ISyE 6664 (Fall 2017)

Stochastic Optimization (Markov Decision Processes Ver.)

Prof. H. Ayhan
Georgia Institute of Technology

LaTeXer: W. KONG
http://wwkong.github.io
Last Revision: March 23, 2018

Table of Contents

These notes are currently a work in progress, and as such may be incomplete or contain errors.

ACKNOWLEDGMENTS:

Special thanks to *Michael Baker* and his LaTeX formatted notes. They were the inspiration for the structure of these notes.

Abstract

The purpose of these notes is to provide the reader with a secondary reference to the material covered in ISyE 6664.

Administrative

1 Markov Decision Processes (MDPs)

We study sequential decision making under uncertainty which takes into account both the outcomes of current decisions and future decisions making opportunities. Here is an outline:

- Decision epochs
  - A set of system states
  - A set of available actions
  - A set of state and action dependent immediate rewards
  - A set of state and action dependent transition probabilities

Examples

Inventory Theory

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 during the month is **random**
- Holding cost
- Revenue
- Penalty for lost sales

Admission Control

Consider a system with m servers, i.e. the capacity is m. One set of calls enter at a Poisson rate with parameter λ_1 and reward r_1 and another set of calls enter at a Poisson rate with parameter λ_2 and reward r_2 with $r_1 > r_2$. Service times are exponential with rate μ.

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

These are the time points where decisions are made.

- T is the set of decision epochs
  - $T = \{1, 2, ..., N\}$ in the finite case, and at time N we do not make decisions
  - $T = \{1, 2, ...\}$ in the infinite case

State and Action Sets

At each depoch (decision epoch), the system is in a certain state $s \in S$. In state s, we can choose an action $a \in A_s$ where A_s is the set of possible actions in state s and we denote

$$A = \bigcup_{s \in S} A_s$$

as the action space. 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 $q(a)$.

Transition probabilities and rewards

We have

$$p_t(\cdot|s,a) : \text{probability distribution at the next decision epoch when action } a \text{ is chosen in state } s \text{ at decision epoch } t$$

$$r_t(s,a) : \text{immediate reward received when action } a \text{ is is chosen in state } s \text{ at time } t$$

The five-tuple

$$\{T, S, A_{s,} p_t(\cdot|s,a), r_t(s,a)\}$$

is called a **Markov decision process** (MDP). We may also use the alternative definition

$$r_t(s,a) : \text{immediate reward received when action } a \text{ is is chosen in state } s \text{ at the decision epoch } t \text{ and the state at the next epoch is } j$$

and define

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

Decision rule: a procedure for action selection in each state

Examples.

Markovian Deterministic Decision Rule

This is a decision $d_t : S \mapsto A$ where $d_t(s) \in A_s$.

Markovian Randomized Decision Rule

This is a decision $d_t : S \mapsto P(A)$ where $q_{d_t}(s) \in P(A_s)$.

History Dependent Deterministic Decision Rule

For a history

$$h_t = (s_1, a_1, ..., a_{t-1}, s_t) = (h_{t-1}, a_{t-1}, s_t)$$

and

$$H_t : \text{ set of all histories}$$

this is a decision $d_t : H_t \mapsto A$.

History Dependent Randomized Decision Rule

This is a decision $d_t : H_t \mapsto P(A)$.

Policies

A policy Π is a sequence of decision of rules

$$\Pi = (d_1, d_2, ..., d_N) \text{ or } \Pi = (d_1, d_2 ...).$$

If $d_t = d$ for all $t \in T$, then $\pi = (d, d, ...)$ is called a **stationary policy**.

Example 1.1. An inventory manager checks his inventory at the end of each month. Depending on the inventory level, he wants to determine how many units to purchase. Assume that raw units arrive overnight. Demand arrives during the month but orders are filled at the end of the month. Assume no backlogs are allowed and the warehouse has a capacity of M. The monthly demand D_t has the following probability mass function:

$$P(D_t = k) = p_k \text{ for } k = 0, 1, 2, ...$$

Assume that if j units are purchased, the purchase cost is $C(j)$. The holding cost for j units is $h(j)$ and the revenue obtained from j units is $f(j)$. Suppose that we are considering an N period problem and it costs the warehouse $g(j)$ if there are j units left at time N. No backlogs are allowed. Model this as an expected profit maximization problem.

Modeling this as a MDP, we have

$$\begin{aligned}
T &= \{1, 2, ..., N\} \\
S &= \{0, 1, ..., M\} \\
A_s &= \{0, 1, ..., M - s\} \\
P(D_t = k) &= p_k, \text{ for } k = 0, 1, ... \\
p_t(j|s, a) &= \begin{cases} 0, & \text{if } j > s + a \\ p_{s+a-j}, & \text{if } 0 < j \leq s + a \\ \sum_{k=s+a}^{\infty} p_k & \text{if } j = 0 \end{cases} \\
r_t(s, a) &= -C(a) - h(s + a) + \sum_{k=0}^{s+a} p_k f(a) + \sum_{k=s+a+1}^{\infty} \text{, for } t = 1, ..., N - 1 \\
r_N(s) &= -g(s).
\end{aligned}$$

Example 1.2. The condition of a piece of equipment used in a manufacturing process deteriorates over time. The condition of the equipment is checked at predetermined discrete decision epochs. Let $S = \{0, 1, ...\}$ represent the condition of the equipment at each decision epoch. The higher the value of s is, the worse the condition of the equipment. At each decision epoch, you can choose either to operate the equipment as it is or replace it with a new one. We assume in each period, the equipment deteriorates by i states with probability $p(i)$. There is a fixed income of R units per period, a state dependent operating cost of $h(s)$, a replacement cost of R units. Again assume that we are interested in a finite horizon of N decision epochs. If the equipment in state s at time N, there is a salvage value of $g(s)$.

Modeling this as a MDP, we have

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

1.2 Finite Horizon MDPs

Let us define

$$\begin{aligned}
V_N^{\Pi}(s) : &\text{ total expected reward for an } N \text{ period problem under} \\
&\text{policy } \pi \text{ when the system state at the first decision} \\
&\text{epoch is } s.
\end{aligned}$$

Suppose Π is a history dependent randomized policy where

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

Then,

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

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

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

We want to find Π^* (among all history dependent randomized policies) such that

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

If an optimal policy does not exist, we look for an epsilon optimal policy such that

$$V_N^{\Pi^*}(s) + \varepsilon > V_N^{\Pi}(s), \text{ for all } \Pi.$$

The value $V_N^*(s)$ is defined as

$$V_N^*(s) = \sup_{\Pi} V_N^{\Pi}(s).$$

Of course, if $\sup$ is attained, then $V_N^*(s) = \max_{\Pi} V_N^{\Pi}(s)$. Going forward, we may interchange the notation

$$V_{\Pi}^N(s) \equiv V_N^{\Pi}.$$

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

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

for $t = 1, ..., N-1$ and $u_N(h_N) = r_N(s_N)$ for all $h_N = (h_{N-1}, a_{N-1}, s_N)$, which is our boundary condition. If Π is Markovian deterministic, then

$$u_t^{\Pi}(S_t) = E\left[\sum_{n=t}^{N-1} r_n(X_n, d_n(X_n)) + r_N(X_n)|X_t = s_t\right].$$

If $h_1 = S$, then

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

from recursively figuring out $V_N^{\Pi}(s)$ by calculating $u_t^{\Pi}(h_t)$. Note that $V_N^{\Pi}(s)$ is not dependent on t. Here is the recursive scheme in detail:

Finite Horizon Policy Evaluation Algorithm

1. Set $t = N$ and $u_N(h_N) = r_N(s_N)$, the terminal reward, for all $h_N = (h_{N-1}, a_{N-1}, s_N)$.

2. If $t = 1$, stop; otherwise go to step 3.

3. Set $t \leftarrow\!\shortmid t - 1$ and compute $u_t^{\Pi}(h_t)$ as

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

4. Return to 2.

For Markovian deterministic Π, we have

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

Theorem 1.1. *Suppose that $\Pi = (d_1, ..., d_{N-1})$ is a history dependent deterministic policy and u_t^{Π} is obtained by the finite horizon policy evaluation algorithm. Then for all $t \leq N$,*

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

and $V_N^{\Pi}(s) = u_1^{\Pi}(h_1)$ *for* $h_1 = s$.

Proof. Clearly the result holds for $t = N$. Suppose the result holds for $n = t_1, ..., N$ and we will prove that it holds for $n = t$.

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

Optimality Equations (Bellman's Equations)

We have

$$u_t^*(h_t) = \sup_u u_t(h_t)$$

where Π belongs to the set of history dependent deterministic policies. □

Lemma 1.1. *Let* w *be a real valued function on an arbitrary discrete set* W *and let* $q(\cdot)$ *be a probability distribution on* W. *Then* $\sup_{u \in W} \geq \sum_{u \in W} q(u)w(u)$

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

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

□

There will be a deterministic rule that performs as well/better than randomized.

Optimality Equations for the N Period Problem

Define

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

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

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

Recall that

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

so by computing u_t^* like this, we will compute $V_N^*(s)$.

Theorem 1.2. *Suppose that* u_T *is a solution to the optimality equations for* $t = 1, ..., N-1$ *with* $u_N(s_N) = r_N(s_N)$. *Then,*

(a) $u_t(h_t) = u_t^*(h_t)$ *for* $t = 1, ..., N-1$

(b) $u_1(s_1) = V_N^*(s_1)$

Proof. See textbook. □

Theorem 1.3. *Suppose that u_t^* for $t = 1, ..., N$ are solutions to the optimality equations subject to the boundary condition and the policy $\Pi^* = (d_1^*, ..., d_{N-1}^*)$ satisfies*

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

Then

(a) $u_t^*(h_t) = u_t^{\Pi^*}(h_t)$

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

Proof. (a) Trivially

$$u_N^*(s_N) = r_N(s_N) = u_N^{\Pi^*}(s_n)$$

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

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

(b) We have

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

□

Hence, the optimal policy $\Pi^* = (d_1^*, ..., d_{N-1}^*)$ is defined as

$$d_t(h_t) \in \operatorname*{argmax}_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(h_t, a, j) \right\}.$$

Theorem 1.4. *Let u_t^* for $t = 1, ..., N$ be the solution to the optimality equations together with the boundary conditions.*

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

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

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

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

Proof. (a) We have

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

Thus, u_N^* depends on h_N only though s_N. The result holds for $n = N$. Let us assume it holds for $n = t+1, ..., N$ and we twill

show that it also holds for $n = t$. Next,

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

and the result holds for $n = t$.

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

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

□

Corollary 1.1. *Let*

Π^{HR} : *set of history dependent randomized policies*
Π^{MD} : *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 1.1. *Assume that S is finite or countable and that*

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

or

(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* $a \in A_s, s \in S$,
$p_t(j|s, a)$ *is continuous in* a *for each* $j \in S, s \in S$

or

(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* $a \in A_s, s \in S$,
$p_t(j|s, a)$ *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

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

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

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

and set

$$A^*_{s_t} = \underset{a\in A_{s_t}}{\operatorname{argmax}} \left\{ r_t(s_t,a) + \sum_{j\in S} p_t(j|s_t,a)u^*_{t+1}(j) \right\}.$$

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

Example 1.3. (Inventory problem revisited)

Consider the setup

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

and

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

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

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

then

$$r(0,1) = -6 - 1 + 8\cdot\frac{3}{4} = -1$$
$$r(0,2) = -12 - 2 + 16\cdot\frac{1}{4} + 8\cdot\frac{1}{2} = -2$$
$$r(0,3) = -10 - 3 + 16\cdot\frac{1}{4} + 8\cdot\frac{1}{2} = -5$$

$$u^*_3(0) = \max\left\{ 0 + 1\cdot 0, \underbrace{-1}_{=r(0,1)} + 0, \underbrace{-2}_{=r(0,2)}, \underbrace{-5}_{=r(0,3)} \right\} = 0, d^*_3(0) = 0$$

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,

$$\begin{aligned} u^*_2(0) &= \max\left\{ 0, -1 + 0\cdot\frac{3}{4} + 5\cdot\frac{1}{4}, -2 + 6\cdot\frac{1}{4} + 5\cdot\frac{1}{2} + 0\cdot\frac{1}{4}, -5 + 5\cdot\frac{1}{4} + 6\cdot\frac{1}{2} + 5\cdot\frac{1}{4} \right\} \\ &= \max\left\{ 0, \frac{1}{4}, 2, \frac{1}{2} \right\} \\ &= 2 \end{aligned}$$

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 Monotone Optimal Policies

Monotonicity of Optimal Policies

Consider

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

Definition 1.1. We say that $g(\cdot, \cdot)$ for $x^+ \geq x^-$ in X and $y^+ \geq y^-$ in Y is **superadditive** if

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

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

Lemma 1.2. *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,*

$$f(x) = \max \left\{ y \in \operatorname*{argmax}_{y \in Y} g(x, y) \right\}$$

is monotone non-decreasing in X.

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

$$g(x^-, f(x^-)) - g(x^-, y) \geq 0.$$

Since g is superadditive,

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

then $f(x^+) \geq f(x^-)$ since

$$g(x^+, f(x^+)) \geq g(x^+, f(x^-)) \text{ and } g(x^+, y) \leq g(x^+, f(x^-))$$

for all $y \leq f(x^-)$ so we must have $f(x^+) \geq f(x^-)$. □

Lemma 1.3. *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 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,

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

□

Theorem 1.5. *Assume that*

(1) $S = \{0, 1, ...\}$

(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* $t = 1, ..., N-1$.

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

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

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

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

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

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

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

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

Theorem 1.6. *Assume that*

(1) $S = \{0, 1, ...\}$

(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* $t = 1, ..., N-1$.

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

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

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

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

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

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

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

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

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

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

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

is superadditive and the result holds. □

Theorem 1.7. *Suppose for* $t = 1, ..., N-1$ *that*

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

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

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

(4) $\sum_{j=0}^{\infty} p_t(j|s,a)$ *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* $t = 1, ..., N-1$.

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

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

superadditive on $S \times A$. □

Monotone Backward Induction

Suppose that $S = \{0, 1, ..., M\}$ 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 $A_0 = A$.

2a) Set

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

2b) Set

$$A^*_{s,t} = \underset{a \in A_s}{\operatorname{argmax}} \left\{ r_t(s,a) + \sum_{j \in S} p_t(j|s,a) u^*_{t+1}(j) \right\}$$

2c) If $s = M$ go to step 3, otherwise set

$$A_{s+1} = \left\{ a \in A : a \geq \max \left\{ a' \in A^*_{s,t} \right\} \right\}$$

2d) Substitute $s+1$ for s and return to 2a).

3) If $t = 1$, stop; otherwise go to 2).

Example 1.4. Given $S = \{0, 1, ...\}$, 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 replace, 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.

We have:

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

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.

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

(2) $\sum_{j=k}^{\infty} p(j|s,a)$ is non-decreasing in s for all $k \in S$ and $a \in A$ since when we replace,

$$\sum_{j=k+1}^{\infty} p(j|s+1,1) - \sum_{j=k}^{\infty} p(j|s,1) = \sum_{j=k}^{\infty} p(j) - \sum_{j=k}^{\infty} p(j) = 0.$$

Now when we do not replace, for $k > s$,

$$\sum_{j=k}^{\infty} p(j|s+1,0) - \sum_{j=k}^{\infty} p(j|s,0) = \sum_{j=k}^{\infty} p(j-s-1) - \sum_{j=k}^{\infty} p(j-s) = p(k-s-1) \geq 0$$

and for $k \leq s$, we have

$$\sum_{j=k}^{\infty} p(j|s+1,0) - \sum_{j=k}^{\infty} p(j|s,0) = \sum_{j=s+1}^{\infty} p(j-s-1) - \sum_{j=s}^{\infty} p(j-s) = 0.$$

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

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

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

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

since u is non-creasing.

1.4 Infinite Horizon MDPs

We assume:

- Transition probabilities and rewards are stationary and $|r(s,a)| \leq M$
- We are given a discount factor
- $\pi = (d_1, d_2, ...)$ is Markovian deterministic

Define

$$\begin{aligned}
v_\lambda^\pi(s) &: \text{total expected discounted reward under policy } \pi \\
&\quad \text{when the initial state is } s \text{ and the discount fact is } \lambda \\
r_d &: \text{vector of rewards under decision rule } d \\
P_d &: \text{probability transition matrix under decision rule } d
\end{aligned}$$

and explicitly

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

Let us denote $v_\lambda^*(s) = \sup_\pi v_\lambda^\pi(s)$. If v_λ^π is the vector of total expected rewards, then

$$\begin{aligned}
v_\lambda^\pi &= r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} r_{d_3} + ... \\
&= r_{d_1} + \lambda P_{d_1} (r_{d_2} + \lambda P_{d_2} r_{d_3} + ...) \\
&= r_{d_1} + \lambda P_{d_1} v_\lambda^{\pi'}
\end{aligned}$$

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

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

Theorem 1.8. *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_λ^∞ can be written as

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

where $L_d(v) = r_d + \lambda P_d v$.

Example 1.5. 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 $p(s_1|s_1, a_{11}) = 0.5$, $p(s_2|s_1, a_{11}) = 0.5$, $p(s_2|s_1, a_{12}) = 1$, and $p(s_2|s_2, a_{21})$. 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

$$v_\lambda^{d^\infty}(s_1) = 5 + \lambda(0.5 v_\lambda^{d^\infty}(s_1) + 0.5 v_\lambda^{d^\infty}(s_2))$$
$$v_\lambda^{d^\infty}(s_2) = -1 + \lambda v_\lambda^{d^\infty}(s_2) \implies v_\lambda^{d^\infty} = \frac{-1}{1-\lambda}$$

and so after substitution,

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

Lemma 1.4. *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,

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

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

Remark 1.1. Given that

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

by taking the limit as $n \to \infty$ on both sides,

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

If v^* is the vector of $v^*(s)$ for $s \in S$, then $v^* = \mathcal{L} v^*$.

Theorem 1.9. *Suppose that there exists a v such that*

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

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

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

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

$$\mathcal{L} v = \sup_\alpha \{ r_\alpha + \lambda P_\alpha v^* \}$$
$$L v = \max_\alpha \{ r_\alpha + \lambda P_\alpha v \}.$$

Then

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

and also since

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

then

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

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

$$-\frac{\epsilon}{2} e \leq \lambda^n P_\pi^n v \leq \frac{\epsilon}{2} e$$

where e is a vector of ones. Hence,

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

and so with can find n sufficiently large so that

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

(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 implies

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

and hence

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

(iii) Trivial. □

Definition 1.2. Let U be a Banach space (complete normed linear space). The operator $T : U \to U$ is a **contraction mapping** if $\exists \lambda$ with $0 \leq \lambda < 1$ such that

$$\|Tv - Tu\| \leq \lambda \|v - u\|.$$

Theorem 1.10. *(Fixed point theorem) Suppose U is Banach space and $T : U \mapsto U$ is a contraction mapping. Then,*

(a) $\exists v^* \in U$ unique such that $Tv^* = v^*$

(b) for arbitrary $v^0 \in U$, the sequence $\{v^n\}$ defined by $v^{n+1} = Tv^n$ converges to v^*.

Proof. (a) Directly,

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

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

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

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 1.2. *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

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

and suppose that

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

Then

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

and we can similarly have $Lu(s) \geq Lv(s)$. 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 1.11. *(i) There exists a v^* satisfying $Lv^* = v^*$ ($\mathcal{L}v^* = v^*$) so $v_\lambda^* = v^*$.*

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

Proof. (ii) If π^* is optimal then $v_\lambda^* = v_\lambda^{\pi^*}$ and hence $Lv_\lambda^{\pi^*} = v_\lambda^{\pi^*}$. If $Lv_\lambda^{\pi^*} = v_\lambda^{\pi^*}$ then $v_\lambda^{\pi^*} = v_\lambda^*$ and hence π^* is optimal. □

Theorem 1.12. *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 $v_\lambda^{(d^*)^\infty} = v_\lambda^*$.*

Theorem 1.13. *Suppose there exists an optimal policy. Then there exists an optimal stationary policy.*

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

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

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

1.5 Algorithms

Theorem 1.14. *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 $p(j|s,a)$ 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 $p(j|s,a)$ is lower semicontinuous in a for each $j \in S$ and $s \in S$.

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

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

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

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

$$d_\epsilon(s) \in \operatorname*{argmax}_{s \in S} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v^{n+1}(j) \right\}.$$

Theorem 1.15. *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 contraction mapping theorem.

(2) We need to show that $\|v_\lambda^{(d_\epsilon)^\infty} - v_\lambda^\epsilon\| \leq \epsilon$. Note that

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

and directly,

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

and

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

Rearranging gives us

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

□

Proposition 1.3.

(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,

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

(2) Directly,

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

and likewise for the $(\leq)$ case. □

So if $v^1 \geq v^0$ in value iteration, then $\{v^n\} \to v_\lambda^*$ is monotone decreasing.

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

$$d_{n+1} \in \operatorname*{argmax}_d \{r_d + \lambda P_d v^n\}$$

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

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

Proof. Directly

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

□

Theorem 1.16. *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 1.6. Recall example with

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

and

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

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

$$d_1(s_1) \in \text{argmax}\left\{5 + \frac{1}{2}v_\lambda^{(d_0)^\infty}(s_1) + \frac{1}{2}v_\lambda^{(d_0)^\infty}(s_2), 10 + v_\lambda^{(d_0)^\infty}(s_2)\right\}$$
$$\implies d_1(s_1) \in \text{argmax}\left\{\frac{(5-5.5\lambda)}{(1-0.5\lambda)(1-\lambda)}, \frac{2(5-5.5\lambda)}{1-\lambda}\right\}$$

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

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

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

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,

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

b. If $\|u_n^0 - v^n\| < \frac{\epsilon(1-\lambda)}{2\lambda}$ 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 $d_\epsilon = d_{n+1}$.

Linear Programming

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

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

where the constraint is equivalent to

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

The dual LP, with dual variables $x(s,a)$, is

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

Theorem 1.17. *(1) For each Markovian randomized decision rule d, let*

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

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 $s\in S$, $\sum_{a\in A_s} x(s,a) > 0$. Define the randomized decision rule 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 $a\in A_s$.

Proposition 1.5. *(1) Let x be a basic feasible solution to the dual LP. then d_X is 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.

Theorem 1.18. *(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 deterministic optimal policy . Then x_{d^*} is a basic optimal solution to the dual LP.

Proposition 1.6. *For any positive vector α the dual LP has the same optimal basis x^*. Hence, $(d_{x^*})^\infty$ does not depend on the choice of α.*

Proof. From sensitivity analysis, changing α only affects feasibility but not optimality. Hence, we will show that x^* r remains feasible as long as α is positive. Now

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

so x^* is feasible only if $\alpha > 0$. □

Example 1.7. Consider our previous example again:

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

and

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

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

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

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

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

Action Elimination

Proposition 1.7. *If for $a' \in A_s$, $r(s, a') + \lambda \sum_{j \in S} p(j|s, a)v^*_\lambda(j) < v^*_\lambda(s)$ then*

$$a' \notin \operatorname*{argmax}_{a \in A_s} \left\{ r(s, a) + \lambda \sum_{j \in S} p(j|s, a)v^*_\lambda(j) \right\}.$$

Proof. We know

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

but we have

$$r(s, a') + \lambda \sum p(j|s, a)v^*_\lambda(j) < v^*_\lambda(s).$$

Clearly a' cannot be optimal in state s. □

Proposition 1.8. *Suppose there exists v^L and v^U such that $v^L \leq v^*_\lambda \leq v^U$. Then if for $a' \in A_s$,*

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

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

Theorem 1.19. *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 $d' \in D^\sigma \cap \operatorname{argmax}_d L_d v$

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

There 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

$$d^* \in D^\sigma \cap \operatorname*{argmax}_d L_d v^*_\lambda.$$

□

Theorem 1.20. *Consider $S = \{0, 1, ...\}$, $A_s = A$ for all $s \in S$. If*

(a) $r(s, a)$ is non-decreasing in s for all $a \in A$,

(b) $\sum_{j=k}^{\infty} p(j|s, a)$ is non-decreasing in s for all $k \in S$ and $a \in A$,

(c) $r(s, a)$ is superadditive on $S \times A$, and

(d) $\sum_{j=k}^{\infty} p(j|s, a)$ is superadditive on $S \times A$,

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

Proof. Let us define

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

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

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

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

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

Thus, $\{v^n\} \in V^\sigma$ and $v^n \to v^*_\lambda \in V^\sigma$. Suppose that $v \in V^\sigma$. Does there exist a $d \in D^\sigma$? For s^-, s^+ and $a^- \leq a^+$ we have

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

and so

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

is superadditive. Hence, there must exist a decision rule

$$d(s) \in \operatorname*{argmax}_{a \in A} \left\{ r(s, a) + \lambda \sum_{j=0}^{\infty} p(j|s, a) v(j) \right\}$$

which is non-decreasing in s from a previous theorem. □

Theorem 1.21. *Consider $S = \{0, 1, ...\}$, $A_s = A$ for all $s \in S$. If*

(a) $r(s, a)$ is non-increasing in s for all $a \in A$,

(b) $\sum_{j=k}^{\infty} p(j|s, a)$ is non-decreasing in s for all $k \in S$ and $a \in A$,

(c) $r(s, a)$ is superadditive on $S \times A$, and

(d) $\sum_{j=k}^{\infty} p(j|s, a)u(j)$ 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-increasing in s.

1.6 Long-Run Average Reward Optimality

Recall that

$$v_{N+1}^{\Pi}(s) = E_S^{\Pi}\left[\sum_{t=1}^{N} r(X_t, Y_t)\right]$$

and define the **gain**

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

and we also define

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

Definition 1.3. 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 1.9. *Let S be countable. Let d^∞ be a stationary Markovian randomized policy and suppose that P_d^* exists.*

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

(b) Furthermore, if S is finite, then $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P_d^{n-1} = P_d^*$ exists and for any d^∞, we have $g^{d^\infty}(s) = P_d^* r_d(s)$.

Definition 1.4. Let P denote the probability transition matrix of a Markov chain $\{X_t : t = 1, 2, ...\}$ and $r(s)$ a reward function for each $s \in S$. We refer to the bivariate stochastic process $\{(X_t, r(X_t)) : t = 1, 2, ...\}$ as a **Markov reward process**.

Remark 1.2. 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 1.10. *Suppose that P^* exists. If j and k are in the same irreducible class, $g(j) = g(k)$. Furthermore, if the Markov chain is irreducible or **unichain** (i.e. a single recurrent class plus some transient states), then $g(s)$ is a constant function.*

Definition. The **bias vector** is defined as

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

where we know that $PP^* = P^*$. Note that

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

and

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

from the fact that

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

and hence

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

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

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

Remark 1.3. 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 the same recurrent class. Then,

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

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

Theorem 1.22. *Let S be finite and let g and h denote the gain and bias 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

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

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

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

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

(b) We first note that adding the first equation plus P^* times the second equation gives us

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

In part (a), we have shown that $h = (I-P+P^*)^{-1}(I-P^*)r$ 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

$$(I-P)(h-h') = 0$$

□

Remark 1.4. Note that if g is a constant vector, then since P is a probability matrix, then $(I-P)g = 0$ trivially.

Corollary 1.2. *Suppose P is unichain. Then the long-run average reward $P^*r = ge$ and it is uniquely determined by solving*

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

Proof. Suppose g and h satisfy the above equation. Then $P^*r = ge$ and $h = (I-P+P^*)^{-1}(I-P^*)r + ke$ for any scalar k. Furthermore, as $P^*h = 0$ then $h = (I-P+P^*)^{-1}(I-P^*)r$. □

Proposition 1.11. *Let g and h represent the gain and bias of a Markov Reward process with finite state space S. Then,*

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

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

Corollary 1.3. *We have*

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

2 Classification of MDPs

Here some classes 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 possibly an empty set of 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 $p_d^n(s|j) > 0$ for some $n \geq 1$.

(d) Weakly communicating: if there exists a closed set of states with each state in this closed set accessible from each in that set under some deterministic stationary policy, plus (possibly empty) set of transient states which is transient under every policy.

(e) Multichain: if the transition matrix corresponding to at least one stationary deterministic policy has two or more closed recurrent classes.

Example 2.1. (Inventory problem revisited) Suppose the warehouse has a capacity of 3 units. We are given

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

Consider a separate policy

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

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

Example 2.2. Given $S = \{s_1, s_2\}$ and $A_{s_1} = \{a_{11}, a_{12}\}$, $A_{s_2} = \{a_{21}\}$, define

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

and $d(s_1) = a_{11}, d(s_2) = a_{12}$, $\delta(s_1) = a_{12}, \delta(s_2) = a_{21}$ and the policies d and δ imply that this is multichain.

Proposition 2.1. *(i) A Markov decision process is communicating if and only if there exists a randomized stationary policy where the chain is irreducible.*

(ii) 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 2.1. *Assume a weakly communicating model and let d be a Markovian deterministic decision rule.*

(a) Let C be a closed irreducible set of recurrent states 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 $s, s' \in S$. 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. Then $\exists s \in S \backslash (T \cup C)$ and $a' \in A_s$ such that

$$\sum_{j \in C} P(j|s, a') > 0.$$

We then set $\delta(s) = a'$ 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

$$\sum_{j \in S \backslash T} P(j|s', a_{s'}) > 0.$$

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

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

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

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

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

□

Theorem 2.2. *(a) Given a Markovian deterministic decision rule d_1 there exists a Markovian deterministic decision rule δ for which g^{δ^∞} is constant and $g^{\delta^\infty} \geq g$.*

(b) If there exists a stationary optimal policy, there exists a stationary optimal policy with constant gain.

2.1 Unichain Markov Decision Processes

Remark 2.1. The Optimality Equations for Unichain MDPs are:

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

This is because, we know that

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

which implies that

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

and if we take $\lambda \uparrow 1$ then

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

Alternatively, since

$$\begin{aligned} v_{N+1} &= Nge + h + o(1) \\ v_N^* &= (N-1) g^* e + h + o(1) \end{aligned}$$

and

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

then

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

and hence

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

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

Theorem 2.3. *(a) If there exists a scalar g and a vector h which satisfy*

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

then $ge \ge g_+^*$.

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

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

then $ge \le g_-^*$.

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

Theorem 2.4. *Suppose S and A_s for each $s \in S$ are finite.*

(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 2.1. A decision rule d_h is called $h-$**improving** if

$$d_h \in \operatorname*{argmax}_d\{r_d + P_d h\}.$$

Theorem 2.5. *Suppose scalar g^* and h vector satisfy the unichain optimality equations. Then if d_h is $h-$improving then $(d_h)^\infty$ is an optimal policy.*

Value Iteration Algorithm

Define

$$sp(v) = \max_s v(s) - \min_s v(s)$$

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

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

3. If $sp(v^{n+1} - v^n) < \epsilon$ go to step 4; otherwise increment n by 1 and return to step 2.

4. For each $s \in S$ choose

$$d_\epsilon(s) \in \operatorname*{argmax}_{a\in A_s}\left\{r(s,a) + \sum_{j\in S} p(j|s,a)v(j)\right\}.$$

Theorem 2.6. *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.*

Theorem 2.7. *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*

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

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

Proof. For any v improving d,

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

and

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

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

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

□

Theorem 2.8. *(i) d_ϵ^∞ is an ϵ-optimal policy where*

$$d_\epsilon(s) \in \operatorname*{argmax}_{a\in A_s} \left\{ r(s,a) + \sum_{j\in S} p(j|s,a) v(j) \right\}.$$

(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 $|g' - g^{(d_\epsilon)^\infty}| < \frac{\epsilon}{2}$.

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

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

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

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

We know that

$$\min_{s\in S}\left(v^{n+1}(s)-v^n(s)\right) \le g^{d_\epsilon^\infty} \max_{s\in S}\left(v^{n+1}(s)-v^n(s)\right)$$

and so

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

An aperiodic transformation

Choose $0 < \tau < 1$ and define

$$\tilde{r}(s,a) = \tau r(s,a)$$
$$\tilde{p}(j|s,a) = (1-\tau)1(j=s) + \tau p(j|s,a).$$

□

Proposition 2.2. *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,

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

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 2.1. *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

$$d_{n+1} \in \operatorname*{argmax}_d \{r_d + P_d h_n\}$$

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 2.3. *Suppose that* $d_{n+1} \in \operatorname{argmax}\{r_d + P_d h_n\}$. *Then,*
(a) $g_{n+1}e = g_n e + P^*_{d_{n+1}}\left[r_{d_{n+1}} - g_n e + (P_{d_{n+1}} - I)h_n\right]$
(b) If $\left[r_{d_{n+1}} - g_n e + (P_{d_{n+1}} - I)h_n\right](s) > 0$ *for some state* s *which is recurrent under* d_{n+1} *then* $g_{n+1} > g_n$.
(c) If $\left[r_{d_{n+1}} - g_n e + (P_{d_{n+1}} - I)h_n\right](s) = 0$ *for all* s *under* d_{n+1} *then* $g_{n+1} = g_n$.

Proof. (a) Directly,

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

□

Corollary 2.2. *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 2.4. *In the unichain models, the iterates of the policy iteration has the following properties:*
(i) $g^{(d_{n+1})^\infty} > g^{(d_n)^\infty}$ *or*
(ii) $g^{(d_{n+1})^\infty} = g^{(d_n)^\infty}$ *but* $h^{(d_{n+1})^\infty}(s) > h^{(d_n)^\infty}(s)$ *for some* $s \in S$ *or*
(iii) $g^{(d_{n+1})^\infty} = g^{(d_n)^\infty}$ *and* $h^{(d_{n+1})^\infty} = h^{(d_n)^\infty}$

Example 2.3. Consider our old example again:

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

and

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

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,

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

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,

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

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

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

Note the policy iteration here does not stop even though the gains are the same.

LP Approach

The LP formulation is

$$\begin{aligned}&\min_{h,g} g\\ &\text{s.t. } g - h(s) + \sum_{j\in S} p(j|s,a)h(j) \geq r(s,a), \forall s \in S \text{ and } \forall a \in A_s\end{aligned}$$

and using dual variables $x(s,a)$, the dual formulation is

$$\begin{aligned}\max \sum_{s\in S}\sum_{a\in A_s} r(s,a)x(s,a)\\ \text{s.t. } \sum_{a\in A_j} x(j,a) - \sum_{s\in S}\sum_{a\in A_s} p(j|s,a)x(s,a) = 0, \forall j \in S\\ \sum_{s\in S}\sum_{a\in A_s} x(s,a) = 1\\ x(s,a) \geq 0, \forall s \in S \text{ and } \forall a \in A_s.\end{aligned}$$

Assume that the Markov decision process is recurrent.

Theorem 2.9. *(a) For each Markovian randomized decision rule d, define*

$$x_d(s,a) = P(d(s) = a)\Pi_d(s)$$

for all $s \in S$, $a \in A_s$. Then $x_d(s,a)$ is a solution to the dual LP.

(b) Let x be a feasible solution to the dual LP. Then for each $s \in S$, $\sum_{a\in A_s} x(s,a) > 0$. Define a randomized decision rule by

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

Then x_{d_x} i s a feasible solution the dual LP.

Theorem 2.10. *Suppose x^* is a basic optimal solution to the dual LP. Then the stationary policy $(d_x^*)^\infty$ in which we choose $d_{x^*}(s) = a$ if $x^*(s,a) > 0$ is an optimal stationary deterministic policy.*

Example 2.4. In our previous example, the dual problem can be reduced to

$$\begin{aligned}\max\ & 5x(s_1,a_{11}) + 10x(s_1,a_{12}) - x(s_2,a_{21})\\ \text{s.t. } & x(s_1,a_{11}) + x(s_1,a_{12}) - 0.5x(s_1,a_{11}) = 0\\ & x(s_1,a_{11}) - x(s_1,a_{12}) - 0.5x(s_1,a_{11}) - x(s_2,a_{22}) = 0\\ & x(s_1,a_{11}) + x(s_1,a_{12}) + x(s_2,a_{21}) = 1\\ & x(s_1,a_{11}), x(s_1,a_{12}), x(s_2,a_{21}) \geq 0\end{aligned}$$

and solving it, we will get

$$x^*(s_1,a_{11}) = x^*(s_1,a_{12}) = 0, x^*(s_2,a_{21}) = 1.$$

Theorem 2.11. *Suppose the Markov decision process is unichain.*

(a) Let d be a Markovian randomized decision rule and R_d be the set of recurrent states under d. Define

$$x_d(s,a) = \begin{cases} P(d(s) = a)\Pi_d(s), & \text{for } s \in R_d\\ 0, & \text{otherwise.}\end{cases}$$

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

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

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

and define $P(d_x(s) = a) = x(s,a) / \left[\sum_{a \in A_s} x(s,a)\right]$ for $s \in S_x$ and arbitrary otherwise. Then $x_{d_x}(s,a) = x(s,a)$ for $a \in A_s$ and $s \in S_x$.

Corollary 2.3. *Let x be a basic feasible solution to the dual LP ans suppose that d_x is defined as in the previous theorem.*
(a) Then for $s \in S_x$, $d_x(s)$ is deterministic and satisfies

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

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

Corollary 2.4. *There exists a bounded optimal basic solution x^* to the dual LP and the policy $(d_{x^*})^\infty$ defined as*

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

is an optimal policy.

Theorem 2.12. *Let V^σ and D^σ be the respective sets of structured values and decision rules. Let $S = \{0, 1, ...\}$. Then if*
(a) for any sequence $\{\lambda_n\}$, $0 \le \lambda_n < 1$ for which $\lambda_n \to 1$,

$$\lim_{n \to \infty} \left[v^*_{\lambda_n} - v^*_{\lambda_n}(0) e \right] \in V^\sigma$$

and,
(b) $h \in V^\sigma$ implies that there exists a d' such that

$$d' \in \operatorname{argmax} \{ r_d + P_d h \} \cap D^\sigma,$$

then $D^\sigma \cap \operatorname{argmax}_{d \in D} \{ r_d + P_d h \} \neq \emptyset$ and

$$d^\sigma \in \operatorname*{argmax}_{d \in D} \{ r_d + P_d h \} \cap D^\sigma$$

is an optimal decision rule if an optimal decision rule exists.

Proof. We have

$$\begin{aligned} v_\lambda &= \frac{ge}{1-\lambda} + h + f(\lambda) \\ v_\lambda(s) &= \frac{g}{1-\lambda} + h(s) + h(\lambda) \\ v_\lambda(0) &= \frac{g}{1-\lambda} + h(0) + h(\lambda) \end{aligned}$$

and

$$[v_\lambda(s) - v_\lambda(0)] = h(s) - h(0) + i(\lambda) \implies \lim_{\lambda \to 1} [v_\lambda(s) - v_\lambda(0)] = h(s).$$

□

Theorem 2.13. *Let $S = \{0, 1, 2, ...\}$ and suppose*
(1) $r(s,a)$ is non-decreasing in s for all $a \in A$,
(2) $\sum_{j=k}^\infty p(j|s,a)$ is non-decreasing in s for all $k \in S$ and $a \in A$,
(3) $r(s,a)$ is superadditive on $S \times A$,

(4) $\sum_{j=k}^{\infty} p(s|j,a)$ *is superadditive on* $S \times A$.

Then if there exists an optimal decision, there exists an optimal decision rule which is non-decreasing in s. *Here,*

$$V^{\sigma} : \textit{set of non-decreasing value functions}$$
$$D^{\sigma} : \textit{set of non-decreasing rules.}$$

Theorem 2.14. *Let* $S = \{0, 1, 2, ...\}$ *and suppose*

(1) $r(s,a)$ *is non-increasing in* s *for all* $a \in A$,

(2) $\sum_{j=k}^{\infty} p(j|s,a)$ *is non-decreasing in* s *for all* $k \in S$ *and* $a \in A$,

(3) $r(s,a)$ *is superadditive on* $S \times A$,

(4) $\sum_{j \in S} p(s|j,a)u(j)$ *is superadditive on* $S \times A$ *for any non-increasing* u.

Then there exists an optimal decision rule which is monotone non-decreasing in s *if there exists an optimal decision rule.*

2.2 Multichain Markov Decision Processes

Example 2.5. Let $S = \{s_1, s_2, s_3\}$, $A_{s_1} = \{a_{11}, a_{12}\}$, $A_{s_2} = \{a_{21}, a_{22}\}$ and $A_{s_3} = \{a_{31}\}$. We have

$$p(s_1|s_1, a_{11}) = 1$$
$$p(s_2|s_1, a_{12}) = 1$$
$$p(s_2|s_2, a_{21}) = 1$$
$$p(s_3|s_2, a_{22}) = 1$$
$$p(s_3|s_3, a_{31}) = 1.$$

Furthermore,

$$r(s_1, a_{11}) = 3$$
$$r(s_1, a_{12}) = 1$$
$$r(s_2, a_{21}) = 0$$
$$r(s_2, a_{22}) = 1$$
$$r(s_3, a_{31}) = 2.$$

The unichain optimality condition is

$$g + h(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} p(j|s,a)h(j) \right\}.$$

This is not sufficient to solve the system.

Multichain Optimality Equations

These are:

$$\max_{a \in A_s} \left\{ \sum_{j \in S} p(j|s,a)g(j) - g(s) \right\} = 0$$

and

$$\max_{a \in B_s} \left\{ r(s,a) - g(s) + \sum_{j \in S} p(j|s,a)h(j) - h(s) \right\} = 0$$

where

$$B_s = \left\{ a \in A_s : \sum_{j \in S} p(j|s,a)g(j) - g(s) = 0 \right\}.$$

As nested optimality equations,

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

and

$$\max_{d\in E}\{r_d - g + (P_d - I)\} = 0 \text{ where } E = \{d \in D : d(s) \in B_s\}.$$

Modified Optimality Equations

These are:

$$\max_{a\in A_s}\left\{\sum_{j\in S} p(j|s,a)g(j) - g(s)\right\} = 0$$

and

$$\max_{a\in A_s}\left\{r(s,a) - g(s) + \sum_{j\in S} p(j|s,a)h(j) - h(s)\right\} = 0.$$

Theorem 2.15. *Suppose S and A_s are finite. Then*

(a) There exists g^* and h for which (g^*, h) satisfy the multichain optimality conditions.

(b) There exists g^* and h' for which (g^*, h') satisfy the modified optimality conditions.

Theorem 2.16. *Suppose S and A_s are finite.*

(a) Suppose g and h satisfy the optimality equations and there exists d^* such that

$$P_{d^*}g = g \text{ and}$$
$$d \in \operatorname{argmax}\{r_d + P_d h\}.$$

Then $(d^*)^\infty$ is long-run average optimal.

(b) Suppose g and h satisfy the modified optimality equations and there exists d^* such that

$$P_{d^*}g = g \text{ and}$$
$$d \in \operatorname{argmax}\{r_d + P_d h\}.$$

Then $(d^*)^\infty$ is long-run average optimal.

The Multichain Policy Iteration

1. Set $n = 0$ and select an arbitrary decision rule d_0

2. (Policy Evaluation) Obtain g_n and h_n such that

$$(I - P_n)g_n = 0$$
$$r_{d_n} - g_n + (P_{d_n} - I)h_n = 0.$$

Solve Step 2 by one of:

[1] $P^*_{d_n}h_n = 0$

[2] Suppose $R_1, ..., R_n$ are recurrent classes under P_{d_n}. Solve the policy evaluation equations by setting $h_n(s_{j_i}) = 0$ where j_i denotes the minimal index such that $s_j \in R_i$ for $i = 1, 2, ..., n$.

[3] $-h_n + (P_d - I)w = 0$

3. (Policy Improvement)

(a) Choose d_{n+1} such that $d_{n+1} \in \operatorname{argmax}_d\{P_d g_n\}$ and setting $d_{n+1} = d_n$ if possible. If $d_{n+1} = d_n$ the go to (b); otherwise increment n by 1 and return to Step 2.

(b) Choose $d_{n+1} \in D$ such that

$$d_{n+1} \in \operatorname*{argmax}_d\{r_d + P_d h_n\}$$

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

4. If $d_{n+1} = d_n$, STOP and set $d^* = d_n$; otherwise increment n by 1 and return to Step 2.

Policy Iteration for Communicating / Weakly Communicating Models

(1) Set $n = 0$. Select a d_0. If P_{d_0} is unichain, set unichain = yes; otherwise, set unichain = no.

(2) If unichain = no, go to (2a), otherwise go to (2b).

(2a) (Policy evaluation) Find vectors g_n and h_n by solving

$$\begin{aligned}(P_{d_n} - I)g_n &= 0\\ r_{d_n} - g_n + (P_{d_n} - I)h_n &= 0\end{aligned}$$

subject to one of [1], [2], [3] in the multichain policy iteration.

(2b) Find scalar g_n and vector h_n by solving

$$r_{d_n} - g_n e + (P_{d_n} - I)h_n = 0$$

subject to one of [1], [2], [3] in the multichain policy iteration.

(3) If g_n is a constant, go (3b), otherwise (3a).

(3a) Let $S_0 = \{s \in S : g_n(s) = \max_{j\in S} g_n(s)\}$ and $d_{n+1}(s) = d_n(s)$ for $s \in S_0$. For $s \in S\backslash S_0$, choose actions that derive the chain to S_0. Set unichain = yes and go to (2).

(3b) Choose $d_{n+1} \in \operatorname{argmax}_{d\in D}\{r_d + P_d h_n\}$, setting $d_{n+1} = d_n$ if possible. If $d_{n+1} = d_n$, go to (4), otherwise set unichain = no, increment n by 1, and go to (2).

(4) Set $d^* = d_n$.

Algorithm for (3a)

For $s \in S_0$, set $d_{n+1}(s) = d_n(s)$ and set $T = S\backslash S_0$.

(i) If $T = \emptyset$, go to (iv)

(ii) Obtain $s' \in T$ and $a \in A_{s'}$ for which $\sum_{j\in S_0} p(j|s', a) > 0$

(iii) Set $T = T\backslash\{s'\}$, $S_0 = S_0 \cup \{s'\}$ and $d_{n+1}(s') = a$ and go to (ii)

(iv) Set unichain = yes and increment n by 1; go to step (2)

Remark 2.2. Given $g \geq P_d g$ and $h + g \geq r_d + P_d h$, suppose that $\alpha(j) > 0$ for all $j \in S$ and $\sum_{j\in S} \alpha(j) = 1$. The primal LP can be written as

$$\begin{aligned}\min \sum_{s\in S} \alpha(s)g(s)&\\ \text{s.t. } g(s) - \sum_{j\in S} p(j|s,a)g(j) \geq 0, \forall a \in A_s, s \in S &\qquad [y(s,a)]\\ g(s) - \sum_{j\in S} p(j|s,a)h(j) + h(s) \geq r(s,a), \forall a \in A_s, s \in S &\qquad [x(s,a)]\end{aligned}$$

and the dual LP is

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

The second set of constraints, summed over $j \in S$, implies that

$$\sum_{j\in S}\sum_{a\in A_j} x(j,a) + \sum_{j\in S}\sum_{a\in A_j} y(j,a) - \sum_{j\in S}\sum_{s\in S}\sum_{a\in A_s} p(j|s,a)y(s,a) = 1$$

and hence

$$\sum_{j\in S}\sum_{a\in A_j} x(j,a) = 1$$

since the last two terms on the LHS are equal.

Remark 2.3. Suppose (x, y) is a feasible solution to the dual LP. Then,

$$P(d_{x,y}(s) = a) = \begin{cases} x(s,a)/\sum_{a\in A_s} x(s,a), & \text{for } s \in S_x \\ y(s,a)/\sum_{a\in A_s} y(s,a), & \text{for } s \notin S_x \end{cases}$$

where $S_x = \{s \in S : \sum_{a\in A_s} x(s,a) > 0\}$.

Proposition 2.5. *If (x, y) is a feasible solution to the dual LP, then S_x is the set of recurrent states and $S\backslash S_x$ si the set of transient states under $(d_{x,y})^\infty$*

Theorem 2.17. *Suppose (x^*, y^*) is an optimal solution to the dual LP. then $(d_{x^*,y^*})^\infty$ is a stationary (long-run average) optimal policy.*

Example 2.6. Consider $S = \{s_1, s_2, s_3, s_4\}$, $A_{s_1} = \{a_{11}\}$, $A_{s_2} = \{a_{21}\}$ and $A_{s_3} = \{a_{31}, a_{32}, a_{33}\}$, and $A_{s_4} = \{a_{41}\}$. We also have

$$\begin{aligned} p(s_3|s_1, a_{11}) &= 1 \\ p(s_3|s_2, a_{21}) &= 1 \\ p(s_1|s_3, a_{31}) &= 1 \\ p(s_2|s_3, a_{32}) &= 1 \\ p(s_4|s_3, a_{33}) &= 1 \\ p(s_4|s_4, a_{41}) &= 1. \end{aligned}$$

Furthermore,

$$\begin{aligned} r(s_1, a_{11}) &= 1 \\ r(s_2, a_{21}) &= 2 \\ r(s_3, a_{31}) &= 4 \\ r(s_3, a_{32}) &= 3 \\ r(s_3, a_{33}) &= 0 \\ r(s_4, a_{41}) &= 2. \end{aligned}$$

The dual LP is

$$\begin{aligned}
\max\ & x(s_1,a_{11}) + 2x(s_2,a_{21}) + 4x(s_3,a_{31}) + 3x(s_3,a_{32}) + 4x(s_4,a_{41}) \\
\text{s.t. } & x(s_1,a_{11}) + y(s_1,a_{11}) - y(s_3,a_{31}) = \frac{1}{4} \\
& x(s_2,a_{21}) + y(s_2,a_{21}) - y(s_3,a_{32}) = \frac{1}{4} \\
& x(s_3,a_{31}) + x(s_3,a_{32}) + y(s_3,a_{33}) + y(s_3,a_{31}) + y(s_3,a_{32}) + y(s_3,a_{33}) \\
& - y(s_1,a_{11}) - y(s_2,a_{21}) = \frac{1}{4} \\
& x(s_4,a_{41}) + y(s_4,a_{41}) - y(s_3,a_{33}) = \frac{1}{4} \\
& x(s_1,a_{11}) - x(s_3,a_{31}) = 0 \\
& x(s_2,a_{21}) - x(s_3,a_{32}) = 0 \\
& x(s_4,a_{41}) - x(s_3,a_{33}) = 0 \\
& x(s_3,a_{31}) + x(s_3,a_{32}) + x(s_3,a_{33}) - x(s_1,a_{11}) - x(s_2,a_{21}) = 0.
\end{aligned}$$

A solution is

$$\begin{aligned}
x(s_1,a_{11}) &= \frac{1}{4} \\
x(s_2,a_{21}) &= \frac{1}{4} \\
x(s_3,a_{31}) &= \frac{1}{8} \\
x(s_3,a_{32}) &= \frac{1}{8} \\
x(s_4,a_{41}) &= \frac{1}{4} \\
x(s_3,a_{33}) &= 0
\end{aligned}$$

and $y(s,a) = 0$ for all s and a.

LP For Weakly Communicating Classes

- Formulate the LP for unichain problem

- Obtain x^* which is an optimal solution of the dual LP for the unichain problem. For $s \in S_{x^*}$ where $S_{x^*} = \left\{s : \sum_{a \in A_s} x(s,a) > 0\right\}$, define $d_{x^*}(s) = a$ for $x^*(s,a) > 0$

- For $s \notin S_{x^*}$, choose an action which drives the chain to S_{x^*} with positive probability

One procedure for this is the algorithm we used in (3a) of the policy iteration for weakly communicating models.

2.3 Uniformization

Uniformization

Let $\{X(t) : t \geq 0\}$ be a continuous time Markov chain with $S = \{0, 1, 2, ...\}$. Let $\lambda(i,j)$ be the rate of going from state i to j and define

$$\lambda(i) = \sum_{j \in S} \lambda(i,j).$$

Assume there exists q such that $\max_{i \in S} \lambda(i) \leq q < \infty$. Let $\{X_n : n \geq 0\}$ be a DTMC with $S = \{0, 1, 2, ...\}$ and

$$P_{ij} = \begin{cases} \frac{\lambda(i,j)}{q}, & \text{for all } j \neq i \\ 1 - \frac{\lambda(i)}{q}, & \text{for all } j = i. \end{cases}$$

Applications

Consider a Markov decision process problem such that $\{X^\pi(t) : t \geq 0\}$ is a continuous time Markov chain under any policy. Let $S = \{0, 1, 2...\}$ and A_s be the set of actions in state s. Let $\lambda(i, j, a)$ be the rate of going from state i to j when action a is chosen in state i where $a \in A_i$.

Let $\lambda(i, a) = \sum_{j \in S} \lambda(i, j, a)$ and suppose there exists q such that

$$\max_{i \in S} \max_{a \in A_i} \lambda(i, a) \leq q < \infty.$$

Using uniformization, we convert the original continuous time problem into discrete time in the following way:

- $S = \{0, 1, 2, ...\}$, A_s remain the same
- $p(j|i, a) = \begin{cases} \frac{\lambda(i,j,a)}{q}, & \text{for } j \neq i \\ 1 - \frac{\lambda(i,a)}{q}, & \text{for } j = i \end{cases}$

Admission Control

Suppose customers come to a system with respect to a Poisson process of rate λ. There is a single server whose service time is exponential with rate μ. At the time of arrival, a gatekeeper may accept or reject the incoming customer.

If he accepts, there is a reward of R. In addition , when there are j customers in the system, there is a per unit time holding cost of $f(j)$.

To model this through uniformization and MDPs, we use the parameters

- $q = \lambda + \mu$
- $\{X^\pi(t), Y^\pi(t) : t \geq 0\}$ where:
  - $X(t)$ is the # of customers at time t
  - $Y(t) \in \{0, 1\}$ where 0 denotes a service completion and 1 denotes an arrival