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: September 13, 2017

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 CS 6505.

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 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.