(s) = \max_{a \in A_s} \left\{ r_d(s,a) + \lambda \sum_{j \in S} p(j \mid s,a) v^n(j) \right\}$$

b. If $\left\| u_n^0 - v^n \right\| < \frac{\epsilon(1-\lambda)}{2\lambda}$ go to step (4). Otherwise go to c.

c. If $k = m_n$ go to e., otherwise compute u_n^{k+1} by

$$u_n^{k+1} = r_{d_{n+1}} + \lambda P_{d_{n+1}} u_n^k = L_{d_{n+1}} u_n^k$$

d. Increment k by 1 and return to c.

e. Set $v^{n+1} = u_n^{m_n}$, increment n by 1 and go to step (2).

(4) Set $d_\epsilon = d_{n+1}$.

2.3 Linear Programming

If $v \geq Lv$ then $v \geq v_\lambda^*$ by Proposition 2.13. For each $j \in S$ pick $\alpha(j) > 0$ and consider the primal LP:

$$\begin{aligned} &\min_v \sum_{j \in S} \alpha(j) v(j) \\ &\text{s.t. } v(s) \geq r(s,a) + \lambda \sum_{j \in S} p(j \mid s,a) v(j), \forall s \in S, \forall a \in A_s \end{aligned}$$

where the constraint is equivalent to

$$v(s) - \lambda \sum_{j \in S} p(j \mid s,a) v(j) \geq r(s,a), \forall s \in S, \forall a \in A_s.$$

Also, please note that the constraint is equivalent to

$$v(s) \geq \max_{a \in A_s} \{r(s,a) + \lambda \sum_{j \in S} p(j \mid s,a) v(j)\} \iff v \geq Lv \implies v \geq v_\lambda^*.$$

The dual LP, with dual variables $x(s,a)$ for each $s \in S, a \in A_s$, is

$$\begin{aligned} &\max \sum_{s \in S} \sum_{a \in A_s} r(s,a) x(s,a) \\ &\text{s.t. } \sum_{a \in A_j} x(j,a) - \lambda \sum_{s \in S} \sum_{a \in A_s} p(j \mid s,a) x(s,a) = \alpha(j), \forall j \in S \\ &\quad x(s,a) \geq 0, \forall a \in A_s, s \in S. \end{aligned}$$

The following theorem shows that there is an "one-to-one" relation between the feasible set of the dual problem above and the set of all Markovian randomzied decision rules.

Theorem 2.18

(1) For each Markovian randomized decision rule d, let

$$x_d(s,a) = \sum_{j\in S}\alpha(j)\sum_{n=1}^{\infty}\lambda^{n-1}P_{d^\infty}(X_n = s, Y_n = a \mid X_1 = j),$$

then $x_d(s,a)$ is a feasible solution to the dual LP.

(2) Suppose that $x(s,a)$ is a feasible solution to the dual LP. Then for each $s \in S$, by $\alpha(s) > 0$, $\sum_{a\in A_s} x(s,a) > 0$. Define the randomized decision ruls d_x^∞ by

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

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

Proof.

(1) Need to show

$$\lambda\sum_{s\in S}\sum_{a\in A_s} p(j \mid s,a)x_d(s,a) = -\alpha(j) + \sum_{a\in A_s} x(j,a).$$

Then

$$\begin{aligned}
&\sum_{s\in S}\sum_{a\in A_s}\lambda p(j \mid s,a)x_d(s,a)\\
=&\sum_{s\in S}\sum_{a\in A_s}\lambda p(j \mid s,a)\sum_{k\in S}\alpha(k)\sum_{n=1}^{\infty}\lambda^{n-1}P_{d^\infty}(X_n = s, Y_n = a \mid X_1 = k)\\
=&\sum_{k\in S}\alpha(k)\sum_{n=1}^{\infty}\lambda^{n}\sum_{a\in A_s}\sum_{s\in S}p(j \mid s,a)P_{d^\infty}(X_n = s, Y_n = a \mid X_1 = k)\\
=&\sum_{k\in S}\alpha(k)\sum_{n=1}^{\infty}\lambda^{n}P_{d^\infty}(X_{n+1} = j \mid X_1 = k) \text{ [by the fact } X_n \text{ is a Markovian Process]}\\
=&\sum_{k\in S}\alpha(k)\left(\sum_{n=2}^{\infty}\lambda^{n-1}P_{d^\infty}(X_n = j \mid X_1 = k) + P(X_1 = j \mid X_1 = k) - P(X_1 = j \mid X_1 = k)\right)\\
=&\sum_{k\in S}\alpha(k)\left(\sum_{n=1}^{\infty}\lambda^{n-1}P_{d^\infty}(X_n = j \mid X_1 = k) - \mathbb{1}\{j = k\}\right)\\
=&\sum_{k\in S}\alpha(k)\sum_{n=1}^{\infty}\lambda^{n-1}P_{d^\infty}(X_n = j \mid X_1 = k) - \alpha(j)\\
=&\sum_{a\in A_j} x_d(j,a) - \alpha(j)
\end{aligned}$$

(2) Let $x(s,a)$ be a feasible solution to the dual LP. Define

$$u(j) := \sum_{a\in A_j} x(j,a).$$

Then

$$
\begin{aligned}
& u(j) - \lambda \sum_{s \in S} \sum_{a \in A_s} P(j \mid s, a) x(s, a) = \alpha(j) \\
\iff & u(j) - \lambda \sum_{s \in S} \sum_{a \in A_s} P(j \mid s, a) x(s, a) \frac{u(s)}{\sum_{a \in A_s} x(s, a)} = \alpha(j) \\
\iff & u(j) - \lambda \sum_{s \in S} \sum_{a \in A_s} P(j \mid s, a) u(s) P(d_x(s) = a) = \alpha(j) \\
\iff & u(j) - \lambda \sum_{s \in S} P_{d_x}(j \mid s) u(s) = \alpha(j) \\
\iff & u^\top [I - \lambda P_{d_x}] = \alpha^\top \\
\iff & u^\top = \alpha^\top [I - \lambda P_{d_x}]^{-1} = \alpha^\top \left(\sum_{n=1}^{\infty} (\lambda P_{d_x})^{n-1} \right).
\end{aligned}
$$

We can then write:

$$
\begin{aligned}
u(s) &= \sum_{k \in S} \alpha(k) \sum_{n=1}^{\infty} \lambda^{n-1} \sum_{a \in A_s} P_{d_x}(X_n = s, Y_n = a \mid X_1 = k) \\
&= \sum_{a \in A_s} \sum_{k \in S} \alpha(k) \sum_{n=1}^{\infty} \lambda^{n-1} P_{d^\infty}(X_n = s, Y_n = a \mid X_1 = k) \\
&= \sum_{a \in A_s} x_{d_x}(s, a) = \sum_{a \in A_j} x(j, a). \\
x_{d_x}(s) &= \sum_{j \in S} \alpha(j) \sum_{n=1}^{\infty} \lambda^{n-1} P_{d_x}(x_n = s, Y_n = a \mid x_1 = j) \\
&= \sum_{j \in S} \alpha(j) \sum_{n=1}^{\infty} \lambda^{n-1} P_{d_x}(x_n = s \mid X_1 = j) P(d_x(s) = a) \\
&= \sum_{j \in S} \alpha(j) \sum_{n=1}^{\infty} \lambda^{n-1} P_{d_x}(x_n = s \mid X_1 = j) P(d_x(s) = a) \\
&= \underbrace{\sum_{j \in S} \alpha(j) \sum_{n=1}^{\infty} \lambda^{n-1} P_{d_x}(x_n = s \mid X_1 = j)}_{\sum_{a \in A_s} x_{d_x}(s, a)} \frac{x(s, a)}{\sum_{a \in A_s} x(s, a)} \\
&= x(s, a) \frac{\sum_{a \in A_s} x_{d_x}(s, a)}{\sum_{a \in A_s} x(s, a)} \\
&= x(s, a)
\end{aligned}
$$

□

Note (From MDP textbook by Martin L. Puterman). Since by definition, different d constructs different $x_d(s, a)$. The above theorem tells us: whenever I have a policy stationary randomized policy d, I can construct an x, and this x can be used to construct another d_x, while this d_x constructs x. Then, since both d_x and d are mapped to the same x, $d_x = d$. Similarly, if two x, x' are mapped to the same d, they are equal (directly from (2) of the above theorem). That is, the mappings we have above are *one-to-one* mappings.

Now since $\alpha(s) > 0 \; \forall s \in S$, without loss of generality, we may assume $\sum_{s \in S} \alpha(s) = 1$. Then, we can consider the $x(s, a)$ defined in (1) of the above theorem as the total discounted joint probability under initial-state distribution

$\{\alpha(j)\}$ that the system visits s and choose action a. To be more specific, if we consider $r(s,a)x(s,a)$, we get

$$\sum_{j\in S}\alpha(j)\sum_{n=1}^{\infty}\lambda^{n-1}r(s,a)P_{d^\infty}(X_n=s,Y_n=a\mid X_1=j),$$

and if we sum over all a and s, then this is

$$\sum_{j\in S}\alpha(j)v_\lambda^{d_x^\infty}(j),$$

where $v_\lambda^{d_x^\infty}(j)$ is the expected discounted reward starting at j. Thus we have

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

which is the expected total discounted reward under policy d_x^∞. Combining with the theorem above, we know for any policy d,

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

Proposition 2.19

(1) Let x be a basic feasible solution to the dual LP, then d_x is a deterministic Markovian decision rule.

(2) Suppose that d is a Markovian deterministic decision rule, then x_d is a basic feasible solution to the dual LP.

Proof.

(1) Since x is a BFS, and $\sum_{a\in A_s}x(s,a)>0$, for each s, there is exactly one $a_s\in A_s$ such that $x(s,a_s)>0$, otherwise, if it has two positive entries, you can perturb them to get two feasible solution such that x is in their convex hull. Then

$$P(d_x(s)=a_s)=1,$$
$$P(d_x(s)=a)=0\ \forall a\in A_s\setminus\{a_s\}.$$

(2) Suppose $x_d(s,a)$ is feasible but not BFS. Then there exists distinct feasible $w(s,a)$, $z(s,a)$, and $0<\beta<1$ such that

$$x_d(s,a)=\beta w(s,a)+(1-\beta)z(s,a).$$

Notice that $\sum_{a\in A_s}w(s,a)>0,\sum_{a\in A_s}z(s,a)>0$. If at least one of them has two nonzero entries, we have $w(s,a)>0$, $z(s,a')>0$ for $a\neq a'$, and hence $P(d_x(s)=a)>0$ and $P(d_x(s)=a')>0$, so d_x is not deterministic. If they both have exactly one nonzero entry, then $x_d(s,a)$ has exactly one nonzero entry for this s, it is a BFS or same to the previous case, it is not deterministic, we are done.

□

Theorem 2.20

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

(2) Suppose that x^* is an optimal solution to the dual LP. Then $d_{x^*}^\infty$ is an optimal policy.

(3) Suppose that x^* is a basic optimal solution to the dual LP. Then $d_{x^*}^\infty$ is a deterministic optimal policy.

(4) Suppose $d^{*\infty}$ is an optimal policy. Then x_{d^*} is an optimal solution to the dual LP.

(5) Suppose $d^{*\infty}$ is a deterministic optimal policy. Then x_{d^*} is a basic optimal solution to the dual LP.

Proof.

(1) Notice that since S, A_s are finite, we always have a feasible solution for the primal, so the dual always has an optimal solution.

(2) Let v^* be an optimal solution of the primal, notice that by the constraints of the primal, $v^* \geq Lv^*$, so $v^* \geq v_\lambda^*$.

$$\begin{aligned}
\sum_{s\in S}\alpha(s)v^*(s) &= \sum_{s\in S}\sum_{a\in A_s} r(s,a)x^*(s,a) \\
&= \sum_{s\in S}\left(\sum_{a\in A_s} x^*(s,a)\right)\left(\sum_{a\in A_s} r(s,a)P(d_{x^*}(s)=a)\right) \\
&= u^\top r_{d^\infty_{x^*}} \\
&= \alpha^\top [I-\lambda P_{d_{x^*}}]^{-1} r_{d^\infty_{x^*}} \\
&= \sum_{s\in S}\alpha(s)v_\lambda^{d^\infty_{x^*}}(s),
\end{aligned}$$

where $u(s) := \left(\sum_{a\in A_s} x^*(s,a)\right)$ and $u^\top = \alpha^\top[I-\lambda P_{d_{x^*}}]^{-1}$. Then by definition $v_\lambda^* \geq v_\lambda^{d_*^\infty}$, and $\alpha(s) > 0$,

$$\sum_{s\in S}\alpha(s)v^*(s) = \sum_{s\in S}\alpha(s)v_\lambda^*(s) = \sum_{s\in S}\alpha(s)v_\lambda^{d_*^\infty}(s) \implies v^*(s) = v_\lambda^*(s) = v_\lambda^{d_*^\infty}(s)\ \forall s\in S,$$

so $d_{x^*}^\infty$ is optimal.

(3) Follows from (2) and the previous proposition.

(4) Let x be an arbitrary feasible solution to the dual LP and build policy d such that $x(s,a) = x_d(s,a)$. Also, let x_{d^*} be the feasible solution built from d^*.

$$\begin{aligned}
\sum_{s\in S}\sum_{a\in A_s} r(s,a)x_{d^*}(s,a) &= \sum_{s\in S}\alpha(s)v_\lambda^{(d^*)^\infty} \\
&\geq \sum_{s\in S}\alpha(s)v_\lambda^{d^\infty} \\
&= \sum_{s\in S}\sum_{a\in A_s} r(s,a)x_d(s,a) \\
&= \sum_{s\in S}\sum_{a\in A_s} r(s,a)x(s,a).
\end{aligned}$$

Since x is arbitrary, x_{d^*} is optimal.

□

Proposition 2.21

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

Proof. Let x^* be the optimal basis so $x^*(s,a) > 0$ for only one $a \in A_s$. From sensitivity analysis, changing α only affects feasibility but not optimality of the basis. Hence, we show that the basis corresponding to x^* remains feasible as long as α is positive. We let x^* be the part corresponding to positive entries, then

$$(x^*)^\top(I-\lambda P_{d_{x^*}}) = \alpha^\top \iff x^* = (I-\lambda P_{d_{x^*}})^{-1}\alpha^\top \geq \alpha^\top > 0.$$

Not that the value of x^* might change as α changes, but the positive entries' positions are not, so the basis stays the same, that is, d_{x^*} does not changes as it's a deterministic policy choosing the unique action a with $x^*(s,a) > 0$ for each s. □

Example 2.3. Consider our previous example again:

$$S = \{s_1, s_2\}, A_{s_1} = \{a_{11}, a_{12}\}, A_{s_2} = \{a_{21}\}$$

and

$$\begin{aligned} p(s_1 \mid s_1, a_{11}) &= \frac{1}{2} \\ p(s_2 \mid s_1, a_{11}) &= \frac{1}{2} \\ p(s_2 \mid s_1, a_{12}) &= 1 \\ p(s_2 \mid s_2, a_{21}) &= 1 \end{aligned}$$

and $\lambda = 0.95$. We also have

$$r(s_1, a_{11}) = 5, r(s_1, a_{12}) = 10, r(s_2, a_{21}) = -1.$$

The primal LP formulation, with $\alpha(s_1) = \alpha(s_2) = \frac{1}{2}$, is

$$\begin{aligned} \min_v \ & \frac{1}{2} v(s_1) + \frac{1}{2} v(s_2) \\ \text{s.t. } & v(s_1) - 0.95\left[0.5v(s_1) + 0.5v(s_2)\right] \geq 5 \\ & v(s_1) - 0.95v(s_2) \geq 10 \\ & v(s_2) - 0.95v(s_2) \geq -1 \end{aligned}$$

and the dual LP is

$$\begin{aligned} \max \quad & 5x(s_1, a_{11}) + 10x(s_1, a_{12}) - x(s_2, a_{21}) \\ \text{s.t.} \quad & x(s_1, a_{11}) + x(s_1, a_{12}) - 0.95\left[0.5x(s_1, a_{11})\right] = \tfrac{1}{2} \\ & x(s_2, a_{21}) - 0.95\left[0.5x(s_1, a_{11}) + x(s_1, a_{12}) + x(s_2, a_{21})\right] = \tfrac{1}{2} \\ & x(s_1, a_{11}) \geq 0 \\ & x(s_1, a_{12}) \geq 0 \\ & x(s_2, a_{21}) \geq 0 \end{aligned}$$

and the dual LP can be solved to get the optimal solution

$$\begin{aligned} x^*(s_1, s_{11}) &= 0.9523 \\ x^*(s_1, s_{12}) &= 0 \\ x^*(s_2, s_{21}) &= 19.0476 \end{aligned}$$

2.4 Action Elimination

Proposition 2.22

If for $a' \in A_s, r(s, a') + \lambda \sum_{j \in S} p(j \mid s, a) v_\lambda^*(j) < v_\lambda^*(s)$ *then*

$$a' \notin \underset{a \in A_*}{argmax} \left\{ r(s, a) + \lambda \sum_{j \in S} p(j \mid s, a) v_\lambda^*(j) \right\}$$

Proof. We know

$$v_\lambda^*(s) = \max_{a \in A,} \left\{ r(s, a) + \lambda \sum p(j \mid s, a) v_\lambda^*(j) \right\}$$

but we have

$$r\left(s, a^{'}\right) + \lambda \sum_{j \in S} p(j \mid s, a') v_\lambda^*(j) < v_\lambda^*(s)$$

Clearly a' cannot be optimal in state s. □

Proposition 2.23

Suppose there exists v^L and v^U such that $v^L \leq v^*_\lambda \leq v^U$. Then if for $a' \in A_s$,

$$r(s,a') + \lambda \sum_{j \in S} p(j \mid s,a) v^u(j) < v^L(s)$$

any stationary policy that uses α' in state s cannot be optimal.

Proof.

$$\begin{aligned} & r(s,a') + \lambda \sum_{j \in S} P(j \mid s,a') v^*_\lambda(j) \\ \leq & r(s,a') + \lambda \sum_{j \in S} P(j \mid s,a') v^u_\lambda(j) \\ < & v^L(s) \leq v^*_\lambda(s) \end{aligned}$$

so a' is not optimal from the previous result. □

Theorem 2.24

Let V^σ be the set of structured values and D^σ be the set of structured decision rules. Suppose that for all v there exists $L_d v = Lv$ and $\|r_d\| \leq M < \infty$ for all d and that

(a) $v \in V^\sigma$ implies that $Lv \in V^\sigma$

(b) $v \in V^\sigma$ implies that there exists $d' \in D^\sigma \cap \arg\max_d L_d v$

(c) for any convergent sequence $\{v^n\} \subseteq V^\sigma, \lim_{n\to\infty} v^n \in V^\sigma$.

There exists an optimal stationary policy $(d^*)^\infty$ where $d^* \in D^\sigma$.

Proof. Choose $v^0 \in V^\sigma$ and set $v^n = Lv^{n-1}$. Then from (a) we know that $v^n \in V^\sigma$ for all $n \in \mathbb{N}$. But from (c) we know that $v^n \to v^*_\lambda \in V^\sigma$. Finally, from (b) we have the existence of $d^* \in D^\sigma$ and

$$d^* \in D^\sigma \cap arg\max_d L_d v^*_\lambda.$$

□

Theorem 2.25

Consider $S = \{0, 1, \ldots\}, A_s = A$ for all $s \in S$. If

1. *$r(s,a)$ is non-decreasing in s for all $a \in A$,*
2. *$\sum_{j=k}^\infty p(j \mid s,a)$ is non-decreasing in s for all $k \in S$ and $a \in A$,*
3. *$r(s,a)$ is super(sub)additive on $S \times A$, and*
4. *$\sum_{j=k}^\infty p(j \mid s,a)$ is super(sub)additive on $S \times A$,*

then there exists an optimal stationary policy $(d^*)^\infty$ for which $d^*(s)$ is non-de(in)creasing in s.

Proof. Let us define

$$\begin{aligned} V^\sigma &= \{v : v(s) \text{ is non-decreasing in } s\} \\ D^\sigma &= \{d : d(s) \text{ is non-decreasing in } s\} \end{aligned}$$

and let $v^0 = 0$. Then $v^1(s) = \max_{a \in A_s, }\{r(s,a)\} \implies v^1 \in V^\sigma$. Assume that $v^n \in V^\sigma$. We will show that $v^{n+1} \in V^\sigma$. We have

$$\begin{aligned} v^{n+1}(s) &= \max_{a \in A_s} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j \mid s,a) v^n(j) \right\} \\ &= r\left(s, a_s^*\right) + \lambda \sum_{j \in S} p\left(j \mid s, a_s^*\right) v^n(j) \end{aligned}$$

and suppose that $s' \geq s$. Then

$$\begin{aligned} v^{n+1}(s) &= r\left(s, a_s^*\right) + \lambda \sum_{j \in S} p\left(j \mid s, a_n^*\right) v^n(j) \\ &\leq r\left(s', a_s^*\right) + \lambda \sum_{j \in S} p\left(j \mid s', a_s^*\right) v^n(j) \\ &\leq \max_{a \in A_i} \left\{ r\left(s', a\right) + \lambda \sum_{j \in S} p\left(j \mid s', a\right) v^n(j) \right\} \\ &= v^{n+1}\left(s'\right) \end{aligned}$$

Thus, $\{v^n\} \in V^\sigma$ and $v^n \to v_\lambda^* \in V^\sigma$ by the fact that pointwise limit of a nondecreasing vector is nondecreasing. Suppose that $v \in V^\sigma$. Does there exist a $d \in D^\sigma$? For $s^-, s^+$and $a^- \leq a^+$we have

$$\sum_{j=0}^{\infty} \left[p\left(j \mid s^+, a^+\right) + p\left(j \mid s^-, a^-\right) \right] v(j) \geq \sum_{j=0}^{\infty} \left[p\left(j \mid s^+, a^-\right) + p\left(j \mid s^-, a^+\right) \right] v(j)$$

and so

$$r(s,a) + \lambda \sum_{j=0}^{\infty} p(j \mid s,a) v(j)$$

is superadditive. Hence, there must exist a decision rule

$$d(s) \in \underset{a \in A}{argmax} \left\{ r(s,a) + \lambda \sum_{j=0}^{\infty} p(j \mid s,a) v(j) \right\} \cap D^\sigma$$

which is non-decreasing in s from the finite case theorem. That is, since

$$r(s,a) + \lambda \sum_{j=0}^{\infty} p(j \mid s,a) v(j)$$

is superadditive, we can always pick the largest a attaining the maximum, which gives a non-decreasing d by Lemma 1.12. □

Theorem 2.26

Consider $S = \{0, 1, \ldots\}$, $A_s = A$ for all $s \in S$. If

1. *$r(s,a)$ is non-increasing in s for all $a \in A$,*
2. *$\sum_{j=k}^{\infty} p(j \mid s,a)$ is non-decreasing in s for all $k \in S$ and $a \in A$,*
3. *$r(s,a)$ is superadditive on $S \times A$, and*
4. *$\sum_{j=k}^{\infty} p(j \mid s,a) u(j)$ is superadditive on $S \times A$ for non-increasing u,*

then there exists an optimal stationary policy $(d^*)^\infty$ for which $d^*(s)$ is non-decreasing in s.

Proof. Similar to the above, notice how Lemma 1.14 is used. □

Monotone Policy Iteration Suppose $S = \{0, \ldots, K\}$.

1. Choose d_0 which is monotone non-decreasing in S. Set $n = 0$.
2. Find v^n by solving $(I - \lambda P_{d_n})v = r_{d_n}$.
3. Set $s = 0$ and $A'_0 = A'$.
   (a) Set
   $$A'^*_s = \arg\max_{a \in A_s} \left\{ r(s,a) + \lambda \sum_{j \in S} p_t(j \mid s, a) v^n(j) \right\}$$
   (b) If $s = K$ go to step 3d), otherwise set
   $$A'_{s+1} = \{a \in A'_s : a \geq \max\{a' \in A^*_s\}\}$$
   (c) Substitute $s + 1$ for s and return to 3a).
   (d) Pick $d^{(s)}_{n+1} \in A^*_s$ setting $d_{n+1} = d_n$ if possible.
4. If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise, substitute $n + 1$ for n and go to step 2

3 Long-Run Average Reward Optimality

3.1 Long-Run Average Reward

Let $\pi = (d_1, d_2, d_3, \ldots)$, $\{X_t : t \geq 0\}$ be the underlying Markov Chain. Recall that if $r_{N+1}(s) = 0$,

$$v_{N+1}^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{N} r\left(X_t, Y_t\right) \middle| X_1 = s\right].$$

Definition 3.1

If $\pi = (d, d, \ldots)$, define the long-run average reward *(gain) under policy π starting from s,*

$$g^{\pi}(s) = \lim_{N\to\infty} \frac{1}{N} v_{N+1}^{\pi}(s) = \lim_{N\to\infty} r_d + P_d r_d + P_d^2 r_d + \ldots + P_d^{N-1} r_d = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P^{n-1} r_{d_n}(s)$$

and we also define

$$P^* = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n$$

Hence, g^{π} exists when $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P^{n-1}$ exists.

Lemma 3.2

Suppose $\lim_{n\to\infty} a_n = a^*$, then the Cesaro limit: $\lim_{N\to\infty} \frac{a_1+\ldots+a_N}{N} = a^*$ but $\lim_{n\to\infty} a_n$ might not exist while the Cesaro limit does.

So by the lemma above, if $\lim_{n\to\infty} P_d^n$ exists, $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P^{n-1}$ exists. If the stationary distribution exists, $\lim_{n\to\infty} P_d^n$ exists. If the Markov Chain has finitely many states, then $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P^{n-1}$ is the long-run time spent in each state.

Proposition 3.3

If S is finite, then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P_d^{n-1} = P_d^*$$

always exists for any d^{∞}, and we have

$$g^{d^{\infty}}(s) = P_d^* r_d(s).$$

Definition 3.4

Define

$$g_+^\pi(s) = \limsup_{N\to\infty} \frac{1}{N} v_{N+1}^*(s)$$
$$g_-^\pi(s) = \liminf_{N\to\infty} \frac{1}{N} v_{N+1}^*(s)$$

A policy π^* is long-run average optimal *if*

$$g_-^{\pi^*}(s) \geq g_+^\pi(s) \text{ for all } \pi$$

A policy π^* is limsup optimal *if*

$$g_+^{\pi^*}(s) \geq g_+^\pi(s) \text{ for all } \pi$$

A policy π^* is liminf optimal *if*

$$g_-^{\pi^*}(s) \geq g_-^\pi(s) \text{ for all } \pi$$

Proposition 3.5

Let S be countable. Let d^∞ be a stationary Markovian randomized policy and suppose that P_d^* exists, then $g^{d^\infty}(s) = P_d^* r_d(s)$.

Definition 3.6

Let P denote the probability transition matrix of a Markov chain $\{X_t : t = 1, 2, \ldots\}$ and $r(s)$ a reward function for each $s \in S$. We refer to the bivariate stochastic process $\{(X_t, r(X_t)) : t = 1, 2, \ldots\}$ as a Markov reward process.

Remark. If P^* exists,

$$g(s) = [P^* r](s) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P^{n-1} r(s)$$

Proposition 3.7

Suppose that P^* exists. If j and k are in the same irreducible class, $g^\pi(j) = g^\pi(k)$. Furthermore, if the Markov chain is irreducible or unichain (i.e. a single recurrent class plus some transient states), then $g^\pi(s)$ is a constant function.

Claim. If $n \geq 1$,$(P - P^*)^n = P^n - P^*$

Proof. When $n = 1$, it's trivial. We do induction on n. Then

$$\begin{aligned}(P - P^*)^{n+1} &= (P^n - P^*)(P - P^*) \\ &= P^{n+1} - P^n P^* - P^* P + P^* P^* \\ &= P^{n+1} - 2P^* + P^* \\ &= P^{n+1} - P^*,\end{aligned}$$

where we know that $PP^* = P^*$. □

Definition 3.8

The bias vector *is defined as*

$$h = (I - P + P^*)^{-1}(I - P^*)\, r.$$

Note that

$$(P^n - P^*)(I - P^*) = P^n - P^* - (P^n - P^*)P^* = P^n - P^* - (P^* - P^*) = P^n - P^*,$$

and

$$(I - P + P^*)^{-1}(I - P^*) = \sum_{n=0}^{\infty}(P^n - P^*)$$

from the fact that

$$(I - P + P^*)^{-1} = \sum_{n=0}^{\infty}(P - P^*)^n = I + \sum_{n=1}^{\infty}(P^n - P^*)$$

and hence

$$\begin{aligned}(I - P + P^*)^{-1}(I - P^*) &= (I - P^*) + \sum_{n=1}^{\infty}(P^n - P^*)(I - P^*)\\ &= (I - P^*) + \sum_{n=1}^{\infty}(P^n - P^*)\\ &= \sum_{n=0}^{\infty}(P^n - P^*)\end{aligned}$$

Therefore, the bias function, can be expressed as

$$h = (I - P + P^*)^{-1}(I - P^*)\, r = \sum_{n=0}^{\infty}(P^n r - P^* r) = \sum_{n=0}^{\infty}(P^n r - g)$$

and we can interpret

$$h(s) = \mathbb{E}_s\left[\sum_{t=1}^{\infty}\left(r(S_t) - g(X_t)\right)\right].$$

In fact, we can also interpret h as capturing the transition behavior of the Markov Chain. If we write $h = \sum_{n=0}^{\infty}(P^n - p^*)r$, it measures the performance of this policy before it reaches stationary.

Remark. Note that since $v_{N+1} = \sum_{t=1}^{N} P^{t-1}r$ then

$$\begin{aligned}h &= \sum_{t=1}^{\infty}\left(P^{t-1}r - g\right)\\ &= \sum_{t=1}^{N}\left(P^{t-1}r - g\right) + \sum_{t=N+1}^{\infty}\left(P^{t-1}r - g\right)\\ &= \sum_{t=1}^{N}P^{t-1}r - Ng + \sum_{t=N+1}^{\infty}\left(P^{t-1} - P^*\right)r\\ &= v_{N+1} - Ng + o(1)\end{aligned}$$

and hence

$$v_{N+1}(s) = h(s) + Ng(s) + o(1)$$

and as $N \to \infty$ we have $v_{N+1}(s) \to h(s) + Ng(s)$. Now suppose that states j and k belong to the same recurrent class. Then, $g(j) = g(k)$ which implies

$$\lim_{N\to\infty}\left[v_{N+1}(j) - v_{N+1}(k)\right] = h(j) - h(k)$$

which is why the bias h is also called the relative value function.

Theorem 3.9

Let S be finite and let g and h denote the gain and bias vectors of a Markov Reward process with transition matrix P and reward vector r. Then

(a) $(I-P)g = 0$ *and* $g + (I-P)h = r$

(b) Suppose that g and h satisfy $(I-P)g = 0$ and $g + (I-P)h = r$. Then $g = P^*r$ and

$$h = \left(I - P + P^*\right)^{-1}\left(I - P^*\right)r + u$$

where $(I-P)u = 0$.

(c) Suppose g and h satisfy the equations in (a) and $P^*h = 0$, then

$$h = (I - P + P^*)^{-1}(I - P^*)r.$$

With (b) and (c) above, we know that h computed in (b) is a "shifted" bias while the one in (c) is the true one.

Proof. (a) Directly $(I-P)P^*r = \left(P^* - P^*\right)r = 0$ and

$$\begin{aligned}
&g + (I-P)h \\
=&P^*r + (I-P)\left(I - P + P^*\right)^{-1}\left(I - P^*\right)r \\
=&P^*r + (I-P)\sum_{n=0}^{\infty}\left(P^n - P^*\right)r \\
=&P^*r + \sum_{n=0}^{\infty}\left(P^n - P^* - P^{n+1} + P^*\right)r \\
=&P^*r + \sum_{n=0}^{\infty}\left(P^n - P^{n+1}\right)r \\
=&P^*r + \left(I - P^*\right)r \\
=&r
\end{aligned}$$

(b) We first note that adding the first equation plus P^* times the second equation. By $P^*(I-P) = P^* - P^*) = 0$, it gives us

$$\begin{aligned}
&P^*g + g - Pg = P^*r \\
\Longrightarrow&\left(I - P + P^*\right)g = P^*r \\
\Longrightarrow&g = \left(I - P + P^*\right)^{-1}P^*r \\
\Longrightarrow&g = \left[I + \sum_{n=1}^{\infty}\left(P^n - P^*\right)\right]r \\
\Longrightarrow&g = P^*r,
\end{aligned}$$

where $[I - (P - P^*)]^{-1} = \sum_{n=0}^{\infty}(P - P^*)^n$.
In part (a), we have shown that $h = \left(I - P + P^*\right)^{-1}\left(I - P^*\right)r$ satisfies $g + (I-P)h = r$. Suppose that h' is another vector satisfying $g + (I-P)h' = r$. Then

$$g + (I-P)h = r \text{ and } g + (I-P)h' = r$$

implies that

$$(I-P)\underbrace{(h - h')}_{-u} = 0.$$

(c) Given $g + (I - P)h = r, P^*h = 0$, we have

$$P^*r + (I - P)h = r, P^*h = 0,$$

which implies

$$P^*r + (I - P + P^*)h = r \implies (I - P + P^*)h = (I - P^*)r \implies h = (I - P + P^*)^{-1}(I - P^*)r.$$

□

Remark. Note that if g is a constant vector, then since P is a probability matrix, then $(I - P)g = 0$ trivially

Corollary 3.10

Suppose P is unichain. Then the long-run average reward $P^*r = ge$ where $g \in \mathbb{R}$ is a constant, and it is uniquely determined by solving

$$ge + (I - P)h = r,$$

where the other equation $(I - P)ge = 0$ is redundant. Furthermore, if $P^*h = 0$ then $h = (I - P + P^*)^{-1}(I - P^*)r$.

Proof. Suppose g and h satisfy the above equation. Then by the previous theorem, $P^*r = ge$ and

$$h = (I - P + P^*)^{-1}(I - P^*)r + ke$$

for any scalar k. Furthermore, as $P^*h = 0$ then $h = (I - P + P^*)^{-1}(I - P^*)$ r. □

Lemma 3.11: Laurent Series Expansion

For $0 < \lambda < 1$, $\rho = \frac{1-\lambda}{\lambda} \implies \lambda = \frac{1}{1+\rho}$. Then the infinite horizon expected value is

$$v_\lambda = \frac{1+\rho}{\rho}g + (1+\rho)h + (1+\rho)\sum_{n=1}^{\infty}\rho^n y_n,$$

for some y_n, which is equivalent to

$$v_\lambda = \frac{1}{1-\lambda}g + \frac{h}{\lambda} + (1-\lambda)\sum_{n=1}^{\infty}\frac{1}{\lambda^n}y_n$$

$$(1-\lambda)v_\lambda = g + \frac{1-\lambda}{\lambda}h + (1-\lambda)^2\sum_{n=1}^{\infty}\frac{1}{\lambda^n}y_n$$

Proposition 3.12

Let g and h represent the gain and bias of a Markov Reward process with finite state space S. Then,

$$v_\lambda = \frac{g}{L\lambda} + h/\lambda + f(\lambda)$$

where $f(\lambda)$ is a vector whose components converge to 0 as $\lambda \uparrow 1$.

Apply the results above,

Corollary 3.13

We have

$$\lim_{\lambda\uparrow 1}(1-\lambda)v_\lambda = g$$

3.2 Classification of MDPs

(a) Recurrent: if the transition matrix corresponding to every stationary deterministic policy yields an irreducible Markov chain.

(b) Unichain: if the transition matrix corresponding to every stationary deterministic policy yields a single recurrent class plus some (possibly none) transient states.

(c) Communicating: if for every pair of states s and j there exists a deterministic stationary policy under which j is accessible from s, that is $p_d^n(s \mid j) > 0$ for some $n \geq 1$.

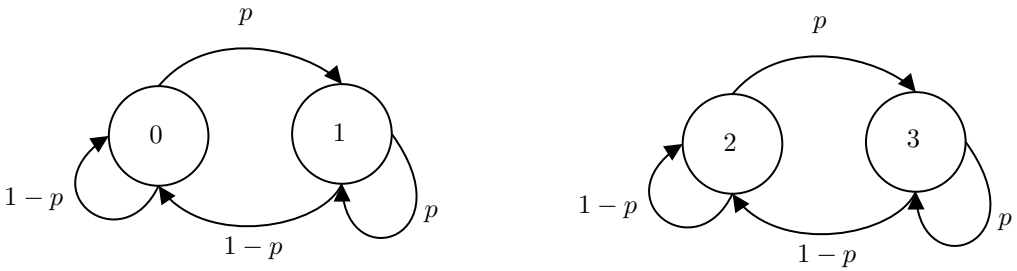
(d) Weakly communicating: if there exists a closed set of states which is a recurrent class under some deterministic stationary policy, plus (possibly empty) set of transient states which is transient under every policy.

(e) Multichain: if there exists a policy under which Markov Chain has multiple recurrent classes.

Example 3.1 (Inventory problem revisited). Suppose the warehouse has a capacity of 3 units. We are given

$$\begin{aligned} P(D_t = 0) &= p \\ P(D_t = 1) &= 1 - p \\ S &= \{0, 1, 2, 3\} \\ A_s &= \{0, 1, \ldots, 3 - s\} \\ d(0) &= 1 \\ d(1) &= 0 \\ d(2) &= 1 \\ d(3) &= 0 \end{aligned}$$

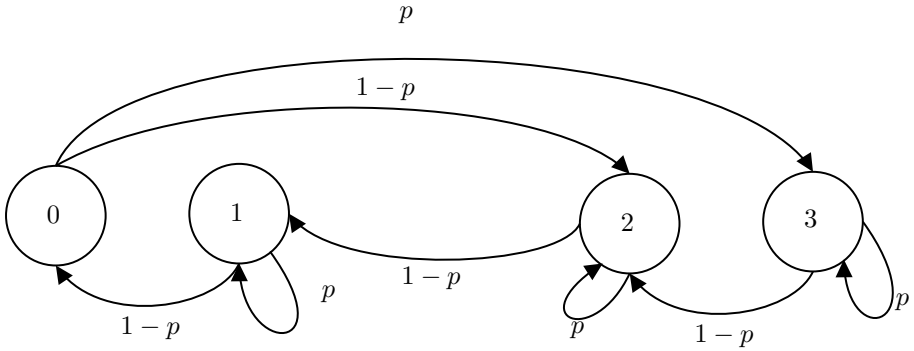
The transition plot is below:



Consider a separate policy

$$\begin{aligned} \delta(0) &= 3 \\ \delta(1) &= 0 \\ \delta(2) &= 0 \\ \delta(3) &= 0. \end{aligned}$$

The transition plot is below:



These two policies, d and δ, imply this is a communicating and multichain MDP.

Example 3.2. Given $S = \{s_1, s_2\}$ and $A_{s_1} = \{a_{11}, a_{12}\}, A_{s_2} = \{a_{21}\}$, define

$$\begin{aligned} p(s_1 \mid s_1, a_{11}) &= 1 \\ p(s_2 \mid s_1, a_{12}) &= 1 \\ p(s_2 \mid s_2, a_{21}) &= 1 \end{aligned}$$

and $d(s_1) = a_{11}, d(s_2) = a_{21}, \delta(s_1) = a_{12}, \delta(s_2) = a_{21}$ and the policies d and δ imply that this is multichain. Note that this is not a weakly-communicating because s_1 is not transient under every policy and even though δ gives a single recurrent class.

Example 3.3. $S = \{s_1, s_2\}$, $A_{s_1} = \{a_{11}\}$, $A_{s_2} = \{a_{21}, a_{22}\}$.

$$P(s_2 \mid s_1, a_{11}) = 1, P(s_1 \mid s_2, a_{22}) = 1, P(s_2 \mid s_2, a_{21}) = 1.$$

Consider $d(s_1) = a_{11}, d(s_2) = a_{21}$ and $\delta(s_1) = a_{11}, \delta(s_2) = a_{22}$. Thus the MDP is unichain and communicating.

Example 3.4. $S = \{s_1, s_2\}$, $A_{s_1} = \{a_{11}, a_{12}\}$, $A_{s_2} = \{a_{21}, a_{22}\}$.

$$P(s_1 \mid s_1, a_{11}) = 1, P(s_2 \mid s_1, a_{12}) = 1, P(s_2 \mid s_2, a_{21}) = 1, P(s_1 \mid s_2, a_{22}) = 1.$$

$$\begin{aligned} d_1(s_1) = a_{11}, d_1(s_2) &= a_{21} \\ d_2(s_1) = a_{11}, d_2(s_2) &= a_{22} \\ d_3(s_1) = a_{12}, d_3(s_2) &= a_{21} \\ d_4(s_1) = a_{12}, d_4(s_2) &= a_{22}. \end{aligned}$$

Thus the MDP is multichain (by d_1) and communicating (by d_2, d_3, d_4).

Proposition 3.14

1. *A Markov decision process is communicating if and only if there exists a randomized stationary policy where the chain is irreducible.*
2. *A Markov decision process is weakly communicating if and only if there exists a randomized stationary policy under which the chain has a single recurrent set with some set of transient states where under any policy, these states must be transient.*

Theorem 3.15

Assume a weakly communicating model and let d be a Markovian deterministic decision rule.

(a) Let C be a recurrent class in the Markov Chain corresponding to d^∞. Then there exists a deterministic decision rule δ with $\delta(s) = d(s)$ for all $s \in C$ and for which the chain generated by d has C as its irreducible set.

(b) Suppose the stationary policy d^∞ has $g^{d^\infty}(s) < g^{d^\infty}(s')$ for some $s \in C, s' \in S$. Then there exists a stationary policy δ^∞ for which

$$g^{\delta^\infty}(s) = g^{\delta^\infty}(s') \geq g^{d^\infty}(s')$$

Proof. (a) Let T be the set of transient states that are transient under all policies. Then $\exists s_0 \in S \backslash (T \cup C)$ and $a'_{s_0} \in A_s$ such that

$$\sum_{j \in C} P(j \mid s_0, a_{s_0}) > 0.$$

If $S = T \cup C$, there exists $s_0 \in T$ should work too because at least one transient state should go to C. We then set $\delta(s_0) = a_{s_0}$ and augment $T \cup C$ with $T \cup C \cup s$ and continue in this fashion until $\delta(s)$ is defined for all $s \in S \backslash T$. By definition of T, there exists $s' \in T$ and $a_{s'} \in A_{s'}$ for which

$$\sum_{j \in S \backslash T} P\left(j \mid s', a_{s'}\right) > 0$$

We then set $\delta\left(s'\right) = a_{s'}$.

(b) If $s' \in C$ then the result follows from (a) with $g^{\delta^\infty}(s) = g^{\delta^\infty}\left(s'\right) = g^{d^\infty}\left(s'\right)$ by the g is constant in the same recurrent class. If s' is transient under d^∞ then there exists a recurrent state s'' for which

$$g^{d^\infty}\left(s''\right) \geq g^{d^\infty}\left(s'\right)$$

since essentially g^{d^∞} is a weighted average of all gains for recurrent states it can end up in. So there exists s'' which yields the largest gain. Then apply (a) when C is the closed set containing s''. We will get

$$g^{\delta^\infty}\left(s''\right) = g^{\delta^\infty}\left(s'\right) = g^{d^\infty}\left(s''\right) \geq g^{d^\infty}\left(s'\right).$$

□

Theorem 3.16

1. *Given a Markovian deterministic decision rule d_1 there exists a Markovian deterministic decision rule δ for which g^{δ^∞} is constant and $g^{\delta^\infty} \geq g$.*
2. *If there exists a stationary optimal policy, there exists a stationary optimal policy with constant gain.*

3.2.1 Unichain Markov Decision Processes

Remark. The Optimality Equations for Unichain MDPs are:

$$\begin{cases} \max_{a \in A_s} \left\{ r(s,a) - g + \sum_{j \in S} p(j \mid s,a) h(j) - h(s) \right\} & = 0 \\ \max_d \left\{ r_d - ge + \left(P_d - I\right) h \right\} & = 0 \\ g + (I - P)h & = r \end{cases}$$

where the first equation is equivalent to

$$\begin{aligned} g + h(s) &= \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j \mid s,a) h(j) \right\} \\ \iff \max_d \left\{ r_d - g^* e + (P_d - I) h \right\} &= 0. \end{aligned}$$

This is because, we know that

$$v_\lambda^* = \frac{1}{1-\lambda} g^* e + \frac{h}{\lambda} + f(\lambda) = \max_{d \in D} \left\{ r_d + \lambda P_d v_\lambda^* \right\}$$

which implies that

$$\begin{aligned} 0 &= \max_{d \in D} \left\{ r_d + \left(\lambda P_d - I\right) v_\lambda^* \right\} \\ &= \max_{d \in D} \left\{ r_d + \left(\lambda P_d - I\right) \left[\frac{1}{1-\lambda} g^* e + \frac{h}{\lambda} + f(\lambda) \right] \right\} \\ &= \max_{d \in D} \left\{ r_d + \left(\lambda P_d - I\right) \frac{1}{1-\lambda} g^* e + \left(\lambda P_d - I\right) \frac{h}{\lambda} + \left(\lambda P_d - I\right) f(\lambda) \right\} \\ &= \max_{d \in D} \left\{ r_d + \frac{\lambda - 1}{1-\lambda} g^* e + \left(\lambda P_d - I\right) \frac{h}{\lambda} + \left(\lambda P_d - I\right) f(\lambda) \right\} \end{aligned}$$

where the second last to the last line is by $\lambda P_d g e^* = \lambda g^* e$. And if we take $\lambda \uparrow 1$ then

$$0 = \max_d \{r_d - g^* e + (P_d - I)\, h\}$$

Alternatively, since

$$v_{N+1} = N g^* e + h + o(1)$$
$$v_N^* = (N-1) g^* e + h + o(1)$$

and

$$v_N^* = \max_{d \in D} \{r_d + \lambda P_d v_N^*\}$$

then

$$N g^* e + h + o(1) = \max_{d \in D} \{r_d + P_d\, ((N-1) g^* e + h + o(1))\}$$

and hence by $P_d e = e$,

$$0 = \max_{d \in D} \{r_d - g^* e + (P_d - I)\, h + o(1)\}$$

and as $N \to \infty, 0 = \max_{d \in D} \{r_d - g^* e + (P_d - I)\, h\}$.

Theorem 3.17

Suppose S is countable,

(a) If there exists a scalar g and a vector h which satisfy

$$\max_{d \in D} \{r_d - g + (P_d - I)\, h\} \leq 0$$

then $ge \geq g_+^*$.

(b) If there exists a scalar g and a vector h with

$$\max_{d \in D} \{r_d - g + (P_d - I)\, h\} \geq 0$$

then $ge \leq g_-^*$.

(c) If there exists a scalar g and a vector h with

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

then $ge = g_+^* = g_-^* = g^*$.

Proof. (a) We can write the condition as $\max_d \{r_d + P_d h\} \leq ge + h$, then we have

$$\begin{aligned}
ge + h &\geq r_d + P_d h \text{ for all } d \\
\pi &= (d_1, d_2, d_3, \ldots) \\
ge &\geq r_{d_2} + (P_{d_2} - I)h \implies ge = P_{d_1} ge \geq P_{d_1} r_{d_2} + P_{d_1}(P_{d_2} - I)h \\
ge &\geq P_{d_1} P_{d_2} r_{d_3} + P_{d_1} P_{d_2}(P_{d_3} - I)h \\
&\vdots \\
ge &\geq P_{d_1} P_{d_2} \ldots P_{d_{N-1}} r_{d_N} + P_{d_1} P_{d_2} \ldots P_{d_{N-1}}(P_{d_N} - I)h
\end{aligned}$$

Add the $ge \geq$ inequalities up, we get

$$\begin{aligned}
Nge \geq \big[&r_{d_1} + P_{d_1} r_{d_2} + \ldots + P_{d_1} P_{d_2} \ldots P_{d_{N-1}} r_{d_N} + (P_d - I)h\big] \\
&+ P_{d_1}(P_{d_2} - I)h + \ldots + P_{d_1} P_{d_2} \ldots P_{d_{N-1}}(P_{d_N} - I)h
\end{aligned}$$

Treat the terms in the square bracket as an expected reward with $r_{N+1} = 0$, then we get

$$ge \geq \frac{V_{N+1}^{\pi}}{N} + \underbrace{\frac{1}{N}\left((P_d - I)h + P_{d_1}(P_{d_2} - I)h + \ldots + P_{d_1}\ldots P_{d_{N-1}}(P_{d_N} - I)h\right)}_{(1)}$$

where (1) equals to $\frac{1}{N}(P_{d_1}P_{d_2}\ldots P_{d_N} - I)h$ and $\|P_{d_1}\ldots P_{d_N}h\| \leq \|h\| < \infty$, and so $\limsup_{N\to\infty}(1) = 0$. Thus

$$ge \geq \limsup_{N\to\infty} \frac{v_{N+1}^{\pi}}{N} \implies ge \geq g_+^*.$$

(b) By $\max_d\{r_d + P_d h\} \geq ge + h$, there exists d such that $ge \leq r_d + (P_d - I)h$. Let $\pi = d^\infty$. Using the argument in part (a), we have

$$ge \leq \liminf_{N\to\infty} \frac{v_{N+1}^{d^\infty}}{N} \leq g_-^*.$$

□

Theorem 3.18

Suppose S and A_s for each $s \in S$ are finite, and the model is unichain

(a) Then there exists a scalar g and a vector h for which

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

(b) If there exists any other solution (g', h') then $g = g'$.

Definition 3.19

A decision rule d_h is called h-improving if

$$d_h \in \underset{d}{argmax}\,\{r_d + P_d h\},$$

or equivalently

$$r_{d_h} + P_{d_h}h = \max_d\{r_d + P_d h\}.$$

is an optimal policy.

Theorem 3.20

Suppose scalar g^* and h vector satisfy the unichain optimality equations. Then if d^* is h-improving then $(d^*)^\infty$ is a long-run average optimal policy.

Proof. Note $0 = \max_d\{r_d - g^*e + (P_d - I)h\}$, so $d^* \in \arg\max\{r_d + P_d h\}$ implies $r_{d^*} + P_{d^*}h = \max_d\{r_d + P_d h\}$. Hence,

$$r_{d^*} + P_{d^*}h - g^*e - h = \max_d\{r_d - g^*e + (P_d - I)h\} = 0,$$

which implies $g^*e + (I - P_{d^*})h = r_{d^*}$ and $g^* = g^{(d^*)^\infty}$. □

Theorem 3.21

Suppose S and A_S for all $s \in S$ are finite, then

(a) There exists a stationary optimal policy.

(b) There exists g^* and h satisfying the optimal equation.

(c) Any stationary policy derived from h-improving decision rule is long-run reward optimal,

(d) $g^* e = g^*_+ = g^*_-$.

Value Iteration Algorithm Define:

$$\mathrm{sp}(v) = \max_s v(s) - \min_s v(s)$$

which is a semi-norm.

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

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

3. If $\mathrm{sp}\left(v^{n+1} - v^n\right) < \epsilon$ go to step 4; otherwise increment n by 1 and return to step 2.
4. For each $s \in S$ choose,

$$d_\epsilon(s) \in \underset{a \in A_s}{argmax} \left\{ r(s,a) + \sum_{j \in S} p(j \mid s,a) v^{n+1}(j) \right\}.$$

Theorem 3.22

Suppose S and A_S are finite for each $s \in S, r(s,a)$ is bounded and the model is unichain. Then for a vector v we have

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

where $d \in \mathrm{argmax}\{r_d + P_d v\}$.

Proof. For any v improving d, by definition,

$$\begin{aligned} g^{d^\infty} e = P_d^* r_d &= P_d^* [\underbrace{r_d + P_d v}_{Lv} - v] \\ &= P_d [Lv - v] \\ &\ge P_d \min_s (Lv(s) - v(s)) e \\ &= \min_s (Lv(s) - v(s)) e \end{aligned}$$

and

$$\min_{s \in S}(Lv(s) - v(s)) \le g^{d^\infty} \le g^*$$

We know that there exists a δ^∞ such that $g^{\delta^\infty} = g^*$. Hence

$$\begin{aligned} g^* e = g^{\delta^\infty} e = P_\delta^* r_\delta &= P_\delta^* [\underbrace{r_\delta + P_\delta v}_{\leq Lv} - v] \\ &\leq P_\delta^* [Lv - v] \\ &\leq P_\delta^* \max_{s \in S} [Lv(s) - v(s)] e \\ &= \max_{s \in S} [Lv(s) - v(s)] e \end{aligned}$$

□

Theorem 3.23

(i) d_ϵ^∞ *is an* ϵ*-optimal policy where*

$$d_\epsilon(s) \in \underset{a \in A_s}{argmax} \left\{ r(s,a) + \sum_{j \in S} p(j \mid s, a) v^{n+1}(j) \right\},$$

where $v^{n+1} = Lv^n$*. Then* $|g^* - g^{(d_\epsilon)^\infty}| \leq \epsilon$

(ii) Define $g' = \frac{1}{2}\left[\max_s \left(v^{n+1}(s) - v^n(s)\right) + \min_s \left(v^{n+1}(s) - v^n(s)\right)\right]$*. Then* $|g' - g^*| < \frac{\epsilon}{2}$ *and* $\left|g' - g^{(d_\epsilon)^\infty}\right| < \frac{\epsilon}{2}$.

Proof.

(i) We need $|g^* - g^{(d_\epsilon)^\infty}| < \epsilon$. Using the previous result

$$\min_{s \in S} \left(Lv^n(s) - v^n(s)\right) \leq g^{d_\epsilon^\infty} \leq g^* \leq \max_{s \in S} \left(Lv^n(s) - v^n(s)\right)$$

and $\mathrm{sp}(v^{n+1} - v^n) \leq \epsilon$.

(ii) Note that if $x \leq y \leq z$ and $z - x < \epsilon$ then

$$-\frac{\epsilon}{2} < \frac{1}{2}(x - z) \leq y - \frac{1}{2}(x + z) \leq \frac{1}{2}(z - x) \leq \frac{\epsilon}{2}.$$

We know that

$$\underbrace{\min_{s \in S} \left(v^{n+1}(s) - v^n(s)\right)}_{x} \leq \underbrace{g^{d_\epsilon^\infty}}_{y} \leq \underbrace{\max_{s \in S} \left(v^{n+1}(s) - v^n(s)\right)}_{z}$$

and so

$$\begin{aligned} -\frac{\epsilon}{2} < -\frac{1}{2}\left(sp\left(v^{n+1} - v^n\right)\right) \leq g^{d_\epsilon^\infty} &- \frac{1}{2}\underbrace{(\min_{s \in S}\left(v^{n+1}(s) - v^n(s)\right) + \max_{s \in S}\left(v^{n+1}(s) - v^n(s)\right))}_{g'} \\ &\leq \frac{1}{2}\left(sp\left(v^{n+1} - v^n\right)\right) \leq \frac{\epsilon}{2}. \end{aligned}$$

The same argument applies by the previous theorem and replacing $g^{(d_\epsilon^*)^\infty}$ by g^*.

□

Theorem 3.24

Suppose that all stationary policies yield unichain Markov chains and that every policy has an aperiodic Markov chain. Then the value iteration converges in a finite number of iterations.

An aperiodic transformation Choose $0 < \tau < 1$ and define

$$\begin{aligned}
\tilde{r}(s,a) &= \tau r(s,a) \\
\tilde{P}(j \mid s,a) &= (1-\tau)\mathbb{1}(j=s) + \tau p(j \mid s,a) \\
\sum_{j\in S} \tilde{P}(j \mid s,a) &= (1-\tau) + \sum_{j\in S} \tau P(j \mid s,a) \\
&= (1-\tau) + \tau \sum_{j\in S} P(j \mid s,a) \\
&= 1 - \tau + \tau = 1 \\
\tilde{r}_d &= \tau r_d \\
\tilde{P}_d &= (1-\tau)I + \tau P_d
\end{aligned}$$

Proposition 3.25

For any decision rule d,

$$\tilde{P}_d^* = P_d \text{ and } \tilde{g}^{d^\infty} = \tau g^{d^\infty}.$$

Proof. We need $P_d^* \tilde{P}_d = \tilde{P}_d P_d^* = P_d^*$. Directly,

$$\begin{aligned}
P_d^* \tilde{P}_d &= P_d^* \left((1-\tau)I + \tau P_d\right) \\
&= (1-\tau)P_d^* + \tau P_d^* P_d \\
&= P_d^* - \tau P_d^* + \tau P_d^* = P_d^*
\end{aligned}$$

and hence

$$\tilde{P}_d P_d^* = (1-\tau)P_d^* + \tau P_d P_d^* = P_d^*$$

Now

$$\tilde{g}_d = \tilde{P}_d \tilde{r}_d = \tilde{P}_d \tau r_d = \tau P_d^* r_d = \tau g^{d^\infty}$$

□

Corollary 3.26

The set of long-run average optimal stationary policies under the original and the transformed model are the same. That is, $\tilde{g}^* = \tau g^*$.

Policy Iteration for Unichain Models

1. Set $n = 0$ and select an arbitrary decision rule d_n.
2. (Policy evaluation) Obtain a scalar g_n and a vector h_n such that

$$r_{d_n} - g_n e + (P_d - I)\, h_n = 0.$$

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

$$d_{n+1} \in \underset{d}{argmax} \{r_d + P_d h_n\}$$

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

4. If $d_{n+1} = d_n$, stop and $d^* = d_n$; otherwise increment n by 1 and go to step 2 .

Doing Policy Evaluation

1. Choose h_n to satisfy $P^*_{d_n} h_n = 0$.
2. Pick a recurrent state s_0 under d_n and set $h_n(s_0) = 0$.
3. Choose h_n to satisfy

$$-h_n + (P_{d_n} - I)\, w = 0$$

for some vector w.

Proposition 3.27

Suppose that $d_{n+1} \in \operatorname{argmax}\{r_d + P_d h_n\}$. *Then,*

(a) $g_{n+1}e = g_n e + P^*_{d_{n+1}} \left[r_{d_{n+1}} - g_n e + \left(P_{d_{n+1}} - I\right) h_n\right]$

(b) If $\left[r_{d_{n+1}} - g_n e + \left(P_{d_{n+1}} - I\right) h_n\right](s) > 0$ *for some state* s *which is recurrent under* d_{n+1} *then* $g_{n+1} > g_n$.

(c) If $\left[r_{d_{n+1}} - g_n e + \left(P_{d_{n+1}} - I\right) h_n\right](s) = 0$ *for all* s *under* d_{n+1} *then* $g_{n+1} = g_n$.

Proof. (a) By computing d_{n+1} and optimality equation, $g_{n+1}e = P^*_{d_{n+1}} r_{d_{n+1}}$. Directly,

$$\begin{aligned} g_{n+1}e &= P^*_{d_{n+1}}[r_{d_{n+1}} - g_n e + g_n e] \\ &= g_n e + P^*_{d_{n+1}} \left[r_{d_{n+1}} - g_n e + \left(P_{d_{n+1}} - I\right) h_n\right] \end{aligned}$$

by $P^*_{d_{n+1}} g_n e = g_n e$ and $P^*_{d_{n+1}} P_{d_{n+1}} = P^*_{d_{n+1}} I$.

(b) Given the assumption, then $P^*_{d_{n+1}} \left[r_{d_{n+1}} - g_n e + \left(P_{d_{n+1}} - I\right) h_n\right](s) > 0$ for recurrent state.

(c) The vector above is zero when the state is transient.

□

Corollary 3.28

Suppose the Markov decision process is recurrent. Assume the set of states and actions are finite. Then the policy iteration converges monotonically in a finite number of iterations to a solution (g^*, h) *and average optimal solution policy* $(d^*)^\infty$.

Proposition 3.29

Suppose $d_{n+1} \in \arg\max_d \{r_d + P_d h_n\}$, *then*

$$h^{(d_{n+1})^\infty} = h^{(d_n)^\infty} - P^*_{d_{n+1}} h^{(d_n)^\infty} + (I - P_{d_{n+1}} + P^*_{d_{n+1}})^{-1}(I - P^*_{d_{n+1}})[r_{d_{n+1}} - g_n e + (P_{d_{n+1}} - I)h_n]$$

Proof.

$$\begin{aligned} h &= (I - P + P^*)^{-1}(I - P^*)r \\ h^{(d_{n+1})^\infty} &= (I - P_{d_{n+1}} + P^*_{d_{n+1}})^{-1}(I - P^*_{d_{n+1}})r_{d_{n+1}} \\ &= (I - P_{d_{n+1}} + P^*_{d_{n+1}})^{-1}(I - P^*_{d_{n+1}})[r_{d_{n+1}} - g_n e + g_n e + (P_{d_{n+1}} - I)h_n - (P_{d_{n+1}} - I)h_n] \end{aligned}$$

Now for a general h and P, P^*, we have

$$[(I-P)+\sum_{n=1}^{\infty}(P^n-P^*)]e=0$$

$$\begin{aligned}&[(I-P^*)+\sum_{n=1}^{\infty}(P^n-P^*)](P-I)\\ =&P-I-P^*+P^*+\sum_{n=1}^{\infty}(P^{n+1}-P^n)\\ =&P-I+P^*-P=P^*-I\end{aligned}$$

Then plug the above two equations into $h^{(d_{n+1})^\infty}$, get

$$h^{(d_{n+1})^\infty}=(I-P_{d_{n+1}}+P^*_{d_{n+1}})^{-1}(I-P^*_{d_{n+1}})[r_{d_{n+1}}-g_ne+(P_{d_{n+1}}-I)h_n]-(P_{d_{n+1}}-I)h_n.$$

Then by the optimality equation and MDP being a unichain, $h^{d_n^\infty}-h_n$ is a constant vector, so by Corollary 3.10

$$(P^*_{d_{n+1}}-I)(h_n-h^{d_n^\infty})=0$$

plug this equation in, we get the required result. □

Proposition 3.30

Suppose $d_{n+1}\in\arg\max_d\{r_d+P_dh_n\}$. *If*

$$[r_{d_{n+1}}-g_ne+(P_{d_{n+1}}-I)h_n](s)=0$$

for all s *that are recurrent under* $(d_{n+1})^\infty$, *and*

$$[r_{d_{n+1}}-g_ne+(P_{d_{n+1}}-I)h_n](s_0)>0$$

for some s_0 *that is transient under* $(d_{n+1})^\infty$ *then* $h^{(d_{n+1})^\infty}(s)>h^{(d_n)^\infty}(s)$ *for some* s *which is transient under* $(d_{n+1})^\infty$.

Proof. If

$$[r_{d_{n+1}}-g_ne+(P_{d_{n+1}}-I)h_n](s)=0$$

for all s that are recurrent under $(d_{n+1})^\infty$, then $d_{n+1}(s)=d_n(s)$ for all recurrent state s.

$$r(s,d_{n+1}(s))+\sum_j P(j\mid s,d_{n+1}(s))h_n(j)=\max_a\left\{r(s,a)+\sum_j P(j\mid s,a)h_n(j)\right\}.$$

Since $P_{d_{n+1}}(j\mid s)=0$ if s is recurrent and j is transient under d_{n+1}, $(d_{n+1})^\infty$ and $(d_n)^\infty$ have the same recurrent states. Then $P^*_{d_{n+1}}=P^*_{d_n}$. □

Note. The above proposition shows that while the gain stays the same under policy iteration, the bias might change. The actions might remain the same but only for the recurrent states, so you can stop because of the gain optimality; even though the value iteration might not when one seeks to find a policy being both gain and bias optimal (which is not covered here).

Proposition 3.31

In the unichain models, the iterates of the policy iteration has the following properties:

1. $g^{(d_{n+1})^\infty} > g^{(d_n)^\infty}$ *or*
2. $g^{(d_{n+1})^\infty} = g^{(d_n)^\infty}$ *but* $h^{(d_{n+1})^\infty}(s) > h^{(d_n)^\infty}(s)$ *for some* $s \in S$ *or*
3. $g^{(d_{n+1})^\infty} = g^{(d_n)^\infty}$ *and* $h^{(d_{n+1})^\infty} = h^{(d_n)^\infty}$

Example 3.5. Consider our old example again:

$$S = \{s_1, s_2\}, A_{s_1} = \{a_{11}, a_{12}\}, A_{s_2} = \{a_{21}\}$$

and

$$\begin{aligned} p(s_1 \mid s_1, a_{11}) &= \frac{1}{2} \\ p(s_2 \mid s_1, a_{11}) &= \frac{1}{2} \\ p(s_2 \mid s_1, a_{12}) &= 1 \\ p(s_2 \mid s_2, a_{21}) &= 1 \end{aligned}$$

and

$$r(s_1, a_{11}) = 5, r(s_1, a_{12}) = 10, r(s_2, a_{21}) = -1$$

We have

$$d_0(s_1) = a_{12}, d_0(s_2) = a_{21}$$

Now,

$$\begin{aligned} 0 &= 10 - g - h(s_1) - h(s_2) \\ 0 &= -1 - g \Longrightarrow g = -1 \end{aligned}$$

and also $h(s_2) = 0, h(s_1) = 11$. Hence,

$$d_1(s_1) \in \operatorname{argmax}\left\{5 + \frac{1}{2} \cdot 11 + \frac{1}{2} \cdot 0, 10 + 1 \cdot 0\right\} = a_{11}$$

and similarly $d_1(s_2) = a_{21}$. Next,

$$\begin{aligned} 0 &= 5 - g - \frac{1}{2}h(s_1) - \frac{1}{2}h(s_2) \\ 0 &= -1 - g \Longrightarrow g = -1 \end{aligned}$$

and also $h(s_2) = 0, h(s_1) = 12$. Hence,

$$d_2(s_1) \in \operatorname{argmax}\left\{5 + \frac{1}{2} \cdot 12 + \frac{1}{2} \cdot 0, 10 + 1 \cdot 0\right\} = a_{11}.$$