MARKOV DECISION PROCESSES

LODEWIJK KALLENBERG

UNIVERSITY OF LEIDEN

Preface

Branching out from operations research roots of the 1950's, Markov decision processes (MDPs) have gained recognition in such diverse fields as ecology, economics, and communication engineering. These applications have been accompanied by many theoretical advances. Markov decision processes, also referred to as stochastic dynamic programming or stochastic control problems, are models for sequential decision making when outcomes are uncertain. The Markov decision process model consists of decision epochs, states, actions, rewards, and transition probabilities. Choosing an action in a state generates a reward and determines the state at the next decision epoch through a transition probability function. Policies or strategies are prescriptions of which action to choose under any eventuality at every future decision epoch. Decision makers seek policies which are *optimal* in some sense.

These lecture notes aim to present a unified treatment of the theoretical and algorithmic aspects of Markov decision process models. It can serve as a text for an advanced undergraduate or graduate level course in operations research, econometrics or control engineering. As a prerequisite, the reader should have some backround in linear algebra, real analysis, probability, and linear programming. Throughout the text there are a lot of examples; at the end of each chapter there is a section with bibliographic notes and a section with exercises (the notes include 118 exercises). A solution manuel is available on request (e-mail to kallenberg@math.leidenuniv.nl).

Chapter 1 introduces the Markov decision process model as a sequential decision model with actions, rewards, transitions and policies. We illustrate these concepts with nine different applications: red-black gambling, how-to-serve in tennis, optimal stopping, replacement problems, maintenance and repair, production control, optimal control of queues, stochastic scheduling, and the multi-armed bandit problem.

Chapter 2 deals with the finite horizon model and the principle of dynamic programming, *backward induction*. We also study under which conditions optimal policies are *monotone*, i.e. nondecreasing or nonincreasing in the ordering of the state space.

In chapter 3 the discounted rewards over an infinite horizion are studied. This results in the *optimality equation* and solution methods to solve this equation: *policy iteration, linear programming, value iteration* and *modified value iteration*.

Chapter 4 discusses the total rewards over an infinite horizion under the assumption that the transition matrices are substochastic. We derive equivalent statements for the property that the model is a contracting dynamic programming model. For contracting dynamic programming the similar results and algorithms can be formulated as for the discounted reward model. Special sections are devoted to positive, negative and convergent models, and to optimal stopping problems.

Chapter 5 discusses the criterion of average rewards over an infinite horizion, in the most general case. Firstly, polynomial algorithms are developed to classify MDPs as irreducible or communicating. The distinction between unichain and multichain turns out to be $\mathcal{NP}$-complete, so there is no hope of a polynomial algorithm. Then, the stationary, the fundamental and the deviation matrices are introduced, and the relations between them and their properties are derived. Next, an extension of a theorem by Blackwell and the Laurent series expansion are presented. These results are fundamental to analyse the relation between discounted, average and more sentitive optimality criteria. With these results, as in the discounted case but via a more complicated analysis, the optimality equation is derived, and solution methods to solve this equation are presented (policy iteration, linear programming and value iteration).

In chapter 6 special cases of the average reward criterion (irreducible, unichain and communicating) are considered. In all these cases the optimality equation and the methods of policy iteration, linear programming, value iteration and modified value iteration can be simplified.

Chapter 7 introduces more sensitive optimality criteria: bias optimality, n-discount optimality and Blackwell optimality. We present a unifying framework, based on the Laurent series expansion, to derive *sensitive discount optimality equations*. Using a lexicographic ordering of the Laurent series, we derive the policy iteration method for n-discount optimality. For bias optimality, an optimal policy can be found with a three-step linear programming approach. When in addition the model is a unichain MDP, the linear programs for bias optimality can be simplified. In the irreducible case, one can derive a sequence of nested linear programs to compute n-discount optimal policies for any n. Also for Blackwell optimality, even in the most general case, linear programming can be applied. However, then the elements are not real numbers, but lie in a much general ordered field, namely in an ordered field of rational functions.

In chapter 8 the applications introduced in chapter 1 are analysed in much more detail. In most cases theoretical and computational (algorithms) results are presented. It turns out that in many cases polynomial algorithms exist, e.g. of order $\mathcal{O}(N^3)$, where N is the number of states.

Chapter 9 deals with some other topics: additional constraints and multiple objectives (both for discounted MDPs as well as for average MDPs), and mean-variance tradeoffs.

In the last chapter (chapter 10) *stochastic games* are treated, particularly the *two-person zero-sum* stochastic game. Then, both players may choose actions from their own action sets, resulting in transitions and rewards determined by both players. Zero-sum means that the reward for player 1 has to be payed by player 2. Hence, there is a conflicting situation: player 1 wants to maximize the rewards, while player 2 tries to mimimize the rewards. We discuss the *value* of the game and the concept of optimal policies for discounted as well as for average rewards. We also derive mathematical programming formulations and iterative methods. In some special cases we can present finite methods.

For these notes I have used a lot of material, collected over the years, from various sourses. It is impossible for me to mention them here. However, there is one exception. I wish to express my gratitude to Arie Hordijk, who introduced me to the topic of MDP, as my teacher and former colleague. Already 20 years ago we started with a first draft of some chapters for a book on MDP. Our work was interrupted and, unfortunately, never continued.

Lodewijk Kallenberg

Leiden, April, 2009.

Contents

1 Introduction **1**

1.1 The MDP model . . . 1

1.2 Policies and optimality criteria . . . 3

1.2.1 Policies . . . 3

1.2.2 Optimality criteria . . . 6

1.3 Examples . . . 15

1.3.1 Red-black gambling . . . 15

1.3.2 Gaming: How to serve in tennis . . . 16

1.3.3 Optimal stopping . . . 17

1.3.4 Replacement problems . . . 18

1.3.5 Maintenance and repair . . . 19

1.3.6 Production control . . . 20

1.3.7 Optimal control of queues . . . 20

1.3.8 Stochastic scheduling . . . 22

1.3.9 Multi-armed bandit problem . . . 23

1.4 Bibliographic notes . . . 24

1.5 Exercises . . . 26

2 Finite Horizon **29**

2.1 Introduction . . . 29

2.2 Backward induction . . . 29

2.3 An equivalent stationary infinite horizon model . . . 32

2.4 Monotone optimal policies . . . 33

2.5 Bibliographic notes . . . 36

2.6 Exercises . . . 36

3 Discounted rewards **39**

3.1 Introduction . . . 39

3.2 Monotone contraction mappings . . . 40

3.3 The optimality equation . . . 43

3.4 Policy iteration . . . 48

3.5 Linear programming . . . 54

3.6 Value iteration . . . 64
3.7 Modified Policy Iteration . . . 74
3.8 Bibliographic notes . . . 83
3.9 Exercises . . . 84

4 Total reward **89**
4.1 Introduction . . . 89
4.2 Equivalent statements for contracting . . . 93
4.3 The contracting model . . . 98
4.4 Positive MDPs . . . 101
4.5 Negative MDPs . . . 105
4.6 Convergent MDPs . . . 110
4.7 Special models . . . 113
4.7.1 Red-black gambling . . . 113
4.7.2 Optimal stopping . . . 117
4.8 Bibliographic notes . . . 121
4.9 Exercises . . . 121

5 Average reward - general case **125**
5.1 Introduction . . . 125
5.2 Classification of MDPs . . . 126
5.2.1 Definitions . . . 126
5.2.2 Classification of Markov chains . . . 126
5.2.3 Classification of Markov decision chains . . . 127
5.3 Stationary, fundamental and deviation matrix . . . 134
5.3.1 The stationary matrix . . . 134
5.3.2 The fundamental matrix and the deviation matrix . . . 137
5.4 Extension of Blackwell's theorem . . . 141
5.5 The Laurent series expansion . . . 142
5.6 The optimality equation . . . 143
5.7 Policy iteration . . . 146
5.8 Linear programming . . . 148
5.9 Value iteration . . . 156
5.10 Bibliographic notes . . . 162
5.11 Exercises . . . 164

6 Average reward - special cases **167**
6.1 The irreducible case . . . 167
6.1.1 Optimality equation . . . 167
6.1.2 Policy iteration . . . 169
6.1.3 Linear programming . . . 171

6.1.4 Value iteration . . . 174
6.1.5 Modified policy iteration . . . 174
6.2 Unichain case . . . 178
6.2.1 Optimality equation . . . 178
6.2.2 Policy iteration . . . 179
6.2.3 Linear programming . . . 181
6.2.4 Value iteration . . . 182
6.2.5 Modified policy iteration . . . 183
6.3 Communicating case . . . 188
6.3.1 Optimality equation . . . 188
6.3.2 Policy iteration . . . 188
6.3.3 Linear programming . . . 191
6.3.4 Value iteration . . . 195
6.3.5 Modified value iteration . . . 195
6.4 Bibliographic notes . . . 199
6.5 Exercises . . . 199

7 More sensitive optimality criteria 203
7.1 Introduction . . . 203
7.2 Equivalence between n-discount and n-average optimality . . . 204
7.3 Stationary optimal policies and optimality equations . . . 206
7.4 Lexicographic ordering of Laurent series . . . 210
7.5 Policy iteration for n-discount optimality . . . 214
7.6 Linear programming and n-discount optimality (irreducible case) . . . 220
7.6.1 Average optimality . . . 220
7.6.2 Bias optimality . . . 221
7.6.3 n-discount optimality . . . 223
7.7 Blackwell optimality and linear programming . . . 224
7.8 Bias optimality and linear programming . . . 231
7.8.1 The general case . . . 231
7.8.2 The unichain case . . . 240
7.9 Overtaking and average overtaking optimality . . . 241
7.10 Bibliographic notes . . . 242
7.11 Exercises . . . 243

8 Special models 247
8.1 Replacement problems . . . 247
8.1.1 A general replacement model . . . 247
8.1.2 A replacement model with increasing deterioration . . . 252
8.1.3 Skip to the right model with failure . . . 254
8.1.4 A separable replacement problem . . . 255

8.2 Maintenance and repair problems . . . 256
8.2.1 A surveillance-maintenance-replacement model . . . 256
8.2.2 Optimal repair allocation in a series system . . . 259
8.3 Production and inventory control . . . 262
8.3.1 No backlogging . . . 262
8.3.2 Backlogging . . . 265
8.3.3 Inventory control and single-critical-number policies . . . 268
8.3.4 Inventory control and (s, S)-policies . . . 271
8.4 Optimal control of queues . . . 276
8.4.1 The single-server queue . . . 276
8.4.2 Parallel queues . . . 281
8.5 Stochastic scheduling . . . 282
8.5.1 Maximizing finite-time returns on a single processor . . . 282
8.5.2 Optimality of the μc-rule . . . 283
8.5.3 Optimality of threshold policies . . . 285
8.5.4 Optimality of join-the-shortest-queue policies . . . 286
8.5.5 Optimality of LEPT and SEPT policies . . . 288
8.5.6 Maximizing finite-time returns on two processors . . . 294
8.5.7 Tandem queues . . . 295
8.6 Multi-armed bandit problems . . . 297
8.6.1 Introduction . . . 297
8.6.2 A single project with a terminal reward . . . 298
8.6.3 Multi-armed bandits . . . 299
8.6.4 Methods for the computation of the Gittins indices . . . 304
8.7 Separable problems . . . 310
8.7.1 Introduction . . . 310
8.7.2 Examples (part 1) . . . 311
8.7.3 Discounted rewards . . . 313
8.7.4 Average rewards - unichain case . . . 315
8.7.5 Average rewards - general case . . . 318
8.7.6 Examples (part 2) . . . 325
8.8 Exercises . . . 328
8.9 Bibliographic notes . . . 330

9 Other topics **333**
9.1 Additional constraints . . . 333
9.1.1 Introduction . . . 333
9.1.2 Infinite horizon and discounted rewards . . . 334
9.1.3 Infinite horizon and average rewards . . . 337
9.2 Multiple objectives . . . 349

9.2.1 Discounted rewards . . . 350
9.2.2 Average rewards . . . 352
9.3 Mean-variance tradeoffs . . . 354
9.3.1 Formulations of the problem . . . 354
9.3.2 A unifying framework . . . 356
9.3.3 Determination of an optimal solution . . . 356
9.3.4 Determination of an optimal policy . . . 361
9.4 Bibliographic notes . . . 363
9.5 Exercises . . . 363

10 Stochastic Games **365**
10.1 Introduction . . . 365
10.1.1 The model . . . 365
10.1.2 Optimality criteria . . . 366
10.1.3 Matrix games . . . 367
10.2 Discounted rewards . . . 370
10.2.1 Value and optimal policies . . . 370
10.2.2 Mathematical programming . . . 373
10.2.3 Iterative methods . . . 374
10.2.4 Finite methods . . . 382
10.3 Average rewards . . . 394
10.3.1 Value and optimal policies . . . 394
10.3.2 The Big Match . . . 394
10.3.3 Mathematical programming . . . 399
10.3.4 Perfect information and irreducible games . . . 402
10.3.5 Finite methods . . . 408
10.4 Bibliographic notes . . . 414
10.5 Exercises . . . 416

Chapter 1

Introduction

In this first chapter we introduce the model of a Markov decision process (for short MDP). We present several optimality criteria and give some examples of problems that can be modelled as an MDP.

1.1 The MDP model

An MDP is a model for sequential decision making under uncertainty, taking into account both the short-term outcomes of current decisions and opportunities for making decisions in the future. While the notion of an MDP may appear quite simple, it encompasses a wide range of applications and has generated a rich mathematical theory. In an MDP model one can distinguish the following seven characteristics.

1. The state space
At any time point at which a decision has to be made, the state of the system is observed by the decision maker. The set of possible states is called the state space and will be denoted by S. The state space may be finite, denumerable, compact or even more general. In a finite state space, the number of states, i.e. $|S|$, will be denoted by N.

2. The action sets
When the decision maker observes that the system is in state i, he (we will refer to the decision maker as 'he') chooses an action from a certain action set that may depend on the observed state: the action set in state i is denoted by $A(i)$. Similarly to the state space the action sets may be finite, denumerable, compact or more general.

3. The decision time points
The time intervals between the decision points may be constant or random. In the first case the model is said to be a *Markov decision process*; when the times between consecutive decision points are random the model is called a *semi-Markov decision process*.

4. The immediate rewards (or costs)

Given the state of the system and the chosen action, an immediate reward (or cost) is earned (there is no essential difference between rewards and costs, because maximizing rewards is equivalent to minimizing costs). These rewards depend on the decision time point, the observed state and the chosen action and not on the history of the process. The immediate reward at decision time point t for an action a in state i will be denoted by $r_i^t(a)$; if the reward is independent of the time t, we will write $r_i(a)$ instead of $r_i^t(a)$.

5. The transition probabilities

Given the state of the system and the chosen action, the state at the next decision time point is determined by a transition law. These transitions only depend on the decision time point t, the observed state i and the chosen action a and not on the history of the process. This property is called the *Markov property*. If the transitions really depend on the decision time point, the problem is said to be *nonstationary*. If the state at time t is i and action a is chosen, we denote the probability that at the next time point the system is in state j by $p_{ij}^t(a)$. If the transitions are independent of the time points, the problem is called *stationary*, and the transition probabilities are denoted by $p_{ij}(a)$.

6. The planning horizon

The process has a planning horizon, i.e. the time points at which the system has to be controlled. This horizon may be finite, infinite or of random length.

7. The optimality criterion

The objective is to determine a policy, i.e. a decision rule for each decision time point and each history (including the present state) of the process, that optimizes the performance of the system. The performance is measured by a utility function. This function assigns to each policy, given the starting state of the process, a value. In the next section we will present several optimality criteria.

Example 1.1 *Inventory model with backlogging*

An inventory has to be managed over a planning horizon of T weeks. The optimization problem is: which inventory strategy minimizes the total expected costs?

At the beginning of each week the manager observes the inventory on hand and decides how many units to order. We assume that orders can be delivered instantaneously, and that there is a finite inventory capacity of B units. We also assume that the demands D_t in week t, $1 \leq t \leq T$, are independent random variables that have nonnegative integer values and that the numbers $p_j(t) := \mathbb{P}\{D_t = j\}$ are known for all $j \in \mathbb{N}_0$ and $t = 1, 2, \ldots, T$. If the demand during a period exceeds the inventory on hand, the shortage is backlogged in the next period.

Let the inventory at the start of week t be i (shortages are modelled as negative inventory), let the number of ordered units be a and let j be the inventory at the end of week t.

Then, the following costs are involved, where we use the notation $\delta(x) = \begin{cases} 1 & \text{if } x \geq 1; \\ 0 & \text{if } x \leq 0. \end{cases}$

ordering costs: $K_t \cdot \delta(a) + k_t \cdot a$;
inventory costs: $h_t \cdot \delta(j) \cdot j$;
backlogging costs: $q_t \cdot \delta(-j) \cdot (-j)$.

The data K_t, k_t, h_t, q_t and $p_j(t),\ j \in \mathbb{N}$, are known for all $t \in \{1, 2, \ldots, T\}$.

If an order is made in week t, there is a fixed cost K_t and a cost k_t for each ordered unit. If at the end of the week there is a positive inventory, then there are inventory costs of h_t per unit; when there is shortage, there are backlogging costs of q_t per unit.

This inventory problem can be modelled as a nonstationary MDP over a finite planning horizon, with a denumerable state space and finite action sets:

$S = \{\ldots, -1, 0, 1, \ldots, B\};\ A(i) = \{a \geq 0 \mid 0 \leq i + a \leq B\};$

$p_{ij}^t(a) = \begin{cases} p_{i+a-j}(t) & j \leq i + a; \\ 0 & B \geq j > i + a; \end{cases}$

$r_i^t(a) = -\{K_t \cdot \delta(a) + k_t \cdot a + \sum_{j=0}^{i+a} p_j(t) \cdot h_t \cdot (i + a - j) + \sum_{j=i+a+1}^{\infty} p_j(t) \cdot q_t \cdot (j - i - a)\}.$

1.2 Policies and optimality criteria

1.2.1 Policies

A *policy* R is a sequence of decision rules: $R = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$, where π^t is the decision rule at time point $t,\ t = 1, 2, \ldots$. The *decision rule* π^t at time point t may depend on all availlable information on the system until time t, i.e. on the states at the time points $1, 2, \ldots, t$ and the actions at the time points $1, 2, \ldots, t-1$. The formal definition of a policy is as follows. Consider the Cartesian product

$$S \times A = \{(i, a) \mid i \in S,\ a \in A(i)\}$$

and let H_t denote the set of the possible *histories* of the system up to time point t, i.e.

$$H_t := \{h_t = (i_1, a_1, \ldots, i_{t-1}, a_{t-1}, i_t) \mid (i_k, a_k) \in S \times A,\ 1 \leq k \leq t-1;\ i_t \in S\}. \tag{1.1}$$

A decision rule π^t at time point t is function on H_t which prescribes the action to be taken at time t as a transition probability from H_t into A, i.e.

$$\pi_{h_t a_t}^t \geq 0 \text{ for every } a_t \in A(i_t) \text{ and } \sum_{a_t} \pi_{h_t a_t}^t = 1 \text{ for every } h_t \in H_t. \tag{1.2}$$

Let C denote the set of all policies. A policy is said to be *memoryless* if the decision rule π^t is independent of $(i_1, a_1, \ldots, i_{t-1}, a_{t-1})$ for every $t \in \mathbb{N}$. For a memoryless policy the decision rule

at time t only depends on the state i_t; therefore the notation $\pi^t_{i_t a_t}$ is used. We call $C(M)$ the set of the memoryless policies. Memoryless policies are also called *Markov policies*.

If a policy is memoryless and the decision rules are independent of the time point t, i.e. $\pi^1 = \pi^2 = \cdots$, then the policy is called *stationary*. Hence, a stationary policy is determined by a nonnegative function π on $S \times A$ such that $\sum_a \pi_{ia} = 1$ for every $i \in S$. The stationary policy $R = (\pi, \pi, \dots)$ is denoted by π^∞. The set of stationary policies is notated by $C(S)$.

If the decision rule π of a stationary policy is nonrandomized, i.e. for every $i \in S$, we have $\pi_{ia} = 1$ for (exactly) one action a_i (consequently $\pi_{ia} = 0$ for every $a \neq a_i$), then the policy is called *deterministic*. A deterministic policy can be described by a function f on S, where $f(i)$ is the chosen action a_i, $i \in S$. A deterministic policy is denoted by f^∞ (and sometimes by f) and the set of deterministic policies by $C(D)$.

A matrix $P = (p_{ij})$ is a *transition matrix* if $p_{ij} \geq 0$ for all (i, j) and $\sum_j p_{ij} = 1$ for all i. For a Markov policy $R = (\pi^1, \pi^2, \dots)$ the transition matrix $P(\pi^t)$ and the reward vector $r(\pi^t)$ are defined by

$$\left\{P(\pi^t)\right\}_{ij} = \sum_a p^t_{ij}(a)\pi^t_{ia} \text{ for every } (i,j) \in S \times S \text{ and } t \in \mathbb{N}; \tag{1.3}$$

$$\left\{r(\pi^t)\right\}_i = \sum_a r^t_i(a)\pi^t_{ia} \text{ for every } i \in S \text{ and } t \in \mathbb{N}. \tag{1.4}$$

Consider an initial distribution β, i.e. β_i is the probability that the system starts in state i, and an policy R. Then, by a theorem of Ionescu Tulcea (cf. Bertsekas and Shreve [14] p.140), there exists a unique probability measure $\mathbb{P}_{\beta,R}$ on H_∞, where

$$H_\infty = \{h_\infty = (i_1, a_1, i_2, a_2, \dots) \mid (i_k, a_k) \in S \times A,\ k = 1, 2, \dots\}. \tag{1.5}$$

Let the random variables X_t and Y_t denote the state and action at time t, $t = 1, 2, \dots$, and let $\mathbb{P}_{\beta,R}\{X_t = j, Y_t = a\}$ be the probability that at time t the state is j and the action is a, given that policy R is used and the initial distribution is β. If $\beta_i = 1$ for some $i \in S$, then we write $\mathbb{P}_{i,R}$ instead of $\mathbb{P}_{\beta,R}$. The expectation operator with respect to the probability measure $\mathbb{P}_{\beta,R}$ or $\mathbb{P}_{i,R}$ is denoted by $\mathbb{E}_{\beta,R}$ or $\mathbb{E}_{i,R}$, repectively.

Lemma 1.1

For any Markov policy $R = (\pi^1, \pi^2, \dots)$, any initial distribution β and $t \in \mathbb{N}$, we have

(1) $\mathbb{P}_{\beta,R}\{X_t = j, Y_t = a\} = \sum_i \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij} \cdot \pi^t_{ja}$, $(j, a) \in S \times A$.

(2) $\mathbb{E}_{\beta,R}\{r^t_{X_t}(Y_t)\} = \sum_i \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t)\}_i$, *where* $P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1}) = I$ *(the identity matrix) for* $t = 1$.

Proof

By induction on t. For $t = 1$,

$$\mathbb{P}_{\beta,R}\{X_t = j, Y_t = a\} = \beta_j \cdot \pi^1_{ja} = \sum_i \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij} \cdot \pi^t_{ja}$$

and

$$\mathbb{E}_{\beta,R}\{r^t_{X_t}(Y_t)\} = \sum_{i,a}\ \beta_i \cdot \pi^1_{ia} r^1_i(a) = \sum_i\ \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t)\}_i.$$

Assume that the results are shown for t; we show that the results also hold for $t+1$:

$$\begin{aligned}
\mathbb{P}_{\beta,R}\{X_{t+1}=j, Y_{t+1}=a\} &= \textstyle\sum_{k,b} \mathbb{P}_{\beta,R}\{X_t=k, Y_t=b\} \cdot p^t_{kj}(b) \cdot \pi^{t+1}_{ja} \\
&= \textstyle\sum_{k,b,i}\ \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ik} \cdot \pi^t_{kb} \cdot p^t_{kj}(b) \cdot \pi^{t+1}_{ja} \\
&= \textstyle\sum_i\ \beta_i \cdot \sum_k \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ik} \cdot \sum_b\ \pi^t_{kb} \cdot p^t_{kj}(b) \cdot \pi^{t+1}_{ja} \\
&= \textstyle\sum_i\ \beta_i \cdot \sum_k \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ik} \cdot \{P(\pi^t)\}_{kj} \cdot \pi^{t+1}_{ja} \\
&= \textstyle\sum_i\ \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^t)\}_{ij} \cdot \pi^{t+1}_{ja}.
\end{aligned}$$

Furthermore, one has

$$\begin{aligned}
\mathbb{E}_{\beta,R}\{r^{t+1}_{X_{t+1}}(Y_{t+1})\} &= \textstyle\sum_{j,a}\ \mathbb{P}_{\beta,R}\{X_{t+1}=j, Y_{t+1}=a\} \cdot r^{t+1}_j(a) \\
&= \textstyle\sum_{j,a,i}\ \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^t)\}_{ij} \cdot \pi^{t+1}_{ja} \cdot r^{t+1}_j(a) \\
&= \textstyle\sum_i\ \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^t)\}_{ij} \cdot \sum_a\ \pi^{t+1}_{ja} \cdot r^{t+1}_j(a) \\
&= \textstyle\sum_i\ \beta_i \cdot \sum_j \{P(\pi^1)P(\pi^2)\cdots P(\pi^t)\}_{ij} \cdot \{r(\pi^{t+1})\}_j \\
&= \textstyle\sum_i\ \beta_i \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^t)r(\pi^{t+1})\}_i.
\end{aligned}$$

□

The next theorem shows that for any initial distribution β, any sequence of policies $R_1, R_2, \ldots$ and any convex combination of the marginal distributions of $\mathbb{P}_{\beta,R_k},\ k \in \mathbb{N}$, there exists a Markov policy R_* with the same marginal distribution.

Theorem 1.1

For any initial distribution β, any sequence of policies $R_1, R_2, \ldots$ and any sequence of nonnegative real numbers $p_1, p_2, \ldots$ satisfying $\sum_k\ p_k = 1$, there exists a Markov policy R_* such that

$$\mathbb{P}_{\beta,R_*}\{X_t=j, Y_t=a\} = \sum_k\ p_k \cdot \mathbb{P}_{\beta,R_k}\{X_t=j, Y_t=a\},\ (j,a) \in S \times A,\ t \in \mathbb{N}. \tag{1.6}$$

Proof

Define the Markov policy $R_* = (\pi^1, \pi^2, \ldots)$ by

$$\pi^t_{ja} = \frac{\sum_k\ p_k \cdot \mathbb{P}_{\beta,R_k}\{X_t=j, Y_t=a\}}{\sum_k\ p_k \cdot \mathbb{P}_{\beta,R_k}\{X_t=j\}},\ t \in \mathbb{N},\ (j,a) \in S \times A. \tag{1.7}$$

In case the denominator is zero, take for $\pi^t_{ja},\ a \in A(j)$ arbitrary nonnegative numbers such that $\sum_a\ \pi^t_{ja} = 1,\ j \in S$.

Take any $(j,a) \in S \times A$. We show the theorem by induction on t. For $t=1$, we obtain

$$\mathbb{P}_{\beta,R_*}\{X_1=j\} = \beta_j \text{ and } \sum_k\ p_k \cdot \mathbb{P}_{\beta,R_k}\{X_1=j\} = \beta_j.$$

If $\beta_j = 0$, then $\mathbb{P}_{\beta,R_*}\{X_1=j, Y_1=a\} = \sum_k\ p_k \cdot \mathbb{P}_{\beta,R_k}\{X_1=j, Y_1=a\} = 0$.

If $\beta_j \neq 0$, then from (1.7) it follows that

$$\begin{aligned}\sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_1 = j, Y_1 = a\} &= \sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_1 = j\} \cdot \pi^1_{ja} = \beta_j \cdot \pi^1_{ja} \\ &= \mathbb{P}_{\beta,R_*}\{X_1 = j, Y_1 = a\}.\end{aligned}$$

Assume that (1.6) holds t. We have to prove that (1.6) also holds for $t+1$.

$$\begin{aligned}\mathbb{P}_{\beta,R_*}\{X_{t+1} = j\} &= \sum_{l,b} \mathbb{P}_{\beta,R_*}\{X_t = l, Y_b = b\} \cdot p^t_{lj}(b) \\ &= \sum_{l,b,k} p_k \cdot \mathbb{P}_{\beta,R_k}\{X_t = l, Y_b = b\} \cdot p^t_{lj}(b) \\ &= \sum_k p_k \cdot \sum_{l,b} \mathbb{P}_{\beta,R_k}\{X_t = l, Y_b = b\} \cdot p^t_{lj}(b) \\ &= \sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1} = j\}.\end{aligned}$$

If $\mathbb{P}_{\beta,R_*}\{X_{t+1} = j\} = 0$, then $\sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1} = j\} = 0$, and consequently,

$$\mathbb{P}_{\beta,R_*}\{X_{t+1} = j, Y_{t+1} = a\} = \sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1} = j, Y_{t+1} = a\} = 0.$$

If $\mathbb{P}_{\beta,R_*}\{X_{t+1} = j\} \neq 0$, then

$$\begin{aligned}\mathbb{P}_{\beta,R_*}\{X_{t+1} = j, Y_{t+1} = a\} &= \mathbb{P}_{\beta,R_*}\{X_{t+1} = j\} \cdot \pi^{t+1}_{ja} = \sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1} = j\} \cdot \pi^{t+1}_{ja} \\ &= \sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1} = j\} \cdot \frac{\sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1}=j, Y_{t+1}=a\}}{\sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1}=j\}} \\ &= \sum_k p_k \cdot \mathbb{P}_{\beta,R_k}\{X_{t+1} = j, Y_{t+1} = a\}.\end{aligned}$$

□

Corollary 1.1

For any starting state i and any policy R, there exists a Markov policy R_* such that

$$\mathbb{P}_{i,R_*}\{X_t = j, Y_t = a\} = \mathbb{P}_{i,R}\{X_t = j, Y_t = a\}, \ t \in \mathbb{N}, \ (j,a) \in S \times A,$$

and

$$\mathbb{E}_{i,R_*}\{r^t_{X_t}(Y_t)\} = \mathbb{E}_{i,R}\{r^t_{X_t}(Y_t)\}, \ t \in \mathbb{N}.$$

1.2.2 Optimality criteria

In this course we consider the following optimality criteria:

1. Total expected reward over a finite horizon.
2. Total expected discounted reward over an infinite horizon.
3. Total expected reward over an infinite horizon.
4. Average expected reward over an infinite horizon.
5. More sensitive optimality criteria over an infinite horizon.

Assumption 1.1

In infinite horizon models we assume that the immediate rewards and the transition probabilities are stationary, and we denote them by $r_i(a)$ and $p_{ij}(a)$, respectively, for all i, j and a.

Total expected reward over a finite horizon

Consider an MDP with a finite planning horizon of T periods. For any policy R and initial state $i \in S$, the *total expected reward* over the planning horizon is defined by:

$$v_i^T(R) = \sum_{t=1}^{T} \mathbb{E}_{i,R}\{r_{X_t}^t(Y_t)\} = \sum_{t=1}^{T} \sum_{j,a} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j^t(a), \ i \in S. \tag{1.8}$$

Interchanging the summation and the expectation in (1.8) is allowed, so $v_i^T(R)$ may also be defined as the expected total reward, i.e.

$$v_i^T(R) = \mathbb{E}_{i,R}\Big\{\sum_{t=1}^{T} r_{X_t}^t(Y_t)\Big\}, \ i \in S.$$

Let

$$v_i^T = sup_{R \in C}\ v_i^T(R), \ i \in S,$$

or in vector notation, $v^T = sup_{R \in C}\ v^T(R)$. The vector v^T is called the *value vector*. From Corollary 1.1 and Lemma 1.1, it follows that

$$v^T = sup_{R \in C(M)}\ v^T(R), \tag{1.9}$$

and

$$v^T(R) = \sum_{t=1}^{T} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t), \text{ for } R = (\pi^1, \pi^2, \cdots) \in C(M). \tag{1.10}$$

A policy R_* is called an *optimal policy* if

$$v^T(R_*) = sup_{R \in C}\ v^T(R).$$

It is nontrivial that there exists an optimal policy: the supremum has to be attained and it has to be attained simultaneously for all starting states. It can be shown (see the next chapter) that an optimal Markov policy $R_* = (f_*^1, f_*^2, \cdots, f_*^T)$ exists, where f_*^t is a deterministic decision rule $1 \le t \le T$.

Total expected discounted reward over an infinite horizon

Assume that an amount r earned at time point 1 is deposited in a bank with *interest rate* ρ. This amount becomes $(1+\rho)\cdot r$ at time point 2, $(1+\rho)^2 \cdot r$ at time point 3, etc. Hence, an amount r at time point 1 is comparable with $(1+\rho)^{t-1} \cdot r$ at time point t, $t = 1, 2, \ldots$.

Let $\alpha = (1+\rho)^{-1}$, called the *discount factor*. Note that $\alpha \in (0, 1)$. Then, conversely, an amount r received at time point t can be considered as equivalent to an amount $\alpha^{t-1} \cdot r$ at time point 1, the so-called *discounted value*.

The reward $r_{X_t}(Y_t)$ at time point t has at time point 1 the discounted value $\alpha^{t-1} \cdot r_{X_t}(Y_t)$. The *total expected α-discounted reward*, given initial state i and policy R, is denoted by $v_i^\alpha(R)$ and defined by

$$v_i^\alpha(R) = \sum_{t=1}^{\infty} \mathbb{E}_{i,R}\{\alpha^{t-1} \cdot r_{X_t}(Y_t)\} = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{j,a} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j(a). \tag{1.11}$$

Another way to consider the discounted reward is by the *expected total α-discounted reward*, i.e.

$$\mathbb{E}_{i,R}\left\{\sum_{t=1}^{\infty} \alpha^{t-1} \cdot r_{X_t}(Y_t)\right\}.$$

Since

$$\left|\sum_{t=1}^{\infty} \alpha^{t-1} \cdot r_{X_t}(Y_t)\right| \leq \sum_{t=1}^{\infty} \alpha^{t-1} \cdot M = (1-\alpha)^{-1} \cdot M,$$

where $M = max_{i,a}|r_i(a)|$, the theorem of dominated convergence (e.g. Bauer [8] p. 71) implies

$$\mathbb{E}_{i,R}\left\{\sum_{t=1}^{\infty} \alpha^{t-1} \cdot r_{X_t}(Y_t)\right\} = \sum_{t=1}^{\infty} \mathbb{E}_{i,R}\{\alpha^{t-1} \cdot r_{X_t}(Y_t)\} = v_i^\alpha(R), \tag{1.12}$$

i.e. the expected total discounted reward and the total expected discounted reward criteria are equivalent.

Let $R = (\pi^1, \pi^2, \dots) \in C(M)$, then

$$v^\alpha(R) = \sum_{t=1}^{\infty} \alpha^{t-1} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t). \tag{1.13}$$

Hence, a stationary policy π^∞ satisfies

$$v^\alpha(\pi^\infty) = \sum_{t=1}^{\infty} \alpha^{t-1} P(\pi)^{t-1} r(\pi). \tag{1.14}$$

Like before, the *value vector* v^α and the concept of optimality for a policy R_* are defined by

$$v^\alpha = sup_R\ v^\alpha(R) \ \text{ and } \ v^\alpha(R_*) = v^\alpha. \tag{1.15}$$

In Chapter 3 we show the existence of an optimal deterministic policy f_*^∞ for this criterion and we show that the value vector v^α is the unique solution of the so-called *optimality equation*

$$x_i = max_{a \in A(i)} \left\{r_i(a) + \alpha \sum_j p_{ij}(a) x_j\right\},\ i \in S. \tag{1.16}$$

Furthermore, we will see that f_*^∞ is an optimal policy if

$$r_i(f_*) + \alpha \sum_j p_{ij}(f_*) v_j^\alpha \geq r_i(a) + \alpha \sum_j p_{ij}(a) v_j^\alpha,\ a \in A(i),\ i \in S. \tag{1.17}$$

Total expected reward over an infinite horizon

The *total expected reward*, given initial state i and policy R, is denoted by $v_i(R)$ and defined by

$$v_i(R) = \sum_{t=1}^{\infty} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\} = \sum_{t=1}^{\infty} \sum_{j,a} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j(a). \tag{1.18}$$

We consider this criterion under the following assumptions.

Assumption 1.2

(1) The model is *substochastic*, i.e. $\sum_j p_{ij}(a) \leq 1$ for all $(i,a) \in S \times A$.

(2) For any initial state i and any policy R the expected total reward $v_i(R)$ is well-defined (possibly $\pm\infty$).

Under the above assumption $v_i(R)$ is well defined for all policies R and intial states i. The *value vector* and the concept of an *optimal policy* are defined in the usual way. Under the additional assumption that every policy R is *transient*, i.e. $\sum_{t=1}^{\infty} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} < \infty$ for all i, j and a, it can be shown (cf. Kallenberg [108], chapter 3) that, with $\alpha = 1$, most properties of the discounted MDP model are valid for total reward.

Average expected reward over an infinite horizon

In the criterion of average reward the limiting behaviour of $\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t)$ is considered for $T \to \infty$. Since $lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t)$ may not exist and interchanging limit and expectation is not allowed, in general, there are four different evaluation measures which can be considered:

1. Lower limit of the average expected reward:
   $\phi_i(R) = liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\},\ i \in S$, with value vector $\phi = sup_R\ \phi(R)$.
2. Upper limit of the average expected reward:
   $\overline{\phi}_i(R) = limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\},\ i \in S$, with value vector $\overline{\phi} = sup_R\ \overline{\phi}(R)$.
3. Expectation of the lower limit of the average reward:
   $\psi_i(R) = \mathbb{E}_{i,R}\{liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\ r_{X_t}(Y_t)\},\ i \in S$, with value vector $\psi = sup_R\ \psi(R)$.
4. Expectation of the upper limit of the average reward:
   $\overline{\psi}_i(R) = \mathbb{E}_{i,R}\{limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\ r_{X_t}(Y_t)\},\ i \in S$, with value vector $\overline{\psi} = sup_R\ \overline{\psi}(R)$.

Lemma 1.2

$\psi(R) \leq \phi(R) \leq \overline{\phi}(R) \leq \overline{\psi}(R)$ for every policy R.

Proof

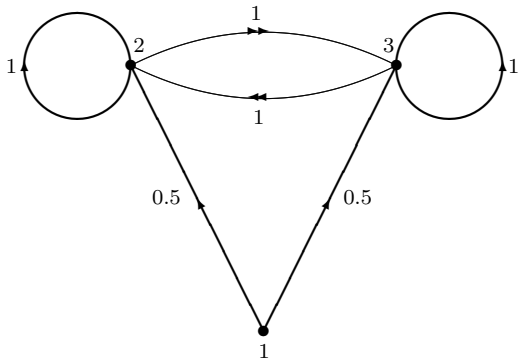
The second inequality is obvious. The first and the last inequality follow from Fatou's lemma (e.g. Bauer [8], p.126):

$$\psi_i(R) = \mathbb{E}_{i,R}\{liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\ r_{X_t}(Y_t)\} \leq liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\} = \phi_i(R)$$

and

$$\overline{\phi}_i(R) = limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\} \leq \mathbb{E}_{i,R}\{limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\ r_{X_t}(Y_t)\} = \overline{\psi}_i(R). \quad \square$$

Example 1.2



Consider the following MDP:
$S = \{1, 2, 3\};\ A(1) = \{1\},\ A(2) = A(3) = \{1, 2\}.$
$p_{11}(1) = 0,\ p_{12}(1) = 0.5;\ p_{13}(1) = 0.5;\ r_1(1) = 0.$
$p_{21}(1) = 0,\ p_{22}(1) = 1;\ p_{23}(1) = 0;\ r_2(1) = 1.$
$p_{21}(2) = 0,\ p_{22}(2) = 0;\ p_{23}(2) = 1;\ r_2(2) = 1.$
$p_{31}(1) = 0,\ p_{32}(1) = 0;\ p_{33}(1) = 1;\ r_3(1) = 0.$
$p_{31}(2) = 0,\ p_{32}(2) = 1;\ p_{33}(2) = 0;\ r_3(2) = 0.$
We use directed graphs to illustrate examples.
The nodes of the graph represent the states.
If the transition probability $p_{ij}(a)$ is positive there is an arc (i, j) from node i to node j;for $a = 1$ we use a simple arc, for $a = 2$ a double arc, etc.; next to the arc the probability $p_{ij}(a)$ is given. The graph of this MDP model is pictured.
If we start in state 1, we never return to that state, but we will remain in state 2 or state 3 for ever. Because of the reward structure, $\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t)$ is the average number of visits to state 2 in the periods $1, 2, \ldots, T$. Let the policy $R = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$ be defined by $\pi^1_{i1} = 1$ for $i = 1, 2, 3$

and for $t \geq 2$: $\pi^t_{h_t a_t} = \begin{cases} \frac{k_t}{k_t+1} & \text{if } a_t = 1 \\ \frac{1}{k_t+1} & \text{if } a_t = 2, \end{cases}$ where for $h_t = (i_1, a_1, i_2, a_2, \ldots, i_{t-1}, a_{t-1}, i_t)$,

$$k_t = max\{k \geq 1 \mid i_{t-1} = i_{t-2} = \cdots i_{t-k+1} = i_t,\ i_{t-k} \neq i_t\},$$

so k_t is the maximum number of periods we consecutively are in state i_t at the time points $t, t-1, \ldots$. This implies that each time we stay in the same state (2 or 3) we have a higher probability to stay there for one more period. The probability to stay in state 2 (or 3) for ever, given that we enter state 2 (or 3), is

$$\mathbb{P}_R\{X_t = 2 \text{ for } t = t_0 + 1, t_0 + 2, \ldots \mid X_{t_0} = 2 \text{ and } X_{t_0-1} \neq 2\} =$$
$$lim_{t\to\infty} \left\{\tfrac{1}{2} \cdot \tfrac{2}{3} \cdot \tfrac{3}{4} \cdots \tfrac{t-1}{t}\right\} = lim_{t\to\infty} \tfrac{1}{t} = 0.$$

Hence, with probability 1, a switch from state 2 to state 3 will occur at some time point; similarly, with probability 1, there is a switch from state 3 to state 2 at some time point. We even can compute the expected number of periods before such a switch occurs:

$$\mathbb{E}_R\{\text{number of consecutive stays in state 2}\} =$$
$$\sum_{k=1}^{\infty} k \cdot \mathbb{P}_R\{X_j = 2 \text{ for } t = t_0 + 1, t_0 + 2, \ldots, t_0 + k - 1;\ X_{t_0+k} \neq 2 \mid X_{t_0} = 2 \text{ and } X_{t_0-1} \neq 2\} =$$
$$\sum_{k=1}^{\infty} k \cdot \left\{\tfrac{1}{2} \cdot \tfrac{2}{3} \cdot \tfrac{3}{4} \cdots \tfrac{k-1}{k} \cdot \tfrac{1}{k+1}\right\} = \sum_{k=1}^{\infty} \tfrac{1}{k+1} = \infty.$$

So, as long as we stay in state 2, we obtain a reward of 1 in each period. The expected numer of stays in state 2 is infinite. Therefore, with probability 1, there is an infinite number of time

points at which the average reward are arbitrary close to 1. Similarly for state 3, with probability 1, there is an infinite number of time points at which the average reward are arbitrary close to 0. This implies for policy R that

$$limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t) = 1 \text{ with probability 1}$$

and

$$liminf_{T}\to\infty\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t) = 0 \text{ with probability 1.}$$

From this we obtain

$\psi_1(R)=\mathbb{E}_{1,R}\{liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t)\}=0$ and $\overline{\psi}_1(R)=\mathbb{E}_{1,R}\{limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t)\}=1$.

If the process starts in state 1, then at any time point t - by symmetry - the probability to be in state 2 and earn 1 will be equal to the probability to be in state 3 and earn 0. So, $\mathbb{E}_{1,R}\{r_{X_t}(Y_t)\} = \frac{1}{2}$ for all $t \geq 2$. Hence,

$$\phi_1(R) = liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i,R}\{r_{X_t}(Y_t)\} = \overline{\phi}_1(R) = limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i,R}\{r_{X_t}(Y_t)\} = \frac{1}{2}.$$

So, this example shows that

$$\psi_1(R) = 0 < \phi_1(R) = \frac{1}{2} = \overline{\phi}_1(R) < \overline{\psi}_1(R) = 1.$$

We also give an example in which $\phi_i(R) < \overline{\phi}_i(R)$ for some policy R and some state i.
Consider the following very simple MDP:
$S = \{1\};\ A(1) = \{1,2\};\ p_{11}(1) = p_{11}(2) = 1;\ r_1(1) = 1,\ r_1(2) = -1.$
Take the policy R that chooses action 1 at $t = 1$; action 2 at $t = 2, 3$; action 1 at $t = 4, 5, 6, 7$; action 2 at $t = 8, 9, \ldots, 15$. In general: action 1 at $t = 2^{2k}, 2^{2k}+1, \ldots, 2^{2k}+2^{2k}-1$ for $k = 0, 1, \ldots$ and action 2 at $t = 2^{2k+1}, 2^{2k+1}+1, \ldots, 2^{2k+1}+2^{2k+1}-1$ for $k = 0, 1, \ldots$.
This gives a deterministic stream of rewards: $+1; -1, -1; +1, +1, +1, +1; -1, -1, -1, -1, -1, -1, -1, -1; \ldots$ with total rewards $+1;\ 0, -1;\ 0, +1, +2, +3, +2, +1,\ 0, -1, -2, -3, -4, -5, -4, -3, -2, -1,\ 0, +1, +2, +3, +4, +5, +6, +7, +8, +9, +10, +11$.
To compute the limsup we take the time points $2^{2k-1} - 1$ voor $k = 1, 2, \ldots$, i.e. the time points $1, 7, 31, 127, \ldots$; for the liminf we consider the time points $2^{2k} - 1$ voor $k = 1, 2, \ldots$, i.e. the time points $3, 15, 63, 255, \ldots$.
Let T_k be the time points when we change the chosen action, i.e. $T_k = 2^k - 1$ for $k = 1, 2, \ldots$, and let A_k denote the total reward at the time points T_k, i.e. $A_1 = +1, A_2 = -1, A_3 = +3, A_4 = -5, A_5 = +11$. It can easily be shown (this is left to the reader) that $|A_k| + |A_{k+1}| = 2^k$ and $|A_{k+1}| = 2|A_k| + (-1)^k$.
This implies that $|A_k| = \frac{1}{3}\{2^k - (-1)^k\}$. Since A_k is positive iff k is odd, we obtain

$$A_k = \frac{1}{3}\{(-1)^{k+1}2^k + 1\},\ k = 1, 2, \ldots.$$

Hence,

$$\phi_1(R) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{1,R}\{r_{X_t}(Y_t)\} = lim_{k\to\infty} \frac{A_{2k}}{2^{2k}-1} = lim_{k\to\infty} \frac{\frac{1}{3}\{-2^{2k}+1\}}{2^{2k}-1} = -\frac{1}{3}$$

and

$$\overline{\phi}_1(R) = limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{1,R}\{r_{X_t}(Y_t)\} = lim_{k\to\infty} \frac{A_{2k-1}}{2^{2k-1}-1} = lim_{k\to\infty} \frac{\frac{1}{3}\{2^{2k-1}+1\}}{2^{2k-1}-1} = +\frac{1}{3}.$$

For these four criteria the *value vector* and the concept of an *optimal policy* can be defined in the usual way. Bierth [21] has shown that

$$\psi(\pi^\infty) = \phi(\pi^\infty) = \overline{\phi}(\pi^\infty) = \overline{\psi}(\pi^\infty) \text{ for every stationary policy } \pi^\infty,$$

and that there exists a deterministic optimal policy which is optimal for all these four criteria. Hence, the four criteria are equivalent in the sense that an optimal deterministic policy for one criterion is also optimal for the other criteria.

More sensitive optimality criteria over an infinite horizon

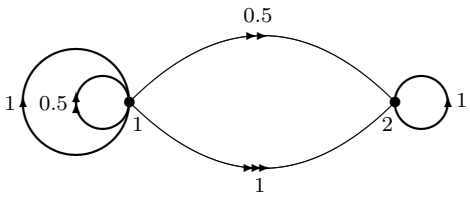
The average reward criterion has the disadvantage that it does not consider rewards earned in a finite number of periods. For example, the streams of rewards $0, 0, 0, 0, 0, \ldots$ and $100, 100, 0, 0, 0, \ldots$ have the same average value although usually the second stream will be preferred. Hence, there is a need for criteria that select policies which are average optimal but also make the right 'early decisions' as well. There are several ways to create more sensitive criteria. One way is to consider discounting for discount factors that tend to 1. Another way is to use more subtle kinds of averaging. We present some of these criteria.

1. Bias optimality
A policy R_* is called bias optimal if $lim_{\alpha\uparrow 1}\ \{v^\alpha(R_*) - v^\alpha\} = 0$.

2. Blackwell optimality
A policy R_* is Blackwell optimal if $v^\alpha(R_*) = v^\alpha$ for all $\alpha \in \{\alpha_0, 1)$ and some $0 < \alpha_0 < 1$.

From this definition it is clear that Blackwell optimality implies bias optimality. The next example shows policies f_1^∞, f_2^∞ and f_3^∞ such that f_1^∞ is average optimal but not bias-optimal, f_2^∞ is bias-optimal but not Blackwell optimal, and f_3^∞ is Blackwell optimal. Therefore, Blackwell optimality is more selective than bias-optimality which in his turn is more selective than average optimality.

Example 1.3
Consider the following MDP:
$S = \{1, 2\};\ A(1) = \{1, 2, 3\},\ A(2) = \{1\}.$
$p_{11}(1) = 1,\ p_{12}(1) = 0;\ p_{11}(2) = \ p_{12}(2) = 0.5;$
$p_{11}(3) = 0,\ p_{12}(3) = 1;\ p_{21}(1) = 0,\ p_{22}(1) = 1;$
$r_1(1) = 0,\ r_1(2) = 1,\ r_1(3) = 2,\ r_2(1) = 0.$
If the system is in state 2, the system stays in state 2 forever and no rewards are earned. In state 1 we have to choose between the actions 1, 2 and 3, which is denoted by the policies $f_1^\infty,\ f_2^\infty$ and f_3^∞, respectively.



All policies have the same average reward (0 for both starting states) and the discounted reward for these policies only differ in state 1 (in state 2 the discounted reward are 0 for every discount factor α).

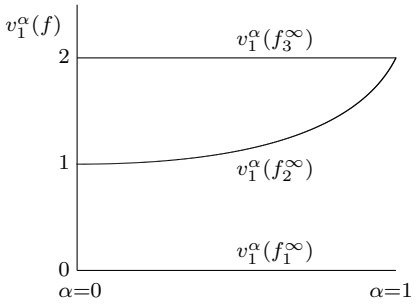
It is easy to see that for all α, we have $v_1^\alpha(f_1^\infty) = 0$ and $v_1^\alpha(f_3^\infty) = 2$. For the second policy, we have an immediate reward 1 and in state one we stay with probability 0.5 and we go to state 2 with probability 0.5. Hence, we obtain $v_1^\alpha(f_2^\infty) = 1 + 0.5\alpha \cdot v_1^\alpha(f_2^\infty) + 0.5\alpha \cdot v_2^\alpha(f_2^\infty)$, so $v_1^\alpha(f_2^\infty) = \frac{2}{2-\alpha}$.

On the right we present a picture of these policies as function of the discount factor α. From this picture it is obvious that $v_1^\alpha = 2$ ($v^\alpha = sup_{f \in C(D)}\, v^\alpha(f)$, this is shown in the next chapter) and that the policy f_3^∞ is the only Blackwell optimal policy. Furthermore, we have $lim_{\alpha\uparrow 1}\, \{v_1^\alpha(f_1^\infty) - v_1^\alpha\} = -2$,

$lim_{\alpha\uparrow 1}\, \{v_1^\alpha(f_2^\infty) - v_1^\alpha\} = lim_{\alpha\uparrow 1}\, \{\frac{2}{2-\alpha} - 2\} = 0$, and

$lim_{\alpha\uparrow 1}\, \{v_1^\alpha(f_3^\infty) - v_1^\alpha\} = lim_{\alpha\uparrow 1}\, \{2 - 2\} = 0$.

Hence, both the policies f_2^∞ and f_3^∞ are bias-optimal.



3. n-discount optimality

For $n = -1, 0, 1, \ldots$ the policy R_* is called n-discount optimal if

$$lim_{\alpha\uparrow 1}\, (1-\alpha)^{-n}\{v^\alpha(R_*) - v^\alpha\} = 0.$$

Obviously, 0-discount optimality is the same as bias-optimality. It can be shown that that (-1)-discount optimality is equivalent to average optimality, and that Blackwell optimality is equivalent to n-discount optimality for all $n \geq |S| - 1 = N - 1$.

4. n-average optimality

Let R be any policy. For $t \in \mathbb{N}$ and $n = -1, 0, 1, \ldots$, we define the vector $v^{n,t}(R)$ by

$$v^{n,t}(R) = \begin{cases} v^t(R) & \text{for } n = -1 \\ \sum_{s=1}^t v^{n-1,s}(R) & \text{for } n = 0, 1, \ldots \end{cases}$$

For $n = -1, 0, 1, \ldots$ a policy R_* is said to be n-average optimal if for all policies R:

$$liminf_{T\to\infty} \frac{1}{T}\{v^{n,T}(R_*) - v^{n,T}(R)\} \geq 0.$$

It can be shown that n-average optimality is equivalent to n-discount optimality.

5. Overtaking optimality

A policy R_* is *overtaking optimal* if $liminf_{T\to\infty}\, \{v^T(R_*) - v^T(R)\} \geq 0$ for all policies R. In contrast with other criteria, an overtaking optimal policy doesn't exist in general.

6. Average overtaking optimality

A policy R_* is *average overtaking optimal* if $liminf_{T\to\infty}\, \frac{1}{T}\sum_{t=1}^T \{v^t(R_*) - v^t(R)\} \geq 0$ for all policies R. It can be shown that average overtaking optimality is equivalent to bias optimality and therefore also equivalent to 0-discount optimality.

1.3 Examples

In Example 1.1 we saw an inventory model with backlogging as an example of an MDP. In this section we introduce other examples of MDPs: gambling, gaming, optimal stopping, replacement, maintenance and repair, production, optimal control of queues, stochastic scheduling and the so-called multi-armed bandit problem. For some models the optimal policy has a special structure. In those cases we will indicate the optimal structure.

1.3.1 Red-black gambling

In red-black gambling a gambler with a fortune of i euro may bet any amount $a \in \{1, 2, \ldots, i\}$. He wins his amount with probability p and he looses with probability $1 - p$. The gambler's goal is to reach a certain fortune N. The gambler continues until either he has reached his goal or he has lost all his money. The problem is to determine a policy that maximizes the probability to reach this goal.

This problem can be modelled as an MDP with the total reward criterion. The fortune of the gambler is the state. Since the gambling problem is over when the gambler has reached his goal or has lost all his money, there are no transitions when the game is in either state N or state 0. Maximizing the probability to reach the amount N is equivalent to assigning a reward 1 to state N and rewards 0 to the other states. The MDP model for the gambling problem is as follows.

$S = \{0, 1, \ldots, N\};\ A(0) = A(N) = \{0\},\ A(i) = \{1, 2, \ldots, min(i, N - i)\},\ 1 \leq i \leq N - 1.$

$$\text{For } 1 \leq i \leq N-1, a \in A(i):\ p_{ij}(a) = \begin{cases} p & ,j = i + a \\ 1 - p & ,j = i - a \\ 0 & ,j \neq i + a, i - a \end{cases} \quad \text{and } r_i(a) = 0.$$

$p_{0j}(0) = p_{Nj}(0) = 0, j \in S;\ r_0(0) = 0, r_N(0) = 1.$

Since under any policy state N or state 0 is reached with probability 1, it is easy to verify that Assumption 1.2 is satisfied. Notice also that

$$v_i(R) = \sum_{t=1}^{\infty} \sum_{j,a} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j(a) = \sum_{t=1}^{\infty} \mathbb{P}_{i,R}\{X_t = N\},$$

i.e. the reward is equal to the probability to reach state N.

It can be shown that an optimal policy has the following structure:

if $p > \frac{1}{2}$, then *timid play*, i.e. always bet the amount 1, is optimal.

if $p = \frac{1}{2}$, then any policy is optimal.

if $p < \frac{1}{2}$, then *bold play*, i.e. betting $min(i, N - i)$ in state i, is optimal.

1.3.2 Gaming: How to serve in tennis

The scoring in tennis is conventionally in steps from 0 to 15 to 30 to 40 to *game*. We simply use the numbers 0 through 4 for these scores. If the score reaches deuce, i.e. 40 - 40, the game is won by the player who has as first two points more than his opponent. Therefore, deuce is equivalent to 30 - 30, and similarly advantage server (receiver) is equivalent to 40 - 30 (30 - 40).

Hence, the scores can be represented by $(i, j),\ 0 \leq i, j \leq 3$, excluding the pair (3,3) which is equivalent to (2,2), where i denotes the score of the server and j of the receiver. Furthermore, we use the states (4) and (5) for the case that the server or the receiver, respectively, wins the game. When the score is (i, j), the server may serve a first service ($s = 1$), or a second service ($s = 2$) in case the first serve is fault. This leads to the following 32 states:

$$\begin{cases} (i,j,s) \quad 0 \leq i,j \leq 3,\ (i,j) \neq (3,3),\ s = 1,2 & : \text{the states in which the game is going on} \\ (4) & : \text{the target state for the server} \\ (5) & : \text{the target state for the reciever} \end{cases}$$

For the sake of simplicity, suppose that the players can choose between two types of services: a *fast service* ($a = 1$) and a *slow service* ($a = 2$). The fast service is more likely to be fault, but also more difficult to return; the slow service is more accurate and easier to return correctly.

Let p_1 (p_2) be the probability that the fast (slow) service is good (lands in the given bounds of the court) and let q_1 (q_2) be the probability of winning the point by the server when the fast (slow) service is good. We make the following obvious assumptions: $p_1 \leq p_2$ and $q_1 \geq q_2$.

Suppose the server chooses for his first service action a. Then, the event that the server serves right and wins the point has probability $p_a q_a$; the event that the server serves right and loses the point has probability $p_a(1 - q_a)$; the event that the server serves a fault and continues with his second service has probability $1 - p_a$. For the second service, the server either wins or loses the point with probabilities $p_a q_a$ and $1 - p_a q_a$, respectively.

In the states where there is no game point, i.e. $i \neq 3$ or $j \neq 3$, we have the following transition probabilities:

$$\begin{cases} p_{(i,j,1)(i+1,j,1)}(a) = p_a q_a; & p_{(i,j,2)(i+1,j,1)}(a) = p_a q_a; \\ p_{(i,j,1)(i,j+1,1)}(a) = p_a(1-q_a); & p_{(i,j,2)(i,j+1,1)}(a) = 1 - p_a q_a; \\ p_{(i,j,1)(i,j,2)}(a) = 1 - p_a. \end{cases}$$

If $i = 3$, we obtain (where $(3, 3, 1)$ has to be replaced by $(2, 2, 1)$)

$$\begin{cases} p_{(3,j,1)(4)}(a) = p_a q_a; & p_{(3,j,2)(4)}(a) = p_a q_a; \\ p_{(3,j,1)(3,j+1,1)}(a) = p_a(1-q_a); & p_{(3,j,2)(3,j+1,1)}(a) = 1 - p_a q_a; \\ p_{(3,j,1)(3,j,2)}(a) = 1 - p_a. \end{cases}$$

Similarly, we obtain for $j = 3$ (also in this case replace $(3, 3, 1)$ by $(2, 2, 1)$)

$$\begin{cases} p_{(i,3,1)(i+1,3,1)}(a) = p_a q_a; & p_{(i,3,2)(i+1,3,1)}(a) = p_a q_a; \\ p_{(i,3,1)(5)}(a) = p_a(1-q_a); & p_{(i,3,2)(5)}(a) = 1 - p_a q_a; \\ p_{(i,3,1)(i,3,2)}(a) = 1 - p_a. \end{cases}$$

When the game is over, i.e. in states (4) and (5), there are no transitions.
The question is: what kind of service to choose, given the score, in order to win the game? To maximize for the server the probability of winning the game, the following reward structure is suitable: all rewards are 0, except in the target state (4) of the server, where the reward in the target state is 1. As utility criterion the total expected reward will be used.

Let $x = \frac{p_1 q_1}{p_2 q_2}$. Then, x is the ratio of serving right and winning the point for the two possible actions $a = 1$ and $a = 2$. It can be shown that the optimal policy has the following structure in each state:

$$\begin{cases} \text{If } x \geq 1 & : \text{always use the fast service} \\ \text{If } 1 - (p_2 - p_1) \leq x < 1 & : \text{use the fast service as first service and the slow service as second} \\ \text{If } x < 1 - (p_2 - p_1) & : \text{use always the slow service} \end{cases}$$

A similar problem is: is it better to use a fast than a slow service to maximize the probability of winning the next point. It turns out that this problem has the same optimal policy. Hence, the optimal policy for winning the game is a *myopic policy*.

1.3.3 Optimal stopping

In an optimal stopping problem there are two actions for every state. The first action is stopping and the second corresponds with continueing. If the stopping action 1 is chosen in state i, then a terminal reward r_i is earned and the process terminates. This termination is modelled by taking all transition probabilities equal to zero. If action 2 is chosen in state i, then a cost c_i is incurred and the probability of being in state j at the next time point is p_{ij}. Hence, the characteristics of the MDP model are:

$$S = \{1, 2, \ldots, N\};\ A(i) = \{1, 2\},\ i \in S;\ r_i(1) = r_i,\ i \in S;\ r_i(2) = -c_i,\ i \in S;$$
$$p_{ij}(1) = 0,\ i, j \in S;\ p_{ij}(2) = p_{ij},\ i, j \in S.$$

We are interested in finding an optimal *stopping policy*. A stopping policy R is a policy such that for any starting state i the process terminates in finite time with probability 1. As optimality criterion the total expected reward is considered.
Notice that for a stopping policy the total expected reward $v(R)$ is well-defined.
Let v be the *value vector* of this model, i.e.

$$v_i = sup\{v_i(R) \mid R \text{ is a stopping policy}\},\ i \in S.$$

A stopping policy R_* is an *optimal policy* if $v(R_*) = v$.
Let

$$S_0 = \left\{ i \in S \;\middle|\; r_i \geq -c_i + \sum_j p_{ij} r_j \right\},$$

i.e. S_0 is the set of states in which immediate stopping is as least as good as continuing for one period and then choosing the stopping action. A *one-step look ahead policy* is a policy which

chooses the stopping action in state i if and only if $i \in S_0$. An optimal stopping problem is called *monotone* if $p_{ij} = 0$ for all $i \in S_0,\ j \notin S_0$, i.e. if S_0 is closed under P. It can be shown that in a monotone optimal stopping problem the one-step look ahead policy is optimal.

Example 1.4 *Selling the house*
Someone wants to sell his house. He receives a proce offer every week. Suppose successive offers are independent and have a value of j euros with probability p_j, for $j = 0, 1, \ldots, N$. We assume that an offer that is not immediately accepted can be accepted at a later time point. When the house remains unsold, then there are maintenance costs c during that week. What is an optimal policy for selling the house?
This problem is an optimal stopping problem: the state space is $S = \{0, 1, \ldots, N\}$, where state i corresponds to the highest offer i so far. In state i there are two actions: accept the offer i (i.e. $r_i = i$) and stop, or continue with costs c and with transition probabilities

$$p_{ij} = \begin{cases} p_j & j > i \\ 1 - \sum_{j>i} p_j & j = i \\ 0 & j < i \end{cases}$$

For this problem

$$S_0 = \left\{ i \in S \;\middle|\; i \geq -c + i \cdot \{1 - \sum_{j>i} p_j\} + \sum_{j>i} j \cdot p_j \right\} = \left\{ i \in S \;\middle|\; c \geq \sum_{j>i} (j-i) p_j \right\}.$$

Notice that $\sum_{j=i+1}^{N} (j-i)p_j = p_{i+1} + 2p_{i+2} + \cdots + (N-i)p_N$ is a monotone nonincreasing function of i. Let

$$i_* = min \left\{ i \;\middle|\; c \geq \sum_{j>i} (j-i) p_j \right\}.$$

Then, $S_0 = \{i \in S \mid i \geq i_*\}$. Since $p_{ij} = 0,\ j < i$, the problem is monotone. An optimal policy accepts the first offer that is at least i_* (such a policy is called a *control-limit policy*). Since $\sum_{j>i}(j-i)p_j$ is the expected additional income above i in the next period, an offer is accepted if the the cost during the next week is at least the expected additional income of the offers next week.

1.3.4 Replacement problems

Consider an item (e.g. a component of an electric system or a truck of a transportation company) that can be in one of a finite number of states, say the states $0, 1, \ldots, N$. Each state may be associated with some parameter, e.g. the age of the item. Suppose that at the beginning of each period the decision has to be made whether or not to replace the item. The motivation for replacing an item is to avoid 'bad' states with high costs.
Action 1 corresponds to replacing the item by a new one (the state of a new item is state 0 and the transition to the new item is instantaneous). For an old item in state i we receive a reward s_i and we have to pay costs c for the new item .

Action 2 is to keep the item for (at least) one more period. Let p_{ij} be the probability that an item of state i is in state j at the beginning of the next period, and suppose that c_i is the maintenance cost for an item of state i during one period.
The characteristics of the MDP model are:

$$S = \{0, 1, \ldots, N\};\ A(0) = \{2\},\ A(i) = \{1, 2\},\ 1 \le i \le N;$$
$$p_{ij}(1) = p_{0j},\ 1 \le i \le N, j \in S;\ p_{ij}(2) = p_{ij},\ 0 \le i \le N, j \in S;$$
$$r_i(1) = s_i - c - c_0,\ 1 \le i \le N;\ r_i(2) = -c_i,\ 0 \le i \le N.$$

Many replacement problems have an optimal *control-limit policy*, i.e. the item is repaced by a new one when the state (age) is at least a given number i_*.

1.3.5 Maintenance and repair

Consider a series system of n unreliable components, maintained by a single repairman. Each of the components may be either working or failed. The state space can be represented by a vector $x = (x_1, x_2, \ldots, x_n)$, where $x_i = 1$ (working) or 0 (failed). The system is functioning if and only if the state is $(1, 1, \ldots, 1)$.
The failure time and repair time of component $i,\ 1 \le i \le n$ are exponentially distributed with rates λ_i and μ_i, respectively, and independently of the state of other components. Notice that, by the memoryless property of the exponential distribution, the elapsed time that a working component operates or a failed component is under repair is not relevant for the description of a state.

It is assumed that the repairman may change instantaneously among failed components. That is, for example, if component i fails while component j is being repaired, the repairman may switch instantaneously from j to i, or to any other failed component.
The objective is to find a policy which assigns the repairman to a failed component in such a way that the average expected time that the system is functioning is maximized. This problem is a finite state *continuous-time Markov decision problem*, i.e. each deterministic and stationary policy f^∞ generates a continuous-time Markov chain. A continuous-time Markov chain is a stochastic process that moves at stochastic time points, where the transitions are according to a discrete-time Markov chain; it remains in a state during an exponentially distributed time before it proceeds to a different state.

Let the deterministic and stationary policy f^∞ assign the repairman to component i in state x. Then, we denote this assignment by $f(x) = i$. We also use the notations:

$$(1_k, x) = (x_1, x_2, \ldots, x_{k-1}, 1, x_{k+1}, \ldots, x_n);\ C_1(x) = \{i \mid x_i = 1\};\ \lambda_1(x) = \textstyle\sum_{i \in C_1(x)}\ \lambda_i;$$
$$(0_k, x) = (x_1, x_2, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n);\ C_0(x) = \{i \mid x_i = 0\}.$$

Given policy f^∞, the Markov chain remains in state x during an exponentially distributed time with rate $\lambda_1(x) + \mu_{f(x)}$. The transition probabilities of the Markov chain satisfy

$$p_{x,(1_{f(x)},x)}(f(x)) = \frac{\mu_{f(x)}}{\lambda_1(x) + \mu_{f(x)}};\ p_{x,(0_k,x)}(f(x)) = \frac{\lambda_k}{\lambda_1(x) + \mu_{f(x)}},\ k \in C_1(x).$$

The following results can be shown:

(1) an optimal policy can be found in the class of deterministic and stationary policies that never leave the repairman idle when there is a failed component.

(2) maximizing the average expected time that the system is functioning, is equivalent to minimizing the time until the functioning state $(1, 1, \ldots, 1)$ is reached.

(3) the optimal policy is irrespective of the repair rates $\mu_i,\ 1 \leq i \leq n$, and is the policy that assigns the repairman to the failed component with the smallest failure rate (*SFR policy*), i.e. the longest expected lifetime.

The results (1) and (2) are intuitively clear; however, the result (3) is counterintuitive.

1.3.6 Production control

Consider a production process of a certain item over a planning horizon of T periods. Let the demand in period t be known and deterministic, say $D_t,\ 1 \leq t \leq T$. The production in period t has a capacity b_t, and let $c_t(a)$ denote the production cost for the production of a units in period $t,\ 1 \leq t \leq T$. In each period the demand has to be fulfilled, so shortages are not allowed and there is no backlogging. There are inventory costs $h_t(i)$ in period t, when the inventory at the end of period t is equal to $i,\ 1 \leq t \leq T$. The inventory at the end of the planning horizon must be zero. The aim is to determine the production in the various periods so as to satisfy the demands at minimum total costs.

The definition of the states is a little tricky for this model. We use a two-dimensional description, namely (i, t) to denote the situation of having i units inventory at the beginning of period t. Actions correspond to production. When in state (i, t) action a is selected, then this action has to satisfy the following three conditions:

(1) $0 \leq a \leq b_t$: the capacity constraint.

(2) $D_t \leq i + a$: the demand requirement.

(3) $i + a \leq \sum_{s=t}^{T} D_s$: the total production may not exceed the total demand for the remaining periods.

Hence, the MDP model for this production problem is:

$S = \{(i, t) \mid 0 \leq i \leq \sum_{s=1}^{T} D_s;\ 1 \leq t \leq T\};\ A[(i, t)] = \{a \mid 0 \leq a \leq b_t;\ D_t \leq i+a \leq \sum_{s=t}^{T} D_s\};$

$p_{(i,t)(i+a-D_t,t+1)}(a) = 1$ for every $(i, t) \in S$ and $a \in A[(i, t)]$;

$r_{(i,t)}(a) = -\{c_t(a) + h_t(i + a - D_t)\}$ for every $(i, t) \in S$ and $a \in A[(i, t)]$.

1.3.7 Optimal control of queues

Consider a single server queueing system where customers arrive according to a Poisson process and where the service time of a customer is exponentially distributed ($M/M/1$ queue). Suppose that the arrival and service rates can be controlled by a finite number of actions. We say that the system is in state i when there are i customers in the system. Action a means that the arrival and

the service rates are $\lambda_i(a)$ and $\mu_i(a)$, respectively. Any customer has waiting cost c per time unit and when a customer enters the system a reward r is incurred. For this model several variations can be considered, changing the assumptions about the decision time points, for example. We discuss two models, *continuous control* and *semi-Markov control*

Continuous control

In continuous control the parameters can be controlled at any time. If the system is in state i and action a is chosen, then the expectation of the interarrival time between new customers is $\lambda_i^{-1}(a)$ and the expectation of the service time equals $\mu_i^{-1}(a)$. One can approach this model by *time discretization*. Then, a discrete MDP approximation scheme can be obtained by using time points $t \cdot h,\ t \in \mathbb{N}_0$, where h is a sufficiently small positive number, called the *step-size*.
Sufficiently small means

$$0 < h < min_{(i,a)} \left\{ min\{\lambda_i^{-1}(a), \mu_i^{-1}(a)\} \right\},$$

in which case the so-called first order approximation of the transition probabilities is allowed. In doing so, we obtain the following MDP model:

$S = \mathbb{N}_0;\ A(i) = \{1, 2, \ldots, m\};\ r_i(a) = \{r \cdot \lambda_i(a) - c \cdot i\} \cdot h;\ p_{i,i+1}(a) = \lambda_i(a) \cdot h;$

$p_{i,i-1}(a) = \delta(i) \cdot \mu_i(a) \cdot h,$ where $\delta(i) = 1$ if $i \geq 1$ and 0 if $i = 0;\ p_{i,i}(a) = 1 - \{p_{i,i+1}(a) + p_{i,i-1}(a)\}.$

Semi-Markov control

Another natural model can be obtained by using the arrival and departure times as the decision time points. Then, we have a semi-Markov decision problem in which the time until the next decision is a random variable which depends only on the current state and the chosen action. In our model the time until the next decision time point is the minimum of two negative exponential distributions. This time has a negative exponential distribution; if i is the current state and action a is chosen, then this exponential distribution has parameter

$$\nu_i(a) = \lambda_i(a) + \delta(i)\mu_i(a).$$

The transition probabilities satisfy

$$p_{i,i-1}(a) = \frac{\delta(i)\mu_i(a)}{\nu_i(a)};\ p_{i,i+1}(a) = \frac{\lambda_i(a)}{\nu_i(a)}.$$

Let $r_i(a)$ be the expected reward until the next decision time point. Then,

$$r_i(a) = \frac{r \cdot \lambda_i(a) - c \cdot i}{\nu_i(a)}.$$

By the technique of *uniformization*, a semi-Markov model can be transformed into an equivalent MDP with equidistant decision epochs. Define $\nu = max_{i,a}\ \nu_i(a)$, and define the transition probabilities p' and the one-step reward r' for the MDP by

$$p'_{ij}(a) = \begin{cases} \frac{\nu_i(a)p_{ij}(a)}{\nu} & j \neq i \\ \frac{\nu - \nu_i(a)}{\nu} & j = i \end{cases} \quad \text{and } r'_i(a) = \frac{r_i(a)\nu_i(a)}{\nu}.$$

It can be shown that these models are equivalent.

1.3.8 Stochastic scheduling

In a scheduling problem, jobs have to be processed on a number of machines. Each machine can only process one job at a time. Each job i has a given processing time T_{ij} on machine j. In stochastic scheduling, these processing times are random variables. At certain time points decisions have to be made, e.g. which job is assigned to which machine. There is a utility function by which different policies can be measured, and we want to find a policy that optimizes the utility function.

This kind of problem is also considered in the control of queues. Then, instead of jobs and machines, the terms customers and servers are used. There are two types of models: *customer assignment models*, in which each arriving customer has to be assigned to one of the queues and *server assignment models*, where servers have to be assigned to one of the queues of customers). We do not explicitly present the MDP model for these stochastic scheduling models. We confine ourselves to the formulation of some variants for which the optimal policy has a nice structure.

One server allocation to parallel queues with preemption: μc-rule
Customers arrive at a system of m parallel queues and one server. The system operates at discrete time points, i.e. arrival times and service times take values in the set $\{1, 2, \dots\}$. Furthermore, the arrival times are arbitrary and the service time T_i, for a customer in queue i, is geometrically distributed with rate μ_i,

$$\mathbb{P}\{T_i = n\} = (1-\mu_i)^{n-1} \cdot \mu_i, \ n \in \mathbb{N}, \text{ with } \mu_i \in (0,1), \ 1 \le i \le m, \text{ and } \mathbb{E}\{T_i\} = \mu_i^{-1}.$$

At any time point $t = 1, 2, \dots$ the server chooses a customer from one of the queues; this is an example of a server assignment model. Services may be interrupted and resumed later on (*preemption*). For each customer in queue i, a cost c_i is charged per unit of time that this customer is in the system. A policy is a rule to assign each server to one of the queues. Which policy minimizes the total cost in T periods?

Let $N_i^t(R)$ be the number of customers in period t in queue i, if policy R is used. Then, the performance measure is

$$min_R \ \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{i=1}^{m} c_i \cdot N_i^t(R)\right\}.$$

It can be shown that the so-called μc-rule is an optimal policy. This rule assigns the server to queue k, where k is a nonempty queue satisfying

$$\mu_k c_k = max_i\{\mu_i c_i \mid \text{queue } i \text{ is nonempty}\}.$$

Note that $\mu_i c_i$ is the expected cost per unit of service for a customer in queue i, and by using the μc-rule, the largest reduction of the expected cost in the next period is obtained.

Poisson arrivals and two servers: threshold policy
Consider a system with two servers where the customers arrive according to a Poisson process with rate λ, and where there is only one queue. The service times are assumed to be exponentially distributed with the respective rates μ_1 (for server 1) and μ_2 (for server 2), where $\mu_1 \geq \mu_2$. When one of the servers becomes available, the decision has to be taken whether or not to send the customer to this server.
This is a customer assignment model. The model is not discrete, but continuous in time. Let $N^t(R)$ be the number of customers in the system at time t. As performance measure the total discounted costs are used, i.e.

$$min_R \ \mathbb{E}\left\{\int_0^\infty \ e^{-\alpha t} N^t(R) dt\right\},$$

where $\alpha > 0$, which is the continuous analogon of the total discounted costs in the discrete case.

For this model an optimal *threshold policy* exists, namely server 1 will always be used when it becomes available, and the slower server, server 2, is only used when the total number of customers in the queue exceeds some threshold number n.

1.3.9 Multi-armed bandit problem

The multi-armed bandit problem is a model for dynamic allocation of a resource to one of n independent alternative projects. The terminology 'multi-armed bandit' comes from the interpretation of the projects as arms of a gambling machine.
Any project may be in one of a finite number of states, say project j in the set $S_j,\ j = 1, 2, \ldots, n$. Hence, the state space S is the Cartesian product

$$S = S_1 \times S_2 \times \cdots \times S_n.$$

Each state $i = (i_1, i_2, \ldots, i_n)$ has the same action set $A = \{1, 2, \ldots, n\}$, where action a means that project a is chosen, $a = 1, 2, \ldots, n$. So, at each stage one can be working on exactly one of the projects.
When project a is chosen in state i - the chosen project is called the active project - the immediate reward and the transition probabilities only depend on the active project, whereas the states of the remaining projects are frozen. As a utility function the total discounted reward is chosen.
There are many applications of this model, e.g. in machine scheduling, in the control of queueing systems and in medicine, when dealing with selection decision trials.

It can be shown that an optimal policy is the policy that selects project a in state $i = (i_1, i_2, \ldots, i_n)$, where a satisfies

$$G_a(i_a) = max_{1 \leq k \leq n}\ G_k(i_k)$$

for certain numbers $G_k(i_k),\ i_k \in S_k,\ 1 \leq k \leq n$. Such a policy is called an *index policy*. Surprisingly, these numbers $G_k(i_k)$ only depend on project k and not on the other projects. This

result is a fundamental contribution made by Gittins and therefore these indices are called the *Gittins indices*.

As a consequence, the multi-armed bandit problem can be solved by a sequence of n one-armed bandit problems. This is a *decomposition* result by which the dimensionality of the problem is reduced considerably. Algorithms with complexity $\mathcal{O}\big(\sum_{k=1}^{n} n_k^3\big)$, where $n_k = |S_k|,\ 1 \leq k \leq n$, do exist for the computation of all indices.

1.4 Bibliographic notes

Bellman's book [11] can be considered as the starting point for the study of Markov decision processes. However, as early as 1953, Shapley's paper [183] on stochastic games includes as a special case a discounted Markov decision process. Around 1960 the basics for solution methods for MDPs were developed in publications as Howard [101], De Ghellinck [39], d'Epenoux [53], Manne [134] and Blackwell [22]. Since the early sixties, many results on MDPs have been published in numerous journals, monographs, books and proceedings. Around 1970 a first series of books was published, e.g. Derman [55], Mine and Osaki [139] and Ross [168]. In 1994, the rather comprehensive book by Puterman was published ([157]).

The result mentioned in Corollary 1.1 on the sufficiency of Markov policies for performance measures, that only depend on the marginal distributions, is due to Derman and Strauch [57]. The extension to Theorem 1.1 was given by Strauch and Veinott [194]. The equivalence between the four criteria for the average reward has been shown by Bierth [21].

In a fundamental paper Blackwell ([22]) introduced the concepts of bias optimality (Blackwell called it *nearly optimal*) and Blackwell optimality. An algorithm for finding a Blackwell optimal policy was constructed by Miller and Veinott [138]. The n-discount optimality criterion was proposed in Veinott [217]. He also showed that Blackwell optimality is equivalent to n-discount optimality for all $n \geq |S| - 1$.

The concept of n-average optimality was announced in Veinott [216], which is an abstract of a preliminary report. This report was never published. In Sladky [186] a proof is given of the equivalence between n-average optimality and n-discount optimality.

The criterion of overtaking optimality was proposed by Denardo and Rothblum [52]. For this criterion no optimal policy may exist. Denardo and Rothblum also provided conditions under which an optimal policy exists. The concept of average overtaking optimality was proposed by Veinott [214], where he used the terminology *optimal*. He presented an algorithm for finding a bias-optimal policy, showed that an average overtaking policy is bias-optimal and conjectured that the converse was also true. This conjecture has been proven by Denardo and Miller [51].

There is also an extensive literature on examples of MDP models. Seminal paper on inventory models are written by Scarf ([172],[173]), Iglehart ([103],[104]) and Veinott [215]. A standard

reference on gambling is Dubins and Savage [61], who have shown for example that the bold policy is optimal if $p \leq \frac{1}{2}$. The optimality of the timid policy for $p \geq \frac{1}{2}$ can be found in Ross ([169] and [170]). The example of the tennis game is due to Norman [143] and Prussing [156].

A dynamic programming approach for optimal stopping problems can be found in Breiman [25], who showed the optimality of control-limit policies. The house selling example comes from Ross [168]. There are a lot of references on replacement models. The survey of Sherif and Smith [185] contains over 500 references. Results on the optimality of control-limit policies can be found in Derman [54], Kolesar [121], Derman [55], Ross [168] and Kao [112]. Our presentation of the n-component series system with exponential distributions is based on Katehakis and Derman [115]. They showed the optimality of the *SFR-policy*. This result was first conjectured by Smith [187].

There is a close relation between production control problems and flows in networks. For more detailed information about this subject we refer the reader to Chapter 5 in Denardo [49]. The literature on optimal control of queues is also quite extensive . Markov decision processes with continuous time parameter were introduced by Bellmann ([11], Chapter 11). The technique of uniformization was already suggested by Howard ([101], page 113). Schweitzer [175] has generalized this idea for general non-exponential mean holding times and has explicitly given the data transformations mentioned at the end of Section 1.3.7.

For reviews on stochastic scheduling we refer to Weiss [227], Walrand ([224], Chapter 8), and Righter [164]. The optimality of the μc-rule is due to Baras, Ma and Makowsky [5], see also Buyukkoc, Varaiya and Walrand [29]. The structural result of an optimal threshold policy in the two server model with Poisson arrivals is from Lin and Kumar [128]. The most fundamental contribution on multi-armed bandit problems has been made by Gittins ([80], [79]). The importance of Gittins' work had not been recognized in the seventies. The re-discovery is due to Whittle [236] who gave an easier and more natural proof. Other proofs are given by Ross [170], Varaiya, Walrand and Buyukkoc [212], Tsitsiklis [197] and Weber [226]. Several methods are developed for the computation of the Gittins indices: Varaiya, Walrand and Buyukkoc [212], Chen and Katehakis [31], Kallenberg [109], Katehakis and Veinott [117], Ben-Israel and Flåm [13], and Liu and Liu [130].

1.5 Exercises

Exercise 1.1 *Inventory model without backlogging*
Consider a finite horizon nonstationary inventory model without backlogging if demands exceed the supply:

$$\begin{aligned} T &= \text{the number of periods in the planning horizon;} \\ p_j(t) &= \text{the probability of demand } j \text{ in period } t,\ j = 0, 1, \ldots; \\ c &= \text{the cost price of an item;} \\ h &= \text{the holding cost of an item that is unsold at the end of a period;} \\ p &= \text{the penalty cost of an item that cannot be delivered during a period;} \\ B &= \text{the finite inventory capacity.} \end{aligned}$$

The optimization problem is: which inventory strategy minimizes the total expected costs?
Formulate this model as a Markov decision model.

Exercise 1.2 *Number of Markov policies*
Let $N = |S|$ and $m_i = |A(i)|,\ i \in S$.
What is the number of nonrandomized Markov policies in this finite horizon MDP with T periods?

Exercise 1.3 *n-discount optimality*
Show that n-discount optimality implies $(n-1)$-discount optimality for $n = 0, 1, \ldots$.

Exercise 1.4 *Red-black gambling with $p = \frac{1}{2}$*
Consider the red-black gambling model with $p = \frac{1}{2}$.
Let f_1^∞ be the deterministic policy betting 1 euro in every round of the game.

a. Show that policy f_1^∞ satisfies $v_i(f_1^\infty) = \frac{i}{N},\ 0 \leq i \leq N$.
   Hint: Derive a recurrence relation and solve it.

b. Show that any deterministic policy f^∞ satisfies $v_i(f^\infty) = \frac{i}{N},\ 0 \leq i \leq N$.

Exercise 1.5 *How to serve in tennis*
Let $v(i, j, s)$ be the probability of winning the next game in tennis when the score is (i, j) and s is the first or secound service the server will play ($s = 1$ or $s = 2$); let $v(4) = 1$ and $v(5) = 0$.
Since winning (loosing) a point will not decrease (increase) the probability of winning a game, we have the relations:

$v(i+1, j, 1) \geq v(i, j, 1);\ v(i+1, j, 1) \geq v(i, j, 2);\ v(i, j, 1) \geq v(i, j, 2);\ v(i, j, 2) \geq v(i, j+1, 1)$.

a. Give the optimality equation for this model.

b. Show the optimality of the policy mentioned in the text as the optimal policy.

Exercise 1.6 *Optimal stopping problem*
Consider a person who wants to sell an asset for which he is offered an amount of money at the beginning of each week. We assume that these offers are independent and that an offer of amount j will be made with probability $p_j,\ 0 \leq j \leq N$. He has to decide immediately either to accept or to reject the offer. If the offer is not accepted, the offer is lost and a cost c is occured. Which policy will maximize the expected total income?
a. Formulate this problem as an optimal stopping problem.
b. Show that this problem has an optimal control-limit policy.

Exercise 1.7 *Automobile replacement problem*
Suppose that we review a car every month and that the decision is made either to keep the present car or to trade in the car for another car of a certain age. The age of a car is measured in months. In order to keep the state space finite, we assume that there is a largest age N, i.e. a car of age N will always be reset by another car. Furthermore, we assume that a car of age i has a probability p_i of a breakdown in which case it ends up in state N.
Suppose that we have the following costs and rewards:

b_i = cost of buying a car of age i;

t_i = trade-in value of a car of age i;

c_i = expected maintenance cost in the next period for a car of age i.

Give the MDP data for this automobile problem.

Exercise 1.8 *Production problem*
Consider the following variant of the production problem formulated above. Let the demand D_t in period t be stochastic with $p_j(t) = \mathbb{P}[D_t = j],\ j = 0, 1, \ldots, N_t$ and $1 \leq t \leq T$. Because of the uncertainties it is no longer possible to satisfy the demands with probability 1. Therefore, we require that the demand in period t has to be satisfied with probability at least $\alpha_t,\ 1 \leq t \leq T$. Formulate the MDP model for this variant of the production problem.

Exercise 1.9 *Queueing problem*
Consider a single server queueing system with a finite capacity N. The service time is a negative exponential distribution with parameter μ. The system manager can control the system by increasing or decreasing the price he charges for the service facility in order to encourage or discourage the arrival of customers.
Assume that the manager must choose one of a finite number of prices, say $p_1, p_2, \cdots, p_m$, where $0 < p_1 < p_2 < \cdots < p_m$. If there are i customers in the system and he chooses p_a, then the arriving process is a Poisson process with parameter λ_a, an arriving customer has to pay p_a and the system manager has c_i as waiting cost per time unit.
It is quite natural to assume that:

(1) $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ (lower prices give more arrivals).
(2) $0 \leq c_0 \leq c_1 \leq \cdots \leq c_N$ (more costs for longer waiting).
(3) $p_m > c_{N-1}$ (possibility for the manager to get a positive net reward for each arriving customer).

a. Give the specifications for the time discretization approach in case of continuous control.
b. Give the specifications for the semi-Markov approach. Apply uniformization to obtain an equivalent MDP model.

Exercise 1.10 *Stochastic scheduling: μc-rule*

Assume that m customers are present at the service station and have to be processed nonpreemptively by one server. Let μ_i^{-1}be the expected service time and c_i the cost per unit time for customer i. Show, by an interchanging argument, that the μc-rule is optimal for scheduling the jobs in order to minimize the total expected costs.

Chapter 2

Finite Horizon

2.1 Introduction

A system with rewards $r_i^t(a)$ and transition probabilities $p_{ij}^t(a)$ has to be controlled over a planning horizon of T periods. As you see in the notation, these rewards and transition probabilities may be nonstationary. As the utility function the total expected reward is considered as defined in (1.8). As will be shown in the next section, an optimal Markov policy with deterministic but in general nonstationary decision rules exists. Furthermore, we show that such an optimal policy can be obtained by *backward induction.*, that is based on the *principle of optimality.*

In section 2.3 an alternative stationary model over an infinite horizon is described. This model is equivalent to the finite horizon nonstationary model in the sense that there is equivalence between the policies in both models such that equivalent policies have the same value of their utility functions. Hence, results of the infinite horizon model, such as the treatment of MDP's with *side constraints*, can be applied to the finite horizon model.
In section 2.4 we study under which conditions optimal policies are *monotone*, i.e. nondecreasing or nonincreasing. Such a concept is worthwhile if there is a natural *ordering* in the state space. Knowlegde about the monotone structure of optimal policies enables us to find such policy with less computational effort than without the monotone structure.

2.2 Backward induction

In this section we see how to compute an optimal policy by backward induction. Backward induction is an iterative procedure. Starting at the end of the planning horizon one computes the values for the previous periods. Then, after T iterations, where T is the number of periods in the planning horizon, an optimal policy is found.
The notation $r(f^t)$ and $P(f^t)$, as defined in (1.4) and (1.3) respectively, is used for the reward vector and transition matrix of a deterministic decision rule f^t at decision time point t. In a finite planning horizon with T periods only the decision rules for the first T decision time points are relevant. Hence, we write $R = (\pi^1, \pi^2, \ldots, \pi^T)$.

Theorem 2.1
Let $x_i^{T+1} = 0,\ i \in S$. *Let for* $t = T, T-1, \ldots, 1$ *consecutively, respectivelely a deterministic decision rule* f^t *and a vector* x^t *be defined as*

$$\left\{r(f^t)\right\}_i + \left\{P(f^t)x^{t+1}\right\}_i = max_{a\in A(i)}\ \left\{r_i^t(a) + \sum_j\ p_{ij}^t(a)x_j^{t+1}\right\},\ i \in S \tag{2.1}$$

and

$$x^t = r(f^t) + P(f^t)x^{t+1}.$$

Then, $R_* = (f^1, f^2, \ldots, f^T)$ *is an optimal policy and* x^1 *is the value vector* v^T.

Proof
We use induction on T. Let $R = (\pi^1, \pi^2, \ldots, \pi^T)$ be an arbitrary policy.

$T = 1:\ v_i^T(R) = \sum_{j,a}\ \mathbb{P}\{X_1 = j, Y_1 = a\} \cdot r_j^1(a) = \sum_a\ r_i^1(a)\pi_{ia}^1$

$$\leq max_{a\in A(i)}\ r_i^1(a) = x_i^1 = v_i^1(R_*),\ i \in S.$$

Assume that the result has been shown for $T = 1, 2, \ldots, t$. Take an arbitrary state i.
From Corollary 1.1 it follows that there exists a Markov policy $\overline{R}$ such that $v_i^{t+1}(\overline{R}) = v_i^{t+1}(R)$. Let $\overline{R} = (\sigma^1, \sigma^2, \ldots, \sigma^{t+1})$. Define the Markov policy $R' = (\rho^1, \rho^2, \ldots, \rho^t)$ by $\rho_{ja}^k = \sigma_{ja}^{k+1}$ for $k = 1, 2, \ldots, t$. From the induction assumption it follows that $v_j^t(R') \leq x_j^2,\ j \in S$, because for a planning horizon of $t+1$ periods x^2 is the same as x^1 for a planning horizon of t periods. Hence,

$$\begin{aligned} v_i^{t+1}(R) &= v_i^{t+1}(\overline{R}) = \textstyle\sum_a\ \sigma_{ia}^1\{r_i^1(a) + \sum_j\ p_{ij}^1(a)v_j^t(R')\} \\ &\leq \textstyle\sum_a\ \sigma_{ia}^1\{r_i^1(a) + \sum_j\ p_{ij}^1(a)x_j^2\} \leq max_a\ \{r_i^1(a) + \sum_j\ p_{ij}^1(a)x_j^2\} = x_i^1. \end{aligned}$$

On the other hand,

$$\begin{aligned} x^1 &= r(f^1) + P(f^1)x^2 = r(f^1) + P(f^1)\{r(f^2) + P(f^2)x^3\} \\ &= \cdots = \textstyle\sum_{s=1}^{t+1}\ \{P(f^1)P(f^2)\cdots P(f^{s-1})r(f^s)\} = v^{t+1}(R_*), \end{aligned}$$

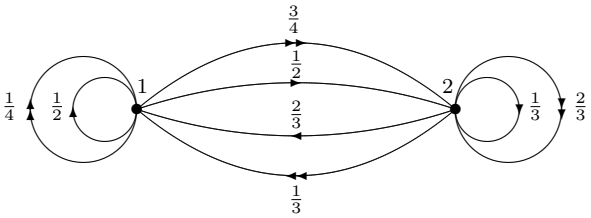
i.e. $v^{t+1}(R_*) = x^1 \geq v^{t+1}(R)$, i.e. R_* is an optimal policy and x^1 is the value vector. □

Algorithm 2.1 *Determining an optimal policy for a nonstationary MDP over* T *periods*

1. $x = 0$.
2. For $t = T, T-1, \ldots, 1$:
   (a) Determine the deterministic decision rule f^t such that
   $$\left\{r(f^t)\right\}_i + \left\{P(f^t)x^{t+1}\right\}_i = max_{a\in A(i)}\ \left\{r_i^t(a) + \sum_j\ p_{ij}^t(a)x_j^{t+1}\right\},\ i \in S.$$
   (b) $x = r(f^t) + P(f^t)x$.
3. $R_* = (f^1, f^2, \ldots, f^T)$ is an optimal policy and x is the value vector.

Example 2.1

Condider an MDP with the following data:



$S = \{1, 2\};\ A(1) = A(2) = \{1, 2\};\ T = 3.$

$p_{11}(1) = \frac{1}{2}; p_{12}(1) = \frac{1}{2}; r_1(1) = 1;$

$p_{11}(2) = \frac{1}{4}; p_{12}(2) = \frac{3}{4}; r_1(2) = 0;$

$p_{21}(1) = \frac{2}{3}; p_{22}(1) = \frac{1}{3}; r_2(1) = 2;$

$p_{21}(2) = \frac{1}{3}; p_{22}(2) = \frac{2}{3}; r_2(2) = 5.$

Start with $x_1 = x_2 = 0.$

$t = 3 : i = 1 : max\{1, 0\} = 1;\ f^3(1) = 1;\ x_1 = 1.$

$\quad i = 2 : max\{2, 5\} = 5;\ f^3(2) = 2;\ x_2 = 5.$

$t = 2 : i = 1 : max\{1 + \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 5, 0 + \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot 5\} = 4;\ f^2(1) = 1 \text{ (or 2)}; x_1 = 4.$

$\quad i = 2 : max\{2 + \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 5, 5 + \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 5\} = \frac{26}{3};\ f^2(2) = 2;\ x_2 = \frac{26}{3}.$

$t = 1 : i = 1 : max\{1 + \frac{1}{2} \cdot 4 + \frac{1}{2} \cdot \frac{26}{3}, 0 + \frac{1}{4} \cdot 4 + \frac{3}{4} \cdot \frac{26}{3}\} = \frac{15}{2};\ f^1(1) = 2;\ x_1 = \frac{15}{2}.$

$\quad i = 2 : max\{2 + \frac{2}{3} \cdot 4 + \frac{1}{3} \cdot \frac{26}{3}, 5 + \frac{1}{3} \cdot 4 + \frac{2}{3} \cdot \frac{26}{3}\} = \frac{109}{9};\ f^1(2) = 2;\ x_2 = \frac{109}{9}.$

$R_* = (f^1, f^2, f^3)$ is an optimal policy and $x = (\frac{15}{2}, \frac{109}{9})$ is the value vector.

Application 2.1 *Scheduling*

Suppose that N jobs have to be processed on one machine. Assume that the machine can process at most one job at a time, that job j has processing time p_j and that $c_j(t)$ is the cost if job j is completed at time t.

A strategy R corresponds to a permutation of the N jobs, say $R = \{i_1, i_2, \ldots, i_N\}$. Given strategy $R = \{i_1, i_2, \ldots, i_N\}$, job i_k has completion time $\sum_{j=1}^{k} p_{i_j}$. Hence, the corresponding cost is $c_{i_k}(\sum_{j=1}^{k} p_{i_j})$ and the total costs of this policy are $\sum_{k=1}^{N} c_{i_k}(\sum_{j=1}^{k} p_{i_j})$. Which order of the jobs minimizes the total costs?

This problem can be modelled as a finite horizon MDP with a layered state space. The states are the 2^N subsets of $\{1, 2, \ldots, N\}$. Layer 1 consists of the single state $\{1, 2, \ldots, N\}$, layer 2 has the N states $\{1, 2, \ldots, N\}\backslash\{j\},\ 1 \le j \le N$, and so on until layer $N + 1$, which consists of the empty state $\emptyset$. Any path from $\{1, 2, \ldots, N\}$ to $\emptyset$ corresponds to a permutation: at each stage, the job which is deleted from the state is chosen as scheduled at this stage on the machine.

When job j is chosen, i.e. deleted from a subset $J \subseteq \{1, 2, \ldots, N\}$, the jobs from $\{1, 2, \ldots, N\}\backslash J$ are already scheduled on the machine. Hence, the completion time of job j is $\sum_{i \notin J} p_i + p_j$ with costs $c_j(\sum_{i \notin J} p_i + p_j)$. Therefor e, this scheduling problem is equivalent to a *layered shortest path problem*, which can be solved as an MDP with finite horizon (see Exercise 2.1).

Formally, in state $J \subseteq \{1, 2, \ldots, N\}$ the action set $A(J)$ satisfies $A(J) = \{j \mid j \in J\}$ and, if action j is chosen, the immediate costs are $c_j(\sum_{i \notin J} p_i + p_j)$ and there is a deterministic transition to state $J\backslash\{j\}$ in the next layer.

2.3 An equivalent stationary infinite horizon model

In this section we present a stationary infinite horizon model which is equivalent to the standard nonstationary finite horizon model. Consider the following stationary MDP with infinite horizon for which the state space, action sets, immediate rewards and transition probabilities $S^*,\ A^*,\ r^*$ and p^* are given by:

$$
\begin{aligned}
S^* &= \{(i,t) \mid i \in S,\ t = 1,2,\ldots,T+1\} \\
A^*\{(i,t)\} &= \begin{cases} A(i) & i \in S,\ t = 1,2,\ldots,T \\ \{1\} & i \in S,\ t = T+1 \end{cases} \\
r^*_{(i,t)}(a) &= \begin{cases} r_i^t(a) & i \in S,\ t = 1,2,\ldots,T,\ a \in A(i) \\ 0 & i \in S,\ t = T+1,\ a = 1 \end{cases} \\
p^*_{(i,t)(j,s)}(a) &= \begin{cases} p_{ij}^t(a) & i \in S,\ t = 1,2,\ldots,T,\ a \in A(i),\ j \in S,\ s = t+1 \\ 0 & \text{elsewhere} \end{cases} \\
p^*_{(i,T+1)(j,s)}(1) &= \begin{cases} 1 & i \in S,\ j = i,\ s = T+1 \\ 0 & \text{elsewhere} \end{cases}
\end{aligned}
$$

This new infinite horizon model has a layered state space with transitions from (i,t) to $(j,t+1)$ until we reach a state in layer $(\cdot, T+1)$. All states of this last layer are absorbing. This infinite horizon model is a so-called *transient MDP*. For initial state (i,t) and policy R the total expected reward over the infinite horizon is denoted by $v_{(i.t)}(R)$. Any Markov policy $R = (\pi^1, \pi^2, \ldots, \pi^T)$ of the finite horizon model corresponds to a stationary policy π^∞ of the infinite horizon model by

$$
\pi_{(i.t)}(a) = \begin{cases} \pi_i^t(a) & i \in S,\ t = 1,2,\ldots,T,\ a \in A(i) \\ 1 & i \in S,\ t = T+1,\ a = 1 \end{cases}
$$

The next Lemma shows that for these corresponding policies the respective utility functions have the same value.

Lemma 2.1
Let $R = (\pi^1, \pi^2, \ldots, \pi^T)$ be a Markov policy of the finite horizon model with corresponding stationary policy π^∞ of the infinite horizon model. Then, $v_i^T(R) = v_{(i,1)}(\pi^\infty)$ for all $i \in S$.

Proof
By induction on t it is easy to show that for all i, j, t

$\left\{\{P^*(\pi)\}^{t-1}\right\}_{(i,1)(j,t)} = \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij}$ and $r^*_{(j,t)}(\pi) = r_j(\pi^t),\ t \le T.$

$\sum_j \left\{\{P^*(\pi)\}^T\right\}_{(i,1)(j,T+1)} = 1$ and $r^*_{(j,T+1)}(\pi) = 0.$

Hence,

$$\begin{aligned} v_{(i,1)}(\pi^\infty) &= \sum_{j=1}^{\infty} [P^*(\pi)]^{t-1} r^* \pi)]_{(i,1)} = \sum_{t=1}^{T} [P^*(\pi)]^{t-1} r^*(\pi)]_{(i,1)} \\ &= \sum_{t=1}^{T} \sum_{j \in S} \{[P^*(\pi)]^{t-1}\}_{(i,1)(j,t)} r^*_{(j,t)}(\pi) \\ &= \sum_{t=1}^{T} \sum_{j \in S} [P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})]_{ij} r_j(\pi^t) \\ &= \sum_{t=1}^{T} [P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})] r(\pi^t)]_i = v_i^T(R), \ i \in S. \end{aligned}$$

□

Since the finite horizon model has an optimal policy in the class of Markov policies with deterministic decision rules, the corresponding infinite horizon transient MDP model has an optimal policy in the class of deterministic policies. By the method of linear programming for MDPs (see the next chapters), MDPs with additional constraints can be handled.

2.4 Monotone optimal policies

In this section we study under which conditions optimal policies are *monotone*, i.e. nondecreasing or nonincreasing, in the ordering of state space. Such concept is worthwhile if there is a natural *ordering* in the state space. Knowlegde about the structure of optimal policies enables us to find such policies with less computational effort. In section 1.3 we have encountered several examples of special models with structured optimal policies, e.g. control-limit policies.

Let X and Y be ordered sets and let $f(x,y)$ a real-valued function on $X \times Y$. The function f is said to be *supermodular* (also called *superadditive*) if for any $x_1, x_2 \in X$ and $y_1, y_2 \in Y$ with $x_1 \geq x_2$ and $y_1 \geq y_2$

$$f(x_1,y_1) + f(x_2,y_2) \geq f(x_1,y_2) + f(x_2,y_1).$$

If the reverse inequality holds, the function is called *submodular* or *subadditive*.
For $x \in X$, $\overline{y}(x) \in argmax_{y \in Y} f(x,y)$ if $f\big(x,\overline{y}(x)\big) \geq f(x,y)$ for all $y \in Y$.

If $X = Y = \mathbb{R}$ and $f(x,y)$ is twice differentiable and supermodular, then $\frac{\partial^2 f(x,y)}{\partial x \partial y} \geq 0$ for all x and y (see Exercise 2.8).

Examples of supermodular functions on $\mathbb{R} \times \mathbb{R}$ are (see Exercise 2.6):
(1) $f(x,y) = xy$.
(2) $f(x,y) = g(x+y)$, where g is convex.

Lemma 2.2
Suppose f is supermodular on $X \times Y$ and for each $x \in X$ $max_y\ f(x,y)$ exists.
Let $y(x) = max\Big\{\overline{y}(x) \in argmax_{y \in Y} f(x,y)\Big\}$, then $y(x)$ is nondecreasing in x.

Proof
Let $x_1 \geq x_2$ and choose $y \leq y(x_2)$. Then, $f\big(x_1, y(x_2)\big) + f(x_2,y) \geq f(x_1,y) + f\big(x_2, y(x_2)\big)$, i.e.

$$f\big(x_1, y(x_2)\big) \geq f(x_1,y) + \{f\big(x_2, y(x_2)\big) - f(x_2,y)\} \geq f(x_1,y),$$

the last inequality by the definition of $y(x_2)$. Hence, $f(x_1, y(x_2)) \geq f(x_1, y)$ for all $y \leq y(x_2)$. Assume that $y(x_1) \leq y(x_2)$. Then, $f\big(x_1, y(x_2)\big) \geq f\big(x_1, y(x_1)\big) \geq f(x_1, y)$ for all y, and, by the definition of $y(x_1)$, $y(x_1) \geq y(x_2)$, i.e. $y(x_1) = y(x_2)$. Hence, $y(x_1) \geq y(x_2)$. □

We show the existence of optimal monotone policies under the following conditions:[1]

Assumption 2.1

(1) $r_i^t(a)$ is nondecreasing in i for all a and t;

(2) $\sum_{j=k}^N p_{ij}^t(a)$ is nondecreasing in i for all k, a and t.

(3) $A(i) = \{1, 2, \ldots, M\},\ i \in S$.

For the proof of the optimality of monotone policies, the next lemma is useful.

Lemma 2.3

Let $y, z : S \to \mathbb{R}_+$ satisfy $\sum_{j=k}^N y_j \geq \sum_{j=k}^N z_j,\ 2 \leq k \leq N$ and $\sum_{j=1}^N y_j = \sum_{j=1}^N z_j$, and let $v : S \to \mathbb{R}$ satisfy $v_{j+1} \geq v_j,\ j = 1, 2, \ldots, N-1$. Then, $\sum_{j=1}^N v_j y_j \geq \sum_{j=1}^N v_j z_j$.

Proof

Let $v_0 = 0$. Then,

$$\begin{aligned}
\textstyle\sum_{j=1}^N v_j y_j = &\ \textstyle\sum_{j=1}^N y_j \Big\{ \sum_{k=1}^j (v_k - v_{k-1}) \Big\} = \sum_{k=1}^N (v_k - v_{k-1}) \Big\{ \sum_{j=k}^N y_j \Big\} \\
= &\ \textstyle\sum_{k=2}^N (v_k - v_{k-1}) \Big\{ \sum_{j=k}^N y_j \Big\} + v_1 \sum_{j=1}^N y_j \\
\geq &\ \textstyle\sum_{k=2}^N (v_k - v_{k-1}) \Big\{ \sum_{j=k}^N z_j \Big\} + v_1 \sum_{j=1}^N z_j = \sum_{j=1}^N v_j z_j.
\end{aligned}$$

□

Theorem 2.2

Under Assumption 2.1 the x_i^t, as defined in Theorem 2.1, is nondecreasing in i, for all t.

Proof

Apply backward induction on t. For $t = T+1$: $x_i^{T+1} = 0$ for all i, so the result holds.

Assume that the result holds fot $t+1$ and consider $x^t = r(f^t) + P(f^t)x^{t+1}$.

Let $i_1 \geq i_2$, and let $y_j = p_{i_1 j}^t(f^t(i_2))$ and $z_j = p_{i_2 j}^t(f^t(i_2))$.

From Assumption 2.1 (2) it follows that $\sum_{j=i_1}^N y_j \geq \sum_{j=i_2}^N z_j$.

Notice that $\sum_{j=1}^N y_j = \sum_{j=1}^N z_j = 1$ and that $x_{i+1}^{t+1} \geq x_i^{t+1}$ for $i = 1, 2, \ldots, N-1$.

Apply Lemma 2.3: $\sum_{j=1}^N p_{i_1 j}^t(f^t(i_2)) x_j^{t+1} \geq \sum_{j=1}^N p_{i_2 j}^t(f^t(i_2)) x_j^{t+1}$.

Hence, using Assumption 2.1 (1),

$$\begin{aligned}
x_{i_1}^t = &\ max_{a \in A} \{ r_{i_1}^t(a) + \textstyle\sum_{j=1}^N\ p_{i_1 j}^t(a) x_j^{t+1} \} \geq r_{i_1}^t(f^t(i_2)) + \sum_{j=1}^N\ p_{i_1 j}^t(f^t(i_2)) x_j^{t+1} \\
\geq &\ r_{i_2}^t(f^t(i_2)) + \textstyle\sum_{j=1}^N\ p_{i_2 j}^t(f^t(i_2)) x_j^{t+1} = x_{i_2}^t.
\end{aligned}$$

□

[1] We consider the nondecreasing case; similar results hold in the nonincreasing case.

Theorem 2.3

Let Assumption 2.1 hold and furthermore assume that

(4) $r_i^t(a)$ *is supermodular on* $S \times A$

(5) $\sum_{j=k}^{N} p_{ij}^t(a)$ *is supermodular on* $S \times A$ *for all* $k \in S$

there exists an optimal policy $R_* = (f^1, f^2, \ldots, f^T)$*, where* $f^t(i)$ *is nondecreasing in* i *for* $t = 1, 2, \ldots, T$*, i.e. the decision rules* $f^t(i)$ $(1 \le t \le T)$ *are monotone.*

Proof

Take any $1 \le t \le T$. We first prove that $s_i^t(a) = r_i^t(a) + \sum_{j=1}^{N} p_{ij}^t(a) x_j^{t+1}$ is supermodular on $S \times A$. Let $i_1 \ge i_2,\ a_1 \ge a_2$, and let $y_j = p_{i_1 j}^t(a_1) + p_{i_2 j}^t(a_2),\ z_j = p_{i_1 j}^t(a_2) + p_{i_2 j}^t(a_1),\ j \in S$. By Assumption (5), for all $k \in S$, $\sum_{j=k}^{N} y_j \ge \sum_{j=k}^{N} z_j$. Since $\sum_{j=1}^{N} y_j = \sum_{j=1}^{N} z_j = 2$, and x_i^{t+1} is nondecreasing in i (see Theorem 2.2), applying Lemma 2.3 yields

$$\sum_{j=1}^{N} \{p_{i_1 j}^t(a_1) + p_{i_2 j}^t(a_2)\} x^{j+1} \ge \sum_{j=1}^{N} \{p_{i_1 j}^t(a_2) + p_{i_2 j}^t(a_1)\} x^{j+1},$$

i.e. $\sum_{j=1}^{N} p_{ij}^t(a) x_j^{t+1}$ is supermodular. Because the sum of supermodular functions is supermodular, $s_i^t(a)$ is also supermodular. If the action $f^t(i)$ in formula (2.1) is not unique, take the largest optimal action. Then, applying Lemma 2.2 yields the result that $f^t(i)$ is nondecreasing in i. □

Algorithm 2.2 *Determining an optimal policy with monotone decision rules for a nonstationary MDP over* T *periods under the above assumptions* $(1), (2), \ldots, (5)$

1. $x^{T+1} = 0;\ t = T.$
2. (a) $i = 1;\ A(i) = \{1, 2, \ldots, M\}.$
   (b) Determine $f^t(i)$ such that
   $$\left\{r(f^t)\right\}_i + \left\{P(f^t)x^{t+1}\right\}_i = max_{a \in A(i)} \left\{r_i^t(a) + \sum_j p_{ij}^t(a) x_j^{t+1}\right\}$$
   (if there is more than one optimizing action, take the largest).
   (c) $x_i^t = r_i^t\big(f^t(i)\big) + \sum_{j=1}^{N} p_{ij}^t\big(f^t(i)\big) x_j^{t+1}.$
   (d) If $i = N$: go to step 3;
   otherwise: $A_{i+1} = \{a \mid a \ge f^t(i)\};\ i := i + 1$ and go to step 2b.
3. If $t = 1$: stop with $R_* = (f^1, f^2, \ldots, f^T)$ as optimal policy and x as value vector.
   otherwise: $t := t - 1$ and go to step 2a.

The advantage of this algorithm is that the maximization can be carried out over action sets which become smaller in the order of the states. If for some state i the action set consists of a singleton no optimization is needed in higher states.

2.5 Bibliographic notes

The principles of optimality and backward induction were presented in Bellman's book [11]. This book had an enormous impact in the field of dynamic programming. Hordijk [93] has shown that the principle of optimality together with the validity of backward induction may be viewed as a consequence of the duality theory of linear programming.

The equivalence between the standard nonstationary finite horizon and a stationary inifinite horizon model was presented in Kallenberg ([106], [107]). A related paper is Derman and Klein [56].
The development of monotone optimal policies is provided by the work of Serfozo [180] and Topkis [196]. Our presentation follows Puterman ([157], section 4.7). Other contributions are given e.g. by Ross [170] and Heyman and Sobel [87].

2.6 Exercises

Exercise 2.1
Consider a layered network: i.e. the set of vertices $V = V_1 \cup V_2 \cup \cdots \cup V_p$, where $V_1 = 1$ and $V_1 = N$, and for all arcs (i, j), $i \in V_k$ and $j \in V_{k+1}$ for some $k = 1, 2, \ldots, p-1$; the arc (i, j) has length l_{ij}.
Show that the problem of finding the shortest path (and its length) from vertex 1 to vertex N can be modelled as an MDP over a finite horizon.

Exercise 2.2
Consider a scheduling problem as in Application 2.1 with the data:
$N = 4;\ p_1 = 1,\ p_2 = 2,\ p_3 = 3,\ p_4 = 4;\ c_1(t) = max(0, t-2),\ c_2(t) = max(0, t-7),$
$c_3(t) = max(0, t-5)$ and $c_4(t) = max(0, t-6)$.
(1) Draw the layered network for this scheduling problem;
(2) Compute an optimal ordering of the jobs by backward induction.

Exercise 2.3
Suppose you have an employee and at the beginning of each month you can decide on his salary for that month: either a low salary ($ 2300) or a high salary ($ 3000). Knowing his salary, the employee can decide to send in his resignation, which is enforced at the end of that month.
The probability that he sends in his resignation depends on his salary: 40% for a low salary and 20% for a high salary. When the employee quits, a temporary employee has to be hired immediately for $ 4000 per month. When you have a temporary employee you will advertise each month for a new permanent employee.

The probability to find a new permanent employee (who can start at the beginning of the following month and will receive the same salary conditions as the original employee) depends on the advertising budget: 70% for advertising budget $ 300 and 90% for advertising budget $ 600. What is an optimal policy for the next six months?

Exercise 2.4

Construct the corresponding infinite horizon model for the finite horizon MDP of Exercise 2.3.

Exercise 2.5

Show that the sum of supermodular functions is supermodular.

Exercise 2.6

Show that the following functions on $\mathbb{R}^1 \times \mathbb{R}^1$ are supermodular:

a. $f(x, y) = xy$.

b. $f(x, y) = g(x + y)$, where g is convex.

Exercise 2.7

Let $f(x, y)$ be a function on $X \times Y$, where $X = Y = \mathbb{Z}_+$, and suppose

$$f(i + 1, a + 1) + f(i, a) \geq f(i, a + 1) + f(i + 1, a) \text{ for all } i \in X \text{ and } a \in Y.$$

Show that $f(x, y)$ is superadditive.

Exercise 2.8

Let $f(x, y)$ be a twice differential function on $\mathbb{R}^1 \times \mathbb{R}^1$. Show that $f(x, y)$ is superadditive if and only if $\frac{\partial^2 f(x,y)}{\partial x \partial y}$ is a nonnegative function.

Hint: Consider $\int_{y_2}^{y_1} \left\{ \int_{x_2}^{x_1} \frac{\partial^2 f(x,y)}{\partial x \partial y} dx \right\} dy$.

Chapter 3

Discounted rewards

3.1 Introduction

This chapter deals with the total expected discounted reward over an infinite planning horizon. We assume that the model is stationary. The criterion of the total expected discounted reward is quite natural when the planning horizon is rather large and returns at the present time are of more value than returns which are earned later in time. We recall that the total expected α-discounted reward, given initial state i, policy R and discount factor $\alpha \in (0,1)$, is denoted by $v_i^\alpha(R)$ and defined by

$$v_i^\alpha(R) = \sum_{t=1}^{\infty} \mathbb{E}_{i,R}\{\alpha^{t-1} \cdot r_{X_t}(Y_t)\} = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{j,a} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j(a). \tag{3.1}$$

As already mentioned, by the theorem of dominated convergence, the expected total α-discounted reward, i.e.

$$\mathbb{E}_{i,R}\Big\{\sum_{t=1}^{\infty} \alpha^{t-1} \cdot r_{X_t}(Y_t)\Big\},$$

gives the same expression as (3.1). Hence, the expected total discounted reward criterion and the total expected discounted reward criterion are equivalent. We also recall that a stationary policy π^∞ satisfies

$$v^\alpha(\pi^\infty) = \sum_{t=1}^{\infty} \alpha^{t-1} P(\pi)^{t-1} r(\pi). \tag{3.2}$$

Since $\Big\{I - \alpha P(\pi)\Big\} \cdot \Big\{I + \alpha P(\pi) + \cdots + \{\alpha P(\pi)\}^{t-1}\Big\} = I - \{\alpha P(\pi)\}^t$ and $\{\alpha P(\pi)\}^t \to 0$ for $t \to \infty$, we obtain

$$\sum_{t=1}^{\infty} \Big\{\alpha P(\pi)\Big\}^{t-1} = \Big\{I - \alpha P(\pi)\Big\}^{-1} \text{ and } v^\alpha(\pi^\infty) = \Big\{I - \alpha P(\pi)\Big\}^{-1} r(\pi).$$

The α-discounted value vector v^α and the optimality of policy R_* are defined by

$$v^\alpha = sup_R\ v^\alpha(R) \quad \text{and} \quad v^\alpha(R_*) = v^\alpha. \tag{3.3}$$

From the mathematical point of view, the discounted reward criterion is good manageable: there is a very complete general theory. In this chapter we only discuss the case of finite state space and finite action sets, but the results can be extended to a much higher level of generality.

In this chapter, we first discuss the theory of monotone contraction mappings in the context of MDPs. Then, the optimality equation, bounds for the value vector and suboptimal actions are considered. Next, the classical methods (policy iteration, linear programming, value iteration) and the hybrid method of modified policy iteration are studied.

3.2 Monotone contraction mappings

To find an optimal policy and the α-discounted value vector v^α, the *optimality equation*

$$x_i = max_{a\in A(i)}\Big\{r_i(a) + \alpha \sum_j\ p_{ij}(a)x_j\Big\},\ i \in S, \tag{3.4}$$

plays a central role. In the next section we will see that v^α is the unique solution of this equation. For the moment, we give the following intuitive argumentation. Suppose that at time point $t = 1$, when the system is in state i, action $a \in A(i)$ is chosen, and that from $t = 2$ on an optimal policy is followed. Then, the total expected α-discounted reward is equal to $r_i(a)+\alpha\sum_j\ p_{ij}(a)v_j^\alpha$. Since any optimal policy obtains at least this amount, we have

$$v_i^\alpha \geq max_{a\in A(i)}\Big\{r_i(a) + \alpha \sum_j\ p_{ij}(a)v_j^\alpha\Big\},\ i \in S.$$

On the other hand, let a_i be the action chosen by an optimal policy in state i. Then,

$$v_i^\alpha = r_i(a_i) + \alpha \sum_j\ p_{ij}(a_i)v_j^\alpha \leq max_{a\in A(i)}\Big\{r_i(a) + \alpha \sum_j\ p_{ij}(a)v_j^\alpha\Big\},\ i \in S.$$

Hence, v^α is a solution of (3.4) and v^α is a fixed-point of the mapping $U : \mathbb{R}^N \to \mathbb{R}^N$, defined by

$$(Ux)_i = max_{a\in A(i)}\Big\{r_i(a) + \alpha \sum_j\ p_{ij}(a)x_j\Big\},\ i \in S. \tag{3.5}$$

We will show that U is a contraction mapping. Hence, by the general theory of contracting mappings, v^α is the unique solution of (3.4) and can be computed by value iteration.

Contraction mappings

Let X be a Banach space [1] with norm $\|\cdot\|$, and let B: $X \to X$. The operator B is called a *contraction mapping* if for some $\beta \in (0, 1)$

$$\|Bx - By\| \leq \beta\|x - y\| \text{ for all } x, y \in X. \tag{3.6}$$

The number β is called the *contraction factor* of B. An element $x \in X$ is said to be a *fixed-point* of B if $Bx^* = x^*$. The next theorem [2] ensures the existence of a unique fixed-point for contraction mappings in a Banach space.

[1] For a definition of Banach space, see textbooks on Functional Analysis or Appendix C in Puterman [157].

[2] For a proof of the theorem, see textbooks on Functional Analysis or Puterman [157], p.150.

Theorem 3.1 *Fixed-point Theorem*
Let X be a Banach space and suppose $B : X \to X$ is a contraction mapping. Then,
(1) $x^* = lim_{n\to\infty} B^n x$ exists for every $x \in X$ and x^* is a fixed-point of B.
(2) x^* is the unique fixed-point of B.

The next theorem gives bounds on the distance between the fixed-point x^* and $B^n x$ for $n = 0, 1, 2, \ldots$.

Theorem 3.2
Let X be a Banach space and suppose $B : X \to X$ is a contraction mapping with contraction factor β and fixed-point x^*. Then,

$$(1) \quad \|x^* - B^n x\| \leq \beta(1-\beta)^{-1} \cdot \|B^n x - B^{n-1} x\| \leq \beta^n (1-\beta)^{-1} \cdot \|Bx - x\| \ \forall x \in X, \ n \in \mathbb{N}.$$
$$(2) \quad \|x^* - x\| \leq (1-\beta)^{-1} \cdot \|Bx - x\| \ \forall x \in X.$$

Proof
(1) For $m > n$, we have

$$\begin{aligned} \|B^m x - B^n x\| &\leq \beta \cdot \|B^{m-1} x - B^{n-1} x\| \\ &\leq \beta \cdot \{\|B^{m-1} x - B^{m-2} x\| + \|B^{m-2} x - B^{m-3} x\| + \cdots + \|B^n x - B^{n-1} x\|\} \\ &\leq \beta \cdot \{\beta^{m-n-1} + \beta^{m-n-2} + \cdots + 1\} \cdot \|B^n x - B^{n-1} x\| \\ &\leq \beta(1-\beta)^{-1} \cdot \|B^n x - B^{n-1} x\|. \end{aligned}$$

Hence, since $B^m x^* = x^*$, we obtain

$$\begin{aligned} \|x^* - B^n x\| &= \|B^m x^* - B^n x\| \leq \|B^m x^* - B^m x\| + \|B^m x - B^n x\| \\ &\leq \beta^m \cdot \|x^* - x\| + \beta(1-\beta)^{-1} \cdot \|B^n x - B^{n-1} x\| \text{ for } m > n. \end{aligned}$$

The first inequality is obtained by letting $m \to \infty$. The second inequality holds because

$$\|B^n x - B^{n-1} x\| \leq \beta^{n-1} \cdot \|Bx - x\|.$$

(2) Apply the triangle inequality and part (1) for $n = 1$:

$$\begin{aligned} \|x^* - x\| &\leq \|x^* - Bx\| + \|Bx - x\| \leq \beta(1-\beta)^{-1} \cdot \|Bx - x\| + \|Bx - x\| \\ &= (1-\beta)^{-1} \cdot \|Bx - x\|. \end{aligned}$$

□

Remark:
The above theorem implies that the convergence rate of $B^n x$ to the fixed-point is at least linear (cf. Stoer and Bulirsch ([192] p.251)). This kind of convergence is also called *geometric convergence*.

Monotonicity

Let X be a partially ordered set and $B : X \to X$. The mapping B is called *monotone* if $x \leq y$ implies $Bx \leq By$.

Theorem 3.3

Let X be a partially ordered Banach space. Suppose that $B: X \to X$ is a monotone contraction mapping with fixed-point x^*. Then,

(1) $Bx \leq x$ implies $x^* \leq Bx \leq x$;

(2) $Bx \geq x$ implies $x^* \geq Bx \geq x$.

Proof

(1) By the monotonicity of B it can easily be verified (with induction on n) that

$x \geq Bx \geq \cdots \geq B^n x,\ n \in \mathbb{N}$. Therefore, we have $x^* = lim_{n\to\infty} B^n x \leq Bx \leq x$.

(2) The proof is similar to the proof of part (1). □

It is easy to verify that $\mathbb{R}^N$ with norm $\|x\|_\infty = max_{1\leq i\leq N}|x_i|$ (*supremum norm*) and with ordering $x \leq y$ if $x_i \leq y_i$ for all $1 \leq i \leq N$ is a partially ordered Banach space. Also, for $x \in \mathbb{R}^N$, we have $x \leq \|x\|_\infty \cdot e$, where e is the vector $(1, 1, \ldots, 1)$.

Lemma 3.1

(1) Let $B: \mathbb{R}^N \to \mathbb{R}^N$ be a monotone contraction mapping with contraction factor β, and let d be a scalar. Then, $x \leq y + d \cdot e$ implies $Bx \leq By + \beta \cdot |d| \cdot e$.

(2) Let $B: \mathbb{R}^N \to \mathbb{R}^N$ be a mapping with the property that $x \leq y + d \cdot e$ implies $Bx \leq By + \beta \cdot |d| \cdot e$ for some $0 \leq \beta < 1$ and for all scalars d. Then, with respect to the supremum norm, B is a monotone contraction with contraction factor β.

Proof

(1) From the monotonicity of B it follows that

$$\begin{aligned} Bx &\leq B(y + d \cdot e) = B(y + d \cdot e) - By + By \leq \|B(y + d \cdot e) - By\|_\infty \cdot e + By \\ &\leq \beta \cdot \|(y + d \cdot e) - y\|_\infty \cdot e + By = \beta \cdot |d| \cdot e + By. \end{aligned}$$

(2) Taking $d = 0$ yields the monotonicity. Since $x - y \leq \|x - y\|_\infty \cdot e$ and $y - x \leq \|x - y\|_\infty \cdot e$, the property of B mentioned in the theorem implies that $Bx - By \leq \beta \cdot \|x - y\|_\infty \cdot e$ and $By - Bx \leq \beta \cdot \|x - y\|_\infty \cdot e$, i.e. $\|Bx - By\|_\infty \leq \beta \cdot \|x - y\|_\infty$. □

Lemma 3.2

Let $B: \mathbb{R}^N \to \mathbb{R}^N$ be a monotone contraction mapping, with respect to the supremum norm, with contraction factor β and fixed-point x^*. Suppose that there exist scalars a and b such that $a \cdot e \leq Bx - x \leq b \cdot e$ for some $x \in \mathbb{R}^N$. Then,

$x - (1-\beta)^{-1}|a| \cdot e \leq Bx - \beta(1-\beta)^{-1}|a| \cdot e \leq x^* \leq Bx + \beta(1-\beta)^{-1}|b| \cdot e \leq x + (1-\beta)^{-1}|b| \cdot e.$

Proof

Since $Bx \leq x + b \cdot e \leq x + |b| \cdot e$, it follows from the monotonicity of B that

$$\begin{aligned} B^2x &\leq B(x + |b| \cdot e) = B(x + |b| \cdot e) - Bx + Bx \leq Bx + \|B(x + |b| \cdot e) - Bx\|_\infty \cdot e \\ &\leq Bx + \beta|b| \cdot e \leq x + (1 + \beta)|b| \cdot e. \end{aligned}$$

Using the same arguments it can be shown (with induction on n) that

$$B^n x \leq Bx + (\beta + \cdots + \beta^{n-1})|b| \cdot e \leq x + (1 + \beta + \cdots + \beta^{n-1})|b| \cdot e, \ n \in \mathbb{N}.$$

By letting $n \to \infty$,

$$x^* \leq Bx + \beta(1-\beta)^{-1}|b| \cdot e \leq x + (1-\beta)^{-1}|b| \cdot e.$$

Because $Bx \geq x + a \cdot e \geq x - |a| \cdot e$, an analogous reasoning shows that

$$x^* \geq Bx - \beta(1-\beta)^{-1}|a| \cdot e \leq x - (1-\beta)^{-1}|a| \cdot e.$$

□

Corollary 3.1

Let B be a monotone contraction in $\mathbb{R}^N$, with respect to the supremum norm, with contraction factor β and fixed-point x^*. Then,

$$\begin{aligned} x - (1-\beta)^{-1} \cdot \|Bx - x\|_\infty \cdot e &\leq Bx - \beta(1-\beta)^{-1} \cdot \|Bx - x\|_\infty \cdot e \leq x^* \\ &\leq Bx + \beta(1-\beta)^{-1} \cdot \|Bx - x\|_\infty \cdot e \leq x + (1-\beta)^{-1}\|Bx - x\|_\infty \cdot e. \end{aligned}$$

Proof

Notice that $-\|Bx - x\|_\infty \cdot e \leq Bx - x \leq \|Bx - x\|_\infty \cdot e$ and apply Lemma 3.2. □

Lemma 3.3

Let $B : \mathbb{R}^N \to \mathbb{R}^N$ be a monotone contraction mapping, with respect to the supremum norm, with contraction factor β, fixed-point x^* and with the property that $B(x + c \cdot e) = Bx + \beta c \cdot e$ for every $x \in \mathbb{R}^N$ and scalar c.

Suppose that there exist scalars a and b such that $a \cdot e \leq Bx - x \leq b \cdot e$ for some $x \in \mathbb{R}^N$. Then,

$$x + (1-\beta)^{-1}a \cdot e \leq Bx + \beta(1-\beta)^{-1}a \cdot e \leq x^* \leq Bx + \beta(1-\beta)^{-1}b \cdot e \leq x + (1-\beta)^{-1}b \cdot e.$$

Proof

By the monotonicity of B it follows from $Bx \leq x + b \cdot e$ that

$$B^2 x \leq B(x + b \cdot e) = Bx + \beta \cdot e \leq x + (1 + \beta)b \cdot e,$$

and by induction on n,

$$B^n x \leq Bx + (\beta + \beta^2 + \cdots + \beta^{n-1})b \cdot e \leq x + (1 + \beta + \beta^2 + \cdots + \beta^{n-1})b \cdot e.$$

Taking the limit for $n \to \infty$ gives,

$$x^* \leq Bx + \beta(1-\beta)^{-1}b \cdot e \leq x + (1-\beta)^{-1}b \cdot e.$$

The proof of the lower bounds is similar. □

3.3 The optimality equation

In this section we discuss the optimality equation (3.4) for the α-discounted value vector v^α. We show that v^α is the unique solution of (3.4). Furthermore, we will derive bounds for the value vector. By these bounds suboptimality tests can be formulated to exclude nonoptimal actions. The results are obtained by applying the theory of monotone contraction mappings, as presented

in Section 3.2. Besides the mapping U defined in (3.5), we introduce a mapping $L_\pi : \mathbb{R}^N \to \mathbb{R}^N$ for any randomized decision rule π, defined by

$$L_\pi x = r(\pi) + \alpha P(\pi)x. \tag{3.7}$$

Let $f_x(i)$ be such that

$$r_i(f_x(i)) + \alpha \sum_j p_{ij}(f_x(i))x_j = max_a \Big\{ r_i(a) + \alpha \sum_j p_{ij}(a)x_j \Big\},\ i \in S.$$

Then,

$$L_{f_x} x = Ux = max_f\ L_f x,$$

where the maximization is taken over all deterministic decision rules f. Let $\|P(\pi)\|_\infty$ be the subordinate matrix norm,[3] then $\|P(\pi)\|_\infty$ satisfies (see e.g. Stoer and Boelirsch [192], p. 178)

$$\|P(\pi)\|_\infty = max_i \sum_j p_{ij}(\pi) = 1.$$

Theorem 3.4

The mapppings L_π and U are monotone contraction mappings (with respect to the supremum norm) with contraction factor α.

Proof

Suppose that $x \geq y$. Let π be any stationary decision rule. Because $P(\pi) \geq 0$,

$$L_\pi x = r(\pi) + \alpha P(\pi)x \geq r(\pi) + \alpha P(\pi)y = L_\pi y, \tag{3.8}$$

i.e. L_π is monotone. U is also monotone, since

$$Ux = max_f\ L_f x \geq L_{f_y} x \geq L_{f_y} y = Uy.$$

Furthermore, we obtain

$$\|L_\pi x - L_\pi y\|_\infty = \|\alpha P(\pi)(x-y)\|_\infty \leq \alpha \cdot \|P(\pi)\|_\infty \cdot \|x-y\|_\infty = \alpha \cdot \|x-y\|_\infty,$$

i.e. L_π is a contraction with contraction factor α. The derivation for operator U is

$$Ux - Uy = L_{f_x} x - L_{f_y} y \leq L_{f_x} x - L_{f_x} y = \alpha \cdot P(f_x)(x-y) \leq \alpha \cdot \|x-y\|_\infty \cdot e. \tag{3.9}$$

Interchanging x and y yields

$$Uy - Ux \leq \alpha \cdot \|x-y\|_\infty \cdot e. \tag{3.10}$$

From (3.9) and (3.10) it follows that $\|Ux - Uy\|_\infty \leq \alpha \cdot \|x-y\|_\infty$, i.e. U is a contraction with contraction factor α. □

The next theorem shows that for any randomized decision rule π, the total expected α-discounted reward of the policy π^∞ is the fixed-point of the mapping L_π.

[3] Given a vector norm $\|x\|$, the corresponding subordinate matrix norm for a square matrix A is defined by $\|A\| = max_{\{x \mid \|x\|=1\}} \|Ax\|$.

Theorem 3.5

$v^\alpha(\pi^\infty)$ is the unique solution of the functional equation $L_\pi x = x$.

Proof

Theorem 3.1 and Theorem 3.4 imply that it is sufficient to show that $L_\pi v^\alpha(\pi^\infty) = v^\alpha(\pi^\infty)$.
We have

$$\begin{aligned} L_\pi v^\alpha(\pi^\infty) - v^\alpha(\pi^\infty) = &\quad r(\pi) - \{I - \alpha P(\pi)\}v^\alpha(\pi^\infty) \\ = &\quad r(\pi) - \{I - \alpha P(\pi)\}\{I - \alpha P(\pi)\}^{-1} r(\pi) = 0. \end{aligned}$$

□

Corollary 3.2

$v^\alpha(\pi^\infty) = lim_{n\to\infty}\ L_\pi^n x$ for any $x \in \mathbb{R}^N$.

The next theorem shows that the value vector v^∞ is the fixed-point of the mapping U.

Theorem 3.6

v^α is the unique solution of the functional equation $Ux = x$.

Proof

It is sufficient to show that $Uv^\alpha = v^\alpha$. Let $R = (\pi^1, \pi^2, \dots)$ be an arbitrary Markov policy. Then,

$$\begin{aligned} v^\alpha(R) = &\quad r(\pi^1) + \textstyle\sum_{t=2}^\infty\ \alpha^{t-1} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t) \\ = &\quad r(\pi^1) + \alpha P(\pi^1) \textstyle\sum_{s=1}^\infty\ \alpha^{s-1} P(\pi^2)P(\pi^3)\cdots P(\pi^s)r(\pi^{s+1}) \\ = &\quad r(\pi^1) + \alpha P(\pi^1) v^\alpha(R_2) = L_{\pi^1} v^\alpha(R_2), \end{aligned}$$

where $R_2 = (\pi^2, \pi^3, \dots)$. From the monotonicity of L_{π^1} and the definition of U,

$$v^\alpha(R) = L_{\pi^1} v^\alpha(R_2) \le L_{\pi^1} v^\alpha \le U v^\alpha,\ \ R \in C(M).$$

Hence, $v^\alpha = sup_{R\in C(M)}\ v^\alpha(R) \le Uv^\alpha$. Take any $\varepsilon > 0$. Since $v^\alpha = sup_{R\in C(M)}\ v^\alpha(R)$, for any $j \in S$ there exists a Markov policy $R_j^\varepsilon = (\pi^1(j), \pi^2(j), \dots)$ such that $v_j^\alpha(R_j^\varepsilon) \ge v_j^\alpha - \varepsilon$.
Let $a_i \in A(i)$ be such that $r_i(a_i) + \alpha \sum_j\ p_{ij}(a_i) v_j^\alpha = max_a\ \{r_i(a) + \alpha \sum_j\ p_{ij}(a) v_j^\alpha\},\ i \in S$.
Consider the policy $R^* = (\pi^1, \pi^2, \dots)$ defined by

$$\pi_{ia}^1 = \begin{cases} 1 & \text{if } a = a_i \\ 0 & \text{otherwise} \end{cases} \quad \text{and } \pi_{i_1 a_1 \cdots i_t a}^t = \pi_{i_t a}^{t-1}(i_2),\ a \in A(i_t),\ t \ge 2,$$

i.e. R^* is the policy that chooses a_i in state i at time point $t = 1$, and if the state at time $t = 2$ is i_2, then the policy follows $R_{i_2}^\varepsilon$, where the process is considered to be originating in state i_2. Therefore,

$$\begin{aligned} v_i^\alpha \quad \ge &\quad v_i^\alpha(R^*) = r_i(a_i) + \alpha \textstyle\sum_j\ p_{ij}(a_i) v_j^\alpha(R_j^\varepsilon)\ \ge\ r_i(a_i) + \alpha \textstyle\sum_j\ p_{ij}(a_i)(v_j^\alpha - \varepsilon) \\ = &\quad max_a\ \{r_i(a) + \alpha \textstyle\sum_j\ p_{ij}(a) v_j^\alpha\} - \alpha\varepsilon = (Uv^\alpha)_i - \alpha\varepsilon,\ i \in S. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrarily chosen, $v^\alpha \ge Uv^\alpha$.

□

Because $v^\alpha = Uv^\alpha = L_{f_{v^\alpha}} v^\alpha$, it follows from Theorem 3.5 that $v^\alpha = v^\alpha(f^\infty_{v^\alpha})$, i.e. $f^\infty_{v^\alpha}$ is an optimal policy. If $f^\infty \in C(D)$ satisfies

$$r_i(f) + \alpha \sum_j p_{ij}(f) v_j^\alpha = max_a \{r_i(a) + \alpha \sum_j p_{ij}(a) v_j^\alpha\}, \ i \in S,$$

then f^∞ is called a *conserving policy*. Conserving policies are optimal. Therefore, the equation $Ux = x$ is called the *optimality equation*.

Corollary 3.3

(1) There exists a deterministic α-discounted optimal policy.
(2) $v^\alpha = lim_{n\to\infty} U^n x$ for any $x \in \mathbb{R}^N$.
(3) Any conserving policy is α-discounted optimal.

As already mentioned, we will derive bounds for the value vector v^α. These bounds can be obtained by using Lemma 3.3. Therefore, we note that the mappings L_π and U satisfy, for any $x \in R^N$ and scalar c, $L_f(x + c \cdot e) = L_f x + \alpha c \cdot e$ and $U(x + c \cdot e) = Ux + \alpha c \cdot e$.

Theorem 3.7

For any $x \in \mathbb{R}^N$, we have

(1) $x - (1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e \le Ux - \alpha(1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e \le v^\alpha(f_x^\infty) \le v^\alpha \le Ux + \alpha(1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e \le x + (1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e.$

(2) $\|v^\alpha - x\|_\infty \le (1-\alpha)^{-1}\|Ux - x\|_\infty.$

(3) $\|v^\alpha - v^\alpha(f_x^\infty)\|_\infty \le 2\alpha(1-\alpha)^{-1}\|Ux - x\|_\infty.$

Proof

Take an arbitrary $x \in \mathbb{R}^N$. By Lemma 3.3, for $a = -\|Ux - x\|_\infty$, $b = \|Ux - x\|_\infty$ and $B = L_{f_x}$, we obtain (notice that $Bx = L_{f_x} x = Ux$),

$$x - (1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e \le Ux - \alpha(1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e \le v^\alpha(f_x^\infty) \le v^\alpha.$$

Next, applying Lemma 3.3 again for $B = U$ the remaining part of (1) implies

$$v^\alpha \le Ux + \alpha(1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e \le x + (1-\alpha)^{-1}\|Ux - x\|_\infty \cdot e.$$

The parts (2) and (3) follow directly from part (1). □

Theorem 3.8

For any $x \in \mathbb{R}^N$, we have

(1) $x + (1-\alpha)^{-1} min_i \ (Ux - x)_i \cdot e \le Ux + \alpha(1-\alpha)^{-1} min_i \ (Ux - x)_i \cdot e \le v^\alpha(f_x^\infty) \le v^\alpha \le Ux + \alpha(1-\alpha)^{-1} max_i \ (Ux - x)_i \cdot e \le x + (1-\alpha)^{-1} max_i \ (Ux - x)_i \cdot e.$

(2) $\|v^\alpha - v^\alpha(f_x^\infty)\|_\infty \le \alpha(1-\alpha)^{-1} span \ (Ux - x)$, *where* $span(y) = max_i \, y_i - min_i \, y_i$.

Proof

Note that $min_i\ (Ux - x)_i \cdot e \leq Ux - x \leq max_i(Ux\ - x)_i \cdot e$. It is easy to verify that for $a = min_i\ (Ux - x)_i$ and $b = max_i\ (Ux - x)_i$ the proof is similar to the proof of Theorem 3.7. □

Remark

Since $-min_i\ (Ux - x)_i \leq \|Ux - x\|_\infty$ and $max_i\ (Ux - x)_i \leq \|Ux - x\|_\infty$, we have $span\ (Ux - x) \leq 2 \cdot \|Ux - x\|_\infty$. Consequently, the bounds given by Theorem 3.8 are stronger than those given by Theorem 3.7.

Next, we discuss the elimination of suboptimal actions. An action $a \in A(i)$ is called *suboptimal* if there doesn't exist an α-discounted optimal policy $f^\infty \in C(D)$ with $f(i) = a$. Because f^∞ is α-discounted optimal if and only if $v^\alpha(f^\infty) = v^\alpha$, and because $v^\alpha = Uv^\alpha$, an action $a \in A(i)$ is suboptimal if and only if

$$v_i^\alpha > r_i(a) + \alpha \sum_j\ p_{ij}(a)v_j^\alpha. \tag{3.11}$$

Suboptimal actions can be disregarded. Notice that formula (3.11) is unuseful, because v^α is unknown. However, by upper and lower bounds on v^α as given in Theorems 3.7 and 3.8, suboptimaly tests can be derived, as illustrated in the following theorem.

Theorem 3.9

Suppose that $x \leq v^\alpha \leq y$. If $r_i(a) + \alpha \sum_j\ p_{ij}(a)y_j < (Ux)_i$, then action $a \in A(i)$ is suboptimal.

Proof

$v_i^\alpha = (Uv^\alpha)_i \geq (Ux)_i > r_i(a) + \alpha \sum_j\ p_{ij}(a)y_j \geq r_i(a) + \alpha \sum_j\ p_{ij}(a)v_j^\alpha$. The first inequality is justified by the monotonicity of U. □

Corollary 3.4

Suppose that for some scalars b and c, we have $x + b \cdot e \leq v^\alpha \leq x + c \cdot e$. If

$$r_i(a) + \alpha \sum_j\ p_{ij}(a)x_j < (Ux)_i - \alpha(c - b), \tag{3.12}$$

then action $a \in A(i)$ is suboptimal.

Proof

$r_i(a) + \alpha \sum_j\ p_{ij}(a)(x_j + c) = r_i(a) + \alpha \sum_j\ p_{ij}(a)x_j + \alpha c < (Ux)_i + \alpha b = \{U(x + b \cdot e)\}_i$. □

Applying Corollary 3.4 on the bounds of v^α, derived in the Theorems 3.7 and 3.8, gives the following tests for the elimination of a suboptimal action $a \in A(i)$:

$$r_i(a) + \alpha \sum_j\ p_{ij}(a)x_j < (Ux)_i - 2\alpha(1 - \alpha)^{-1}\|Ux - x\|_\infty. \tag{3.13}$$

$$r_i(a) + \alpha \sum_j p_{ij}(a)(Ux)_j < (U^2x)_i - 2\alpha^2(1-\alpha)^{-1}\|Ux - x\|_\infty. \tag{3.14}$$

$$r_i(a) + \alpha \sum_j p_{ij}(a)x_j < (Ux)_i - \alpha(1-\alpha)^{-1} span\ (Ux - x). \tag{3.15}$$

$$r_i(a) + \alpha \sum_j p_{ij}(a)(Ux)_j < (U^2x)_i - \alpha^2(1-\alpha)^{-1} span\ (Ux - x). \tag{3.16}$$

A suboptimality test T_1 is said to be *stronger* than a suboptimality test T_2 if every action that is excluded as being suboptimal by test T_2 is also excluded as suboptimal by test T_1. The following theorem is intuitively obvious.

Theorem 3.10

Suboptimality tests based on stronger bounds yield stronger tests.

Proof

Suppose that $x^1 \leq x^2 \leq v^\alpha \leq y^2 \leq y^1$. Assume that an action $a \in A(i)$ is suboptimal by a test based on x^1 and y^1. Then,

$$r_i(a) + \sum_j p_{ij}(a)y_j^2 \leq r_i(a) + \sum_j p_{ij}(a)y_j^1 < (Ux^1)_i \leq (Ux^2)_i,$$

i.e. a is also suboptimal by the test based on x^2 and y^2. □

Corollary 3.5

Suboptimality test (3.16) is stronger than any other test; both the tests (3.15) and (3.14) are stronger than test (3.13), but are not mutually comparable.

Remark

In order to apply the tests (3.14) and (3.16) we need U^2x. However, in that case it is better to use the tests (3.13) and (3.15) with Ux instead of x, since $\|U^2x - Ux\|_\infty \leq \alpha \cdot \|Ux - x\|_\infty$ and $span\ (U^2x - Ux) \leq \alpha \cdot span\ (Ux - x)$ (see Exercise 3.9).

3.4 Policy iteration

In the method of *policy iteration* a sequence of deterministic policies $f_1^\infty, f_2^\infty, \ldots$ is constructed such that

$$v^\alpha(f_{k+1}^\infty) > v^\alpha(f_k^\infty) \text{ for } k = 1, 2, \ldots \tag{3.17}$$

where $x > y$, for $x, y \in \mathbb{R}^N$, means that $x_i \geq y_i$ for every i and $x_i > y_i$ for at least one i. Because $C(D)$ is finite, the method of policy iteration is also finite. Furthermore, we will show that the method is finite and gives an α-discounted optimal policy upon termination.

For every $i \in S$ and $f^\infty \in C(D)$, the action set $A(i, f)$ is defined by

$$A(i, f) = \{a \in A(i) \mid r_i(a) + \alpha \sum_j p_{ij}(a)v_j^\alpha(f^\infty) > v_i^\alpha(f^\infty)\}. \tag{3.18}$$

The intuitive idea of the policy iteration method is that if action $f(i)$ is replaced by an action $a \in A(i,f)$, the resulting policy improves the α-discounted reward. Therefore, $A(i,f)$ are called the set of *improving actions*. In the next theorem we show the correctness of this notion.

Theorem 3.11

(1) If $A(i,f) = \emptyset$ for every $i \in S$, then f^∞ is an α-discounted optimal policy.

(2) If $A(i,f) \neq \emptyset$ for some $i \in S$, then $v^\alpha(g^\infty) > v^\alpha(f^\infty)$ for any $g^\infty \in C(D)$ with $g \neq f$ and $g(i) \in A(i,f)$ when $g(i) \neq f(i)$.

Proof

(1) Since $A(i,f) = \emptyset$, we have for every $i \in S$, $L_g v^\alpha(f^\infty) = r(g) + \alpha P(g) v^\alpha(f^\infty) \leq v^\alpha(f^\infty)$ for every deterministic decision rule g. By Theorem 3.3, this implies that $v^\alpha(g^\infty) \leq L_g v^\alpha(f^\infty) \leq v^\alpha(f^\infty)$ for every $g^\infty \in C(D)$, i.e. f^∞ is optimal.

(2) Take any $g \neq f$ such that $g(i) \in A(i,f)$ if $g(i) \neq f(i)$. Then, if $g(i) \neq f(i)$,

$$r_i(g) + \alpha \sum_j p_{ij}(g) v_j^\alpha(f^\infty) > v_i^\alpha(f^\infty). \tag{3.19}$$

If $g(i) = f(i)$,

$$r_i(g) + \alpha \sum_j p_{ij}(g) v_j^\alpha(f^\infty) = r_i(f) + \alpha \sum_j p_{ij}(f) v_j^\alpha(f^\infty) = v_i^\alpha(f^\infty), \tag{3.20}$$

the last equation by Theorem 3.5. From (3.19) and (3.20) it follows that $L_g v^\alpha(f^\infty) = r(g) + \alpha P(g) v^\alpha(f^\infty) > v^\alpha(f^\infty)$. Hence, again by Theorem 3.3, we have $v^\alpha(g^\infty) \geq L_g v^\alpha(f^\infty) > v^\alpha(f^\infty)$. □

Algorithm 3.1 *Policy iteration algorithm*

1. Start with any $f^\infty \in C(D)$.
2. Compute $v^\alpha(f^\infty)$ as the unique solution of the linear system $L_f\, x = x$.
3. a. Compute $s_{ia}(f) = r_i(a) + \alpha \sum_j p_{ij}(a) v_j^\alpha(f^\infty) - v_i^\alpha(f^\infty)$ for every $(i,a) \in S \times A$.

   b. Determine $A(i,f) = \{a \in A(i) \mid s_{ia}(f) > 0\}$ for every $i \in S$.
4. If $A(i,f) = \emptyset$ for every $i \in S$: go to step 6.

   Otherwise: take any $g \neq f$ with $g(i) \in A(i,f)$ when $g(i) \neq f(i)$.
5. $f := g$ and return to step 2.
6. f^∞ is an α-discounted optimal policy (STOP).

Remark

There is some freedom concerning the choice of g in step 4. A usual choice is to take g such that

$$s_{ig(i)}(f) = max_a\ s_{ia}(f),\ i \in S. \tag{3.21}$$

Then, for each $i \in S$: $g(i) = f(i)$ when $A(i,f) = \emptyset$ and $g(i) \in A(i,f)$ when $A(i,f) \neq \emptyset$.

Example 3.1

$\alpha = \frac{1}{2};\ S = \{1,2,3\},\ A(1) = A(2) = A(3) = \{1,2,3\};\ r_1(1) = 1, r_1(2) = 2, r_1(3) = 3;$
$r_2(1) = 6, r_2(2) = 4, r_2(3) = 5;\ r_3(1) = 8, r_3(2) = 9, r_3(3) = 7.$

$p_{11}(1) = 1;\quad p_{12}(1) = 0;\quad p_{13}(1) = 0;\quad p_{11}(2) = 0;\quad p_{12}(2) = 1;\quad p_{13}(2) = 0;$

$p_{11}(3) = 0;\quad p_{12}(3) = 0;\quad p_{13}(3) = 1;\quad p_{21}(1) = 1;\quad p_{22}(1) = 0;\quad p_{23}(1) = 0;$

$p_{21}(2) = 0;\quad p_{22}(2) = 1;\quad p_{23}(2) = 0;\quad p_{21}(3) = 0;\quad p_{22}(3) = 0;\quad p_{23}(3) = 1;$

$p_{31}(1) = 1;\quad p_{32}(1) = 0;\quad p_{33}(1) = 0;\quad p_{31}(2) = 0;\quad p_{32}(2) = 1;\quad p_{33}(2) = 0;$

$p_{31}(3) = 0;\quad p_{32}(3) = 0;\quad p_{33}(3) = 1.$

Start with the policy f, with $f(1) = 3, f(2) = 2$ and $f(3) = 1$.

Iteration 1

The system $L_f\, x = x$ becomes:

$$\begin{array}{rcrcl} x_1 & & - & \frac{1}{2}x_3 = & 3 \\ & \frac{1}{2}x_2 & & = & 4 \\ -\frac{1}{2}x_1 & & + & x_3 = & 8 \end{array} \quad \rightarrow \quad \begin{array}{l} \text{solution: } v^\alpha(f^\infty) = (\frac{28}{3}, 8, \frac{38}{3}). \\ s_{11}(f) = -\frac{11}{3},\ s_{12}(f) = -\frac{10}{3},\ s_{13}(f) = 0. \\ s_{21}(f) = \frac{8}{3},\ s_{22}(f) = 0,\ s_{23}(f) = \frac{10}{3}. \\ s_{31}(f) = 0,\ s_{32}(f) = \frac{1}{3},\ s_{33}(f) = \frac{2}{3}. \end{array}$$

$g(1) = g(2) = g(3) = 3$, which becomes the new policy: $f(1) = f(2) = f(3) = 3$.

Iteration 2

The system $L_f\, x = x$ becomes:

$$\begin{array}{rcrcl} x_1 & & - & \frac{1}{2}x_3 = & 3 \\ & x_2 & - & \frac{1}{2}x_3 = & 5 \\ & & & \frac{1}{2}x_3 = & 7 \end{array} \quad \rightarrow \quad \begin{array}{l} \text{solution: } v^\alpha(f^\infty) = (10, 12, 14). \\ s_{11}(f) = -4,\ s_{12}(f) = -2,\ s_{13}(f) = 0. \\ s_{21}(f) = -1,\ s_{22}(f) = -2,\ s_{23}(f) = 0. \\ s_{31}(f) = -1,\ s_{32}(f) = 1,\ s_{33}(f) = 0. \end{array}$$

$g(1) = g(2) = 3$ and $g(3) = 2$, which becomes the new policy: $f(1) = f(2) = 3$ and $f(3) = 2$.

Iteration 3

The system $L_f x = x$ becomes:

$$\begin{array}{rcrcl} x_1 & & - & \frac{1}{2}x_3 = & 3 \\ & x_2 & - & \frac{1}{2}x_3 = & 5 \\ - & \frac{1}{2}x_2 & + & x_3 = & 9 \end{array} \quad \rightarrow \quad \begin{array}{l} \text{solution: } v^\alpha(f^\infty) = (\frac{32}{3}, \frac{38}{3}, \frac{46}{3}). \\ s_{11}(f) = -\frac{11}{3},\ s_{12}(f) = -\frac{7}{3},\ s_{13}(f) = 0. \\ s_{21}(f) = -\frac{4}{3},\ s_{22}(f) = -\frac{7}{3},\ s_{23}(f) = 0. \\ s_{31}(f) = -2,\ s_{32}(f) = 0,\ s_{33}(f) = -\frac{2}{3}. \end{array}$$

f^∞ with $f(1) = f(2) = 3$ and $f(3) = 2$ is an optimal policy.

We now discuss the elimination of suboptimal actions with test (3.15) and $x = v^\alpha(f^\infty)$. Since

$$(Ux - x)_i = max_a\{r_i(a) + \sum_j p_{ij}(a)v_j^\alpha(f^\infty)\} - v_i^\alpha(f^\infty) = max_a\, s_{ia}(f),\ i \in S.$$

and $span(Ux - x) = max_i max_a\ s_{ia}(f) - min_i max_a\ s_{ia}(f)$, (3.15) becomes

$$s_{ia}(f) < max_a\ s_{ia}(f) - \alpha(1-\alpha)^{-1}\{max_i max_a\ s_{ia}(f) - min_i max_a\ s_{ia}(f)\},$$

resulting in the following theorem.

Theorem 3.12 *(Suboptimality test)*
If $s_{ia_i}(f) < max_a\ s_{ia}(f) - \alpha(1-\alpha)^{-1}\{max_i max_a\ s_{ia}(f) - min_i max_a\ s_{ia}(f)\}$, *then action* $a_i \in A(i)$ *is a suboptimal action.*

Remark
Since $s_{if(i)}(f) = 0,\ i \in S$, we have $max_i max_a\ s_{ia}(f) \geq min_i max_a\ s_{ia}(f) \geq min_i\, s_{if(i)}(f) = 0$.

Algorithm 3.2 *Policy iteration algorithm with suboptimality test and using (3.21)*

1. Start with any $f^\infty \in C(D)$.
2. Compute $v^\alpha(f^\infty)$ as the unique solution x of the linear system $L_f x = x$.
3. a. Compute $s_{ia}(f) = r_i(a) + \alpha \sum_j\ p_{ij}(a)v_j^\alpha(f^\infty) - v_i^\alpha(f^\infty)$ for every $(i,a) \in S \times A$.
   b. Determine $A(i,f) = \{a \in A(i) \mid s_{ia}(f) > 0\}$ for every $i \in S$.
4. If $A(i,f) = \emptyset$ for every $i \in S$: go to step 7.
   Otherwise: take g such that $s_{ig(i)}(f) = max_a\, s_{ia}(f),\ i \in S$.
5. $A(i) = \left\{a \middle| s_{ia}(f) \geq max_a\, s_{ia}(f) - \alpha(1-\alpha)^{-1}\{max_i max_a\, s_{ia}(f) - min_i max_a\, s_{ia}(f)\}\right\},\ i \in S$.
6. $f := g$ and return to step 2.
7. f^∞ is an α-discounted optimal policy (STOP).

Example 3.1 (continued)

Iteration 1

$\alpha(1-\alpha)^{-1}\{max_i max_a\ s_{ia}(f) - min_i max_a\ s_{ia}(f)\} = \frac{10}{3}$.

In state 1, action 1 is excluded, because $-\frac{11}{3} = s_{11}(f) < max_a\ s_{1a}(f) - \frac{10}{3} = -\frac{10}{3}$.

Iteration 2

$\alpha(1-\alpha)^{-1}\{max_i max_a\ s_{ia}(f) - min_i max_a\ s_{ia}(f)\} = 1$.

In state 1, action 2 is excluded, because $-2 = s_{12}(f) < max_a\ s_{1a}(f) - 1 = -1$.

In state 2, action 2 is excluded, because $-2 = s_{22}(f) < max_a\ s_{2a}(f) - 1 = -1$.

In state 3, action 1 is excluded, because $-1 = s_{31}(f) < max_a\ s_{3a}(f) - 1 = 0$.

Next, we show that the policy iteration algorithm 3.2 is equivalent to *Newton's method* for solving the optimality equation $Ux = x$. Furthermore, we can make a statement about the convergence rate. The choice (3.21) implies that $r(g) + \alpha P(g)v^\alpha(f^\infty) - v^\alpha(f^\infty) = Uv^\alpha(f^\infty) - v^\alpha(f^\infty)$, i.e.

$$L_g v^\alpha(f^\infty) = Uv^\alpha(f^\infty) \text{ and } g = f_{v^\alpha(f^\infty)}. \tag{3.22}$$

Define the operator F by

$$F : \mathbb{R}^N \to \mathbb{R}^N \text{ by } Fx = Ux - x. \tag{3.23}$$

Hence, v^α is the unique solution of the equation $Fx = 0$. Since $L_{f_x}x = Ux$, it follows that

$$Fx = L_{f_x}x - x = r(f_x) + \alpha P(f_x)x - x. \tag{3.24}$$

Suppose that Newton's method is applied to solve the equation $Fx = 0$. This method works as follows: starting with the vector x^1, successive values $x^2, x^3, \ldots$ are computed by the formula

$$x^{n+1} = x^n - \{\nabla Fx^n\}^{-1}Fx^n. \tag{3.25}$$

where ∇F is the *Jacobian* of F, i.e. ∇Fx^n is an $N \times N$ matrix defined by

$$\{\nabla Fx^n\}_{ij} = \left\{\frac{\partial(Fx)_i}{\partial x_j}\right\}_{x=x^n}.$$

From (3.24) it follows that $\nabla Fx^n = \alpha P(f_{x^n}) - I$, where we assume that $r(f_x)$ and $P(f_x)$ are constant in a small neighbourhood of x^n. Hence, (3.25) can be written as

$$x^{n+1} = x^n + \{I - \alpha P(f_{x^n})\}^{-1}\Big\{r(f_{x^n}) - \{I - \alpha P(f_{x^n})\}x^n\Big\} = x^n + v^\alpha(f^\infty_{x^n}) - x^n = v^\alpha(f^\infty_{x^n}). \tag{3.26}$$

Theorem 3.13

Suppose that $f_1^\infty, f_2^\infty, \ldots, f_p^\infty$ are the policies obtained by algorithm 3.2 and, on the other hand, suppose that Newton's method is applied in order to solve the equation $Fx = 0$ with starting vector $x^1 = v^\alpha(f_1^\infty)$. Then,

(1) $x^n = v^\alpha(f_n^\infty),\ n = 1, 2, \ldots, p.$

(2) $\|v^\alpha - v^\alpha(f_{n+1}^\infty)\|_\infty \leq 2\alpha(1-\alpha)^{-1}\|v^\alpha - v^\alpha(f_n^\infty)\|_\infty,\ n = 1, 2, \ldots, p-1.$

Proof

(1) We apply induction on n (the result is obvious for $n = 1$). Suppose that $x^n = v^\alpha(f_n^\infty)$, then we have to show that $x^{n+1} = v^\alpha(f_{n+1}^\infty)$. By (3.26) and the induction hypothesis,

$$x^{n+1} = v^\alpha(f^\infty_{x^n}) = v^\alpha(f^\infty_{v^\alpha(f_n^\infty)}). \tag{3.27}$$

It follows from (3.22) that $f_{n+1} = f_{v^\alpha(f_n^\infty)}$. Hence, by (3.27),

$$x^{n+1} = v^\alpha(f^\infty_{v^\alpha(f_n^\infty)}) = v^\alpha(f_{n+1}^\infty). \tag{3.28}$$

(2) $0 \leq v^\alpha - v^\alpha(f_{n+1}^\infty) = v^\alpha - x^{n+1} = v^\alpha - x^n - \{I - \alpha P(f_{x^n})\}^{-1} F x^n$, and

$$\begin{aligned} Fx^n &= Ux^n - x^n = Ux^n - x^n - Uv^\alpha + v^\alpha \geq L_{f_{v^\alpha}} x^n - x^n - Uv^\alpha + v^\alpha \\ &= L_{f_{v^\alpha}} x^n - x^n - L_{f_{v^\alpha}} v^\alpha + v^\alpha = \{I - \alpha P(f_{v^\alpha})\}(v^\alpha - x^n). \end{aligned}$$

Hence,

$$\begin{aligned} 0 &\leq v^\alpha - v^\alpha(f_{n+1}^\infty) \leq v^\alpha - x^n - \{I - \alpha P(f_{x^n})\}^{-1}\{I - \alpha P(f_{v^\alpha})\}(v^\alpha - x^n) \\ &= \{I - \alpha P(f_{x^n})\}^{-1}\{I - \alpha P(f_{x^n})\}(v^\alpha - x^n) - \{I - \alpha P(f_{x^n})\}^{-1}\{I - \alpha P(f_{v^\alpha})\}(v^\alpha - x^n) \\ &= \{I - \alpha P(f_{x^n})\}^{-1}\{\alpha P(f_{v^\alpha}) - \alpha P(f_{x^n})\}(v^\alpha - x^n) \\ &= \alpha\{I - \alpha P(f_{x^n})\}^{-1}\{P(f_{v^\alpha}) - P(f_{x^n})\}(v^\alpha - x^n). \end{aligned}$$

Consequently,

$$\begin{aligned} \|v^\alpha - v^\alpha(f_{n+1}^\infty)\|_\infty &\leq \alpha\|\{I - \alpha P(f_{x^n})\}^{-1}\|_\infty \cdot \|P(f_{v^\alpha}) - P(f_{x^n})\|_\infty \cdot \|v^\alpha - v^\alpha(f_n^\infty)\|_\infty \\ &= \alpha\|\textstyle\sum_{t=0}^\infty [\alpha P(f_{x^n})]^t\|_\infty \cdot \|P(f_{v^\alpha}) - P(f_{x^n})\|_\infty \cdot \|v^\alpha - v^\alpha(f_n^\infty)\|_\infty \\ &\leq \alpha(1-\alpha)^{-1}\|P(f_{v^\alpha}) - P(f_{x^n})\|_\infty \cdot \|v^\alpha - v^\alpha(f_n^\infty)\|_\infty \\ &= 2\alpha(1-\alpha)^{-1}\|v^\alpha - v^\alpha(f_n^\infty)\|_\infty. \end{aligned}$$

□

Remark

In the last line of the proof the inequality $\|P(f_{v^\alpha}) - P(f_{x^n})\|_\infty \leq 2$ is used. This is a theoretical bound. Usually, $\|P(f_{v^\alpha}) - P(f_{x^n})\|_\infty$ is much smaller and for large n this norm tends to zero.

In general, the solution of the linear system $L_f x = x$ by Gauss elimination in step 2 of the policy iteration algorithm needs $\mathcal{O}(N^3)$ operations (cf. Stoer and Bulirsch ([192] pp. 169-172). However, by applying the next theorem, this evaluation can be done in $\mathcal{O}(mN^2)$ operations, where m is the number of states i in which $g(i) \neq f(i)$ with g the decision rule in step 4 of the algorithm.

Theorem 3.14

$(B + UV^t)^{-1} = B^{-1} - B^{-1}U(I + V^t B^{-1} U)^{-1} V^t B^{-1}$, *assuming each of the inverses exists and that the matrices have the appropriate dimensions. In this expression,* V^t *denotes the transpose of matrix* V.

Proof

Let $T = (I + V^t B^{-1} U)^{-1}$, then

$$\begin{aligned} (B + UV^t)(B^{-1} - B^{-1}UTV^tB^{-1}) &= I - UTV^tB^{-1} + UV^tB^{-1} - UV^tB^{-1}UTV^tB^{-1} \\ &= I - U(T - I)V^tB^{-1} - U(T^{-1} - I)TV^tB^{-1} \\ &= I - UTV^tB^{-1} + UV^tB^{-1} - UV^tB^{-1} + UTV^tB^{-1} \\ &= I. \end{aligned}$$

□

Let $\{i \mid g(i) \neq f(i)\} = \{i_1, i_2, \ldots, i_m\}$, U the $N \times m$ matrix $\{e_{i_1}, e_{i_2}, \ldots, e_{i_m}\}$, where e_{i_k} is the i_k-th unit vector in $\mathbb{R}^N$, and V is the $N \times m$ matrix $\{v^1, v^2, \ldots, v^m\}$, where $v_l^k = -\alpha\{p_{i_k l}(g) - p_{i_k l}(f)\},\ 1 \leq k \leq m,\ 1 \leq l \leq N$. Then,

$$(UV^t)_{kj} = \textstyle\sum_l\ u_{kl} v_{jl} = \begin{cases} v_j^k = -\alpha\{p_{i_k j}(g) - p_{i_k j}(f)\} & k = i_1, i_2, \ldots, i_m,\ j \in S; \\ 0 & k \neq i_1, i_2, \ldots, i_m,\ j \in S. \end{cases}$$

Hence, $I - \alpha P(g) = I - \alpha P(f) + UV^t$. Applying Theorem 3.14 yields the next result.

Theorem 3.15
If $I + V^t\{I - \alpha P(f)\}^{-1}U$ is nonsingular, then

$$\{I-\alpha P(g)\}^{-1} = \{I-\alpha P(f)\}^{-1} - \{I-\alpha P(f)\}^{-1}U\{I+V^t\{I-\alpha P(f)\}^{-1}U\}^{-1}V^t\{I-\alpha P(f)\}^{-1}.$$

Corollary 3.6
If $\{I - \alpha P(f)\}^{-1}$ is known, then $\{I - \alpha P(g)\}^{-1}$ can be computed in $\mathcal{O}(mN^2)$ operations.

Proof
$\{I - \alpha P(g)\}^{-1}$ can be computed as follows:

1. $Y_1 = V^t\{I - \alpha P(f)\}^{-1}$: mN^2 operations;
2. $Y_2 = Y_1U$: m^2N operations;
3. $Y_3 = I + Y_2$: m operations;
4. $Y_4 = Y_3^{-1}$: m^3 operations;
5. $Y_5 = Y_4Y_1$: m^2N operations;
6. $Y_6 = \{I - \alpha P(f)\}^{-1}U$: mN^2 operations;
7. $Y_7 = Y_6Y_4$: m^2N operations;
8. $Y_8 = \{I - \alpha P(f)\}^{-1} - Y_7$: N^2 operations.

Hence, the overall complexity is $\mathcal{O}(mN^2)$. □

Remark
The computation of one $s_{ia}(f)$ in step 3 of the algorithm requires $\mathcal{O}(N)$ computations. Since $M = \#(S \times A)$ $s_{ia}(f)$-values must be computed, the complexity of one iteration of the policy iteration algorithm is $\mathcal{O}\big(N(mN + M)\big)$.

3.5 Linear programming

The value vector v^α is the unique solution of the optimality equation (3.4), i.e.

$$v_i^\alpha = max_{a\in A(i)}\ \{r_i(a) + \alpha\textstyle\sum_j\ p_{ij}(a)v_j^\alpha\},\ i \in S.$$

Hence, v^α satisfies

$$v_i^\alpha\ \geq\ r_i(a) + \alpha\sum_j\ p_{ij}(a)v_j^\alpha \text{ for all } (i,a) \in S \times A. \tag{3.29}$$

Intuitively it is clear that v^α is the smallest vector satisfying (3.29). This is the key property for the linear programming approach towards computing the value vector. It turns out that an optimal policy can be obtained from the dual linear program. We also show a one-to-one correspondence between the stationary policies and the feasible solutions of the dual program, such that the extreme points correspond to deterministic policies. Furthermore, we show that the linear programming method for discounted MDPs can be considered equivalent to the policy iteration method, and that exclusion of suboptimal actions can be included in the linear programming method.

A vector $v \in \mathbb{R}^N$ is said to be *α-superharmonic* if

$$v_i \;\geq\; r_i(a) + \alpha \sum_j \; p_{ij}(a) v_j \text{ for all } (i,a) \in S \times A. \tag{3.30}$$

Theorem 3.16

v^α is the smallest α-superharmonic vector (componentwise).

Proof

Since

$$v_i^\alpha = max_{a \in A(i)} \{ r_i(a) + \alpha \textstyle\sum_j \; p_{ij}(a) v_j^\alpha \} \geq r_i(a) + \alpha \textstyle\sum_j \; p_{ij}(a) v_j^\alpha \text{ for all } (i,a) \in S \times A,$$

v^α is α-superharmonic. Suppose that $v \in \mathbb{R}^N$ is also α-superharmonic. Then,

$$v \geq r(f) + \alpha P(f) v \text{ for every } f^\infty \in C(D),$$

which implies $\{I - \alpha P(f)\} v \geq r(f)$. Since $\{I - \alpha P(f)\}^{-1} = \sum_{t=0}^{\infty} \; \alpha^t P^t(f) \geq 0$, we obtain

$$v \geq \{I - \alpha P(f)\}^{-1} r(f) = v^\alpha(f^\infty), \;\; f^\infty \in C(D).$$

Hence, $v_i^\alpha = max_f \; v^\alpha(f^\infty) \leq v$, i.e. v^α is the smallest α-superharmonic vector. □

Corollary 3.7

v^α is the unique optimal solution of the linear programming problem

$$min \left\{ \sum_j \; \beta_j v_j \;\middle|\; \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j \geq r_i(a), \;\; (i,a) \in S \times A \right\} \tag{3.31}$$

where β_j is any strictly positive number for every $j \in S$.

Proof

From Theorem 3.16 it follows that v^α is a feasible solution of (3.31) and that $v^\alpha \leq v$ for every feasible solution v of (3.31). Hence, v^α is the unique optimal solution of (3.31). □

By Corollary 3.7, the value vector v^α can be found as optimal solution of the linear program (3.31). This program does not give an optimal policy. However, an optimal policy can be obtained from the solution of the dual program

$$max \left\{ \sum_{(i,a)} \; r_i(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) & = & \beta_j, \;\; j \in S \\ x_i(a) & \geq & 0, \;\; (i,a) \in S \times A \end{array} \right\}. \tag{3.32}$$

Theorem 3.17

(1) Any feasible solution x of (3.32) satisfies $\sum_a \; x_j(a) > 0, \;\; j \in S$.

(2) The dual program (3.32) has a finite optimal solution, say x^*.

(3) Any $f_*^\infty \in C(D)$ with $x_i^*\big(f_*(i)\big) > 0$ for every $i \in S$ is an α-discounted optimal policy.

Proof

(1) Let x be a feasible solution of (3.32). From the constraints of (3.32) it follows that

$$\sum_a x_j(a) = \beta_j + \alpha \sum_{(i,a)} p_{ij}(a) x_i(a) \geq \beta_j > 0, \ j \in S.$$

(2) Since the primal program (3.31) has a finite optimal solution, namely the value vector v^α, it follows from the theory of linear programming that the dual program (3.32) also has a finite optimal solution.

(3) Take any $f_*^\infty \in C(D)$ with $x_i^*\big(f_*(i)\big) > 0$ for every $i \in S$ (such policy exists by part (1)). Because $x_i^*\big(f_*(i)\big) > 0, \ i \in S$, the complementary slackness property of linear programming implies

$$\sum_j \{\delta_{ij} - \alpha p_{ij}(f_*)\} v_j^\alpha = r_i(f_*), \ i \in S.$$

Hence, in vector notation,

$$\{I - \alpha P(f_*)\} v^\alpha = r(f_*), \text{ which implies } v^\alpha = \{I - \alpha P(f_*)\}^{-1} r(f_*) = v^\alpha(f_*^\infty),$$

i.e. f_*^∞ is an α-discounted optimal policy. □

If the simplex method is used, then the programs (3.31) and (3.32) are solved simultaneously. Hence, by the simplex method both the value vector v^α and an optimal policy are computed.

Next, we show the one-to-one correspondence between the feasible solutions of (3.32) and the set $C(S)$ of stationary policies. For $\pi^\infty \in C(S)$ the vector x^π with components $x_i^\pi(a), \ (i,a) \in S \times A$, is defined by

$$x_i^\pi(a) = \Big\{ \beta^T \{I - \alpha P(\pi)\}^{-1} \Big\}_i \cdot \pi_{ia}, \ (i,a) \in S \times A. \tag{3.33}$$

Define, for any $t \in \mathbb{N}$ and $(i,a) \in S \times A$, a random variable $n_{ia}^{(t)}$ by

$$n_{ia}^{(t)} = \begin{cases} 1 \text{ if } (X_t, Y_t) = (i,a) \\ 0 \text{ otherwise} \end{cases}$$

Then, the total discounted number of times that $(X_t, Y_t) = (i,a)$ equals $\sum_{t=1}^\infty \alpha^{t-1} \ n_{ia}^{(t)}$.

Lemma 3.4

Given initial distribution β, i.e. $\mathbb{P}\{X_1 = j\} = \beta_j$ for all $j \in S$, and a stationary policy π^∞, $x_i^\pi(a)$ satisfies $x_i^\pi(a) = \mathbb{E}_{\beta,\pi}\Big\{ \sum_{t=1}^\infty \alpha^{t-1} n_{ia}^{(t)} \Big\}, \ (i,a) \in S \times A$.

Proof

Since $\{I - \alpha P(\pi)\}^{-1} = \sum_{t=1}^\infty \alpha^{t-1} P^{t-1}(\pi)$, we have

$$\begin{aligned} x_i^\pi(a) &= \textstyle\sum_j \beta_j \cdot \{\sum_{t=1}^\infty \alpha^{t-1} P^{t-1}(\pi)\}_{ji} \cdot \pi_{ia} = \sum_{t=1}^\infty \alpha^{t-1} \Big\{ \sum_j \beta_j \, \mathbb{P}_\pi\{X_t = i \mid X_1 = j\} \Big\} \cdot \pi_{ia} \\ &= \textstyle\sum_{t=1}^\infty \alpha^{t-1} \Big\{ \sum_j \beta_j \, \mathbb{P}_\pi\{X_t = i, Y_t = a \mid X_1 = j\} \Big\} = \sum_{t=1}^\infty \alpha^{t-1} \cdot \mathbb{E}_{\beta,\pi}\left\{ n_{ia}^{(t)} \right\} \\ &= \mathbb{E}_{\beta,\pi}\Big\{ \textstyle\sum_{t=1}^\infty \alpha^{t-1} n_{ia}^{(t)} \Big\}. \end{aligned}$$

□

Conversely, for a feasible solution x of (3.32), define π^x with elements π^x_{ia} by

$$\pi^x_{ia} = \frac{x_i(a)}{\sum_a x_i(a)}, \ (i,a) \in S \times A. \tag{3.34}$$

Theorem 3.18

The mapping (3.33) is a one-to-one mapping of the set of stationary policies onto the set of feasible solutions of the dual program (3.32) with (3.34) as the inverse mapping; furthermore, the set of extreme feasible solutions of (3.32) corresponds to the set $C(D)$ of deterministic policies.

Proof

First, we show that x^π is a feasible solution of (3.32).

$$\begin{aligned}
\textstyle\sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x^\pi_i(a) &= \textstyle\sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} \{\beta^T \{I - \alpha P(\pi)\}^{-1}\}_i \cdot \pi_{ia} \\
&= \textstyle\sum_i \left\{\beta^T \{I - \alpha P(\pi)\}^{-1}\right\}_i \cdot \sum_a \{\delta_{ij} - \alpha p_{ij}(a)\} \\
&= \textstyle\sum_i \left\{\beta^T \{I - \alpha P(\pi)\}^{-1}\right\}_i \cdot \{I - \alpha P(\pi)\}_{ij} \\
&= \left\{\beta^T \{I - \alpha P(\pi)\}^{-1} \cdot \{I - \alpha P(\pi)\}\right\}_j = \beta_j, \ j \in S.
\end{aligned}$$

Since $\{I - \alpha P(\pi)\}^{-1} = \sum_{t=0}^\infty \{\alpha P(\pi)\}^t \geq 0, \ x^\pi_i(a) \geq 0$ for every $(i,a) \in S \times A$.

Next, we prove the one-to-one correspondence. Let x be a feasible solution of (3.32).

Then, (3.34) yields $x_i(a) = \pi^x_{ia} \cdot x_i$, where $x_i = \sum_a x_i(a), \ i \in S$. Therefore, we can write

$$\begin{aligned}
\beta_j &= \textstyle\sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) = \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} \cdot \pi^x_{ia} \cdot x_i \\
&= \textstyle\sum_i \{\delta_{ij} - \alpha p_{ij}\big(\pi(x)\big)\} \cdot x_i, \ j \in S.
\end{aligned}$$

Hence, in vector notation, $\beta^T = x^T \{I - \alpha P\big(\pi(x)\big)\}$, i.e. $x^T = \beta^T \{I - \alpha P\big(\pi(x)\big)\}^{-1} = \{x\big(\pi(x)\big)\}^T$.
Conversely,

$$\pi^{x(\pi)}_{ia} = \frac{x^\pi_i(a)}{\sum_a x^\pi_i(a)} = \pi_{ia}, \ (i,a) \in S \times A. \tag{3.35}$$

Therefore, we have shown the one-to-one correspondence and the fact that (3.34) is the inverse of (3.33). Finally, we show the correspondence between the extreme points of (3.32) and the set $C(D)$. Let $f^\infty \in C(D)$. Then, for every $i \in S$,

$$x^f_i(a) = \begin{cases} \left\{\beta^T \{I - \alpha P(f)\}^{-1}\right\}_i & , \ a = f(i) \\ 0 & , \ a \neq f(i) \end{cases}$$

Suppose x^f is not an extreme feasible solution. Then, there exist feasible solutions x^1 and x^2 of (3.32) and a real number $\lambda \in (0,1)$ such that $x^1 \neq x^2$ and $x^f = \lambda x^1 + (1-\lambda) x^2$.

Since $x^f_i(a) = 0, a \neq f(i), i \in S$, we have $x^1_i(a) = x^2_i(a) = 0, \ a \neq f(i), i \in S$.

Hence, the N-vectors $x^1 = \big(x^1_i(f(i))\big)$ and $x^2 = \big(x^2_i(f(i))\big)$ are solutions of the linear system $x^T \{I - \alpha P(f)\} = \beta^T$. However, this linear system has a unique solution $x^T = \beta^T \{I - \alpha P(f)\}^{-1}$. This implies $x^1 = x^2 = \beta^T \{I - \alpha P(f)\}^{-1}$, which contradicts $x^1 \neq x^2$. Hence, we have shown that

x^f is an extreme solution.
Conversely, let x be an extreme feasible solution of program (3.32). Since (3.32) has N constraints, x has at most N positive components. On the other hand, Theorem 3.17, part (1), implies that in each state there is at least one positive component. Consequently, x has exactly one positive component in each state i, i.e. the corresponding stationary policy is deterministic. $\square$

Algorithm 3.3 *Linear programming algorithm*

1. Take any vector β, where $\beta_j > 0,\ j \in S$.
2. Use the simplex method to compute optimal solutions v^* and x^* of the dual pair of linear programs:

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j \geq r_i(a),\ (i,a) \in S \times A \right\}$$

and

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) & = \beta_j,\ j \in S \\ x_i(a) & \geq 0,\ (i,a) \in S \times A \end{array} \right\}.$$

3. Take $f_*^\infty \in C(D)$ such that $x_i^*(f_*(i)) > 0$ for every $i \in S$.

   v^* is the value vector v^α and f_*^∞ is an α-discounted optimal policy (STOP).

Example 3.2
Consider the model of Example 3.1 and let $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$.
The dual linear program (3.32) becomes:

$$max\ x_1(1) + 2x_1(2) + 3x_1(3) + 6x_2(1) + 4x_2(2) + 5x_2(3) + 8x_3(1) + 9x_3(2) + 7x_3(3)$$

subject to

$$\begin{array}{rcl} \frac{1}{2}x_1(1) + x_1(2) + x_1(3) - \frac{1}{2}x_2(1) - \frac{1}{2}x_3(1) & = & \frac{1}{3} \\ \frac{1}{2}x_1(2) + x_2(1) + \frac{1}{2}x_2(2) + x_2(3) - \frac{1}{2}x_3(2) & = & \frac{1}{3} \\ - \frac{1}{2}x_1(3) - \frac{1}{2}x_2(3) + x_3(1) + x_3(2) + \frac{1}{2}x_3(3) & = & \frac{1}{3} \end{array}$$

$$x_1(1), x_1(2), x_1(3), x_2(1), x_2(2), x_2(3), x_3(1), x_3(2), x_3(3) \geq 0$$

We start with phase I of the simplex method to obtain a first feasible basic solution corresponding to policy f^∞ where $f(1) = 3,\ f(2) = 2$ and $f(3) = 1$. Therefore, we take the columns of $x_1(3), x_2(2)$ and $x_3(1)$ as pivot columns in the first three iterations. The pivot element is the bold number in the tableau. Next, in phase II, the usual choice of the pivot column is taken, i.e. the column with the most negative element in the transformed objective function (last row in the tableau, also called the row of the *reduced costs*). We write the linear programming tableaus in the so-called contracted form (cf. [241]).

Iteration 1

		$x_1(1)$	$x_1(2)$	$x_1(3)$	$x_2(1)$	$x_2(2)$	$x_2(3)$	$x_3(1)$	$x_3(2)$	$x_3(3)$
z_1	$\frac{1}{3}$	$\frac{1}{2}$	1	**1**	$-\frac{1}{2}$	0	0	$-\frac{1}{2}$	0	0
z_2	$\frac{1}{3}$	0	$-\frac{1}{2}$	0	1	$\frac{1}{2}$	1	0	$-\frac{1}{2}$	0
z_3	$\frac{1}{3}$	0	0	$-\frac{1}{2}$	0	0	$-\frac{1}{2}$	1	1	$-\frac{1}{2}$
I	-1	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$
II	0	-1	-2	-3	-6	-4	-5	-8	-9	-7

Iteration 2

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	$x_2(2)$	$x_2(3)$	$x_3(1)$	$x_3(2)$	$x_3(3)$
$x_1(3)$	$\frac{1}{3}$	$\frac{1}{2}$	1	1	$-\frac{1}{2}$	0	0	$-\frac{1}{2}$	0	0
z_2	$\frac{1}{3}$	0	$-\frac{1}{2}$	0	1	$\frac{\mathbf{1}}{\mathbf{2}}$	1	0	$-\frac{1}{2}$	0
z_3	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	$-\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{2}$
I	$-\frac{5}{6}$	$-\frac{1}{4}$	0	$\frac{1}{2}$	$-\frac{3}{4}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{3}{4}$	$-\frac{1}{2}$	$-\frac{1}{2}$
II	1	$\frac{1}{2}$	1	3	$-\frac{15}{2}$	-4	-5	$-\frac{19}{2}$	-9	-7

Iteration 3

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	z_2	$x_2(3)$	$x_3(1)$	$x_3(2)$	$x_3(3)$
$x_1(3)$	$\frac{1}{3}$	$\frac{1}{2}$	1	1	$-\frac{1}{2}$	0	0	$-\frac{1}{2}$	0	0
$x_2(2)$	$\frac{2}{3}$	0	-1	0	2	2	2	0	-1	0
z_3	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	$-\frac{1}{2}$	$\frac{\mathbf{3}}{\mathbf{4}}$	1	$\frac{1}{2}$
I	$-\frac{1}{2}$	$-\frac{1}{4}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	1	$\frac{1}{2}$	$-\frac{3}{4}$	-1	$-\frac{1}{2}$
II	$\frac{11}{3}$	$\frac{1}{2}$	-3	3	$\frac{1}{2}$	8	3	$-\frac{19}{2}$	-13	-7

Iteration 4

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	z_2	$x_2(3)$	z_3	$x_3(2)$	$x_3(3)$
$x_1(3)$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$-\frac{2}{3}$	0	$-\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$
$x_2(2)$	$\frac{2}{3}$	0	-1	0	2	2	**2**	0	-1	0
$x_3(1)$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$	0	$-\frac{2}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$\frac{2}{3}$
I	0	0	0	1	0	1	0	1	0	0
II	10	$\frac{11}{3}$	$\frac{10}{3}$	$\frac{28}{3}$	$-\frac{8}{3}$	8	$-\frac{10}{3}$	$\frac{38}{3}$	$-\frac{1}{3}$	$-\frac{2}{3}$

Iteration 5

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	z_2	$x_2(2)$	z_3	$x_3(2)$	$x_3(3)$
$x_1(3)$	$\frac{7}{9}$	$\frac{2}{3}$	$\frac{7}{6}$	$\frac{4}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{3}$
$x_2(3)$	$\frac{1}{3}$	0	$-\frac{1}{2}$	0	1	1	$\frac{1}{2}$	0	$-\frac{1}{2}$	0
$x_3(1)$	$\frac{8}{9}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{4}{3}$	**1**	$\frac{2}{3}$
II	$\frac{100}{9}$	$\frac{11}{3}$	$\frac{5}{3}$	$\frac{28}{3}$	$\frac{2}{3}$	$\frac{34}{3}$	$\frac{5}{3}$	$\frac{38}{3}$	-2	$-\frac{2}{3}$

Iteration 6

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	z_2	$x_2(2)$	z_3	$x_3(1)$	$x_3(3)$
$x_1(3)$	$\frac{1}{3}$	$\frac{1}{2}$	1	1	$-\frac{1}{2}$	0	0	0	$-\frac{1}{2}$	0
$x_2(3)$	$\frac{7}{9}$	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{7}{6}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{3}$
$x_3(2)$	$\frac{8}{9}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{4}{3}$	1	$\frac{2}{3}$
II	$\frac{116}{9}$	$\frac{13}{3}$	$\frac{7}{3}$	$\frac{32}{3}$	$\frac{4}{3}$	$\frac{38}{3}$	$\frac{7}{3}$	$\frac{46}{3}$	2	$\frac{2}{3}$

The last tableau is an optimal simplex tableau corresponding to the optimal solution:

$x_1^*(1) = 0, x_1^*(2) = 0, x_1^*(3) = \frac{1}{3}$; $x_2^*(1) = 0, x_2^*(2) = 0, x_2^*(3) = \frac{7}{9}$; $x_3^*(1) = 0, x_3^*(2) = \frac{8}{9}, x_3^*(3) = 0$.

The optimal solution of the primal problem is: $v_1^* = \frac{32}{3}, v_2^* = \frac{38}{3}$ and $v_3^* = \frac{46}{3}$.

Hence, the value vector $v^\alpha = (\frac{32}{3}, \frac{38}{3}, \frac{46}{3})$ and the α-discounted optimal policy is f_*^∞ with

$f_*(1) = 3, f_*(2) = 3$ and $f_*(3) = 2$.

We now show the equivalence between the policy iteration method and the linear programming method. Consider a deterministic policy f^∞. We have seen that $x(f)$ is an extreme point of (3.32) and that $x_i^f(f(i)) > 0$ for every $i \in S$. By introducing slack variables $y_i(a)$, $(i,a) \in S \times A$ in the primal problem (3.31), this program becomes

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{rcll} \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j - y_i(a) & = & r_i(a), & (i,a) \in S \times A \\ y_i(a) & \geq & 0, & (i,a) \in S \times A \end{array} \right\}. \tag{3.36}$$

Let $\big(v(f), y(f)\big)$ be the dual solution corresponding to $x(f)$. Then, by the complementary slackness property of linear programming, we have

$$x_i^f(a) \cdot y_i^f(a) = 0 \text{ for every } (i,a) \in S \times A.$$

Since $x_i^f(f(i)) > 0$ for every $i \in S$, $y_i^f(f(i)) = 0$ for every $i \in S$. Hence, from the constraints of (3.36), we obtain in vector notation $\{I - \alpha P(f)\} v(f) = r(f)$, implying that

$$v(f) = \{I - \alpha P(f)\}^{-1} r(f) = v^\alpha(f^\infty),$$

and

$$y_i^f(a) = \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j^\alpha(f^\infty) - r_i(a) = -s_{ia}(f), \; (i,a) \in S \times A,$$

where $s_{ia}(f)$ is defined in the policy iteration algorithms 3.1 and 3.2.

In any simplex tableau, possible choices for the pivot column are those columns of nonbasic $x_i(a)$-variables which have negative reduced costs: $y_i^f(a) < 0$, i.e. $s_{ia}(f) > 0$. Hence, the possible pivot columns in state i are exactly the columns corresponding to the actions of $A(i,f)$, where $A(i,f)$ is defined in (3.18).

Consider in the policy iteration method two subsequent policies, say f^∞ and g^∞, and let

$$E(f,g) = \{i \in S \mid f(i) \neq g(i)\}.$$

If we exchange in the simplex method in one iteration the nonbasic variables $x_i^f(a)$ and $x_i^g(a)$ for every $i \in E(f,g)$, then we obtain a linear programming algorithm in which (in general) more than one pivot step is executed in one iteration, and in which subsequent basic solutions correspond to subsequent policies of the policy iteration method. An algorithm in which in one iteration more than one pivot step can be executed is called a *block-pivoting simplex algorithm* (cf. [38] p. 201).

On the other hand, supppose that the usual simplex algorithm is applied with only one pivot step in one iteration and that we choose as entering variable a nonbasic variable $x_i^f(a)$ corresponding to a variable $s_{ia}(f) > 0$, i.e. $a \in A(i,f)$. Since such a choice is allowed in the policy iteration method, the usual simplex method is a special implementation of the policy iteration method. We summarize the above statements in the following theorem.

Theorem 3.19

(1) Any policy iteration algorithm is equivalent to a block-pivoting simplex algorithm.

(2) Any simplex algorithm is equivalent to a particular policy iteration algorithm.

Example 3.2 (continued)

Start with the simplex tableau corresponding to the first feasible solution. Consider an iteration and let the basic solution corresponds to policy f^∞. Then, choose in each state i for which $y_i^f\big(g(i)\big) = min_a\, y_i^f(a) < 0$ as pivot column the column corresponding to $x_i^f\big(g(i)\big)$. In subsequent tableaus we execute the block-pivoting algorithm where in each iteration the pivot steps correspond to the nonbasic variables $x_i^f\big(g(i)\big)$ for which $y_i^f\big(g(i)\big) < 0,\ i \in S$. The pivots are again indicated by bold numbers.

Iteration 1

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	z_2	$x_2(3)$	z_3	$x_3(2)$	$x_3(3)$
$x_1(3)$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$-\frac{2}{3}$	0	$-\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$
$x_2(2)$	$\frac{2}{3}$	0	-1	0	2	2	**2**	0	-1	0
$x_3(1)$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{1}{3}$	0	$-\frac{2}{3}$	$\frac{4}{3}$	$\frac{4}{3}$	$\mathbf{\frac{2}{3}}$
	10	$\frac{11}{3}$	$\frac{10}{3}$	$\frac{28}{3}$	$-\frac{8}{3}$	8	$-\frac{10}{3}$	$\frac{38}{3}$	$-\frac{1}{3}$	$-\frac{2}{3}$

Iteration 2

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	z_2	$x_2(2)$	z_3	$x_3(2)$	$x_3(1)$
$x_1(3)$	$\frac{1}{3}$	$\frac{1}{2}$	1	1	$-\frac{1}{2}$	0	0	0	0	$-\frac{1}{2}$
$x_2(3)$	$\frac{1}{3}$	0	$-\frac{1}{2}$	0	1	1	$\frac{1}{2}$	0	$-\frac{1}{2}$	0
$x_3(3)$	$\frac{4}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	1	$\frac{1}{2}$	2	$\mathbf{\frac{3}{2}}$	$\frac{3}{2}$
	$\frac{108}{9}$	4	2	10	1	12	2	14	-1	1

Iteration 3

		$x_1(1)$	$x_1(2)$	z_1	$x_2(1)$	z_2	$x_2(2)$	z_3	$x_3(3)$	$x_3(1)$
$x_1(3)$	$\frac{1}{3}$	$\frac{1}{2}$	1	1	$-\frac{1}{2}$	0	0	0	0	$-\frac{1}{2}$
$x_2(3)$	$\frac{7}{9}$	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{7}{6}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{2}$
$x_3(2)$	$\frac{8}{9}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	1
	$\frac{116}{9}$	$\frac{13}{3}$	$\frac{7}{3}$	$\frac{32}{3}$	$\frac{4}{3}$	$\frac{38}{3}$	$\frac{7}{3}$	$\frac{46}{3}$	$\frac{2}{3}$	2

This tableau gives the value vector $v^\alpha = (\frac{32}{3}, \frac{38}{3}, \frac{46}{3})$ and the α-discounted optimal policy is f_*^∞ with $f_*(1) = 3, f_*(2) = 3$ and $f_*(3) = 2$.

Next, we discuss the *elimination of suboptimal actions*. Since the linear programming method is equivalent to the policy iteration method, we can copy the results of section 3.4, in particular Theorem 3.12. Instead of the numbers $s_{ia}(f)$, we use in linear programming the dual slack variables $y_i^f(a)$, where $y_i^f(a) = -s_{ia}(f),\ (i,a) \in S \times A$. The following theorem holds.

Theorem 3.20
If $y_i^f(a_i) > min_a\ y_i^f(a) - \alpha(1-\alpha)^{-1}\{min_i min_a\ y_i^f(a) - max_i min_a\ y_i^f(a)\}$, then action $a_i \in A(i)$ is suboptimal.

Example 3.2 (continued)

We consider the usual simplex method without block-pivoting and start with the first feasible tableau (iteration 4).

Iteration 4

$\alpha(1-\alpha)^{-1}\{min_i min_a\ y_i^f(a) - max_i min_a\ y_i^f(a)\} = -\frac{10}{3}$.

In state 1, action 1 is excluded, because $\frac{11}{3} = y_1^f(1) > min_a\ y_1^f(a) + \frac{10}{3} = \frac{10}{3}$.

Iteration 5

$\alpha(1-\alpha)^{-1}\{min_i min_a\ y_i^f(a) - max_i min_a\ y_i^f(a)\} = -2$.

In this iteration, no suboptimal actions are found.

A second method for eliminating suboptimal actions is based on a general LP property, due to Cheng [32], presented in the next theorem.

Theorem 3.21
Let B be a nondegenerate basis of the linear program $max\{p^T x \mid Ax = b;\ x \geq 0\}$, and x_j a nonbasic variable with reduced cost $z_j = p_B^T B^{-1} a_{\bullet j} - p_j > 0$, where $a_{\bullet j}$ is the column of A corresponding to x_j. Then, x_j will remain a nonbasic variable of an optimal basic solution if either one of the following conditions is satisfied:
(1) $B^{-1}a_{\bullet j} \geq 0$.
(2) $B^{-1}a_{\bullet j} \ngeq 0$ and $z_j - \theta\{p_B B^{-1} b - M\} > 0$, where $\theta = min_i \frac{\{B^{-1}a_{\bullet j}\}_i}{\{B^{-1}b\}_i}$ and M is an upper bound of the optimum.

In order to apply this property in the linear programming method for MDPs, we need an upper bound for the optimum of program (3.32). Such a bound is provided by the next lemma.

Lemma 3.5
$v^\alpha(f^\infty) - (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a) \cdot e$ is an upper bound of the value vector v^α.

Proof
Take any deterministic policy g^∞. Since $y_i^f(a) = \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j^\alpha(f^\infty) - r_i(a),\ (i,a) \in S \times A$, we obtain

$$min_{(i,a)}\ y_i^f(a) \cdot e \leq \{I - \alpha P(g)\} v^\alpha(f^\infty) - r(g).$$

Hence,

$$v^\alpha(f^\infty) \geq \{I - \alpha P(g)\}^{-1} r(g) + (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a) \cdot e = v^\alpha(g^\infty) + (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a) \cdot e.$$

Let g^∞ be an optimal policy. Then, $v^\alpha = v^\alpha(g^\infty) \leq v^\alpha(f^\infty) - (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a) \cdot e.$ □

Corollary 3.8
$\sum_j \beta_j \{v_j^\alpha(f^\infty) - (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a)\}$ is an upper bound of the optimum of program (3.32).

Proof
The optimal value of (3.32) is equal to the optimum of program (3.31) which is $\sum_j \beta_j v_j^\alpha$.
By Lemma 3.5, $\sum_j \beta_j v_j^\alpha \leq \sum_j \beta_j \{v_j^\alpha(f^\infty) - (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a)\}$. □

Theorem 3.22
Let a^* and b^* be the columns of the nonbasic variable $x_i^f(a)$ and the right-hand-side, respectively, in the simplex tableau corresponding to policy f^∞, and let $y_i^f(a) > 0$.
Then, action $a \in A(i)$ is suboptimal if either one of the following conditions is satisfied:
(1) $a^* \geq 0$.
(2) $a^* \not\geq 0$ and $y_i^f(a) - min_j \frac{a_j^*}{b_j^*} \cdot \left\{ \sum_j \beta_j \right\} \cdot (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a) > 0$.

Proof
We first remark that, by part (1) of Theorem 3.17, each basis of program (3.32) is nondegenerate.
(1) This result follows immediately from Theorem 3.21.
(2) Notice that the value $p_B B^{-1} b$ of the LP solution in the tableau of policy f^∞ is $\sum_j \beta_j v_j^\alpha(f^\infty)$. Hence, by the second condition in Theorem 3.21 and Corollary 3.8, action $a \in A(i)$ is suboptimal if

$$y_i^f(a) - min_j \frac{a_j^*}{b_j^*} \cdot \left\{ \sum_j \beta_j \right\} \cdot (1-\alpha)^{-1} \cdot min_{(i,a)}\ y_i^f(a).$$

□

Example 3.2 (continued)
Again, we start with the first feasible tableau (iteration 4).
Iteration 4
$(1-\alpha)^{-1} min_{(i,a)}\ y_i^f(a) = -\frac{20}{3}$.
In state 1, action 1 is excluded, because condition (1) of Theorem 3.22 is satisfied.
Iteration 5
$(1-\alpha)^{-1} min_{(i,a)}\ y_i^f(a) = -4$.
In state 2, action 2 is excluded, because condition (1) of Theorem 3.22 is satisfied.

3.6 Value iteration

In the method of *value iteration* the value vector v^α is successively approximated, starting with some guess v^1, by a sequence $\{v^n\}_{n=1}^\infty$, which converges to v^α. This method is also called *successive approximation*. In this method a *nearly optimal policy* is determined. When applying the policy iteration method (and also in principle in the linear programming method) one has to solve a system of N linear equations in each iteration. For a very large state space this might be prohibitive. The method of value iteration does not have this disadvantage. An iteration of this method is quite simple. In addition, sometimes this method can also be used to prove properties of the structure of optimal policies. On the other hand, especially for discount factors close to 1, the convergence can be very slow.

In this section we discuss the basic value iteration method including suboptimality tests. Most of the properties of the value iteration method are based on the theory of monotone contraction mappings and on the optimality equation (see the sections 3.2 and 3.3).

For $\delta > 0$ we call a vector $v \in \mathbb{R}^N$ a *δ-approximation* of v^α if $\|v^\alpha - v\|_\infty \leq \delta$; for $\varepsilon > 0$ a policy R is an *ε-optimal policy* if $\|v^\alpha - v^\alpha(R)\|_\infty \leq \varepsilon$.

From Corollary 3.3, part (2), it follows that $v^\alpha = lim_{n\to\infty} U^n x$ for every $x \in \mathbb{R}^N$.
Define the sequence $v^1, v^2, \ldots$ by

$$\begin{cases} v^1 \in \mathbb{R}^N & \text{arbitrarily chosen} \\ v^{n+1} = Uv^n & n = 1, 2, \ldots \end{cases} \tag{3.37}$$

with corresponding sequence $f_1^\infty, f_2^\infty, \ldots$ of policies, where $f_n = f_{v^n}$ for every $n \in \mathbb{N}$. Then, we have

$$v^{n+1} = Uv^n = L_{f_n} v^n = r(f_n) + \alpha P(f_n) v^n,\ n \in \mathbb{N}. \tag{3.38}$$

The next lemma shows that f_n^∞ is an ε-optimal policy for n sufficiently large.

Lemma 3.6
$\|v^\alpha(f_n^\infty) - v^\alpha\|_\infty \leq 2\alpha^n (1-\alpha)^{-1} \cdot \|v^2 - v^1\|_\infty,\ n \in \mathbb{N}.$

Proof

From Theorem 3.7, part (3), it follows that

$$\begin{aligned}\|v^\alpha(f_n^\infty) - v^\alpha\|_\infty &\leq 2\alpha(1-\alpha)^{-1} \cdot \|Uv^n - v^n\|_\infty = 2\alpha^n(1-\alpha)^{-1} \cdot \|Uv^n - Uv^{n-1}\|_\infty \\ &\leq 2\alpha^2(1-\alpha)^{-1} \cdot \|v^n - v^{n-1}\|_\infty \\ &\leq \cdots \leq 2\alpha^n(1-\alpha)^{-1} \cdot \|v^2 - v^1\|_\infty, \ n \in \mathbb{N}.\end{aligned}$$

□

Algorithm 3.4 *Value iteration (version 1)*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.
2. a. Compute $y_i = max_a\{r_i(a) + \alpha \sum_j p_{ij}(a)x_j\}, \ i \in S$.

   b. Let $f(i) = argmax_a\{r_i(a) + \alpha \sum_j p_{ij}(a)x_j\}, \ i \in S$.
3. If $\|y - x\|_\infty \leq \frac{1}{2}(1-\alpha)\alpha^{-1}\varepsilon$: f^∞ is an ε-optimal policy and y is a $\frac{1}{2}\varepsilon$-approximation of the value vector v^α (STOP);

   Otherwise: $x := y$ and return to step 2.

Theorem 3.23

Algorithm 3.4 is finite and correct.

Proof

Since the sequence $\{U^n x\}_{n=1}^\infty$ converges to v^α, the algorithm is finite. Suppose that the algorithm terminates with x, y and f, where $y = Ux$ and $f = f_x$. From the proof of Lemma 3.6 it follows that $\|v^\alpha(f^\infty) - v^\alpha\|_\infty \leq 2\alpha(1-\alpha)^{-1} \cdot \|y - x\|_\infty \leq \varepsilon$, i.e. f^∞ is an ε-optimal policy. Furthermore, $\|v^\alpha - y\|_\infty = \|Uv^\alpha - Ux\|_\infty \leq \alpha \cdot \|v^\alpha - x\|_\infty \leq \alpha(1-\alpha)^{-1} \cdot \|y - x\|_\infty \leq \frac{1}{2}\varepsilon$, the second last inequality by Theorem 3.7, part (2). □

Example 3.3

Consider the model of Example 3.1 and start with $x = (4, 4, 4)$ and $\varepsilon = 0.2$. The results of the computation are summarized below. The algorithm terminates as soon as the norm of the difference of two subsequent y-vectors is at most 0.1.

	Iteration						
	1	2	3	4	5	6	7
y_1	5.00	8.50	9.50	10.13	10.38	10.53	10.59
y_2	8.00	10.50	11.50	12.13	12.38	12.53	12.59
y_3	11.00	13.00	14.25	14.75	15.06	15.19	15.27
f_1	3	3	3	3	3	3	3
f_2	1	3	3	3	3	3	3
f_3	2	2	2	2	2	2	2

Hence, f^∞ with $f(1) = 3$, $f(2) = 3$ and $f(3) = 2$ is a 0.2-optimal policy and $(10.59, 12.59, 15.27)$ is a 0.1-approximation of v^α.

Remark

We see in the example that already after one iteration the optimal policy is found, although the approximation y is far away from v^α. This phenomenon occurs often when using the method of value iteration.

We now present an algorithm with a test for the *exclusion of suboptimal actions*, based on (3.15).

Algorithm 3.5 *Value iteration (version 2)*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.
2. a. Compute $y_i = max_a y_i(a),\ i \in S$, where $y_i(a) = r_i(a) + \alpha \sum_j p_{ij}(a)x_j,\ (i,a) \in S \times A$.

   b. Let $f(i) = argmax_a\{r_i(a) + \alpha \sum_j p_{ij}(a)x_j\},\ i \in S$.
3. If $\|y - x\|_\infty \leq \frac{1}{2}(1-\alpha)\alpha^{-1}\varepsilon$: f^∞ is an ε-optimal policy and y is a $\frac{1}{2}\varepsilon$-approximation of the value vector v^α (STOP).

   Otherwise: go to step 4.
4. a. Compute $span = max_i\,(y_i - x_i) - min_i\,(y_i - x_i)$.

   b. For all $(i,a) \in S \times A$: if $y_i(a) < y_i - \alpha(1-\alpha)^{-1} \cdot span$, then $A(i) := A(i) - \{a\}$.

   c. If $\#A(i) = 1$ for every $i \in S$, then f^∞ is an optimal policy (STOP).
5. $x := y$ and return to step 2.

Theorem 3.24

Algorithm 3.5 is finite and correct.

Proof

Let $A^{(n)}(i)$ be the action set in state i in iteration n. Define the operator $U^{(n)} : \mathbb{R}^N \to \mathbb{R}^N$ by

$$\{U^{(n)}\}_i = max_{a \in A^{(n)}(i)} \Big\{ r_i(a) + \alpha \sum_j p_{ij}(a)x_j \Big\},\ i \in S.$$

Algorithm 3.5 computes the sequence $v^1, v^2, \ldots$ where $v^{n+1} = U^{(n)}v^n$. Since the operator depends on n, we cannot simply use the general theory for contracting operators. We first show the finiteness of the algorithm. Let the actions $b, c \in A(i)$ be such that

$b \in argmax_{a \in A^{(n)}(i)}\{r_i(a) + \alpha \sum_j p_{ij}(a)v_j^n\}$ and $c \in argmax_{a \in A^{(n-1)}(i)}\{r_i(a) + \alpha \sum_j p_{ij}(a)v_j^{n-1}\}$.

Since $A^{(n)}(i) \subseteq A^{(n-1)}(i)$ and $b \in A^{(n-1)}(i)$, we can write

$$\begin{aligned} v_i^{n+1} - v_i^n &\leq \{r_i(b) + \alpha \textstyle\sum_j p_{ij}(b)v_j^n\} - \{r_i(b) + \alpha \textstyle\sum_j p_{ij}(b)v_j^{n-1}\} \\ &= \alpha \textstyle\sum_j p_{ij}(b)\{v_j^n - v_j^{n-1}\} \leq \alpha \textstyle\sum_j p_{ij}(b) \cdot \|v^n - v^{n-1}\|_\infty = \alpha \cdot \|v^n - v^{n-1}\|_\infty. \end{aligned}$$

On the other hand, because $v_i^n = r_i(c) + \alpha \sum_j p_{ij}(c)v_j^{n-1}$, i.e. in the algorithm we have $y_i(c) = y_i$, action c is not excluded in step 4b of the algorithm. Hence, $c \in A^{(n)}(i)$ and we obtain

$$\begin{aligned} v_i^n - v_i^{n+1} &\leq \{r_i(c) + \alpha \textstyle\sum_j p_{ij}(c) v_j^{n-1}\} - \{r_i(c) + \alpha \textstyle\sum_j p_{ij}(c) v_j^n\} \\ &= \alpha \textstyle\sum_j p_{ij}(c)\{v_j^{n-1} - v_j^n\} \leq \alpha \textstyle\sum_j p_{ij}(c) \cdot \|v^n - v^{n-1}\|_\infty = \alpha \cdot \|v^n - v^{n-1}\|_\infty. \end{aligned}$$

Now, it follows that $\|v^{n+1} - v^n\|_\infty \leq \alpha \cdot \|v^n - v^{n-1}\|_\infty \leq \cdots \leq \alpha^{n-1} \cdot \|v^2 - v^1\|_\infty$, i.e. the algorithm is finite.

Next, we show by induction on n that the suboptimality test is correct. The first iteration is correct. Suppose that the elimination is correct during the iterations $1, 2, \ldots, n-1$ and consider iteration n. Above, it was shown that $U^{(n)}$ is a contraction with contraction factor α. Since no optimal actions are excluded, v^α is the fixed-point of $U^{(n)}$. Hence, by taking $U^{(n)}$ and $A^{(n)}(i)$ instead of U and $A(i)$, it follows from the general theory derived in Section 3.3 that the suboptimality test is correct.

Finally, we show that the algorithm terminates with an ε-optimal policy f^∞ and an ε-approximation of v^α. Let m be the last iteration of the algorithm. If $\#A(i) = 1$ for every $i \in S$, then f^∞ is optimal. Otherwise, let f_x be such that

$$y = U^{(m)}x = L_{f_x}x.$$

Since v^α and $v^\alpha(f^\infty)$ are the fixed-points of $U^{(m)}$ and L_{f_x}, it follows (see Theorem 3.7) that

$$\|v^\alpha - v^\alpha(f_x^\infty)\|_\infty \leq 2\alpha(1-\alpha)^{-1} \cdot \|U^{(m)}x - x\|_\infty = 2\alpha(1-\alpha)^{-1} \cdot \|y - x\|_\infty \leq \varepsilon$$

and

$$\|v^\alpha - y\|_\infty = \|U^{(m)}v^\alpha - U^{(m)}x\|_\infty \leq \alpha\|v^\alpha - x\|_\infty \leq \alpha(1-\alpha)^{-1} \cdot \|y - x\|_\infty \leq \tfrac{1}{2}\varepsilon. \qquad \square$$

Remarks

1. If the algorithm terminates in step 4c, then an optimal policy f^∞ is obtained, but the value vector v^α is unknown. Also it is unknown how good the approximation y is. In order to compute the exact value of v^α we have to solve the linear system $x = L_f x$.
2. It is not necessary to execute step 4 in each iteration; it can be done, for instance, periodically.

Example 3.3 (continued)

Iteration 1

$y_1(1) = 3, y_1(2) = 4, y_1(3) = 5 : y_1 = 5.$ $y_2(1) = 8, y_2(2) = 6, y_2(3) = 7 : y_2 = 8.$
$y_3(1) = 10, y_3(2) = 11, y_3(3) = 9 : y_3 = 11.$ $f(1) = 3, f(2) = 1, f(3) = 2.$ $span = 6.$
No actions can be excluded.

Iteration 2

$y_1(1) = 3.5, y_1(2) = 6, y_1(3) = 8.5 : y_1 = 8.5.$
$y_2(1) = 7.5, y_2(2) = 8, y_2(3) = 10.5 : y_2 = 10.5.$
$y_3(1) = 10.5, y_3(2) = 13, y_3(3) = 12.5 : y_3 = 13.$ $f(1) = 3, f(2) = 3, f(3) = 2.$ $span = 1.5.$
In state 1 the actions 1 and 2 are excluded; in state 2 the actions 1 and 2 and in state 3 action 1.

Iteration 3
$y_1(3) = 9.5 : y_1 = 9.5$. $y_2(3) = 11.5 : y_2 = 11.5$. $y_3(2) = 14.25, y_3(3) = 13.5 : y_3 = 14.25$.
$f(1) = 3, f(2) = 3, f(3) = 2$. $span = 0.25$.
In state 3 actions 3 is excluded. Hence, f^∞ with $f(1) = 3$, $f(2) = 3$ and $f(3) = 2$ is an optimal policy.

The method of value iteration is an iterative procedure to solve the functional equation $Ux = x$. In this section we discuss two variants of the standard procedure, the *Pre-Gauss-Seidel* and the *Gauss-Seidel* variant, respectively. These variants are based on contraction mappings with fixed-point v^α and with contraction factor at most α. Hence, they may be considered as accelerations of the basic algorithm.

Variant 1 (Pre-Gauss-Seidel)
In the basic method (3.37), v^{n+1} is computed from v^n by the formula

$$v_i^{n+1} = max_a \Big\{ r_i(a) + \alpha \sum_{j=1}^{N} p_{ij}(a) v_j^n \Big\}, \ i = 1, 2, \ldots, N. \tag{3.39}$$

Since, in general, v^{n+1} is a better approximation of v^α than v^n, it seems favourable to use the values $v_1^{n+1}, v_2^{n+1}, \ldots, v_{i-1}^{n+1}$ in the computation of v_i^{n+1} instead of $v_1^n, v_2^n, \ldots, v_{i-1}^n$. So, the following formula is used:

$$v_i^{n+1} = max_a \Big\{ r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a) v_j^{n+1} + \alpha \sum_{j=i}^{N} p_{ij}(a) v_j^n \Big\}, \ i = 1, 2, \ldots, N. \tag{3.40}$$

This is the so-called *Pre-Gauss-Seidel* variant. Similar to the mappings L_π and U for the standard procedure, the Pre-Gauss-Seidel variant can be described by the operators $\overline{L}_\pi$ and $\overline{U}$, which are mappings from $\mathbb{R}^N$ to $\mathbb{R}^N$, defined by

$$\{\overline{L}_\pi x\}_i = r_i(\pi) + \alpha \sum_{j=1}^{i-1} p_{ij}(\pi) \{\overline{L}_\pi x\}_j + \alpha \sum_{j=i}^{N} p_{ij}(\pi) x_j, \ i = 1, 2, \ldots, N, \tag{3.41}$$

and

$$\{\overline{U} x\}_i = max_a \Big\{ r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a) \{\overline{U} x\}_j + \alpha \sum_{j=i}^{N} p_{ij}(a) x_j \Big\}, \ i = 1, 2, \ldots, N. \tag{3.42}$$

For every $x \in \mathbb{R}^N$ the policy $\overline{f}_x^\infty$ is the policy that satisfies $\overline{L}_{\overline{f}_x} x = \overline{U} x$.

Theorem 3.25
The operators $\overline{L}_\pi$ and $\overline{U}$ are monotone contracting mappings with fixed-points $v^\alpha(\pi^\infty)$ and v^α, respectively, with contraction factor α.

Proof

We apply Lemma 3.1, part (2). Therefore, suppose that $x \leq y \leq d \cdot e$ for some scalar d. With induction on state i we will show that $\{\overline{L}_\pi x\}_i \leq \{\overline{L}_\pi y\}_i + \alpha \cdot |d|,\ i = 1, 2, \ldots, N$. For $i = 1$, we have

$$\begin{aligned}\{\overline{L}_\pi x\}_1 &= \{L_\pi x\}_1 = r_1(\pi) + \alpha \textstyle\sum_{j=1}^N\ p_{1j}(\pi)x_j \\ &\leq r_1(\pi) + \alpha \textstyle\sum_{j=1}^N\ p_{1j}(\pi)y_j + \alpha \cdot |d| \sum_{j=1}^N\ p_{1j}(\pi) = \{\overline{L}_\pi y\}_1 + \alpha \cdot |d|.\end{aligned}$$

Suppose that $\{\overline{L}_\pi x\}_j \leq \{\overline{L}_\pi y\}_j + \alpha \cdot |d|$ for $j = 1, 2, \ldots, i-1$. Then, we can write

$$\begin{aligned}\{\overline{L}_\pi x\}_i &= r_i(\pi) + \alpha \textstyle\sum_{j=1}^{i-1}\ p_{ij}(\pi)\{\overline{L}_\pi x\}_j + \alpha \sum_{j=i}^N\ p_{ij}(\pi)x_j \\ &\leq r_i(\pi) + \alpha \textstyle\sum_{j=1}^{i-1}\ p_{ij}(\pi)\{\overline{L}_\pi y\}_j + \alpha^2 \cdot |d| \sum_{j=1}^{i-1}\ p_{ij}(\pi) \\ &\qquad + \alpha \textstyle\sum_{j=i}^N\ p_{ij}(\pi)y_j + \alpha \cdot |d| \sum_{j=i}^N\ p_{ij}(\pi) \\ &= \{\overline{L}_\pi y\}_i + \alpha^2 \cdot |d| \textstyle\sum_{j=1}^{i-1}\ p_{ij}(\pi) + \alpha \cdot |d| \sum_{j=i}^N\ p_{ij}(\pi) \\ &\leq \{\overline{L}_\pi y\}_i + \alpha \cdot |d| \textstyle\sum_{j=1}^N\ p_{ij}(\pi) = \{\overline{L}_\pi y\}_i + \alpha \cdot |d|.\end{aligned}$$

Hence, by Lemma 3.1, part (2), $\overline{L}_\pi$ is a monotone contraction with contraction factor α. Again by induction on state i, one can easily show that v_π^α satisfies (3.41), i.e. v_π^α is the unique fixed-point of $\overline{L}_\pi$. The proof for $\overline{U}$ is similar, and is left to the reader. □

Lemma 3.7

(1) $\overline{U}x = sup_\pi\ \overline{L}_\pi x$ *for every* $x \in \mathbb{R}^N$.

(2) $\overline{f}_{v^\alpha}^\infty$ *is an* α*-discounted optimal policy.*

(3) $x - (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e \leq \overline{U}x - \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e \leq v^\alpha(\overline{f}_x^\infty) \leq v^\alpha \leq \overline{U}x + \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e \leq x + (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e.$

(4) $\|v^\alpha - x\| \leq (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty.$

(5) $\|v^\alpha(\overline{f}_x^\infty) - v^\alpha\| \leq 2\alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty.$

Proof

(1) By induction on i, it can be shown that $(\overline{L}_\pi x)_i \leq (\overline{U}x)_i$ for $i = 1, 2, \ldots, N$.

For $i = 1$ the result is obvious and the induction step is

$$\begin{aligned}\{\overline{L}_\pi x\}_i &= r_i(\pi) + \alpha \textstyle\sum_{j=1}^{i-1}\ p_{ij}(\pi)(\overline{L}_\pi x\}_j + \alpha \sum_{j=i}^N\ p_{ij}(\pi)x_j \\ &\leq r_i(\pi) + \alpha \textstyle\sum_{j=1}^{i-1}\ p_{ij}(\pi)\{\overline{U}x\}_j + \alpha \sum_{j=i}^N\ p_{ij}(\pi)x_j \\ &\leq max_a\Big\{r_i(a) + \alpha \textstyle\sum_{j=1}^{i-1}\ p_{ij}(a)\{\overline{U}x\}_j + \alpha \sum_{j=i}^N\ p_{ij}(a)x_j\Big\} = \{\overline{U}x\}_i.\end{aligned}$$

Because $\overline{L}_{\overline{f}_x} = \overline{U}x$, it follows that $\overline{U}x = sup_\pi\ \overline{L}_\pi x$.

(2) Because $\overline{L}_{\overline{f}_{v^\alpha}} v^\alpha = \overline{U}v^\alpha = v^\alpha,\ v^\alpha$ is the fixed-point of $\overline{L}_{\overline{f}_{v^\alpha}}$, i.e. $v^\alpha = v^\alpha(\overline{f}_{v^\alpha}^\infty) : \overline{f}_{v^\alpha}^\infty$ is an α-discounted optimal policy.

The parts (3), (4) and (5) can be shown in a way analogous to the proof of Theorem 3.7. □

Lemma 3.8

(1) $\overline{U}(x + c \cdot e) \leq \overline{U}x + \alpha c \cdot e$ *for every* $x \in \mathbb{R}^N$ *and every* $c \geq 0$.

(2) $\overline{U}(x + c \cdot e) \geq \overline{U}x + \alpha c \cdot e$ *for every* $x \in \mathbb{R}^N$ *and every* $c < 0$.

Proof

Using induction on the state i, the proof is straightforward. □

Theorem 3.26

If $r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)\{\overline{U}x\}_j + \alpha \sum_{j=i}^{N} p_{ij}(a)x_j < \{\overline{U}x\}_i - 2\alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty$,
then action a *is suboptimal.*

Proof

Using Lemma 3.7 part (3), it follows that

$$\begin{aligned} x - (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e &\leq \overline{U}x - \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e \leq v^\alpha \leq \\ &\leq \overline{U}x + \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e \leq x + (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot e. \end{aligned}$$

Therefore, we can write

$$\begin{aligned} v_i^\alpha &= max_a\{r_i(a) + \alpha \sum_j p_{ij}(a)v_j^\alpha\} \\ &\geq max_a\Big\{r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)\big\{\{\overline{U}x\}_j - \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty\big\} \\ &\qquad + \alpha \sum_{j=i}^{N} p_{ij}(a)\{x_j - (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty\}\Big\}. \end{aligned}$$

Since

$$\begin{aligned} &\alpha \sum_{j=1}^{i-1} p_{ij}(a)\alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty + \alpha \sum_{j=i}^{N} p_{ij}(a)(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty] \leq \\ &\alpha \sum_{j=1}^{i-1} p_{ij}(a)(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty + \alpha \sum_{j=i}^{N} p_{ij}(a)(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty = \\ &\alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \cdot \sum_{j=1}^{N} p_{ij}(a) = \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty, \end{aligned}$$

we obtain

$$\begin{aligned} v_i^\alpha &\geq max_a\Big\{r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)\{\overline{U}x\}_j + \alpha \sum_{j=i}^{N} p_{ij}(a)x_j - \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty\Big\} \\ &= (\overline{U}x)_i - \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \\ &> r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)\{\overline{U}x\}_j + \alpha \sum_{j=i}^{N} p_{ij}(a)x_j + \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty \\ &= r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)\big\{\{\overline{U}x\}_j + (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty\big\} \\ &\qquad + \alpha \sum_{j=i}^{N} p_{ij}(a)\{x_j + (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty\} \\ &\geq r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)\big\{\overline{U}x\big\}_j + \alpha(1-\alpha)^{-1}\|\overline{U}x - x\|_\infty\big\} \\ &\qquad + \alpha \sum_{j=i}^{N} p_{ij}(a)\{x_j + (1-\alpha)^{-1}\|\overline{U}x - x\|_\infty\} \\ &= r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)v_j^\alpha + \alpha \sum_{j=i}^{N} p_{ij}(a)v_j^\alpha = r_i(a) + \alpha \sum_{j=1}^{N} p_{ij}(a)v_j^\alpha. \end{aligned}$$ □

From the previous results it follows that the following algorithm computes an ε-optimal policy whithin a finite number of iterations.

Algorithm 3.6 *Value iteration (Pre-Gauss-Seidel)*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.
2. For $i = 1, 2, \ldots, N$ do:
   a. Compute $y_i(a) = r_i(a) + \alpha\{\sum_{j=1}^{i-1} p_{ij}(a)y_j + \sum_{j=i}^{N} p_{ij}(a)x_j\},\ a \in A(i)$.
   b. Compute y_i and $f(i)$ such that $y_i = max_a\, y_i(a) = y_i\big(f(i)\big)$.
3. If $\|y - x\|_\infty \leq \frac{1}{2}(1-\alpha)\alpha^{-1}\varepsilon$: f^∞ is an ε-optimal policy and y is a $\frac{1}{2}\varepsilon$-approximation of the value vector v^α (STOP);.
4. a. For all $(i, a) \in S \times A$: if $y_i(a) < y_i - 2\alpha(1-\alpha)^{-1}\|y - x\|_\infty$, then $A(i) := A(i) - \{a\}$.
   b. If $\#A(i) = 1$ for every $i \in S$, then f^∞ is an optimal policy (STOP).
5. $x := y$ and return to step 2.

Example 3.3 (continued)

Start with $x = (4, 4, 4)$. The computations can be represented by the following scheme:

$y_1(1) = 1 + \frac{1}{2}x_1;\ y_1(2) = 2 + \frac{1}{2}x_2;\ y_1(3) = 3 + \frac{1}{2}x_3;\ y_1 = max\{y_1(1), y_1(2), y_1(3)\}$.

$y_2(1) = 6 + \frac{1}{2}y_1;\ y_2(2) = 4 + \frac{1}{2}x_2;\ y_2(3) = 5 + \frac{1}{2}x_3;\ y_2 = max\{y_2(1), y_2(2), y_2(3)\}$.

$y_3(1) = 8 + \frac{1}{2}y_1;\ y_3(2) = 9 + \frac{1}{2}y_2;\ y_3(3) = 7 + \frac{1}{2}x_3;\ y_3 = max\{y_3(1), y_3(2), y_3(3)\}$.

Iteration 1

$y_1(1) = 3, y_1(2) = 4, y_1(3) = 5 : y_1 = 5;\ f(1) = 3$.

$y_2(1) = 8.5, y_2(2) = 6, y_2(3) = 7 : y_2 = 8.5;\ f(2) = 1$.

$y_3(1) = 10.5, y_3(2) = 13.25, y_3(3) = 9 : y_3 = 13.25;\ f(3) = 2$.

$x = (5, 8.5, 13.25)$.

Iteration 2

$y_1(1) = 3, y_1(2) = 6.25, y_1(3) = 9.61 : y_1 = 9.61;\ f(1) = 3$.

$y_2(1) = 10.81, y_2(2) = 8.25, y_2(3) = 11.61 : y_2 = 11.61;\ f(2) = 3$.

$y_3(1) = 12.81, y_3(2) = 14.81, y_3(3) = 13.61 : y_3 = 14.81;\ f(3) = 2$.

$x = (9.61, 11.61, 14.81)$.

Iteration 3

$y_1(1) = 5.81, y_1(2) = 7.81, y_1(3) = 10.41 : y_1 = 10.41;\ f(1) = 3$.

$y_2(1) = 11.20, y_2(2) = 9.81, y_2(3) = 12.41 : y_2 = 12.41;\ f(2) = 3$.

$y_3(1) = 13.20, y_3(2) = 15.20, y_3(3) = 14.41 : y_3 = 15.20;\ f(3) = 2$.

$i = 1$: the actions 1 and 2 are excluded.

$i = 2$: action 2 is excluded.

$i = 3$: the action 1 is excluded.

$x = (10.41, 12.41, 15.20)$.

Iteration 4

$y_1(1) = 10.60 : y_1 = 10.60;\ f(1) = 3.$

$y_2(1) = 11.30, y_2(3) = 12.60 : y_2 = 12.60;\ f(2) = 3.$

$y_3(2) = 15.30, y_3(3) = 14.60 : y_3 = 15.30;\ f(3) = 2.$

$i = 2$: action 1 is excluded; $i = 3$: action 3 is excluded.

f^∞ with $f(1) = 3,\ f(2) = 3$ and $f(3) = 2$ is an optimal policy.

Remarks

1. The convergence to the value vector is faster in the Pre-Gauss-Seidel variant than in the standard version. On the other side, the exclusion of suboptimal actions is, in general, not so successful.
2. The performance of the Pre-Gauss-Seidel variant depends on the ordering of the states. Therefore, it is worthwhile to apply the following scheme, in which the states are ordered in the usual way first and then reversed:

   $y_i = max_a\{r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)y_j + \alpha \sum_{j=i}^{N} p_{ij}(a)x_j\},\ i = 1, 2, \ldots, N.$

   $z_i = max_a\{r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)y_j + \alpha \sum_{j=i}^{N} p_{ij}(a)z_j\},\ i = N, N-1, \ldots, 1.$

Variant 2 (Gauss-Seidel)

The idea of the Pre-Gauss-Seidel variant can be extended to the term with $j = i$. Then, formula (3.41) becomes (with L^* instead of $\overline{L}$):

$$(L^*_\pi x)_i = r_i(\pi) + \alpha \sum_{j=1}^{i} p_{ij}(\pi)(L^*_\pi x)_j + \alpha \sum_{j=i+1}^{N} p_{ij}(\pi)x_j\},\ i = 1, 2, \ldots, N,$$

i.e.

$$(L^*_\pi x)_i = \{1-\alpha p_{ii}(\pi)\}^{-1}\Big\{r_i(\pi)+\alpha \sum_{j=1}^{i-1} p_{ij}(\pi)(L^*_\pi x)_j+\alpha \sum_{j=i+1}^{N} p_{ij}(\pi)x_j\Big\},\ i = 1, 2, \ldots, N. \quad (3.43)$$

The corresponding operator U^* and the maximizing decision rule f^*_x are defined by

$$(U^*x)_i = max_a \{1 - \alpha p_{ii}(a)\}^{-1}\Big\{r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)(U^*x)_j + \alpha \sum_{j=i+1}^{N} p_{ij}(a)x_j\Big\},\ i = 1, 2, \ldots, N. \quad (3.44)$$

and $L^*_{f^*_x} x = U^*x$.

Theorem 3.27

(1) The operator L^*_π is a monotone contraction with fixed-point $v^\alpha(\pi^\infty)$ and with contraction factor $\beta_\pi = \alpha \cdot max_i \frac{1-p_{ii}(\pi)}{1-\alpha p_{ii}(\pi)} \le \alpha$.

(2) The operator U^* is a monotone contraction with fixed-point v^α and with contraction factor $\beta = \alpha \cdot max_{i,a} \frac{1-p_{ii}(a)}{1-\alpha p_{ii}(a)}$.

Proof

The proof is similar to the proof of Theorem 3.25 and is left to the reader (Exercise 3.22).

Lemma 3.9

(1) $U^*x = sup_\pi\ L^*_\pi x$ *for every* $x \in \mathbb{R}^N$.

(2) $f^{*\infty}_{v^\alpha}$ *is an* α*-discounted optimal policy.*

(3) $x - (1-\beta)^{-1}\|U^*x - x\|_\infty \cdot e \le U^*x - \beta(1-\beta)^{-1}\|U^*x - x\|_\infty \cdot e \le v^\alpha(f^{*\infty}_x) \le v^\alpha \le U^*x + \beta(1-\beta)^{-1}\|U^*x - x\|_\infty \cdot e \le x + (1-\beta)^{-1}\|U^*x - x\|_\infty \cdot e.$

(4) $\|v^\alpha - x\| \le (1-\beta)^{-1}\|U^*x - x\|_\infty$.

(5) $\|v^\alpha(f^{*\infty}_x) - v^\alpha\| \le 2\beta(1-\beta)^{-1}\|U^*x - x\|_\infty$.

Proof

The proof is similar to the proof of Lemma 3.7. □

Lemma 3.10

(1) $U^*(x + c\cdot e) \le U^*x + \beta c \cdot e$ *for every* $x \in \mathbb{R}^N$ *and every* $c \ge 0$;

(2) $U^*(x + c\cdot e) \ge U^*x + \beta c \cdot e$ *for every* $x \in \mathbb{R}^N$ *and every* $c < 0$.

Proof

The proof is similar to the proof of Lemma 3.8 □

Theorem 3.28

If $\{1 - \alpha p_{ii}(a)\}^{-1}\Big\{r_i(a) + \alpha\sum_{j=1}^{i-1} p_{ij}(a)(U^*x)_j + \alpha\sum_{j=i+1}^{N} p_{ij}(a)x_j\Big\} < (U^*x)_i - 2\beta(1-\beta)^{-1}\cdot\|U^*x - x\|_\infty$, *then action* a *is suboptimal.*

Proof

The proof is similar to the proof of Theorem 3.26 and left to the reader (Exercise 3.23). □

Algorithm 3.7 *Value iteration (Gauss-Seidel)*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary; $\beta = \alpha \cdot max_{i,a}\ \frac{1-p_{ii}(a)}{1-\alpha p_{ii}(a)}$.

2. For $i = 1, 2, \ldots, N$ do:

   a. Compute $y_i(a) = \{1-\alpha p_{ii}(a)\}^{-1}\Big\{r_i(a)+\alpha\sum_{j=1}^{i-1} p_{ij}(a)y_j+\alpha\sum_{j=i+1}^{N} p_{ij}(a)x_j\Big\}$, $a \in A(i)$.

   b. Compute y_i and $f(i)$ such that $y_i = argmax_a\, y_i(a) = y_i\big(f(i)\big)$.

3. If $\|y - x\|_\infty \le \frac{1}{2}(1-\beta)\beta^{-1}\varepsilon$: f^∞ is an ε-optimal policy and y is a $\frac{1}{2}\varepsilon$-approximation of the value vector v^α (STOP);.

4. a. For all $(i,a) \in S \times A$: if $y_i(a) < y_i - 2\beta(1-\beta)^{-1}\|y - x\|_\infty$, then $A(i) := A(i) - \{a\}$.

   b. If $\#A(i) = 1$ for every $i \in S$, then f^∞ is an optimal policy (STOP).

5. $x := y$ and return to step 2.

Example 3.3 (continued)

Start with $x = (4, 4, 4)$; $\beta = 0.5$. The computations are given by the following scheme:

$y_1(1) = 2;\ y_1(2) = 2 + \frac{1}{2}x_2;\ y_1(3) = 3 + \frac{1}{2}x_3;\ y_1 = max\{y_1(1), y_1(2), y_1(3)\};$

$y_2(1) = 6 + \frac{1}{2}y_1;\ y_2(2) = 8;\ y_2(3) = 5 + \frac{1}{2}x_3;\ y_2 = max\{y_2(1), y_2(2), y_2(3)\};$

$y_3(1) = 8 + \frac{1}{2}y_1;\ y_3(2) = 9 + \frac{1}{2}y_2;\ y_3(3) = 14;\ y_3 = max\{y_3(1), y_3(2), y_3(3)\};$

Iteration 1

$y_1(1) = 2, y_1(2) = 4, y_1(3) = 5 : y_1 = 5;\ f(1) = 3.$

$y_2(1) = 8.5, y_2(2) = 8, y_2(3) = 7 : y_2 = 8.5;\ f(2) = 1.$

$y_3(1) = 10.5, y_3(2) = 13.25, y_3(3) = 14 : y_3 = 14;\ f(3) = 3.$

$x = (5, 8.5, 14).$

Iteration 2

$y_1(1) = 2, y_1(2) = 6.25, y_1(3) = 10 : y_1 = 10;\ f(1) = 3.$

$y_2(1) = 11, y_2(2) = 8, y_2(3) = 12 : y_2 = 12;\ f(2) = 3.$

$y_3(1) = 13, y_3(2) = 15, y_3(3) = 14 : y_3 = 15;\ f(3) = 2.$

$x = (10, 12, 15).$

Iteration 3

$y_1(1) = 2, y_1(2) = 7, y_1(3) = 10.5 : y_1 = 10.5;\ f(1) = 3.$

$y_2(1) = 11.25, y_2(2) = 8, y_2(3) = 12.5 : y_2 = 12.5;\ f(2) = 3.$

$y_3(1) = 13.25, y_3(2) = 15.25, y_3(3) = 14 : y_3 = 15.25;\ f(3) = 2.$

$i = 1$: the actions 1 and 2 are excluded.

$i = 2$: the actions 1 and 2 are excluded.

$i = 3$: the actions 1 and 3 are excluded.

$x = (10.41, 12.41, 15.20).$

f^∞ with $f(1) = 3,\ f(2) = 3$ and $f(3) = 2$ is an optimal policy.

3.7 Modified Policy Iteration

In step 2 of the policy iteration method (see Algorithm 3.1) we determine $v^\alpha(f\infty)$ as unique solution of the linear system $L_f\, x = x$, i.e.

$$\{I - \alpha P(f)\}x = x. \tag{3.45}$$

In a model with N states, this requires $\mathcal{O}(N^3)$ elementary operations (e.g. multiplications). Hence, for large N, obtaining an exact solution of (3.45) may be computationally prohibitive. In section 3.4 we have shown (see (3.25)) that, for consecutive policies f^∞ and g^∞ in Algorithm 3.1, and with $x = v^\alpha(f^\infty)$ and $y = v^\alpha(g^\infty)$,

$$y = x + \{I - \alpha P(f)\}^{-1}\{Ux - x\} = x + \sum_{i=0}^{\infty} \{\alpha P(f)\}^i \{Ux - x\}. \tag{3.46}$$

In the *modified policy iteration* method the matrix

$$A := \sum_{i=0}^{\infty} \{\alpha P(f)\}^i$$

is truncated by

$$A^{(k)} = \sum_{i=0}^{k-1} \{\alpha P(f)\}^i \text{ for some } 1 \leq k \leq \infty.$$

For $k = 1$, $A^{(1)} = I$, and formula (3.46) becomes $y = x + (Ux - x) = Ux$, i.e. the modified policy iteration method is value iteration; for $k = \infty$, $A^{(\infty)} = A$, and formula (3.46) is the policy iteration method. For $1 < k < \infty$, the modified policy iteration method may be considered as a combination of policy ieration and value iteration. Policy iteration may be viewed as Newton's method for the solution of the optimality equation $Ux = x$. Similarly, the modified policy iteration method can be considered as an inexact Newton method.

We allow different values of k to be chosen in each ieration and we denote by $k(n)$ the value of k in iteration n. Hence, we obtain

$$\begin{aligned} x^{n+1} &= x^n + A^{(k(n))}\{Ux^n - x^n\} \\ &= x^n + \textstyle\sum_{i=0}^{k(n)-1} \{\alpha P(f_n)\}^i \{r(f_n) + \alpha P(f_n)x^n - x^n\} \\ &= r(f_n) + \alpha P(f_n) r(f_n) + \cdots + \left\{\alpha P(f_n)\right\}^{k(n)-1} r(f_n) + \left\{\alpha P(f_n)\right\}^{k(n)} x^n \\ &= \left\{L_{f_n}\right\}^{k(n)} x^n. \end{aligned}$$

The modified policy iteration method is presented in the following algorithm.

Algorithm 3.8 *Modified policy iteration*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.
2. a. Choose any k with $1 \leq k \leq \infty$.
   b. Determine f such that $L_f\, x = Ux$.
   c. If $\|Ux - x\|_\infty \leq \frac{1}{2}(1-\alpha)\alpha^{-1}\varepsilon$: f^∞ is an ε-optimal policy (STOP).
3. a. $y := \{L_f\}^k\, x$.
   b. $x := y$ and return to step 2.

Example 3.4

Consider the model of Example 3.1, start with $x = (\frac{28}{3}, 8, \frac{28}{3})$, let $\varepsilon = 0.2$ (implying that $\frac{1}{2}(1-\alpha)\alpha^{-1}\varepsilon = 0.1$) and take $k = 2$ in each iteration.

Iteration 1

$Ux = (\frac{28}{3}, \frac{34}{3}, \frac{40}{3})$; $f(1) = f(2) = f(3) = 3$.

$y = (\frac{29}{3}, \frac{35}{3}, \frac{41}{3})$; $x = (\frac{29}{3}, \frac{35}{3}, \frac{41}{3})$.

Iteration 2

$Ux = (9.833, 11.833, 14.833);\ f(1) = f(2) = 3,\ f(3) = 2.$

$y = (10.417, 12.417, 14.917);\ x = (10.417, 12.417, 14.917).$

Iteration 3

$Ux = (10.459, 12.459, 15.209);\ f(1) = f(2) = 3,\ f(3) = 2.$

$y = (10.604, 12.604, 15.229);\ x = (10.604, 12.604, 15.229).$

Iteration 4

$Ux = (10.615, 12.615, 15.302);\ f(1) = f(2) = 3,\ f(3) = 2.$

$y = (10.651, 12.651, 15.308);\ x = (10.651, 12.651, 15.308).$

Iteration 5

$Ux = (10.654, 12.654, 15.309);\ f(1) = f(2) = 3,\ f(3) = 2.$

f^∞ is an ε-optimal policy.

Let $x^1, x^2, \ldots$ be subsequent approximations of v^α, obtained by Algorithm 3.8. Then,

$$x^{n+1} = \{L_{f_n}\}^{k(n)}\, x^n,\ n = 1, 2, \ldots. \tag{3.47}$$

Since the operator depends on n, it is not obvious from the general theory that this operator is monotone and/or contracting. The next example shows that, in general, the operator $\{L_{f_n}\}^{k(n)}$ is neither a contraction nor is it monotone.

Example 3.5

Let $S = \{1, 2\};\ A(1) = A(2) = \{1, 2\};\ \alpha = \frac{3}{4}.\ r_1(1) = 1,\ r_1(2) = 0,\ r_2(1) = 1,\ r_2(2) = 0.$

$p_{11}(1) = 0,\ p_{12}(1) = 1;\ p_{11}(2) = 1,\ p_{12}(2) = 0;\ p_{21}(1) = 0,\ p_{22}(1) = 1;\ p_{21}(2) = 1,\ p_{22}(2) = 1.$

In an iteration, x will be transformed to $\{L_{f_x}\}^k$ for some k, where f_x satisfies $Ux = L_{f_x}\, x$.

Let $x = (3, 0)$, then $(Ux)_1 = max\{1 + \alpha \cdot 0, 0 + \alpha \cdot 3\} = \frac{9}{4},\ (Ux)_2 = max\{1 + \alpha \cdot 0, 0 + \alpha \cdot 3\} = \frac{9}{4}$.

Hence, $f_x(1) = f_x(2) = 2$, and consequently, $r(f_x) = \binom{0}{0}$ and $P(f_x) = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, so $\{P(f_x)\}^i = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ for all $i \geq 1$. Therefore, $\{L_{f_x}\}^k\, x = \left(\frac{3}{4}\right)^k \{P(f_x)\}^k x = \left(\frac{3}{4}\right)^k \binom{3}{3}$.

Next, let $y = (0, 0)$, then $(Uy)_1 = max\{1 + \alpha \cdot 0, 0 + \alpha \cdot 0\} = 1,\ (Uy)_2 = max\{1 + \alpha \cdot 0, 0 + \alpha \cdot 0\} = 1$.

Hence, $f_y(1) = f_y(2) = 1$, and consequently, $r(f_y) = \binom{1}{1}$ and $P(f_y) = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, so $\{P(f_x)\}^i = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ for all $i \geq 1$. Therefore, $\{L_{f_y}\}^k\, y = \binom{1}{1} + \alpha\binom{1}{1} + \cdots + \alpha^{k-1}\binom{1}{1} + \alpha^k\binom{0}{0} = \frac{1-\alpha^k}{1-\alpha}\binom{1}{1} = \left\{1 - \left(\frac{3}{4}\right)^k\right\}\binom{4}{4}$.

Notice that $x = \binom{3}{0} \geq \binom{0}{0} = y$, but not $\left(\frac{3}{4}\right)^k \binom{3}{3} \geq \left\{1 - \left(\frac{3}{4}\right)^k\right\}\binom{4}{4}$ for all k, since for $k \to \infty$, $\left(\frac{3}{4}\right)^k \binom{3}{3} \to \binom{0}{0}$ and $\left\{1 - \left(\frac{3}{4}\right)^k\right\}\binom{4}{4} \to \binom{4}{4}$, i.e. the mapping is not monotone.

Suppose that the operator is a contraction. Then, $\|\{L_{f_x}\}^k\, x - \{L_{f_y}\}^k\, y\|_\infty \leq \beta \cdot \|x - y\|_\infty$ for some $0 < \beta < 1$ and for all k. Since $\|\{L_{f_x}\}^\infty\, x - \{L_{f_y}\}^\infty\, y\|_\infty = 4 > 3 = \|x - y\|_\infty$, the operator is not a contraction.

Although the operator $\{L_{f_n}\}^{k(n)}$ is neither a contraction nor monotone, it can be shown that $\{L_{f_n}\}^{k(n)}\, x^n$ converges to the value vector v^α for any starting vector x^1. In order to prove this result we need the following lemma.

Lemma 3.11

Let $x^{n+1} = \{L_{f_n}\}^{k(n)}\, x^n$ *and* $\beta(n) = \alpha^{k(n)}\beta(n-1),\ n \in \mathbb{N}$, *with* $\beta(0) = 1$.
Assume that $Ux^1 - b \cdot e \le x^1 \le Ux^1 + d \cdot e$ *for some* $b, d \ge 0$.
Then, for $n = 0, 1, \ldots,$

(1) $x^{n+1} \le Ux^{n+1} + \beta(n)d \cdot e.$
(2) $x^{n+1} \le v^\alpha + (1-\alpha)^{-1}\beta(n)d \cdot e.$
(3) $x^{n+2} \ge Ux^{n+1} - \alpha(1-\alpha)^{-1}\beta(n)d \cdot e.$
(4) $x^{n+2} \ge v^\alpha - \alpha^{n+1}(1-\alpha)^{-1}\{(n+1)d + b\} \cdot e.$

Proof

(1) We prove this result by induction on n (for $n = 0$ the result is obvious).
Assume that $x^n \le Ux^n + \beta(n-1)d \cdot e$. Since for any fixed $f^\infty \in C(D)$ and any fixed $k \in \mathbb{N}$ the operator $\{L_f\}^k$ is a monotone contraction with factor α^k and with the additional property that $\{L_f\}^k(x + c \cdot e) = \{L_f\}^k\, x + \alpha^k c \cdot e$ for any $x \in \mathbb{R}^N$ and any scalar c, we obtain

$$\begin{aligned} x^{n+1} &= \{L_{f_n}\}^{k(n)}\, x^n \le \{L_{f_n}\}^{k(n)} \left\{Ux^n + \beta(n-1)d \cdot e\right\} \\ &= \{L_{f_n}\}^{k(n)}\{Ux^n\} + \alpha^{k(n)}\beta(n-1)d \cdot e\big\} = \{L_{f_n}\}^{k(n)}\{L_{f_n}x^n\} + \beta(n)d \cdot e \\ &= \{L_{f_n}\}^{k(n)+1}\, x^n + \beta(n)d \cdot e = \{L_{f_n}\}\{L_{f_n}^{k(n)}x^n\} + \beta(n)d \cdot e \\ &= L_{f_n}\, x^{n+1} + \beta(n)d \cdot e \le Ux^{n+1} + \beta(n)d \cdot e. \end{aligned}$$

(2) Iterating the inequality of part (1) gives for any $m \ge 1$

$$\begin{aligned} x^{n+1} &\le Ux^{n+1} + \beta(n)d \cdot e \\ &\le U\big\{Ux^{n+1} + \beta(n)d \cdot e\big\} + \beta(n)d \cdot e = U^2x^{n+1} + \alpha\beta(n)d \cdot e + \beta(n)d \cdot e \le \cdots \\ &\le U^m x^{n+1} + (1 + \alpha + \cdots + \alpha^{m-1})\beta(n)d \cdot e. \end{aligned}$$

Therefore, by letting $m \to \infty$, we obtain $x^{n+1} \le v^\alpha + (1-\alpha)^{-1}\beta(n)d \cdot e$.

(3) Also by part (1), we obtain

$$\begin{aligned} x^{n+2} &= \{L_{f_{n+1}}\}^{k(n+1)}\, x^{n+1} = \{L_{f_{n+1}}\}^{k(n+1)-1}\{L_{f_{n+1}}\, x^{n+1}\} = \{L_{f_{n+1}}\}^{k(n+1)-1}\{Ux^{n+1}\} \\ &\ge \{L_{f_{n+1}}\}^{k(n+1)-1}\{x^{n+1} - \beta(n)d \cdot e\} = \{L_{f_{n+1}}\}^{k(n+1)-1}\, x^{n+1} - \alpha^{k(n+1)-1}\beta(n)d \cdot e. \end{aligned}$$

Iterating the inequality $\{L_{f_{n+1}}\}^{k(n+1)}\, x^{n+1} \ge \{L_{f_{n+1}}\}^{k(n+1)-1}\, x^{n+1} - \alpha^{k(n+1)-1}\beta(n)d \cdot e$, gives

$$\begin{aligned} x^{n+2} &= \{L_{f_{n+1}}\}^{k(n+1)}\, x^{n+1} \ge L_{f_{n+1}}x^{n+1} - \{\alpha + \alpha^2 + \cdots + \alpha^{k(n+1)-1}\}\beta(n)d \cdot e \\ &\ge L_{f_{n+1}}x^{n+1} - \alpha(1-\alpha)^{-1}\beta(n)d \cdot e. \end{aligned}$$

(4) Since $Ux^1 \le x^1 + b \cdot e$, it follows that $U^{n+2}x^1 \le U^{n+1}x^1 + \alpha^{n+1}b \cdot e,\ n = -1, 0, \ldots.$
Hence, by iterating, $U^m\{U^{n+2}x^1\} \le U^{n+1}x^1 + (1 + \alpha + \cdots + \alpha^m)\alpha^{n+1}b \cdot e,\ m = 0, 1, \ldots.$
Therefore, $v^\alpha = lim_{m\to\infty}U^m\{U^{n+2}x^1\} \le U^{n+1}x^1 + (1-\alpha)^{-1}\alpha^{n+1}b \cdot e.$
For part (4), it is sufficient to show that $U^{n+1}x^1 \le x^{n+2} + (1-\alpha)^{-1}(n+1)\alpha^{n+1}d \cdot e.$

From part (3) it follows that

$$\begin{aligned} U^{n+1-j}x^{j+1} &= U^{n-j}\{Ux^{j+1}\} \leq U^{n-j}\{x^{j+2} + \alpha(1-\alpha)^{-1}\beta(j)d \cdot e\} \\ &= U^{n-j}x^{j+2} + \alpha^{n+1-j}(1-\alpha)^{-1}\beta(j)d \cdot e, \ j = 0, 1, \ldots. \end{aligned}$$

Summing up the above inequality over $j = 0, 1, \ldots, n$ gives

$$U^{n+1}x^1 \leq x^{n+2} + (1-\alpha)^{-1}\{\textstyle\sum_{j=0}^{n} \alpha^{n+1-j}\beta(j)\}d \cdot e.$$

Since $\beta(j) = \alpha^{k(j)+k(j-1)+\cdots+k(1)} \leq a^j$ for $j = 0, 1, \ldots, n$, the above inequality implies that

$$U^{n+1}x^1 \leq x^{n+2} + (1-\alpha)^{-1}(n+1)\alpha^{n+1}d \cdot e.$$

□

Theorem 3.29

Let $x^{n+1} = \{L_{f_n}\}^{k(n)}\, x^n, \ = 1, 2, \ldots.$ *Then,* $v^\alpha = lim_{n\to\infty}\, x^n$.

Proof

We apply Lemma 3.11 with $b = d = \|Ux^1 - x^1\|_\infty$. Since $lim_{n\to\infty}\, \beta(n) = 0$ and $lim_{n\to\infty}\, n\alpha^n = 0$, we obtain

$$\begin{aligned} limsup_{n\to\infty}\, x^n &\leq limsup_{n\to\infty}\, \{v^\alpha + (1-\alpha)^{-1}\beta(n-1)d \cdot e\} = v^\alpha \\ &= lim_{n\to\infty}\, \{v^\alpha - \alpha^{n-1}(1-\alpha)^{-1}\{(n-1)d + b\} \cdot e\} \\ &\leq liminf_{n\to\infty}\, x^n, \end{aligned}$$

i.e. $v^\alpha = lim_{n\to\infty}\, x^n$. □

Theorem 3.30

Algorithm 3.8 terminates in a finite number of iterations with an ε*-optimal policy.*

Proof

Since v^α is the fixed-point of U, we have

$$\begin{aligned} \|Ux^n - x^n\|_\infty &\leq \|Ux^n - Uv^\alpha\|_\infty + \|Uv^\alpha - x^n\|_\infty = \|Ux^n - Uv^\alpha\|_\infty + \|v^\alpha - x^n\|_\infty \\ &\leq \alpha \cdot \|x^n - v^\alpha\|_\infty + \|v^\alpha - x^n\|_\infty = (1+\alpha) \cdot \|v^\alpha - x^n\|_\infty. \end{aligned}$$

Because $v^\alpha = lim_{n\to\infty}\, x^n$, the stop criterion of step 2c in Algorithm 3.8 is satisfied after a finite number of iterations. From Theorem 3.7 part (3) and the stop criterion of Algorithm 3.8 it follows that

$$\|v^\alpha - v^\alpha(f_x^\infty)\|_\infty \leq 2\alpha(1-\alpha)^{-1}\|Ux - x\|_\infty \leq \varepsilon,$$

i.e. Algorithm 3.8 terminates with an ε-optimal policy. □

Convergence rate

We may assume that $Ux^1 \geq x^1$, because for $x^1 = (1-\alpha)^{-1}min_i\{max_a\, r_i(a)\} \cdot e$ this property is satisfied, namely:

$$\begin{aligned} \{Ux^1\}_i &= max_a\{r_i(a) + \alpha \textstyle\sum_j\, p_{ij}(a)x_j^1\} = max_a\, r_i(a) + \alpha(1-\alpha)^{-1}min_i\{max_a\, r_i(a)\} \\ &\geq min_i\{max_a\, r_i(a)\} + \alpha(1-\alpha)^{-1}min_i\{max_a\, r_i(a)\} \\ &= (1-\alpha)^{-1}min_i\{max_a\, r_i(a)\} = x_i^1, \ i \in S. \end{aligned}$$

We will show that the convergence of x^n to v^α is at least linear, i.e. for some $0 < c < 1$,

$$\|v^\alpha - x^{n+1}\|_\infty \leq c \cdot \|v^\alpha - x^n\|_\infty \text{ for } n = 0, 1, \ldots.$$

Since the operator of the modified policy is neither a contraction nor is it monotone, we cannot rely on general theorems. Therefore, we present a special treatment. Consider the related operator $U^{(k)} : \mathbb{R}^N \to \mathbb{R}^N$, defined by

$$U^{(k)}\, x = max_f\, L_f^k\, x. \tag{3.48}$$

Theorem 3.31
$U^{(k)}$ is a monotone contraction with contraction factor α^k and with fixed-point v^α.

Proof
Suppose that $x \geq y$. From the monotonicity of $L_f^k,\ f \in C(D)$, we obtain

$$U^{(k)}\, x = max_f\, L_f^k\, x \geq max_f\, L_f^k\, y = U^{(k)}\, y.$$

Consider a fixed state $i \in S$ and let $f_{x,i}$ be such that $\{U^{(k)}\, x\}_i = \{L_{f_{x,i}}^k\, x\}_i,\ x \in \mathbb{R}^N$.
Then, for each $i \in S$, we have

$$\{U^{(k)}\, x - U^{(k)}\, y\}_i \leq \{L_{f_{x,i}}^k\, x - L_{f_{x,i}}^k\, y\}_i = \alpha^k \{P^k(f_{x,i})(x - y)\}_i \leq \alpha^k \|x - y\|_\infty$$

and

$$\{U^{(k)}\, y - U^{(k)}\, x\}_i \leq \{L_{f_{y,i}}^k\, y - L_{f_{y,i}}^k\, x\}_i = \alpha^k \{P^k(f_{y,i})(x - y)\}_i \leq \alpha^k \|x - y\|_\infty.$$

Hence, $\|U^{(k)}\, x - U^{(k)}\, y\| \leq \alpha^k \|x - y\|_\infty$, i.e. $U^{(k)}$ is a monotone contraction with contraction factor α^k. Finally, we show that v^α is the fixed-point. Let $f_* \in C(D)$ be an α-optimal policy. Since $v^\alpha = L_{f_*}^k\, v^\alpha \geq L_f^k\, v^\alpha$ for every $f \in C(D)$, we obtain

$$v^\alpha \geq max_f\, L_f^k\, v^\alpha = U^{(k)}\, v^\alpha \geq L_{f_*}^k\, v^\alpha = v^\alpha,$$

i.e. v^α is the fixed-point of $U^{(k)}$. □

Consider, besides the sequence $\{x^n\}_{n=1}^\infty$ defined by (3.47), the sequence $\{y^n\}_{n=1}^\infty$ and $\{z^n\}_{n=1}^\infty$, defined by

$$y^1 = z^1 = x^1;\ y^{n+1} = Uy^n,\ z^{n+1} = U^{(k(n))} z^n,\ n \in \mathbb{N}.$$

Lemma 3.12
Under the assumption $Ux^1 \geq x^1$, we have, $Ux^n \geq x^n$ and $v^\alpha \geq z^n \geq x^n \geq y^n$ for every $n \in \mathbb{N}$.

Proof
We apply induction on n. Since $Ux^1 \geq x^1$, we have $v^\alpha = lim_{n\to\infty} U^n\, x^1 \geq U\, x^1 \geq x^1 = y^1 = z^1$, the result is true for $n = 1$. Assume that $Ux^n \geq x^n$ and $v^\alpha \geq z^n \geq x^n \geq y^n$. Then,

$$\begin{aligned} Ux^{n+1} - x^{n+1} &= U\big\{\{L_{f_n}\}^{k(n)}\, x^n\big\} - \{L_{f_n}\}^{k(n)}\, x^n \geq \{L_{f_n}\}^{k(n)+1}\, x^n - \{L_{f_n}\}^{k(n)}\, x^n \\ &= \{L_{f_n}\}^{k(n)}\, \{Ux^n\} - \{L_{f_n}\}^{k(n)}\, x^n = \alpha^{k(n)} \big\{P(f_n)\big\}^{k(n)} \{Ux^n - x^n\} \geq 0 \end{aligned}$$

and

$$v^\alpha = U^{(k(n))}v^\alpha \geq U^{(k(n))}z^n = z^{n+1} = max_f\, L_f^{k(n)}\, z^n \geq L_{f_n}^{k(n)}\, z^n \geq L_{f_n}^{k(n)}\, x^n = x^{n+1}.$$

Since $x^{n+1} = x^n + A^{(k(n))}\{Ux^n - x^n\} = x^n + \sum_{i=0}^{k(n)-1}\{\alpha P(f_n)\}^i\{Ux^n - x^n\}$, we have

$$x^{n+1} = Ux^n + \sum_{i=1}^{k(n)-1}\{\alpha P(f_n)\}^i\{Ux^n - x^n\} \geq Ux^n \geq Uy^n = y^{n+1}.$$ □

The next corollary shows that the convergence is geometric, i.e. $\|v^\alpha - x^{n+1}\|_\infty \leq \alpha\|v^\alpha - x^n\|_\infty$.

Corollary 3.9

Under the assumption $Ux^1 \geq x^1$, we have $\|v^\alpha - x^{n+1}\|_\infty \leq \alpha\|v^\alpha - x^n\|_\infty$.

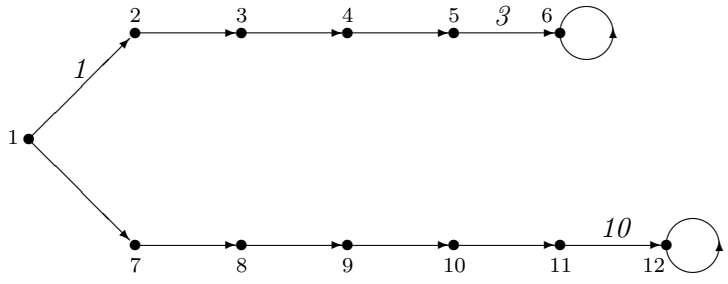
Proof

From the last line of the proof of Lemma 3.12 it follows that $x^{n+1} \geq Ux^n$. Hence, also by Lemma 3.12, $0 \leq v^\alpha - x^{n+1} \leq v^\alpha - Ux^n = Uv^\alpha - Ux^n$, implying $\|v^\alpha - x^{n+1}\|_\infty \leq \alpha\|v^\alpha - x^n\|_\infty$. □

Lemma 3.12 shows that, under the assumption $Ux^1 \geq x^1 = y^1$, the iterates x^n of modified policy iteration always exceed the iterates y^n of value iteration, which is modified policy iteration with $k = 1$. One might conjecture that the iterates of modified policy iteration with fixed order $k+m$ ($m \geq 1$) always dominates those of modified policy iteration with fixed k. The following intricate example shows that this conjecture is false.

Example 3.6

Let $S = \{1, 2, \ldots, 12\}$; $A\{1\} = \{1, 2\}$, $A\{i\} = \{1\}$, $1 \leq i \leq 12$; $r_1(1) = 1$, $r_5(1) = 3$, $r_{11}(1) = 10$ (all other rewards are 0). The transitions are deterministic and from state i to state $i+1$, $2 \leq i \leq 5$ and $7 \leq i \leq 11$. The states 6 and 12 are absorbing. In state 1, action 1 gives a transition to state 2 and action 2 to state 7. Let $\frac{1}{81} < \alpha < 1$. Below is a picture of this model.



Firstly, consider the modified policy iteration method with $k = 3$ and starting vector $x^1 = 0$.

$x^2 = (1, 0, 3\alpha^2, 3\alpha, 3, 0, 0, 0, 10\alpha^2, 10\alpha, 10, 0)$.

$x^3 = (1 + 10\alpha^4, 3\alpha^3, 3\alpha^2, 3\alpha, 3, 0, 10\alpha^4, 10\alpha^3, 10\alpha^2, 10\alpha, 10, 0)$.

Secondly, take $k = 4$ and obtain (call the iterates y^2 and y^3):

$y^2 = (1, 3\alpha^3, 3\alpha^2, 3\alpha, 3, 0, 0, 10\alpha^3, 10\alpha^2, 10\alpha, 10, 0)$.

$y^3 = (3\alpha^4, 3\alpha^3, 3\alpha^2, 3\alpha, 3, 0, 10\alpha^4, 10\alpha^3, 10\alpha^2, 10\alpha, 10, 0)$.

Notice that $y^2 > x^2$ and $x^3 > y^3$.

Exclusion of suboptimal actions

In order to exclude suboptimal actions we need bounds on the value vector v^α. The next theorem provides appropriate bounds.

Theorem 3.32

$x^n + (1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \le Ux^n + \alpha(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \le$
$x^{n+1} + \alpha^{k(n)}(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \le v^\alpha \le x^n + (1-\alpha)^{-1} max_i\,(Ux^n - x^n)_i \cdot e.$

Proof

We start with the upper bound. Let $f^\infty = f^\infty_{v^\alpha}$. Then, we have

$$\begin{aligned} Ux^n - x^n &\ge L_f\, x^n - x^n = r(f) + \alpha P(f)x^n - x^n = r(f) + \alpha P(f)v^\alpha + \alpha P(f)(x^n - v^\alpha) - x^n \\ &= L_f\, v^\alpha + \alpha P(f)(x^n - v^\alpha) - x^n = (v^\alpha - x^n) + \alpha P(f)(x^n - v^\alpha) \\ &= \{I - \alpha P(f)\}(v^\alpha - x^n). \end{aligned}$$

Hence,

$$\begin{aligned} v^\alpha - x^n &\le \{I - \alpha P(f)\}^{-1} Ux^n - x^n \le \{I - \alpha P(f)\}^{-1} max_i\,(Ux^n - x^n)_i \cdot e \\ &= (1-\alpha)^{-1} max_i\,(Ux^n - x^n)_i \cdot e. \end{aligned}$$

For the lower bounds, we can write

$$\begin{aligned} v^\alpha - x^n &= Uv^\alpha - x^n \ge L_{f_n}\, v^\alpha - x^n = r(f_n) + \alpha P(f_n)v^\alpha - x^n \\ &= r(f_n) + \alpha P(f_n)x^n - x^n + \alpha P(f_n)(v^\alpha - x^n) = L_{f_n}\, x^n - x^n + \alpha P(f_n)(v^\alpha - x^n). \end{aligned}$$

Therefore,

$$\begin{aligned} v^\alpha - x^n &\ge \{I - \alpha P(f_n)\}^{-1}\{L_{f_n}\, x^n - x^n\} = \{I - \alpha P(f_n)\}^{-1}\{U\, x^n - x^n\} \\ &= \textstyle\sum_{i=0}^\infty \{\alpha P(f_n)\}^i (Ux^n - x^n). \end{aligned}$$

Since

$$x^{n+1} = x^n + \textstyle\sum_{i=0}^{k(n)-1} \{\alpha P(f_n)\}^i (Ux^n - x^n) = Ux^n + \sum_{i=1}^{k(n)-1} \{\alpha P(f_n)\}^i (Ux^n - x^n),$$

we obtain

$$\begin{aligned} v^\alpha &\ge x^n + \textstyle\sum_{i=0}^{k(n)-1} \{\alpha P(f_n)\}^i (Ux^n - x^n) + \sum_{i=k(n)}^\infty \{\alpha P(f_n)\}^i (Ux^n - x^n) \\ &= x^{n+1} + \textstyle\sum_{i=k(n)}^\infty \{\alpha P(f_n)\}^i (Ux^n - x^n) \\ &\ge x^{n+1} + \alpha^{k(n)}(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \\ &= Ux^n + \textstyle\sum_{i=1}^{k(n)-1} \{\alpha P(f_n)\}^i (Ux^n - x^n) + \alpha^{k(n)}(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \\ &\ge Ux^n + \textstyle\sum_{i=1}^{k(n)-1} \alpha^i\, min_i\,(Ux^n - x^n)_i \cdot e + \alpha^{k(n)}(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \\ &= Ux^n + \alpha(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \\ &= x^n + (Ux^n - x^n) + \alpha(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \\ &\ge x^n + min_i\,(Ux^n - x^n)_i \cdot e + \alpha(1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e \\ &= x^n + (1-\alpha)^{-1} min_i\,(Ux^n - x^n)_i \cdot e. \end{aligned}$$

Theorem 3.33

If

$$r_i(a) + \alpha \sum_j p_{ij}(a) x_j^n < x_i^{n+1} + \alpha^{k(n)} (1-\alpha)^{-1} \, min_k \, (Ux^n - x^n)_k - \alpha (1-\alpha)^{-1} \, max_k \, (Ux^n - x^n)_k \tag{3.49}$$

then action a is suboptimal.

Proof

$$\begin{aligned} r_i(a) + \alpha \textstyle\sum_j \, p_{ij}(a) v_j^\alpha &\leq r_i(a) + \alpha \textstyle\sum_j \, p_{ij}(a) \{x_j^n + (1-\alpha)^{-1} \, max_k \, (Ux^n - x^n)_k\} \\ &= r_i(a) + \alpha \textstyle\sum_j \, p_{ij}(a) x_j^n + \alpha (1-\alpha)^{-1} \, max_k \, (Ux^n - x^n)_k \\ &< x_i^{n+1} + \alpha^{k(n)} (1-\alpha)^{-1} \, min_k \, (Ux^n - x^n)_k \leq v_i^\alpha \end{aligned}$$

□

Algorithm 3.9 *Modified policy iteration with exclusion of suboptimal actions*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.
2. a. Choose any k with $1 \leq k \leq \infty$.
   b. Compute $y_i(a) = r_i(a) + \alpha \sum_j \, p_{ij}(a) x_j, \;\; (i,a) \in S \times A$.
   c. Determine f such that $L_f \, x = Ux$.
   d. If $\|Ux - x\|_\infty \leq \frac{1}{2}(1-\alpha)\alpha^{-1}\varepsilon$: f^∞ is an ε-optimal policy (STOP).
   e. $max = max_k \, (Ux - x)_k$ and $min = min_k \, (Ux - x)_k$.
3. a. $y := \{L_f\}^k \, x$.
   b. $A(i) = \{a \mid y_i(a) \geq y_i + \alpha^k (1-\alpha)^{-1} \, min - \alpha (1-\alpha)^{-1} \, max\}, \;\; i \in S$.
   c. If $|Ai)| = 1, \;\; i \in S$, then f^∞ is an optimal poliy (STOP).
   d. $x := y$ and return to step 2.

Example 3.4 (continued)

Iteration 1

$y_1(1) = \frac{17}{3}, \;\; y_1(2) = 6, \;\; y_1(3) = \frac{28}{3}; \;\; y_2(1) = \frac{32}{3}, \;\; y_2(2) = 8, \;\; y_2(3) = \frac{34}{3};$
$y_3(1) = \frac{38}{3}, \;\; y_3(2) = 13, \;\; y_3(3) = \frac{40}{3}. \; Ux = (\frac{28}{3}, \frac{34}{3}, \frac{40}{3}); \;\; f(1) = f(2) = f(3) = 3.$
$max = 4, \;\; min = 0; \;\; y = (\frac{29}{3}, \frac{35}{3}, \frac{41}{3}); \;\; A\{1\} = \{3\}, \;\; A\{2\} = \{1,3\}, \;\; A\{3\} = \{1,2,3\}.$
$x = (\frac{29}{3}, \frac{35}{3}, \frac{41}{3}).$

Iteration 2

$y_1(1) = 9.833; \;\; y_2(1) = 10.833, \;\; y_2(3) = 11.833; \;\; y_3(1) = 12.833, \;\; y_3(2) = 14.633, \;\; y_3(3) = 13.833.$
$Ux = (9.833, 11.833, 14.833); \;\; f(1) = f(2) = 3, \;\; f(3) = 2; \;\; max = 1.166, \;\; min = 0.166.$
$y = (10.417, 12.417, 14.917). \; A\{1\} = \{3\}, \;\; A\{2\} = \{3\}, \;\; A\{3\} = \{2,3\}.$
$x = (10.417, 12.417, 14.917).$

Iteration 3
$y_1(1) = 10.459;\ y_2(3) = 12.459;\ y_3(2) = 15.209,\ y_3(3) = 14.459;\ Ux = (10.459, 12.459, 15.209).$
$f(1) = f(2) = 3,\ f(3) = 2;\ max = 0.292,\ min = 0.042;\ y = (10.604, 12.604, 15.229).$
$A\{1\} = \{3\},\ A\{2\} = \{3\},\ A\{3\} = \{2\};\ f^\infty$ is an optimal policy.

3.8 Bibliographic notes

The principle of optimality is credited to Bellman [11]. Discounted models appear to have been first analyzed in generality by Howard [101]. Blackwell [22] and Denardo [44] have provided fundamental theoretical papers on discounted models.

The proof of Theorem 3.6 is drawn from Ross [168]. Shapiro [182] made the observation that Brouwer's fixed-point theorem can also be used to prove that the mapping U has a fixed-point (see Exercise 3.7). The concept of conserving policy was proposed by Dubins and Savage [61]. Bather [6] was the first to use the span semi-norm for analyzing Markov decision processes. The idea to use bounds for the value vector in order to derive a suboptimality test is due to MacQueen ([131], [132]). These ideas were extended and improved by Porteus ([152], [154]) and others.

Policy iteration is usually attributed to Howard [101]. We follow the more mathematically treatment of Blackwell [22]. The equivalence between policy iteration and Newton's method was shown in Puterman and Brumelle [158]. Hastings [84] (see Exercise 3.12) has proposed a method to accelerate the policy iteration method.

The linear programming method for discounted MDPs was proposed by d'Epenoux [53]. The equivalence between block-pivoting and policy iteration was mentioned by De Ghellinck [39]. The one-to-one correspondence between the feasible solutions of the dual program and the set of stationary policies can be found in De Ghellinck and Eppen [40]. For an extensive study of linear programming and Markov decision models see Kallenberg [108].

The use of value iteration originates in the work of Shapley [183], who applied it in stochastic games. Hastings ([83], [84]) and Kushner and Kleinman [123] independently suggested the use of (pre)-Gauss-Seidel iteration to accelerate value iteration. Other accelerations were proposed by Kushner and Kleinman [124] and Reetz [163], both papers on *overrelaxation*, Wessels [229] and Van Nunen and Wessels ([209], [210]), these last papers based on *stopping times*.

The method of modified policy iteration was suggested in Morton [140] and formalized by Van Nunen [207], Puterman and Shin ([159], [160]). The example showing that the operator of this method is in general neither a contraction nor monotone is due to Van Nunen [208]. The observation that modified policy iteration method can be viewed as an inexact Newton method to solve the optimality equation $Ux = x$ was observed by Dembo and Haviv [43]. The exclusion of suboptimal actions is developed by Puterman and Shin [160]. Example 3.6 is due to Van der Wal and Van Nunen [204].

3.9 Exercises

Exercise 3.1

Let $\mathcal{M} = \{\mu \in \mathbb{R}^N \mid \mu_i > 0,\ i \in S \text{ and } \sum_j p_{ij}(a)\mu_j \leq \mu_i \text{ for all } (i,a) \in S \times A\}$.

Define $\|x\|_\mu = max_i\ \mu_i^{-1} \cdot |x_i|$.

a. Show that $\|x\|_\mu$ is a norm in $\mathbb{R}^N$.

b. Formulate the generalisations of Lemma 3.2 and Lemma 3.3 with respect to the norm $\|x\|_\mu$.

c. Give the key property by which the proofs of these new Lemmata follow from the versions with the supremum norm.

Exercise 3.2

Suppose that B is a monotone contraction with contraction factor β and fixed-point x^*.

Show that for any $n \in \mathbb{N}$ the mapping B^n is a monotone contraction with contraction factor β^n and fixed-point x^*.

Exercise 3.3

Let X be a Banach space and $B : X \to X$.

Suppose that B is *nonexpanding*, i.e. $\|Bx - By\| \leq \|x - y\|$ for every $x, y \in X$, and suppose furthermore that B^n is a contraction mapping with contraction factor β and fixed-point x^*, for some $n \in \mathbb{N}$. Then, show that

(1) x^* is the unique fixed-point of B.

(2) $\|x^* - x\| \leq n(1-\beta)^{-1} \cdot \|Bx - x\|$ for every $x \in X$.

Exercise 3.4

Let $B : \mathbb{R}^N \to \mathbb{R}^N$ be such that for all $x, y \in \mathbb{R}^N$ and for some $\beta \in [0, 1)$

$max_i\ (Bx - By)_i\ \leq \beta \cdot max_i\ (x - y)_i$ and $min_i\ (Bx - By)_i \geq \beta \cdot min_i\ (x - y)_i$.

Show that:

(1) B is a monotone contraction mapping with contraction factor β;

(2) $x + (1-\beta)^{-1} min_i\ (Bx - x)_i \cdot e \leq x^* \leq x + (1-\beta)^{-1} max_i\ (Bx - x)_i \cdot e$, where x^* is the fixed-point of B and x is an arbitrary point of $\mathbb{R}^N$.

Exercise 3.5

Let $R = (\pi^1, \pi^2, \dots)$ be any Markov policy and let $x \in \mathbb{R}^N$.

Show that $v^\alpha(R) = lim_{n\to\infty}\ L_{\pi^1} L_{\pi^2} \cdots L_{\pi^n}\, x$.

Exercise 3.6

Give the optimality equation of the following model:

$S = \{1, 2\};\ A(1) = A(2) = \{1, 2\};\ \alpha = 0.9;\ r_1(1) = 4, r_1(2) = 2, r_2(1) = 0, r_2(2) = 2;$

$p_{11}(1) = \frac{1}{3}, p_{12}(1) = \frac{2}{3}, p_{11}(2) = \frac{2}{3}, p_{12}(2) = \frac{1}{3}, p_{21}(1) = 1, p_{22}(1) = p_{21}(2) = 0, p_{22}(2) = 1.$

Exercise 3.7

Brouwer's fixed-point theorem is:

Suppose that G is a continuous function which maps a compact convex set $X \subseteq \mathbb{R}^N$ into itself. Then G has a fixed-point.

Show by Brouwer's theorem that U has a fixed-point.

Exercise 3.8

For any $x \in \mathbb{R}^N$ and $\mu \in \mathcal{M}$, defined in Exercise 3.1, we define

$$b_1 = min_i \ \tfrac{(Ux-x)_i}{\mu_i};\ \beta_1 = \begin{cases} \alpha \cdot min_{i,a} \ \frac{1}{\mu_i} \sum_j \ p_{ij}(a)\mu_j \text{ if } b_1 > 0 \\ \alpha \cdot max_{i,a} \ \frac{1}{\mu_i} \sum_j \ p_{ij}(a)\mu_j \text{ if } b_1 \leq 0 \end{cases}$$

$$b_2 = max_i \ \tfrac{(Ux-x)_i}{\mu_i};\ \beta_2 = \begin{cases} \alpha \cdot max_{i,a} \ \frac{1}{\mu_i} \sum_j \ p_{ij}(a)\mu_j \text{ if } b_2 > 0 \\ \alpha \cdot min_{i,a} \ \frac{1}{\mu_i} \sum_j \ p_{ij}(a)\mu_j \text{ if } b_2 \leq 0 \end{cases}$$

Show that:

(1) $\beta_1 b_1 \cdot \mu_i \leq \alpha \cdot b_1 \sum_j \ p_{ij}(a)\mu_j$ and $\beta_2 b_2 \cdot \mu_i \geq \alpha \cdot b_2 \sum_j \ p_{ij}(a)\mu_j$ for every $(i, a) \in S \times A$.

(2) $x + (1-\beta_1)^{-1} b_1 \cdot \mu \leq Ux + \beta_1(1-\beta_1)^{-1} b_1 \cdot \mu \leq v^\alpha(f_x^\infty) \leq v^\alpha \leq Ux + \beta_2(1-\beta_2)^{-1} b_2 \cdot \mu \leq x + (1-\beta_2)^{-1} b_2 \cdot \mu.$

(3) If $r_i(a) + \alpha \sum_j \ p_{ij}(a)x_j < (Ux)_i + \beta_1(1-\beta_1)^{-1} b_1 \cdot \mu_i - \beta_2(1-\beta_2)^{-1} b_2 \cdot \mu_i$, then action $a \in A(i)$ is suboptimal.

(4) If $r_i(a) + \alpha \sum_j \ p_{ij}(a)(Ux)_j < (Ux)_i + \beta_1(1-\beta_1)^{-1} b_1 \cdot \mu_i - \beta_2^2(1-\beta_2)^{-1} b_2 \cdot \mu_i$, then action $a \in A(i)$ is suboptimal.

(5) Test (4) is stronger than test (3).

Exercise 3.9

Show that $span\ (U^2x - Ux) \leq \alpha \cdot span(Ux - x)$.

Exercise 3.10

Consider the following MDP:

$S = \{1, 2\};\ A(1) = A(2) = \{1, 2\};\ \alpha = \frac{1}{2};\ r_1(1) = 1,\ r_1(2) = 0;\ r_2(1) = 2,\ r_2(2) = 2.$

$p_{11}(1) = \frac{1}{2},\ p_{12}(1) = \frac{1}{2};\ p_{11}(2) = \frac{1}{4},\ p_{12}(2) = \frac{3}{4}.$

$p_{21}(1) = \frac{2}{3},\ p_{22}(1) = \frac{1}{3};\ p_{21}(2) = \frac{1}{3},\ p_{22}(2) = \frac{2}{3}.$

Use the policy iteration algorithm 3.2 to find an α-discounted optimal policy for this model (start with $f(1) = f(2) = 1$).

Exercise 3.11

Show that $Fy \geq Fx$ implies that $y \leq x$, where F is defined by $Fx = Ux - x$.

Exercise 3.12

Consider the following modification of the policy iteration method:

1. Start with any $f \in C(D)$.
2. Compute $v^{\alpha}(f^{\infty})$ as unique solution of the linear system $L_f x = x$.
3. For $i = 1$ to N do
   a. compute $d_{ia}(f) = r_i(a) + \alpha \sum_{j=1}^{i-1} p_{ij}(a)x_j + \alpha \sum_{j=i}^{N} p_{ij}(a)v_j^{\alpha}(f^{\infty}),\ a \in A(i)$;
   b. if $d_{ia}(f) \leq v_i^{\alpha}(f^{\infty})$ for every $a \in A(i)$, then $x_i = v_i^{\alpha}(f^{\infty})$ and $g(i) = f(i)$;
   c. if $d_{ia}(f) > v_i^{\alpha}(f^{\infty})$ for some $a \in A(i)$, then $x_i = max_a\ d_{ia}(f)$ and choose $g(i)$ such that $d_{ig(i)} = x_i$.
4. If $g(i) = f(i)$ for every $i \in S$ then go to step 6.
5. $f := g$ and go to step 2.
6. f^{∞} is an α-discounted optimal policy (STOP).

Prove the correctness of this method by showing the following steps:

a. (i) $x \geq v^{\alpha}(f^{\infty})$; (ii) $x = v^{\alpha}(f^{\infty})$ if and only if $f = g$.

b. If $f = g$, then $f^{\infty})$ is an α-discounted optimal policy.

c. If $f \neq g$, then $v^{\alpha}(g^{\infty}) \geq x \geq v^{\alpha}(f^{\infty})$.

Exercise 3.13

Apply the method of Exercise 3.12 to the MDP model of Example 3.1.

Exercise 3.14

Show that for a given initial distribution β and a stationary policy π^{∞},

$$\sum_j \beta_j v_j^{\alpha}(\pi^{\infty}) = \sum_{(i,a)} r_i(a) x_i^{\pi}(a).$$

Exercise 3.15

Use the linear programming method to compute the value vector and an optimal policy for the model of Exercise 3.10 (take $\beta_1 = \beta_2 = \frac{1}{2}$).

Exercise 3.16

Show the following optimality properties:

(1) If $\pi^{\infty} \in C(S)$ is an α-discounted optimal policy, then $x(\pi)$ is an optimal solution of (3.32).

(2) If x is an optimal solution of (3.32), then $\pi^{\infty}(x)$ is an α-discounted optimal policy.

Exercise 3.17
Apply the suboptimality tests of the Theorems 3.20 and 3.22 to the model of Exercise 3.10 (take $\beta_1 = \beta_2 = \frac{1}{2}$).

Exercise 3.18
Use algorithm 3.4 to compute an ε-optimal policy for the model of Exercise 3.10. Take $\varepsilon = 0.2$ and start with $x = (2, 2)$.

Exercise 3.19
Use algorithm 3.5 to compute an ε-optimal policy for the model of Exercise 3.10. Take $\varepsilon = 0.2$ and start with $x = (2, 2)$.

Exercise 3.20
Use algorithm 3.6 to compute an ε-optimal policy for the model of Exercise 3.10. Take $\varepsilon = 0.2$ and start with $x = (2, 2)$.

Exercise 3.21
Use algorithm 3.7 to compute an ε-optimal policy for the model of Exercise 3.10. Take $\varepsilon = 0.2$ and start with $x = (2, 2)$.

Exercise 3.22
Prove Theorem 3.27

Exercise 3.23
Prove Theorem 3.28

Exercise 3.24
Use algorithm 3.8 to compute an ε-optimal policy for the model of Exercise 3.10. Take $\varepsilon = 0.2$, start with $x = (2, 2)$ and take $k = 2$ in each iteration.

Exercise 3.25
Show that $\|v^\alpha - x^{n+1}\|_\infty \leq \beta\|v^\alpha - x^n\|_\infty$, where x^n is the x in iteration n of algorithm 3.8, $\beta = min\{\alpha, \alpha^{k(n)} + (1-\alpha)^{-1}\big(\alpha - \alpha^{k(n)}\big)\|P(f_n) - P(f_\alpha)\|_\infty\}$ with f_n and f_α are f_{x^n} and f_{v^α} respectively. Assume that $Ux^1 \geq x^1$.

Exercise 3.26
Consider the following modified policy algorithm.

Algorithm 3.10

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary, and let $\overline{y} = \infty$.
2. a. Choose any k with $1 \leq k \leq \infty$.
   b. Determine f such that $L_f x = Ux$.
   c. $min = min_i\,(Ux - x)_i$ and $max = max_i\,(Ux - x)_i$.
   d. $\underline{y} = x + (1-\alpha)^{-1}\,min \cdot e;\ \overline{y} = x + (1-\alpha)^{-1}\,max \cdot e$.
   c. If $\|\overline{y} - \underline{y}\|_\infty \leq \varepsilon$: $y = \frac{1}{2}(\overline{y} - \underline{y})$ is an $\frac{1}{2}\varepsilon$-approximation of v^α and f^∞ is an ε-optimal policy (STOP).
3. a. $y := \{L_f\}^k\, x$.
   b. $x := y$ and return to step 2.

(1) Apply this algorithm to the MDP model of exercise 3.24.
(2) Show the following properties for this algorithm under the assumption $Ux^1 \geq x^1$.
($x^n,\ f_n,\ \underline{y}^n,\ \overline{y}^n$ are the $x,\ f,\ \underline{y},\ \overline{y}$ in iteration n):
a. $x^n \leq Ux^n \leq x^{n+1} \leq v^\alpha(f_n^\infty)$.
b. $\underline{y}^n \leq v^\alpha(f_n^\infty) \leq v^\alpha \leq \overline{y}^n$.
c. $\underline{y}^n \uparrow v^\alpha$ and $\overline{y}^n \downarrow v^\alpha$.
d. $\|v^\alpha - y\|_\infty \leq \frac{1}{2}\varepsilon$ and $\|v^\alpha - v^\alpha(f^\infty)\|_\infty \leq \varepsilon$, when the algorithm terminates.

Exercise 3.27
Show that if

$$r_i(a) + \alpha \sum_j p_{ij}(a) x_j^n < x_i^n + (1-\alpha)^{-1}\,min_k\,(Ux^n - x^n)_k - \alpha(1-\alpha)^{-1}\,max_k\,(Ux^n - x^n)_k,$$

then action a is suboptimal.

Chapter 4

Total reward

4.1 Introduction

.

Alternatives to the expected total discounted reward criterion in infinite-horizon models include the total expected reward and the average expected reward criteria. This chapter deals with the total expected reward criterion. We have to make some assumptions on the rewards and/or the transition probabilities, without which the total expected reward may be unbounded or not even well defined. When these assumptions are not fulfilled, the average reward and more sensitive optimality criteria can be applied. These last models will be discussed in the chapters 5 and 6.

Let $v_i^+(R) = \sum_{t=1}^{\infty} \sum_{(j,a)} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j^+(a),\ i \in S$, where $r_j^+(a) = max\{0, r_j(a)\}$, and $v_i^-(R) = \sum_{t=1}^{\infty} \sum_{(j,a)} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j^-(a),\ i \in S$, where $r_j^-(a) = max\{0, -r_j(a)\}$.
The total reward $v_i(R)$ is well defined, possibly $\pm\infty$, if $min\{v_i^+(R), v_i^-(R)\} < \infty$.

Throughout this chapter we assume the following.

Assumption 4.1
(1) The model is *substochastic*, i.e. $\sum_j p_{ij}(a) \leq 1$ for all $(i, a) \in S \times A$.
(2) For any initial state i and any policy R the expected total reward $v_i(R)$ exists (possibly $\pm\infty$).

In the next lemma is shown that the total expected reward is the limit of the discounted expected reward when the discount factor α tends to 1.

Lemma 4.1
For any initial state i and any policy R the expected total reward $v_i(R) = lim_{\alpha \uparrow 1}\, v_i^{\alpha}(R)$.

Proof

Take any initial state i and any policy R. We distinguish the following cases.

Case 1: $-\infty < v_i(R) < +\infty$.

Let $v_i^{(t)}(R)$ be the expected reward in period t: $v_i^{(t)}(R) = \sum_{j,a} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j(a)$.

Take any $\varepsilon > 0$. Then, there exists a T_* such that $|v_i(R) - \sum_{t=1}^{T} v_i^{(t)}(R)| < \varepsilon$ for every $T \geq T_*$.

Since $|v_i^{(t)}(R)|$ is bounded by $M = max_{(i,a)} |r_i(a)|$, the two power series $v_i^{\alpha}(R) = \sum_{t=1}^{\infty} \alpha^{t-1} v_i^{(t)}(R)$ and $\sum_{s=1}^{\infty} \alpha^{s-1}$ have radius of convergence (at least) 1. Hence, for any $\alpha \in [0,1)$, we may write

$$(1-\alpha)^{-1} v_i^{\alpha}(R) = \Big\{ \sum_{s=1}^{\infty} \alpha^{s-1} \Big\} \Big\{ \sum_{t=1}^{\infty} v_i^{(t)}(R) \Big\} = \sum_{t=1}^{\infty} \Big\{ \sum_{s=1}^{t} v_i^{(s)}(R) \Big\} \alpha^{t-1}.$$

Therefore,

$$|(1-\alpha)^{-1}\{v_i^{\alpha}(R) - v_i(R)\}| \;\leq\; \sum_{t=1}^{\infty} |\sum_{s=1}^{t} v_i^{(s)}(R) - v_i(R)| \cdot \alpha^{t-1} \;=$$
$$\sum_{t=1}^{T_*} |\sum_{s=1}^{t} v_i^{(s)}(R) - v_i(R)| \cdot \alpha^{t-1} + \sum_{t=T_*}^{\infty} |\sum_{s=1}^{t} v_i^{(s)}(R) - v_i(R)| \cdot \alpha^{t-1}.$$

Let $A = max_{1 \leq t \leq T_*} |\sum_{s=1}^{t} v_i^{(s)}(R) - v_i(R)|$. Then, we obtain

$$\begin{aligned} |(1-\alpha)^{-1}\{v_i^{\alpha}(R) - v_i(R)\}| &\leq \sum_{t=1}^{T_*} A \cdot \alpha^{t-1} + \sum_{t=T_*}^{\infty} \varepsilon \cdot \alpha^{t-1} \\ &\leq A \cdot \tfrac{1-\alpha^{T_*}}{1-\alpha} + \varepsilon \cdot \sum_{t=1}^{\infty} \alpha^{t-1} < 2\varepsilon(1-\alpha)^{-1} \end{aligned}$$

for α sufficiently close to 1. Hence, $lim_{\alpha \uparrow 1} v_i^{\alpha}(R) = v_i(R)$.

Case 2: $v_i(R) = +\infty$.

Choose $M > 0$ arbitrary. Then, there exists an integer T_* such that $\sum_{t=1}^{T} v_i^{(t)}(R) > M$ for all $T > T^*$. Similarly as in Lemma 4.1 we can write

$$\begin{aligned} (1-\alpha)^{-1} v_i^{\alpha}(R) &= \sum_{t=1}^{\infty} \Big\{ \sum_{s=1}^{t} v_i^{(s)}(R) \Big\} \alpha^{t-1} \\ &= \sum_{t=1}^{T^*} \Big\{ \sum_{s=1}^{t} v_i^{(s)}(R) \Big\} \alpha^{t-1} + \sum_{t=T^*+1}^{\infty} \Big\{ \sum_{s=1}^{t} v_i^{(s)}(R) \Big\} \alpha^{t-1}. \end{aligned}$$

Let $m = min_{1 \leq t \leq T^*} \sum_{s=1}^{t} v_i^{(s)}(R)$, then $(1-\alpha)^{-1} v_i^{\alpha}(R) > m \cdot \frac{1-\alpha^{T^*}}{1-\alpha} + M \cdot \frac{\alpha^{T^*}}{1-\alpha}$, i.e. $v_i^{\alpha}(R) > m \cdot (1-\alpha^{T^*}) + M \cdot \alpha^{T^*}$. For $\alpha \uparrow 1$, we have $m \cdot (1-\alpha^{T^*}) + M \cdot \alpha^{T^*} \to M$.

Hence, since M was arbitrarily chosen, $lim_{\alpha \uparrow 1} v_i^{\alpha}(R) = +\infty = v_i(R)$.

Case 3: $v_i(R) = -\infty$.

The proof is similar to the proof of case 2 and left to the reader (see Exercise 4.3). □

Next, we show the existence of a policy $f_0^{\infty} \in C(D)$ such that $v^{\alpha}(f_0^{\infty}) = v^{\alpha}$ for all $\alpha \in [\alpha_0, 1)$ for some $0 \leq \alpha_0 < 1$. Such a policy is called a *Blackwell optimal policy*.

Theorem 4.1

There exists a policy $f_0^{\infty} \in C(D)$ such that $v^{\alpha}(f_0^{\infty}) = v^{\alpha}$ for all $\alpha \in [\alpha_0, 1)$ and for some $\alpha_0 \in [0,1)$.

Proof

For any $f^\infty \in C(D)$, $v^\alpha(f^\infty)$ is the unique solution of the linear system $\{I - \alpha P(f)\}x = r(f)$. By Cramer's rule, $v_i^\alpha(f^\infty)$ is a rational function in α for each component i.

Suppose there a no deterministic and stationary Blackwell optimal policy. For each $\alpha \in [0.1)$ there exists a discounted optimal policy. Hence, there is a sequence $\{\alpha_k,\ k = 1, 2, \dots\}$ and a sequence $\{f_k,\ k = 1, 2, \dots\}$ such that

$$\alpha_k \uparrow 1 \text{ and } v^\alpha = v^\alpha(f_k^\infty) > v^\alpha(f_{k-1}^\infty) \text{ for } \alpha = \alpha_k,\ k = 2, 3, \dots.$$

Since $C(D)$ is finite, there are different policies f^∞ and g^∞ such that for any increasing subsequence $\alpha_{k_n},\ n = 1, 2, \dots$ with $lim_{n\to\infty}\, \alpha_{k_n} = 1$,

$$\begin{cases} v^\alpha(f^\infty) > v^\alpha(g^\infty) & \text{for } \alpha = \alpha_{k_1}, \alpha_{k_3}, \dots \\ v^\alpha(f^\infty) < v^\alpha(g^\infty) & \text{for } \alpha = \alpha_{k_2}, \alpha_{k_4}, \dots \end{cases} \tag{4.1}$$

Let $h(\alpha) = v^\alpha(f^\infty) - v^\alpha(g^\infty)$, then for each $i \in S$ the function $h_i(\alpha)$ is a continuous rational function in α on $[0, 1)$. From (4.1) it follows that $h_i(\alpha)$ has an infinite number of zeros, which contradicts the rationality of $h_i(\alpha)$, unless $h_i(\alpha) \equiv 0$. Hence, there exists a Blackwell optimal policy $f_0^\infty \in C(D)$. □

Lemma 4.2

There exists an optimal policy $f_*^\infty \in C(D)$.

Proof

By Theorem 4.1, there exists a policy f_0^∞ such that $v^\alpha(f_0^\infty) = v^\alpha$ for all $\alpha \in [\alpha_0, 1)$ for some $\alpha_0 \in [0, 1)$. Hence, by Lemma 4.1, we obtain

$$v(f_0^\infty) = lim_{\alpha\uparrow 1}\, v^\alpha(f_0^\infty) = lim_{\alpha\uparrow 1}\, v^\alpha \geq lim_{\alpha\uparrow 1}\, v^\alpha(R) = v(R) \text{ for every policy } R,$$

i.e. $f_*^\infty = f_0^\infty$ is an optimal policy. □

Remark

In Theorem 4.1 and consequently also in Lemma 4.2, results from discounted MDPs are used. However, in a discounted MDP the transition probabilities sum up to 1 and the results are perhaps not valid in our substochastic model. Therefore, we introduce the following *extended* stochastic model (S^*, A^*, p^*, r^*):

$$S^* = S \cup \{0\};\ A^*(i) = \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases};\ p_{ij}^*(a) = \begin{cases} p_{ij}(a) & i \neq 0,\ j \neq 0,\ a \in A^*(i) \\ 1 - \sum_{j\in S}\, p_{ij}(a) & i \neq 0,\ j = 0,\ a \in A^*(i) \\ 1 & i = 0,\ j = 0,\ a \in A^*(i) \\ 0 & i = 0,\ j \neq 0,\ a \in A^*(i) \end{cases}$$

$$r_i^*(a) = \begin{cases} r_i(a) & i \neq 0,\ a \in A^*(i) \\ 0 & i \neq 0,\ a \in A^*(i) \end{cases}$$

It is straighforward to verify that in this extended stochastic model the total expected reward satisfies $v_i^*(R) = v_i(R),\ i \in S;\ v_0^*(R) = 0$ for every policy R. Hence, the result of Lemma 4.2 is valid.

Theorem 4.2

The value vector v satisfies the optimality equation $x_i = max_a \{r_i(a) + \sum_j p_{ij}(a)x_j\},\ i \in S$.

Proof

Lemma 4.2 implies that $v = v(f^\infty)$ for some $f^\infty \in C(D)$. Hence,

$$v_i = v_i(f^\infty) = r_i(f) + \sum_j p_{ij}(f)v_j(f^\infty) \leq max_a \{r_i(a) + \sum_j p_{ij}(a)v_j(f^\infty)\},\ i \in S. \qquad (4.2)$$

Let $a_i \in A(i),\ i \in S$, be such that $r_i(a_i) + \sum_j p_{ij}(a_i)v_j = max_a \{r_i(a) + \sum_j p_{ij}(a)v_j\},\ i \in S$.

Take policy $R = (\pi^1, \pi^2, \dots) \in C(M)$ such that $\pi^1_{ia} = \begin{cases} 1, & a = a_i \\ 0, & a \neq a_i \end{cases}$; $\pi^t_{ia} = \begin{cases} 1, & a = f(i) \\ 0, & a \neq f(i) \end{cases}$ $t \geq 2$.

Then, we can write

$$v_i \geq v_i(R) = r_i(a_i) + \sum_j p_{ij}(a_i)v_j(f^\infty) = max_a \{r_i(a) + \sum_j p_{ij}(a)v_j(f^\infty)\},\ i \in S. \qquad (4.3)$$

From (4.2) and (4.3) it follows that $v_i = max_a \{r_i(a) + \sum_j p_{ij}(a)v_j(f^\infty)\},\ i \in S$. □

Corollary 4.1

For any $g^\infty \in C(D)$, the total reward $v(g^\infty)$ satisfies the equation $x = r(g) + P(g)x$.

Proof

Take any $g^\infty \in C(D)$ and consider the model with in state i only action $g(i),\ i \in S$. By applying Theorem 4.2 to this model, we obtain

$$v_i(g^\infty) = v_i = r_i(g) + \{P(g)v\}_i = r_i(g) + \{P(g)v(g^\infty)\}_i,\ i \in S. \qquad \square$$

Opmerking:

Unfortunaltely, in contrast to discounted models, the solutions of the equations are not unique, in general. For instance, in the pure stochastic case ($\sum_j p_{ij}(a) = 1$ for all $(i,a) \in S \times A$), if x is a solution, also $x + c \cdot e$ is a solution for any scalar c. The reason is that the monotone mappings $L_f\, x := r(f) + P(f)x$ and $(Ux)_i := max_a \{r_i(a) + \sum_j p_{ij}(a)x_j\},\ i \in S$ are no contractions (the monotonicity is easy to verify).

A policy $f^\infty \in C(D)$ is called *conserving* if $r(f) + P(f)v = v$. From Corollary 4.1 it follows that an optimal policy is conserving. The reverse statement is not true as the next example shows.

Example 4.1

$S = \{1,2\};\ A(1) = \{1,2\};\ A(1) = \{1\};\ p_{11}(1) = 1;\ p_{12}(2) = 1;\ p_{21}(1) = p_{22}(1) = 0;$

$r_1(1) = 0;\ r_1(2) = 2;\ r_2(1) = -1$. It is easy to verify that $v = (1, -1)$.

The policy $f_1(1) = 1,\ f_1(2) = 1$ has $v(f_1^\infty) = (0, -1)$: this policy is nonoptimal.

The policy $f_2(1) = 2,\ f_2(2) = 1$ has $v(f_2^\infty) = (1, -1)$: this policy is optimal.

The policy f_1^∞ is conserving, because $r(f_1) + P(f_1)v = \binom{0}{-1} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\binom{1}{-1} = \binom{0}{-1} + \binom{1}{0} = \binom{1}{-1} = v$.

4.2 Equivalent statements for contracting

An MDP is said to be *contracting* if there exists a vector $\mu \in \mathbb{R}^N$ with $\mu_i > 0,\ i \in S$, and a scalar $\alpha \in [0,1)$ such that $\sum_j p_{ij}(a)\mu_j \leq \alpha \cdot \mu_i$ for all $(i,a) \in S \times A$.

Remark

Discounted MDPs can be considered as contracting MDPs with the total reward criterion. Redefine the transition probabilities by $p'_{ij}(a) = \alpha \cdot p_{ij}(a)$ for all i, j and a, and take $\mu_i = 1$ for all i. Then, $\sum_j p'_{ij}(a)\mu_j = \alpha \cdot \sum_j p_{ij}(a) = \alpha = \alpha \cdot \mu_i$ for all $(i,a) \in S \times A$. So, the model is contracting. Furthermore, for the total reward $v'(R)$ in the redefined model, we have $v'(R) = v^\alpha(R)$ for every policy R.

Introduce a vector $y^N \in \mathbb{R}^N$ inductively by

$$\begin{cases} y_i^0 = 1,\ i \in S \\ y_i^t = max_a \sum_j p_{ij}(a) y_j^{t-1},\ i \in S,\ t = 1, 2, \ldots, N \end{cases} \tag{4.4}$$

The next theorem provides five equivalent descriptions for the assumption that every policy is transient.

Theorem 4.3

The following five statements are equivalent:

(1) Every policy $f^\infty \in C(D)$ is transient.

(2) Every policy R is transient.

(3) $max_{i \in S}\, y_i^N < 1$, where y^N is defined by (4.2).

(4) The MDP model is contracting.

(5) The linear program $max \left\{ \sum_{(i,a)} x_i(a) \;\middle|\; \begin{array}{lcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & \leq & 1,\ j \in S \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}$

has a finite optimum.

Proof

(1) $\Rightarrow$ (2):

Let k be an arbitrary state and take as rewards $r_i(a) = \begin{cases} 1 & i = k,\ a \in A(i); \\ 0 & i \neq k,\ a \in A(i). \end{cases}$

Then, for any starting state i and policy R, we have

$$v_i(R) = \sum_{t=1}^\infty \sum_{j,a} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j(a) = \sum_{t=1}^\infty \mathbb{P}_{i,R}\{X_t = k\}.$$

Let $f_*^\infty \in C(D)$ be an optimal policy (f_*^∞ exists by Lemma 4.2). Since we have assumed that all policies are transient, we obtain

$$\sum_{t=1}^\infty \mathbb{P}_{i,R}\{X_t = k\} = v_i(R) \leq v_i(f_*^\infty) = \sum_{t=1}^\infty \mathbb{P}_{i,f_*^\infty}\{X_t = k\} < \infty,\ i \in S,$$

i.e. every policy R is transient.

(2) $\Rightarrow$ (3):

This (long) proof will be given by four propositions.

Proposition 1: $y_i^t = sup_R \sum_j \mathbb{P}_{i,R}\{X_{t+1} = j\}$ for all i and t.

Proof Proposition 1:

From Corollary 1.1 it follows that it is sufficient to show that $y_i^t = sup_{R \in C(M)} \sum_j \mathbb{P}_{i,R}\{X_{t+1} = j\}$.

We apply induction on t. For $t = 0$: $\sum_j \mathbb{P}_{i,R}\{X_1 = j\} = \mathbb{P}_{i,R}\{X_1 = i\} = 1 = y_i^0,\ i \in S$.

Suppose that the result is correct for t. Take any $i \in S$ and policy $R = (\pi^1, \pi^2, \pi^3, \dots) \in C(M)$.

$\sum_j \mathbb{P}_{i,R}\{X_{t+2} = j\} = \sum_j \big\{ \sum_k p_{ik}(\pi^1)\, \mathbb{P}_{k,R'}\{X_{t+1} = j\}\big\} = \sum_k p_{ik}(\pi^1)\big\{ \sum_j \mathbb{P}_{k,R'}\{X_{t+1} = j\}\big\}$,

where $R' = (\pi^2, \pi^3, \dots)$. Hence, by the induction hypothesis,

$$\sum_j \mathbb{P}_{i,R}\{X_{t+2} = j\} \le \sum_k p_{ik}(\pi^1) y_k^t \le max_a \sum_k p_{ik}(a) y_k^t = y_i^{t+1}.$$

Hence, $sup_R \sum_j \mathbb{P}_{i,R}\{X_{t+1} = j\} \le y_i^t$ for all i and t.

On the other hand, also by induction on t, we show that $\sum_j \mathbb{P}_{i,R_t}\{X_{t+1} = j\} = y_i^t$ for some deterministic Markov policy $R_t = (f_t, f_{t-1}, \dots, f_1, f_1, \dots)$.

For $t = 0$ we have shown above that $\sum_j \mathbb{P}_{i,R}\{X_1 = j\} = y_i^0 = 1$ for any policy R.

Let f_t be such that $y_i^t = \sum_j p_{ij}\big(f_t(i)\big) y_j^{t-1},\ i \in S$. Then, $R_t = (f_t, R_{t-1})$ and

$$\begin{aligned} y_i^t &= \sum_j p_{ij}\big(f_t(i)\big) y_j^{t-1} = \sum_j p_{ij}\big(f_t(i)\big)\big\{ \sum_k \mathbb{P}_{j,R_{t-1}}\{X_t = k\}\big\} \\ &= \sum_k \sum_j p_{ij}\big(f_t(i)\, \mathbb{P}_{j,R_{t-1}}\{X_t = k\} = \sum_k \mathbb{P}_{i,R_t}\{X_{t+1} = k\}. \end{aligned}$$

□

From the proof also follows: $y_i^t = sup_R\big\{ \sum_j \mathbb{P}_{i,R}\{X_{t+1} = j\} \mid R \text{ is a deterministic Markov policy } \big\}$.

Therefore, to prove (3), it is sufficient to show that $\sum_j \mathbb{P}_{i,R}\{X_{N+1} = j\} < 1,\ i \in S$ for all deterministic Markov policies R. Take any $i \in S$ and any deterministic Markov policy $R = (f_1, f_2, \dots)$.

Consider the extended model with state space $S^* = S \cup \{0\}$ and let R^* the policy in the extended model which corresponds to R. Define the following subsets of S^*:

$$T_1 = \{i\} \text{ and } T_k = \{j \in S^* \mid \mathbb{P}_{i,R^*}\{X_k = j\} > 0\} \text{ for } k = 2, 3, \dots.$$

Proposition 2: If, for all $1 \le n \le N$, $0 \notin \cup_{l=1}^n T_l$ implies $T_{n+1} \not\subseteq \cup_{l=1}^n T_l$: statement (3) is true.

Proof Proposition 2:

Suppose $0 \notin \cup_{l=1}^N T_l$. Since the state 0 is absorbing, this implies that $0 \notin \cup_{l=1}^n T_l,\ 1 \le n \le N$.

Then, by the assumption of the proposition, $\cup_{l=1}^{n+1} T_l$ has at least one state more than $\cup_{l=1}^n T_l$ for all $n = 1, 2, \dots, N$. Consequently, $\cup_{l=1}^{N+1} T_l = S^*$ and $0 \in T_{N+1}$, i.e. $\mathbb{P}_{i,R^*}\{X_{N+1} = 0\} > 0$, and therefore $\sum_{j \in S} \mathbb{P}_{i,R}\{X_{N+1} = j\} < 1$.

□

Proposition 3: Let, for some $1 \le n \le N$, $0 \notin \cup_{l=1}^n T_l$ and $T_{n+1} \subseteq \cup_{l=1}^n T_l$.

Take f_* such that $f_*(j) = \begin{cases} f_k(j) & \text{if } j \in T_k \setminus \cup_{l=1}^{k-1} T_l; \\ \text{arbitrarily chosen} & j \notin \cup_{l=1}^n T_l. \end{cases}$

Define $T_1^* = \{i\}$ and $T_k^* = \{j \in S^* \mid \mathbb{P}_{i,f_*^\infty}\{X_k = j\} > 0\}$ for $k = 2, 3, \dots$.

Then, $T_k^* \subseteq \cup_{l=1}^n T_l$ for $k = 1, 2, \dots$.

Proof Proposition 3:

The proof is by induction on k. For $k=1$: $T_1^* = T_1 \subseteq \cup_{l=1}^n T_l$ for all $n \geq 1$. Suppose that $T_k^* \subseteq \cup_{l=1}^n T_l$ for $k = 1, 2, \ldots, m$. Take any $j \in T_{m+1}$. Then, there exists a state $s \in T_m^*$ such that $p_{sj}\big(f_*(s)\big) > 0$. Since $s \in \cup_{l=1}^n T_l$, we have $f_*(s) = f_k(s)$ for some k satisfying $s \in T_k \setminus \cup_{l=1}^{k-1} T_l$. Since $s \in T_k$ and $f_*(s) = f_k(s)$, we obtain $\mathbb{P}_{i,R^*}\{X_{k+1} = j\} \geq \mathbb{P}_{i,R^*}\{X_k = s\} \cdot p_{sj}\big(f_*(s)\big) > 0$. Hence, $j \in T_{k+1} \subseteq \cup_{l=1}^{n+1} T_l = \cup_{l=1}^n T_l$, which completes the proof that $T_{m+1}^* \subseteq \cup_{l=1}^n T_l$. □

Proposition 4: Suppose that we have the same assumptions as in Proposition 3.

Let f^∞ the policy in the substochastic model corresponding with policy f_*^∞ of the extended model, defined in Proposition 3. Then, f^∞ is a nontransient policy.

Proof Proposition 4:

Since $0 \notin \cup_{l=1}^n T_l$ and $T_k^* \subseteq \cup_{l=1}^n T_l$ for all $k \in \mathbb{N}$, we have $\mathbb{P}_{i,f_*^\infty}\{X_k = 0\},\ k \in \mathbb{N}$, i.e. $\sum_{j \in S} \mathbb{P}_{i,f^\infty}\{X_k = j\} = 1,\ k \in \mathbb{N}$. Hence, $\sum_{t=1}^\infty \sum_{j \in S} \mathbb{P}_{i,f^\infty}\{X_t = j\} = +\infty$, implying that f^∞ is nontransient. □

We can complete the proof of statement (3) as follows. Since every policy is transient (by (2)), Proposition 3 implies that $0 \notin \cup_{l=1}^n T_l$ and $T_{n+1} \subseteq \cup_{l=1}^n T_l$ is impossible for all n. Hence, $0 \notin \cup_{l=1}^n T_l$ implies $T_{n+1} \not\subseteq \cup_{l=1}^n T_l,\ 1 \leq n \leq N$. Then, by Proposition 1, statement (3) holds.

(3) $\Rightarrow$ (4):

Let $a = max_i\, y_i^N < 1$ and $b = a^{1/(N+1)}$, Then, $a < b < 1$. Take α such that $b < \alpha < 1$ and define $\mu \in \mathbb{R}^N$ by $\mu_i = sup_R \sum_{t=1}^\infty (1/\alpha)^{t-1} \sum_j \mathbb{P}_{i,R}\{X_t = j\},\ i \in S$. From Proposition 1 it follows that

$$\begin{aligned} a &= max_i\, sup_R \sum_j \mathbb{P}_{i,R}\{X_{N+1} = j\} = max_i\, max_{R \in C(M)} \sum_j \mathbb{P}_{i,R}\{X_{N+1} = j\} \\ &= max_{R \in C(M)}\, max_i \sum_j \mathbb{P}_{i,R}\{X_{N+1} = j\}. \end{aligned}$$

For $R = (\pi^1, \pi^2, \ldots) \in C(M)$, let $P^t(R) = P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})$. Then,

$$\begin{aligned} a &= max_{R \in C(M)}\, max_i \sum_j \mathbb{P}_{i,R}\{X_{N+1} = j\} = max_{R \in C(M)}\, max_i \sum_j p_{ij}^{N+1}(R) \\ &= max_{R \in C(M)} \|P^{N+1}(R)\|_\infty. \end{aligned}$$

For any $t \in \mathbb{N}$, we also have $\|P^t(R)\|_\infty \leq \|P^{\lfloor t/(N+1) \rfloor \cdot (N+1)}(R)\|_\infty \leq a^{\lfloor t/(N+1) \rfloor} \leq a^{-1} \cdot b^t$.

Consequently,

$$\sum_{t=1}^\infty (1/\alpha)^{t-1} \sum_j \mathbb{P}_{i,R}\{X_t = j\} \leq \sum_{t=1}^\infty (1/\alpha)^{t-1} \|P^t(R)\|\infty \leq a^{-1} \cdot b \cdot \sum_{t=1}^\infty (b/\alpha)^{t-1} = \tfrac{\alpha b}{a(\alpha - b)}.$$

Therefore, μ is well-defined. In Chapter 3 it was shown that $v_i^\alpha = max_a \{r_i(a) + \alpha \sum_j p_{ij}(a) v_j^\alpha\}$, where $v_i^\alpha = sup_R \sum_{t=1}^\infty \alpha^{t-1}\, \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j(a)$. Similarly, with $(1/\alpha)$ instead of α and with $r_j(a) = 1$ for all (j, a), it can be shown for $\mu_i = sup_R \sum_{t=1}^\infty (1/\alpha)^{t-1} \sum_j \mathbb{P}_{i,R}\{X_t = j\}$ that $\mu_i = max_a \{1 + (1/\alpha) \sum_j p_{ij}(a)\mu_j\},\ i \in S$. Then, for all $(i, a) \in S \times A$, we have $\alpha\mu_i \geq \alpha + \sum_j p_{ij}(a)\mu_j \geq \sum_j p_{ij}(a)\mu_j$, i.e. the model is contracting.

(4) $\Rightarrow$ (5):
Suppose that the linear program has no finite solutions. Since the linear program is feasible ($x = 0$ is a feasible solution) it follows from the theory of linear programming that there exists a $s \neq 0$ such that

$$s_i(a) \geq 0 \text{ for all } (i,a) \in S \times A, \text{ and } \textstyle\sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}s_i(a) \leq 0,\ j \in S.$$

Define the stationary policy π^∞ by $\pi_{ia} = \begin{cases} \frac{s_i(a)}{\sum_a s_i(a)} & \text{if } \sum_a s_i(a) > 0,\ a \in A(i); \\ \text{arbitrary} & \text{if } \sum_a s_i(a) = 0,\ a \in A(i). \end{cases}$

Hence, we may write with $s_i = \sum_a s_i(a)$,

$$\begin{array}{rcl} 0 & \leq & \sum_a s_j(a) \leq \sum_i \sum_a p_{ij}(a)s_i(a) = \sum_i \sum_a p_{ij}(a)\{\pi_{ia} \cdot s_i\} \\ & = & \sum_i\{\sum_a p_{ij}(a)\pi_i(a)\} \cdot s_i = \sum_i p_{ij}(\pi) \cdot s_i,\ j \in S, \end{array},$$

or, in vector notation $0 \leq s \leq sP(\pi)$. Iterating this inequality gives $0 \leq s \leq sP^n(\pi)$ for all n.
Since the model is contracting, there exists a vector μ with $\mu_i > 0,\ i \in S$, and a scalar $\alpha \in [0,1)$ such that $\sum_j p_{ij}(a)\mu_j \leq \alpha \cdot \mu_i$ for all $(i,a) \in S \times A$. Hence, $0 \leq P(\pi)\mu \leq \alpha \cdot \mu$, and consequently, $0 \leq P^n(\pi)\mu \leq \alpha^n \cdot \mu$ for all n, implying that $P^n(\pi) \to 0$ for $n \to \infty$.
Hence, $0 \leq s \leq sP^n(\pi)$ for all n implies $s = 0$, which gives a contraction.
(5) $\Rightarrow$ (1):
Suppose that there exists a polify $f^\infty \in C(D)$ which is nontransient: $\sum_{t=1}^\infty \mathbb{P}_{i,f^\infty}\{X_t = j\} = \infty$ for some $i, j \in S$. Then, we obtain,

$$\textstyle\sum_{t=1}^\infty e^T P^{t-1}(f)e = \sum_{t=1}^\infty \sum_k \sum_l p_{kl}^{t-1}(f) \geq \sum_{t=1}^\infty p_{ij}^{t-1}(f) = +\infty.$$

Consider the sequence $\{x^n,\ n = 1, 2, \dots\}$ defined by $x_i^n(a) = \begin{cases} \sum_{t=1}^n \sum_k p_{ki}^{t-1}(f) & a = f(i); \\ 0 & a \neq f(i). \end{cases}$

$x_i^n(a) \geq 0$ for all (i,a) and satisfies

$$\begin{array}{rcl} \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}x_i^n(a) & = & \sum_i \{\delta_{ij} - p_{ij}(f)\}\{\sum_{t=1}^n \sum_k p_{ki}^{t-1}(f)\} \\ & = & \sum_k \sum_{t=1}^n \sum_i p_{ki}^{t-1}(f)\{\delta_{ij} - p_{ij}(f)\} \\ & = & \sum_k \sum_{t=1}^n \{p_{kj}^{t-1}(f) - p_{kj}^t(f)\} \\ & = & \sum_k\{\delta_{kj} - p_{kj}^n(f)\} = 1 - \sum_k p_{kj}^n(f) \leq 1,\ j \in S. \end{array}$$

and

$$\begin{array}{rcl} \sum_{(i,a)} x_i^n(a) & = & \sum_i \{\sum_{t=1}^n \sum_k p_{ki}^{t-1}(f)\} = \sum_{t=1}^n \sum_k \sum_i p_{ki}^{t-1}(f) \\ & = & \sum_{t=1}^n e^T P^{t-1}(f)e \ \to\ \infty \text{ for } n \to \infty \end{array}$$

We have a sequence $\{x^n,\ n = 1, 2, \dots\}$ of feasible solutions with objective functions tending to $+\infty$. This contradicts the assumption of (5) that the optimum of the linear program is finite. □

The characterizations (3) and (5) provide finite algorithms for checking the contraction property of a given MDP model. Below we present these algorithms explicitly.

Algorithm 4.1 *Checking the contracting property (iterative approach)*

1. $y_i^0 = 1,\ i \in S,\ t = 1.$
2. $y_i^t = max_a \sum_j p_{ij}(a) y_j^{t-1},\ i \in S.$
3. If $max_i\, y_i^t < 1$: the model is contracting (STOP);

   else, go to step 4.
4. If $t = N$: the problem is not contracting (STOP);

   else, $t := t + 1$ and return to step 2.

Algorithm 4.2 *Checking the contracting property (linear programming approach)*

1. Solve the following linear program

$$max \left\{ \sum_{(i,a)} x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & \leq & 1,\ j \in S \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}.$$

2. If the program has a finite optimum: the problem is contracting (STOP);

   else, the problem is not contracting (STOP).

Remark

If it happens in Algorithm 4.2 that the model is contracting, we obtain from the dual program the μ and α, namely:

The dual linear program

$$min \left\{ \sum_j \mu_j \;\middle|\; \begin{array}{rcl} \sum_j \{\delta_{ij} - p_{ij}(a)\} \mu_j & \geq & 1,\ (i,a) \in S \times A \\ \mu_j & \geq & 0,\ j \in S \end{array} \right\}$$

has also a finite solution. From the constraints it follows that $\mu_i \geq 1 + \sum_j p_{ij}(a)\mu_j \geq 1,\ i \in S$. Let $\alpha = 1 - \frac{1}{max_i\, \mu_i}$. Then, $\sum_j p_{ij}(a)\mu_j \leq \mu_i - 1 \leq \mu_i - \frac{\mu_i}{max_i\, \mu_i} = \alpha \cdot \mu_i,\ (i,a) \in S \times A$.

A discounted model is a contracting model with $p'_{ij}(a) = \alpha p_{ij}(a)$ for all i, j, and a and with $\mu_i = 1,\ i \in S$. In fact the discounting and the contracting models are equivalent: a contracting substochastic model can be transformed into a stochastic discounted model such that for any policy R the total expected reward in the original model differs a multiplicative factor with the total discounted reward in the transformed model.

To prove this equivalence, we introduce the following transformed model $(\overline{S}, \overline{A}, \overline{p}, \overline{r})$:

$$\overline{S} = S \cup \{0\};\ \overline{A}(i) = \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases};\ \overline{r}_i(a) = \begin{cases} \frac{1}{\mu_i} r_i(a) & i \neq 0,\ a \in A^*(i) \\ 0 & i \neq 0,\ a \in A^*(i) \end{cases}$$

$$\overline{p}_{ij}(a) = \begin{cases} \frac{1}{\alpha\mu_i} p_{ij}(a)\mu_j & i \neq 0,\ j \neq 0,\ a \in A^*(i) \\ 1 - \frac{1}{\alpha\mu_i} \sum_{k \in S}\ p_{ik}(a) & i \neq 0,\ j = 0,\ a \in A^*(i) \\ 1 & i = 0,\ j = 0,\ a \in A^*(i) \\ 0 & i = 0,\ j \neq 0,\ a \in A^*(i) \end{cases}$$

$$\begin{array}{lrcl} \text{For } i = 0: & \sum_{j \in \overline{S}}\ p_{ij}(a) & = & \sum_{j \in \overline{S}}\ p_{0j}(1) = 1. \\ \text{For } i \neq 0: & \sum_{j \in \overline{S}}\ p_{ij}(a) & = & \sum_{j \in S}\ p_{ij}(a) + p_{i0}(a) \\ & & = & \sum_{j \in S}\ \frac{1}{\alpha\mu_i} p_{ij}(a)\mu_j + \left\{1 - \frac{1}{\alpha\mu_i} \sum_{k \in S}\ p_{ik}(a)\right\} = 1 \text{ for all } a \in \overline{A}(i). \end{array}$$

Hence, the transformed model is stochastic. In order to analyze the rewards, we may restrict ourselves to Markov policies. Let $R = (\pi^1, \pi^2, \dots)$ be a Markov policy. Then, by induction on t, it is straightforward to show that

$$\{\overline{P}(\pi^1)\overline{P}(\pi^2)\cdots\overline{P}(\pi^t)\}_{ij} = \left(\tfrac{1}{\alpha}\right)^t \cdot \tfrac{1}{\mu_i} \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^t)\}_{ij} \cdot \mu_j \text{ for all } i, j \in S \text{ and } t \in \mathbb{N}.$$

Therefore, we can write,

$$\begin{array}{rcl} \overline{v}_i^{\alpha}(R) & = & \sum_{t=1}^{\infty}\ \alpha^{t-1} \cdot \{\overline{P}(\pi^1)\overline{P}(\pi^2)\cdots\overline{P}(\pi^{t-1})\overline{r}(\pi^t)\}_i \\ & = & \sum_{t=1}^{\infty}\ \sum_{j \in S}\ \frac{1}{\mu_i} \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij} \cdot \mu_j \cdot \overline{r}_j(\pi^t) \\ & = & \sum_{t=1}^{\infty}\ \sum_{j \in S}\ \frac{1}{\mu_i} \cdot \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij} \cdot \mu_j \cdot \frac{1}{\mu_j} r_j(\pi^t) \\ & = & \frac{1}{\mu_i} \cdot \sum_{t=1}^{\infty}\ \sum_{j \in S} \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ij} \cdot r_j(\pi^t) \\ & = & \frac{1}{\mu_i} \cdot \sum_{t=1}^{\infty} \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t)\}_i,\ i \in S \\ & = & \frac{1}{\mu_i} \cdot v_i(R),\ i \in S. \end{array}$$

4.3 The contracting model

Throughout this section we assume that the model is contracting. We have seen that the discounting and contracting models are equivalent. In fact, results for discounted MDPs as the optimality equation, policy iteration, linear programming and value iteration can directly be applied to contracting MDPs. We will summarize this result in the following theorem and algorithms.

Theorem 4.4

(1) The value vector v is the unique solution of the optimality equation

$$x_i = max_a \{r_i(a) + \sum_j\ p_{ij}(a)x_j\},\ i \in S.$$

(2) The value vector is the (componentwise) smallest vector which satisfies

$$x_i \geq r_i(a) + \sum_j\ p_{ij}(a)x_j,\ (i,a) \in S \times A.$$

Algorithm 4.3 *Policy iteration algorithm*

1. Start with any $f^\infty \in C(D)$.

2. Compute $v(f^\infty)$ as the unique solution of the linear system $x = r(f) + P(f)x$.

3. a. Compute $s_{ia}(f) = r_i(a) + \sum_j p_{ij}(a)v_j(f^\infty) - v_i(f^\infty)$ for every $(i,a) \in S \times A$.

   b. Determine $A(i,f) = \{a \in A(i) \mid s_{ia}(f) > 0\}$ for every $i \in S$.

4. If $A(i,f) = \emptyset$ for every $i \in S$: go to step 6.

   Otherwise: take g such that $s_{ig(i)}(f) = max_a\, s_{ia}(f),\ i \in S$.

5. $f := g$ and return to step 2.

6. $v(f^\infty)$ is the value vector and f^∞ an optimal policy (STOP).

Algorithm 4.4 *Linear programming algorithm*

1. Take any vector β, where $\beta_j > 0,\ j \in S$.

2. Use the simplex method to compute optimal solutions v^* and x^* of the dual pair of linear programs:

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \sum_j \{\delta_{ij} - p_{ij}(a)\}v_j \geq r_i(a),\ (i,a) \in S \times A \right\}$$

   and

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & = \beta_j,\ j \in S \\ x_i(a) & \geq 0,\ (i,a) \in S \times A \end{array} \right\}.$$

3. Take $f_*^\infty \in C(D)$ such that $x_i^*(f_*(i)) > 0$ for every $i \in S$.

   v^* is the value vector and f_*^∞ an optimal policy (STOP).

In value iteration, we iterate: $v_i^{n+1} = (Uv^n)_i = max_a \{r_i(a) + \sum_j p_{ij}(a)v_j^n\},\ i \in S$ for $n = 1, 2, \ldots$. For a stop criterion we consider $\|v^{n+1} - v^n\|$ for some norm. In the contracting case we use the $\|\cdot\|_\mu$-norm, defined by $\|x\|_\mu = max_i \frac{1}{\mu_i} \cdot |x_i|$ (see also Exercise 3.1, in which the reader was asked to show that this is a correct norm). For this norm we obtain with operator L_f, defined by $L_f\, x = r(f) + P(f)x$,

$$\begin{array}{rcl} \|L_f\, x - L_f\, y\|_\mu & = & max_i \frac{1}{\mu_i} \cdot |\sum_j p_{ij}(f)(x_j - y_j)| \ \leq\ max_i \frac{1}{\mu_i} \cdot \sum_j p_{ij}(f)|x_j - y_j| \\ & = & max_i \frac{1}{\mu_i} \cdot \sum_j p_{ij}(f)\mu_j \cdot \frac{1}{\mu_j} \cdot |x_j - y_j| \\ & \leq & max_i \frac{1}{\mu_i} \cdot \sum_j p_{ij}(f)\mu_j \cdot \|x - y\|_\mu \leq \alpha \cdot \|x - y\|_\mu, \end{array}$$

i.e. L_f is a contraction with respect to the $\|\cdot\|_\mu$-norm with contraction factor α. Similarly, it can be shown that U also is a contraction with respect to the $\|\cdot\|_\mu$-norm with contraction factor α. The value iteration algorithm in the transient case is similar to the discounted case.

Algorithm 4.5 *Value iteration*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.
2. a. Compute $y_i = max_a\{r_i(a) + \sum_j p_{ij}(a)x_j\},\ i \in S$.
   b. Let $f(i) = argmax_a\{r_i(a) + \sum_j p_{ij}(a)x_j\},\ i \in S$.
3. If $\|y - x\|_\mu \leq \frac{1}{2}(1-\alpha)\alpha^{-1}\varepsilon$: f^∞ is an ε-optimal policy and y is a $\frac{1}{2}\varepsilon$-approximation of the value vector (STOP);
   Otherwise: $x := y$ and return to step 2.

Example 4.2

Consider the transient model with $S = \{1,2\};\ A(1) = A(2) = \{1,2\};\ p_{11}(1) = \frac{1}{2},\ p_{12}(1) = 0;$ $p_{11}(2) = 0,\ p_{12}(2) = \frac{1}{2};\ p_{21}(1) = \frac{1}{2},\ p_{22}(1) = \frac{3}{4};\ p_{21}(2) = \frac{1}{2},\ p_{22}(2) = 0;\ r_1(1) = 2,\ r_1(2) = 3$ $r_2(1) = 1,\ r_2(2) = 4$.

The optimality equation is:

$x_1 = max\{2 + \frac{1}{2}x_1, 3 + \frac{1}{2}x_2\},\ x_2 = max\{1 + \frac{3}{4}x_2, 4 + \frac{1}{2}x_1\}$ with solution $x_1 = \frac{20}{3},\ x_2 = \frac{22}{3}$.

If we apply policy iteration, stating with $f(1) = f(1) = 1$, we obtain:

Iteration 1:

$x_1 = 2 + \frac{1}{2}x_1,\ x_2 = 1 + \frac{3}{4}x_2 \ \rightarrow\ v_1(f^\infty) = 4,\ v_2(f^\infty) = 4.$

$s_{11}(f) = 0,\ s_{12}(f) = 1;\ s_{21}(f) = 0,\ s_{22}(f) = 2 \ \rightarrow\ A(1,f) = A(2,f) = \{2\}.$

$g(1) = g(2) = 2.$

Iteration 2:

$x_1 = 3 + \frac{1}{2}x_2,\ x_2 = 4 + \frac{1}{2}x_1 \ \rightarrow\ v_1(f^\infty) = \frac{20}{3},\ v_2(f^\infty) = \frac{22}{3}.$

$s_{11}(f) = -\frac{4}{3},\ s_{12}(f) = 0;\ s_{21}(f) = -\frac{5}{6},\ s_{22}(f) = 0 \ \rightarrow\ A(1,f) = A(2,f) = \emptyset.$

$(\frac{20}{3}, \frac{22}{3})$ is the value vector and f^∞ with $f(1) = f(2) = 2$ is an optimal policy.

The dual linear program with $\beta_1 = \beta_2 = \frac{1}{2}$ becomes:

$$max\ 2x_1(1) + 3x_1(2) + x_2(1) + 4x_2(2)$$

subject to the constraints:

$$\begin{array}{rcrcrcl} \frac{1}{2}x_1(1) & + & x_1(2) & & - \frac{1}{2}x_2(2) & = & \frac{1}{2} \\ -\frac{1}{2}x_1(2) & + & \frac{1}{4}x_2(1) & + & x_2(2) & = & \frac{1}{2} \\ x_1(1), x_1(2), x_2(1), x_2(2) & & & & & \geq & 0 \end{array}$$

The optimal solution of this dual program is: $x_1(1) = 0,\ x_1(2) = 1,\ x_2(1) = 0,\ x_2(2) = 1.$

The primal problem has the solution: $v_1 = \frac{20}{3},\ v_2 = \frac{22}{3}$.

Hence, the value vector is $\left(\frac{20}{3}, \frac{22}{3}\right)$ and the optimal solution takes in both states action 2.

Finally, we present value iteration for this model. Let $v^0 = (4,4)$ and $\varepsilon = 0.2$.

The iteration is: $y_1 = max\{2 + \frac{1}{2}x_1, 3 + \frac{1}{2}x_2\},\ x_2 = max\{1 + \frac{3}{4}x_2, 4 + \frac{1}{2}x_1\}$.

There are several possibilities for μ and α. If we use Algorithm 4.2 then we obtain (by the solution of the dual): $\mu_1 = 3,\ \mu_2 = 4,\ \alpha = \frac{3}{4}$. The algorithm terminates if the μ-norm of the difference

of two subsequent y-vectors is at most $\frac{1}{6}\varepsilon = \frac{1}{30}$. Since $\|y - x\|_\mu = max\{\frac{1}{3}|y_1 - x_1|, \frac{1}{4}|y_2 - x_2|\}$, the procedure is terminated as soon as $|y_1 - x_1| \leq \frac{1}{10}$ and $|y_2 - x_2| \leq \frac{2}{15}$.

The results of the computation are summarized below.

	Iteration				
	1	2	3	4	5
y_1	5.00	6.00	6.25	6.50	6.57
y_2	6.00	6.50	7.00	7.13	7.25
f_1	2	2	2	2	2
f_2	2	2	2	2	2

Hence, f^∞ with $f(1) = f(2) = 2$ is a 0.2-optimal policy and $(6.57, 7.25)$ is a 0.1-approximation of the value vector.

4.4 Positive MDPs

Throughout this section we assume the following.

Assumption 4.2 *$r_i(a) \geq 0$ for all $(i, a) \in S \times A$.*

A vector $x \in \mathbb{R}^N$ is said to be *superharmonic* if $v_i \geq r_i(a) + \sum_j p_{ij}(a)v_j$ for all $(i, a) \in S \times A$.

Theorem 4.5

The value vector v is the (componentwise) smallest nonnegative superharmonic vector.

Proof

From Theorem 4.2 and Assumption 4.2 it follows that v is a nonnegative superharmonic vector. Suppose that x is also a nonnegative superharmonic vector. It is sufficient to show that $x \geq v(f^\infty)$ for every $f^\infty \in C(D)$. Take an arbitray $f^\infty \in C(D)$. Then, the superharmonicity of x implies $x \geq r(f) + P(f)x$. By iterating this inequality, we obtain

$$x \geq \sum_{t=1}^{n} P^{t-1}(f)r(f) + P^n(f)x \geq \sum_{t=1}^{n} P^{t-1}(f)r(f), \ n \in \mathbb{N}.$$

Hence, let $n \to \infty$, $x \geq \sum_{t=1}^{\infty} P^{t-1}(f)r(f) = v(f^\infty)$. □

Theorem 4.5 implies that the value vector is the unique optimal solution of the linear program

$$min \left\{ \sum_j \beta_j x_j \;\middle|\; \begin{array}{rcll} \sum_j \{\delta_{ij} - p_{ij}(a)\}x_j & \geq & r_i(a), & (i,a) \in S \times A \\ x_j & \geq & 0, & j \in S \end{array} \right\}, \tag{4.5}$$

where $\beta_j > 0, \ j \in S$. The dual program is

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rcll} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & \leq & \beta_j, & j \in S \\ x_i(a) & \geq & 0, & (i,a) \in S \times A \end{array} \right\}. \tag{4.6}$$

The dual program (4.6) is feasible ($x = 0$ is a feasible solution). Therefore, (4.6) either has a finite optimal solution or the there is an infinite solution. We consider both cases separately.

Case 1: (4.6) has a finite optimal solution.
In this case there is also an extreme finite optimal solution x^*, which is computed for instance by the simplex method. The next theorem shows how an optimal policy is obtained from x^*.

Theorem 4.6
Let x^* be an extreme optimal solution of (4.6). Then, any f_*^∞ such that $x_i^*\big(f_*(i)\big) > 0$ for each i with $\sum_a x_i^*(a) > 0$ is an optimal policy.

Proof
By introducing slack variables we can write the constraints of the problem (4.6) as

$$\begin{cases} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) + y_j &= \beta_j, \quad j \in S \\ x_i(a) &\geq 0, \quad (i,a) \in S \times A \\ y_j &\geq 0, \quad j \in S \end{cases}$$

par It follows from the theory of linear programming that the optima of the dual linear programs are equal, i.e. $\sum_j \beta_j v_j = \sum_{(i,a)} r_i(a) x_i^*(a)$. Since x^* is an extreme point and the dual program (4.6) has N constraints, the optimal extreme solution (x^*, y^*) has at most N positive components. Because

$$\sum_a x_j^*(a) + y_j^* = \beta_j + \sum_{(i,a)} p_{ij}(a) x_i^*(a) \geq \beta_j > 0, \ j \in S,$$

for each $j \in S$, either $\sum_a x_j^*(a) > 0$ or $y_j^* > 0$, implying that for each j with $\sum_a x_j^*(a) > 0$ there is exactly one action $f_*(j)$ for which $x_j^*\big(f_*(j)\big) > 0$.
Furthermore, we have $(x^*)^T = (\beta - y^*)^T + (x^*)^T P(f_*)$. By iterating this equality, we obtain

$$(x^*)^T = (\beta - y^*)^T \sum_{t=1}^n P^{t-1}(f_*) + (x^*)^T P^n(f_*) \text{ for all } n \in \mathbb{N}.$$

Consequently,

$$(x^*)^T r(f_*) = (\beta - y^*)^T \sum_{t=1}^n P^{t-1}(f_*) r(f_*) + (x^*)^T P^n(f_*) r(f_*) \text{ for all } n \in \mathbb{N}.$$

Since $v(f_*^\infty) = \sum_{t=1}^\infty p^{t-1}(f_*) r(f_*) \leq v$ and v is finite, we have $lim_{n\to\infty} P^n(f_*) r(f_*) = 0$.
Therefore, by letting $n \to \infty$,

$$\beta^T v = \sum_j \beta_j v_j = \sum_{(i,a)} r_i(a) x_i^*(a) = (x^*)^T r(f_*) = (\beta - y^*)^T v(f_*^\infty) \leq \beta^T v(f_*^\infty),$$

implying that f_*^∞ is an optimal policy. □

Case 2: (4.6) has an infinite optimal solution.
If we solve the problem by the simplex method, we obtain after a finite number of iterations a simplex tableau corresponding to an extreme feasible solution (x^*, y^*) in which one of the columns is nonpositive. In this column, the coefficient of the transformed objective function is strictly negative. This column provides a direction vector $s^* \neq 0$ such that
(1) $x^*(\lambda) := x^* + \lambda s^*$ is feasible for all $\lambda \geq 0$.
(2) $\sum_{(i,a)} r_i(a) x_i^*(a)(\lambda) \to +\infty$ for $\lambda \to +\infty$.

From (1) and (2) it follows that

$$\sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} s_i^*(a) \leq 0,\ j \in S;\ s_i^*(a) \geq 0,\ (i,a) \in S \times A;\ \sum_{(i,a)} r_i(a) s_i^*(a) > 0. \tag{4.7}$$

As we have seen in the proof of Theorem 4.6 the basis of the simplex tableau corresponding to (x^*, y^*) contains for each state $j \in S$ at most one positive $x_j^*(a)$. Let $S_* = \{j \mid \sum_a x_j^*(a) > 0\}$ and let $x_j^*(a_j) > 0,\ j \in S_*$. We first show that if the nonpositive column corresponds to the nonbasic variable $x_k(a_*)$, then $k \notin S_*$. Assume the contrary, i.e. $k \in S_*$ and $x_k^*(a_k) > 0$.

Let a^* be the nonpositive column of $x_k(a_*)$ and let i_j be the rowindex of the basis variable $x_j(a_j),\ j \in S_*$. Then, the direction s^* satisfies $s_j^*(a) = \begin{cases} -a_{i_j}^* & j \in S_*,\ a = a_j; \\ 1 & j = k,\ a = a_*; \\ 0 & \text{elsewhere.} \end{cases}$

Define the policies $f_1^\infty, f_2^\infty \in C(D)$ by $f_1(i) = \begin{cases} a_i & i \in S_*,\ a = a_i \\ a_i & i \notin S_*, \text{ where } a_i \text{ is an arbitrary action} \end{cases}$

and $f_2(i) = \begin{cases} f_1) & i \neq k; \\ a_i* & i = k. \end{cases}$ Furthermore, let $s_i^* = \sum_a s_i^*(a),\ i \in S$, and define the stationary policy π^∞ by $\pi_{ia} = \begin{cases} 1 & i \in S_*,\ i \neq k,\ a = a_i \\ \delta & i \in S_*,\ i = k,\ a = a_k \\ 1-\delta & i \in S_*,\ i = k,\ a = a_* \\ 1 & i \notin S_*,\ a = a_i \\ 0 & \text{elsewhere} \end{cases}$ where $\delta = \frac{-a_{i_j}^*}{1-a_{i_j}^*}$.

Then, $s_i^*(a) = s_i^* \cdot \pi_{ia}$ for all $(i,a) \in S \times A$ and $P(\pi) = \delta \cdot P(f_1) + (1-\delta) \cdot P(f_2)$.

Hence, by (4.7), we obtain

$$0 \geq \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} s_i^*(a) = \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} \pi_{ia} \cdot s_i^* = s_j^* - \sum_i p_{ij}(\pi) s_i^*,\ j \in S, \tag{4.8}$$

implying $0 < \sum_j s_j^* \leq \sum_i \{\sum_j p_{ij}(\pi)\} s_i^* \leq \sum_i s_i^*$. Therefore,

$$(s^*)^T e = (s^*)^T P(\pi) e \text{ and } \textstyle\sum_j p_{ij}(\pi) = 1 \text{ for every } i \text{ with } s_i^* > 0. \tag{4.9}$$

Hence, also $\sum_j p_{ij}(f_1) = \sum_j p_{ij}(f_2) = 1$ for every i with $s_i^* > 0$. From (4.8) and (4.9) it follows that $(s^*)^T \leq (s^*)^T P(\pi)$ and $(s^*)^T e = (s^*)^T P(\pi) e$, and consequently $(s^*)^T = (s^*)^T P(\pi)$. Let $S_0 = \{i \mid s_i^* > 0\}$. Then, from the theory of Markov chains, it is well known that $S_0 \subseteq R(\pi)$, where $R(\pi)$ is the set of states that are recurrent in the Markov chain induced by $P(\pi)$, and S_0 is closed under $P(\pi)$. Therefore,

$$S_0 \text{ is closed under } P(f_1) \text{ and } \sum_j p_{ij}^{(n)}(f_1) = \sum_{j \in S_0} p_{ij}^{(n)}(f_1) = 1,\ i \in S_0,\ n \in \mathbb{N}. \tag{4.10}$$

Since (x^*, y^*) is an extreme feasible solution and since $S_0 \subseteq S_*$, we have on the other hand

$$x_j^*(a_j) = \beta_j + \sum_{(i,a)} p_{ij}(a) x_i(a) \geq \beta_j + \sum_{i \in S_0} p_{ij}(f_1) x_i(a_i),\ j \in S_0. \tag{4.11}$$

Notice that S_0 is closed under $P(f_1)$ and define the vectors $\overline{x}, \overline{\beta}$ and the matrix $\overline{P}$ as the restrictions of the vectors $x^*,\ \beta$ and the matrix $P(f_1)$ to the states of S_0. Then, (4.11) is in vector notation $\overline{x}^T \geq \overline{\beta}^T + \overline{x}^T\overline{P}$. By iterating this equality, we obtain $\overline{x}^T \geq \sum_{t=1}^{n} \overline{\beta}^T\overline{P}^{t-1} + \overline{x}^T\overline{P}^n$.
Consequently, since $\beta_j > 0$ for all j, we have $\sum_{t=1}^{\infty} p_{ij}^{(t-1)}(f_1) < \infty,\ i, j \in S_0$, implying that $p_{ij}^{(n)}(f_1) \to 0$ for $n \to \infty,\ i, j \in S_0$. Hence, $\sum_{j \in S_0} p_{ij}^{(n)}(f_1) \to 0$ for $n \to \infty,\ i \in S_0$.
This contradicts (4.10) and concludes the proof that if the nonpositive column corresponds to the nonbasic variable $x_k(a_*)$, then $k \notin S_*$. Hence, the direction s^* induces a deterministic and stationary policy f_1^∞.

We next show that $v_j(f_1^\infty) = +\infty$ for at least one state j. From the above proof it follows that $S_0 \subseteq R(f_1^\infty)$ and that S_0 is closed under $P(f_1)$. From (4.7) it follows that $(s^*)^T r(f_*) > 0$.
Hence, there is a state $j \in S_0$ such that $r_j(f_1) > 0$. For all states i in the same ergodic set as j, we have

$$v_i(f_1^\infty) = \sum_{t=1}^{\infty} \{P^{t-1}(f_1)r(f_1)\}_i = lim_{n\to\infty}\, n \cdot \tfrac{1}{n} \cdot \sum_{t=1}^{n} \{P^{t-1}(f_1)r(f_1)\}_i$$

and

$$lim_{n\to\infty}\, \tfrac{1}{n} \cdot \sum_{t=1}^{n} \{P^{t-1}(f_1)r(f_1)\}_i = \{P^*(f_1)r(f_1)\}_i \geq p_{ij}^*(f_1)r_j(f_1) > 0,$$

where $P^*(f_1)$ is the stationary matrix of the Markov chain $P(f_1)$. This concludes the proof that $v_j(f_1^\infty) = +\infty$.

We construct in the following way an optimal policy f_*^∞. We first determine the ergodic sets in S_0 which have a state j such that $r_j(f_1) > 0$. For any state in these ergodic sets we define $f_*(i) = f_1(i)$. Outside these ergodic sets, we choose actions which lead to these ergodic sets, if possible. Then, f_*^∞ has for certain initial states, say the states $S_1 \subseteq S$, a total reward $+\infty$. The states $S \backslash S_1$ are closed under every policy and we repeat the same approach to the model of the states $S \backslash S_1$. The method is summarized in the following algorithm.

Algorithm 4.6 *Positive MDPs*

1. Take any vector β, where $\beta_j > 0,\ j \in S$.

2. Use the simplex method to solve the linear program

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & \leq & \beta_j,\ j \in S \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}.$$

   If a finite optimal solution x^* is obtained: go to step 3.

   If an infinite optimum is discovered: go to step 4.

3. Take any $f_*^\infty \in C(D)$ such that $x_i^*(f_*(i)) > 0$ for every i such that $\sum_a x_i^*(a) > 0$.

   The policy f_*^∞ an optimal policy (STOP).

4. Let a^* be the nonpositive column in in the simplex tableau in which the infinite optimum is discovered. Suppose that this column corresponds to the nonbasic variable $x_k(a_*)$.

   Let i_j be the rowindex of the basis variable $x_j(a_j),\ j \in S_* = \{j \mid \sum_a x_j(a) > 0\}$ and let the direction s^* be defined by $s_j^*(a) = \begin{cases} -a_{i_j}^* & j \in S_*,\ a = a_j; \\ 1 & j = k,\ a = a_*; \\ 0 & \text{elsewhere.} \end{cases}$

5. Take $f_*(i)$ such that $s_i^*\big(f_*(i)\big) > 0,\ i \in S_0 = \{i \mid \sum_a s_i^*(a) > 0\}$.

6. Determine in the Markov chain induced by $P(f_*)$ the ergodic sets on S_0.

7. Determine S_1 as the union of the ergodic sets which contain a state j for which $r_j(f_*) > 0$.

8. If $S_1 = S$: then f_*^∞ is an optimal policy (STOP).

   Otherwise: go to step 9

9. If there is a triple $(i,\ a_i,\ j)$ with $i \in S\backslash S_1,\ a_i \in A(i),\ j \in S_1$ and $p_{ij}(a_i) > 0$:

   $f_*(i) = a_i,\ S_1 := S_1 \cup \{i\}$ and go to step 8.

   Otherwise: go to step 10.

10. $S := S\backslash S_1$ and return to step 2.

4.5 Negative MDPs

Throughout this section we assume the following.

Assumption 4.3 *$r_i(a) \le 0$ for all $(i,a) \in S \times A$.*

In this case the total expected reward $v_i(R)$ exists for all $i \in S$ and all policies R, possibly $-\infty$. If there exists a transient policy R, then we have $-\infty < v_i(R) \le v_i \le 0,\ i \in S$. The next theorem shows how the existence of a transient policy can be verified.

Theorem 4.7
There exists a transient policy if and only if the linear program

$$max \left\{ \sum_{(i,a)} x_i(a) \;\middle|\; \begin{array}{lcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 1,\ j \in S \\ x_i(a) & \ge & 0,\ (i,a) \in S \times A \end{array} \right\}$$

has a feasible solution.

Proof
Let x be a feasible solution of the linear program and let $x_i = \sum_a x_i(a),\ i \in S$. Since $\sum_a x_j(a) = 1 + \sum_{(i,a)} p_{ij}(a) x_i(a) \ge 1,\ j \in S$, the stationary policy π^∞, defined by $\pi_{ia} = \frac{x_i(a)}{x_i},\ i \in S,\ a \in A(i)$ is well defined. Because $x_i(a) = \pi_{ia} \cdot x_i,\ i \in S,\ a \in A(i)$,

we can write $\sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}\pi_{ia} \cdot x_i = 1,\ j \in S$. Hence, we have in vector notation, $x^T = e^T + x^T P(\pi)$. By iterating this equality, we obtain

$$x^T = \sum_{t=1}^{n} e^T P^{t-1}(\pi) + x^T P^n(\pi) \geq \sum_{t=1}^{n} e^T P^{t-1}(\pi),\ n \in \mathbb{N},$$

implying $\sum_{t=1}^{\infty} e^T P^{t-1}(\pi) < \infty$, i.e. $\sum_{t=1}^{\infty}\sum_i \{P^{t-1}(\pi)\}_{ij} < \infty,\ j \in S$. Consequently, $\sum_{t=1}^{\infty} \mathbb{P}_{i,\pi^\infty}\{X_j = j\} < \infty$ for all $i,\ j \in S$: π^∞ is a transient policy.

Conversely, let R be a transient policy. Define $x(R)$ by

$$x_{ja}(R) = \sum_{t=1}^{\infty}\sum_i \mathbb{P}_{i,R}\{X_t = j,\ Y_t = a\},\ (j,a) \in S \times A. \tag{4.12}$$

Since R is transient, $x_{ja}(R)$ is well defined and finite for all (j,a). Furthermore, by Corollary 1.1, we may assume that R is a Markov policy, say $R = (\pi^1, \pi^2, \dots)$. Then, we obtain

$$\begin{aligned}
\sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}x_{ia}(R) &= \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\} \cdot lim_{n\to\infty} \sum_{t=1}^{n}\sum_k \mathbb{P}_{k,R}\{X_t = i,\ Y_t = a\} \\
&= \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\} \cdot \sum_k lim_{n\to\infty} \sum_{t=1}^{n} \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ki} \cdot \pi^t_{ia} \\
&= \sum_k lim_{n\to\infty} \sum_{t=1}^{n}\sum_i \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})\}_{ki} \cdot \{I - P(\pi^t)\}_{ij} \\
&= \sum_k lim_{n\to\infty} \sum_{t=1}^{n} \left\{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1}) \cdot \{I - P(\pi^t)\}\right\}_{kj} \\
&= \sum_k lim_{n\to\infty} \{I - P(\pi^1)P(\pi^2)\cdots P(\pi^n)\}_{kj} \\
&= \sum_k \delta_{kj} = 1,\ j \in S.
\end{aligned}$$

□

Consider a policy $f^\infty \in C(D)$. It is intuitively clear that if $\phi_i(f^\infty)$, the average reward with starting state i, is strictly negative, the total reward $v_i(f^\infty) = -\infty$; if $\phi_i(f^\infty) = 0$ and state i is recurrent in the Markov chain induced by f^∞, then $v_i(f^\infty) = 0$. In the next theorem we show this property. In the proof we use some results from average reward MDPs, which are shown in the next chapter.

Theorem 4.8

Let f^∞ be an arbitrary stationary and deterministic policy.
(1) If $\phi_i(f^\infty) < 0$, then $v_i(f^\infty) = -\infty$.
(2) If $\phi_i(f^\infty) = 0$ and i is recurrent in the Markov chain induced by f^∞, then $v_i(f^\infty) = 0$.

Proof

(1) In the next chapter we show that $v^\alpha(f^\infty) = lim_{\alpha\uparrow 1}\left\{\frac{\phi(f^\infty)}{1-\alpha} + u(f)\right\}$ for some vector $u(f)$. Hence, if $\phi_i(f^\infty) < 0$, then $v_i(f^\infty) = lim_{\alpha\uparrow 1} v^\alpha(f^\infty) = -\infty$.

(2) In the next chapter we also show that $\phi(f^\infty) = P^*(f)r(f)$, where $P^*(f)$ is the stationary matrix of $P(f)$. Let R_k be the ergodic set which contains state i. Then, since the rows of R_k $P^*(f)$ are identical for the states of R_k, $\phi_j(f^\infty) = 0$ for all $j \in R_k$.
Furthermore, we have $p^*_{ij}(f) > 0,\ j \in R_k,\ p^*_{ij}(f) = 0,\ j \notin R_k$, and $p^t_{ij}(f) = 0,\ j \notin R_k,\ t \in \mathbb{N}_0$.
From $0 = \phi_i(f^\infty) = \sum_j p^*_{ij}(f)r_j(f) = \sum_{j\in R_k} p^*_{ij}(f)r_j(f)$, we have $r_j(f) = 0,\ j \in R_k$.
Hence, $v_i(f^\infty) = \sum_{t=1}^{\infty}\sum_j p^{t-1}_{ij}(f)r_j(f) = \sum_{j\in R_k} p^{t-1}_{ij}(f)r_j(f) = 0$.

□

In the model for the average reward we have stochastic MDPs ($\sum_j p_{ij}(a) = 1,\ (i,a) \in S \times A$). Therefore we have to use the extended model as introduced in Section 4.1.

Corollary 4.2
Let $f_1^\infty \in C(D)$ be an average optimal policy.
(1) If $\phi_i(f_1^\infty) < 0$, then $v_i = -\infty$.
(2) If $\phi_i(f_1^\infty) = 0$ and i is recurrent in the Markov chain induced by f_1^∞, then $v_i = 0$.

Proof
(1) Since $\phi_i(f^\infty) < 0$ for every $f^\infty \in C(D)$, $v_i(f^\infty) = -\infty$ for every $f^\infty \in C(D)$, i.e. $v_i = -\infty$.
(2) From Theorem 4.8 it follows that $v_i(f_1^\infty) = 0$, implying that $v_i = 0$. □

We can construct an optimal policy f_*^∞ for negative MDPs in the following way.
Firstly, we determine an average optimal policy, say f_1^∞.
Let $S_0 = \{i \mid \phi_i(f^\infty) < 0\}$.
For $i \in S_0$: $v_i = -\infty,\ f_*(i) = f_1(i)$ is optimal in state i and remove state i from the model.
For $i \notin S_0$: if there are actions a such that $\sum_{j \in S_0} p_{ij}(a) > 0$: remove action a from $A(i)$.

In the resulting model, we have $\phi_j(f_1^\infty) = 0$ for all states j, and there is at least one recurrent class. We determine the recurrent states $R(f_1)$ in the Markov chain of $P(f_1)$. From Corollary 4.2 we know that $f_*(i) = f_1(i)$ is optimal in the states $i \in R(f_1)$.
If there are states left, then we try to find an ergodic set with repect to another average optimal policy, say f_2^∞. Therefore, we first change the model in the follwing way:

$$S := S \backslash R(f_1) \cup \{0\};\ A(i) = \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases};\ r_i(a) = \begin{cases} r_i(a) & i \neq 0 \\ -1 & i = 0 \end{cases}$$

$$p_{ij}(a) = \begin{cases} p_{ij}(a) & i \neq 0,\ j \neq 0,\ a \in A(i); \\ \sum_{k \in R(f_1)} p_{ik}(a) & i \neq 0,\ j = 0,\ a \in A(i); \\ 1 & i = 0,\ j = 0,\ a \in A(i); \\ 0 & i = 0,\ j \neq 0,\ a \in A(i). \end{cases}$$

In this reduced model, we compute an average optimal policy, say f_2^∞.
Then, there are two possible situations:
Case 1: $\phi_i(f_2^\infty) = 0$ for at least one state i.
Determine in $\{i \mid \phi_i(f_2^\infty) = 0\}$ the states which are recurrent under $P(f_2)$, say $R(f_2)$. Then, $v_i(f_2^\infty) = 0,\ i \in R(f_2)$, and consequently, $f_*(i) = f_2(i)$ are optimal actions for the states of $R(f_2)$. We remove the states of $R(f_2)$ and repeat this procedure.
Case 2: $\phi_i(f_2^\infty) < 0$ for all states i.
In this case there is an optimal transient policy and we compute such an optimal transient policy, e.g. by Algorithm 4.4.

Every time we encounter Case 1, the state space decreases with at least one state. Hence, after a finite number of iterations either we encounter Case 2 or we have an average optimal policy such that all states i for which $\phi_k(i) = 0$ are recurrent in the Markov chain induced by this policy. Below we present the algorithm.

Algorithm 4.7 *Negative MDPs*

1. If $\sum_j p_{ij}(a) < 1$ for at least one pair $(i,a) \in S \times A$, then construct the extended model:

$$S := S \cup \{0\};\ A(i) = \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases};\ r_i(a) = \begin{cases} r_i(a) & i \neq 0 \\ 0 & i = 0 \end{cases}$$

$$p_{ij}(a) = \begin{cases} p_{ij}(a) & i \neq 0,\ j \neq 0,\ a \in A(i); \\ 1 - \sum_{k \neq 0} p_{ik}(a) & i \neq 0,\ j = 0,\ a \in A(i); \\ 1 & i = 0,\ j = 0,\ a \in A(i); \\ 0 & i = 0,\ j \neq 0,\ a \in A(i). \end{cases}$$

2. a. Compute an average optimal policy f^∞ (see Chapter 5).
   b. Let $S_0 = \{i \mid \phi_i(f^\infty) < 0\};\ f_*(i) = f(i),\ i \in S_0$.
   c. If $S_0 = S$: go to step 7;
   d. $S_1 = \emptyset$.
   e. For every $(i,a) \in (S \backslash S_0) \times A$: if $\sum_{j \in S_0} p_{ij}(a) > 0$: $A(i) := A(i) \backslash \{a\}$.
   f. $S := S \backslash S_0$.

3. a. Determine the set $R(f)$ of the recurrent states in S in the Markov chain $P(f)$.
   b. $f_*(i) = f(i),\ i \in R(f)$.
   c. If $R(f) = S$: go to step 4g.
   d. Constructed the reduced model:

$$S := S \backslash R(f) \cup \{0\};\ A(i) = \begin{cases} A(i) & i \neq 0 \\ \{1\} & i = 0 \end{cases};\ r_i(a) = \begin{cases} r_i(a) & i \neq 0 \\ -1 & i = 0 \end{cases}$$

$$p_{ij}(a) = \begin{cases} p_{ij}(a) & i \neq 0,\ j \neq 0,\ a \in A(i); \\ \sum_{k \in R(f)} p_{ik}(a) & i \neq 0,\ j = 0,\ a \in A(i); \\ 1 & i = 0,\ j = 0,\ a \in A(i); \\ 0 & i = 0,\ j \neq 0,\ a \in A(i). \end{cases}$$

4. a. Compute an average optimal policy f^∞ (see Chapter 5).
   b. $S_2 = \{i \mid \phi_i(f^\infty) < 0\}$.
   c. If $S = S_2$: $S_1 := S_1 \cup (S_2 \backslash \{0\})$ and go to step 4g.
   d. $S_1 := S_1 \cup (S_2 \backslash \{0\})$.
   e. For every $(i,a) \in (S \backslash S_2) \times A$: if $\sum_{j \in S_2} p_{ij}(a) > 0$: $A(i) := A(i) \backslash \{a\}$.

f. $S := S \backslash S_2$ and return to step 3a.

g. If $S_1 = \emptyset$: go to step 7.

5. Constructed the following transient model:

$S := S_1;\ A(i) := A(i),\ i \in S_1;\ r_i(a) := r_i(a),\ i \in S_1,\ a \in A(i);$

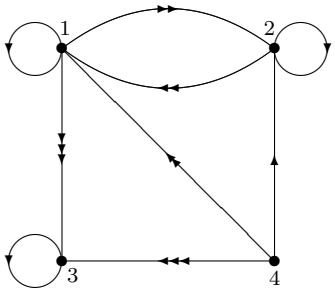
$$p_{ij}(a) = \begin{cases} p_{ij}(a) & i \in S_1,\ j \in S_1,\ a \in A(i); \\ 0 & \text{elsewhere} \end{cases}$$

6. Compute an optimal transient policy f_*^∞, e.g. by Algorithm 4.4.

7. f_*^∞ is an optimal policy (STOP).

Example 4.3

Let $S = \{1,2,3,4\};\ A(1) = \{1,2,3\},\ A(2) = \{1,2\},$
$A(3) = \{1\},\ A(4) = \{1,2,3\}.$
The nonzero transition probabilities are:
$p_{11}(1) = 0.5;\ p_{12}(2) = 1;\ p_{13}(3) = 1;\ p_{22}(1) = 0.5;$
$p_{21}(2) = 1;\ p_{33}(1) = 0.5;\ p_{42}(1) = 1;\ p_{41}(2) = 1;\ p_{43}(3) = 1.$
The rewards are:
$r_1(1) = -1;\ r_1(2) = 0;\ r_1(3) = 1;\ r_2(1) = -1;$
$r_2(2) = 0;\ r_3(1) = -1;\ r_4(1) = 0;\ r_4(2) = -1;\ r_4(3) = 0.$
The graph at the right hand side presents the model (partly).



The algorithm has the following result:

1. $S = \{0,1,2,3,4\};\ A(0) = \{1\};\ A(1) = \{1,2,3\},\ A(2) = \{1,2\},\ A(3) = \{1\},\ A(4) = \{1,2,3\}.$
   The nonzero probabilities are: $p_{00}(1) = 1;\ p_{11}(1) = 0.5;\ p_{10}(1) = 0.5;\ p_{12}(2) = 1;$
   $p_{13}(3) = 1;\ p_{22}(1) = 0.5;\ p_{20}(1) = 0.5;\ p_{21}(2) = 1;\ p_{33}(1) = 0.5;\ p_{30}(1) = 0.5;\ p_{42}(1) = 1;$
   $p_{41}(2) = 1;\ p_{43}(3) = 1.$
   The rewards are: $r_0(1) = 0;\ r_1(1) = -1;\ r_1(2) = 0;\ r_1(3) = 1;\ r_2(1) = -1;\ r_2(2) = 0;$
   $r_3(1) = -1;\ r_4(1) = 0;\ r_4(2) = -1;\ r_4(3) = 0.$
2. An optimal policy of the extended model is: $f(0) = f(1) = f(2) = f(3) = f(4) = 1.$
   Average reward: $\phi_0(f^\infty) = \phi_1(f^\infty) = \phi_2(f^\infty) = \phi_3(f^\infty) = \phi_4(f^\infty) = 0.$
   $S_0 = \emptyset;\ S_1 = \emptyset;\ S = \{0,1,2,3,4\}.$
3. $R(f) = \{0\};\ f_*(0) = 1.$
   The reduced model is the same as the first extended model, except that $r_0(1) = -1.$
4. An optimal policy of the reduced model is: $f(0) = 1;\ f(1) = f(2) = 2;\ f(3) = f(4) = 1.$
   Average reward: $\phi_0(f^\infty) = -1;\ \phi_1(f^\infty) = \phi_2(f^\infty) = 0;\ \phi_3(f^\infty) = -1;\ \phi_4(f^\infty) = 0.$
   $S_2 = \{0,3\};\ S_1 = \{3\};\ A(1) = \{2\};\ A(2) = \{2\};\ A(4) = \{1,2\};\ S = \{1,2,4\}.$
3. $R(f) = \{1,2\};\ f_*(1) = f_*(2) = 2.$
   The reduced model is: $S = \{0,4\};\ A(0) = \{1\};\ A(4) = \{1,2\};$
   $p_{00}(1) = 1;\ p_{40}(1) = 1;\ p_{40}(1) = 1;\ r_0(1) = -1;\ r_4(1) = 0;\ r_4(2) = -1.$

4. An optimal policy of the reduced model is: $f(0) = f(4) = 1;\ \phi_0(f^\infty) = \phi_4(f^\infty) = -1.$
   $S_2 = \{0, 4\} = S;\ S_1 = \{3, 4\}.$
5. $S = \{3, 4\};\ a(3) = \{1\};\ A(4) = \{1, 2\};\ r_3(1) = -1;\ r_4(1) = 0;\ r_4(2) = -1.$
   $p_{33} = 0.5$ (the other transition probabilities are 0).
6. $f_*(3) = f_*(4) = 1;\ v_3(f_*^\infty) = -2;\ v_4(f_*^\infty) = 0.$
7. $f_*(1) = f_*(2) = 2;\ f_*(3) = f_*(4) = 1;\ v_1 = v_2 = 0;\ v_3 = -2;\ v_4 = 0.$

Theorem 4.9

Algorithm 4.7 terminates with an optimal policy.

Proof

We have already mentioned the following properties of the algorithm:

(1) the algorithm terminates since Case 1 can occur only a finite number of times, because in each iteration the number of states decreases with $|R(f)|$;

(2) $v_i(f_*^\infty) = v_i = -\infty$ for all $i \in S_0$;

(3) $v_i(f_*^\infty) = v_i = 0$ for all $i \in S\backslash(S_0 \cup S_1)$.

Hence, it is sufficient to show that S_1 has an optimal transient policy and that f_*^∞ is optimal for the states in S_1.

Firstly, suppose that there exists an optimal nontransient policy, say g^∞, in the model of step 5. Since g^∞ is nontransient, $R(g) \cap S_1 \neq \emptyset$. From the construction of S_1 (see the steps 4c and 4d) it follows that $\phi_i(g^\infty) < 0,\ i \in R(g)$, implying that $v_i(g^\infty) = -\infty,\ i \in R(g)$, which contradicts that g^∞ is optimal.

Next, we will prove that f_*^∞ is an optimal policy. By the properties (2) and (3) it is sufficient to show that $v_i(f_*^\infty) \geq v_i(g^\infty)$ for $i \in S_1$ and for all policies g^∞. Since $v_j(f_*^\infty) = 0$ for all $j \in S\backslash(S_0 \cup S_1)$, we have $r_j(f_*) = 0$ for all $j \in S\backslash(S_0 \cup S_1)$. Hence, for $i \in S_1$, the total reward $v_i(f_*^\infty)$ is equal to the total reward in the transient model. □

4.6 Convergent MDPs

An MDP is *convergent* if $max\{v_i^+(R), v_i^-(R)\} < \infty$ for all policies R and all $i \in S$, i.e. the total absolute reward is finite for each policy. Hence, the value vector v is also finite. In this section we make the following assumption.

Assumption 4.4 *The MDP is convergent.*

A vector x has the property *anne* (asymptotic nonnegative expectation) if for every policy R with $v_i^-(R) < \infty,\ i \in S$, we have $lim_{t\to\infty} \mathbb{P}_{i,R}\{X_t = j\}x_j^- = 0$ for all $i, j \in S$. Hence, any nonnegative vector has the property anne.

Theorem 4.10

The value vector v is the smallest superharmonic vector with the property anne.

Proof

Theorem 4.2 implies that v is superharmonic. Let R be an arbitrary policy. Notice that $v_i^- = max\{-v_i, 0\} \leq max\{-v_i(R), 0\} = \{v_i(R)\}^-$. Since $r_j^-(a) = max\{0, -r_j(a)\} \geq -r_j(a)$, we have

$$\begin{aligned} v_i^- &\leq \{v_i(R)\}^- = max\Big\{ \textstyle\sum_{t=1}^{\infty} \sum_{(j,a)} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot \{-r_j(a)\}, 0 \Big\} \\ &\leq max\Big\{ \textstyle\sum_{t=1}^{\infty} \sum_{(j,a)} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j^-(a), 0 \Big\} \\ &= \textstyle\sum_{t=1}^{\infty} \sum_{(j,a)} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j^-(a) = v_i^-(R) \text{ for all policies } R. \end{aligned}$$

Let $R = (\pi^1, \pi^2, \dots)$ an arbitrary Markov policy with $v_i^-(R) < \infty,\ i \in S$, and let $R_n = (\pi^n, \pi^{n+1}, \dots)$. Then, for any $t \in \mathbb{N}$, we obtain

$$\begin{aligned} \textstyle\sum_j \mathbb{P}_{i,R}\{X_t = j\} v_j^- &\leq \textstyle\sum_j \mathbb{P}_{i,R}\{X_t = j\} v_j^-(R_t) \\ &= \textstyle\sum_j \mathbb{P}_{i,R}\{X_t = j\} \Big\{ \sum_{s=1}^{\infty} \sum_{(k,a)} \mathbb{P}_{j,R_t}\{X_s = k, Y_s = a\} \cdot r_k^-(a) \Big\} \\ &= \textstyle\sum_{s=1}^{\infty} \sum_{(k,a)} \Big\{ \sum_j \mathbb{P}_{i,R}\{X_t = j\}\, \mathbb{P}_{j,R_t}\{X_s = k, Y_s = a\} \cdot r_k^-(a) \Big\} \\ &= \textstyle\sum_{s=1}^{\infty} \sum_{(k,a)} \mathbb{P}_{i,R}\{X_{t+s-1} = k, Y_{t+s-1} = a\} \cdot r_k^-(a) \\ &= \textstyle\sum_{m=t}^{\infty} \sum_{(k,a)} \mathbb{P}_{i,R}\{X_m = k, Y_m = a\} \cdot r_k^-(a). \end{aligned}$$

Let $A_t = \sum_{m=t}^{\infty} \sum_{(k,a)} \mathbb{P}_{i,R}\{X_m = k, Y_m = a\} \cdot r_k^-(a),\ t \in \mathbb{N}$. Since, $v_i^-(R) < \infty$, we have $lim_{t\to\infty} A_t = 0$, implying $lim_{t\to\infty} \sum_j \mathbb{P}_{i,R}\{X_t = j\} v_j^- = 0,\ i \in S$, for any Markov policy R. By Corollary 1.1, $lim_{t\to\infty} \sum_j \mathbb{P}_{i,R}\{X_t = j\} v_j^- = 0$ for all policies, so $lim_{t\to\infty} \mathbb{P}_{i,R}\{X_t = j\} v_j^- = 0$ for all $i, j \in S$ and all policies R. Therefore, we have shown that v has the property anne.

Finally, suppose that w is also a superharmonic vector with the property anne.

Let $C^* = \{R \mid R \text{ is a Markov policy and } v_i^-(R) < \infty\}$. Since $-\infty < v_i < +\infty,\ i \in S$, we have $v_i = sup_{R \in C^*} v_i(R),\ i \in S$. Hence, it is sufficient to show that $v_i(R) \leq w_i,\ i \in S,\ R \in C^*$.

Define by induction: $x_i^0 = w_i$ and $x_i^{n+1} = max\{r_i(a) + \sum_j p_{ij}(a) x_j^n\},\ i \in S$.

Since w is superharmonic it can easily be verified (induction on n) that $x^n \leq w$ for $n = 0, 1, \dots$.

We will show that for any $i \in S$ and Markov policy R: $v_i^t(R) + \sum_j \mathbb{P}_{i,R}\{X_{t+1} = j\} w_j \leq x_i^t,\ t \in \mathbb{N}$.

Choose any Markov policy $R = (\pi^1, \pi^2, \dots)$ (by Corollary 1.1, we may restrict ourselves to Markov policies) and any $i \in S$. The proof is by induction on t. For $t = 1$:

$v_i^1(R) + \sum_j \mathbb{P}_{i,R}\{X_2 = j\} w_j = \sum_a \pi_{ia}^1 \{r_i(a) + \sum_j p_{ij}(a) w_j\} \leq max\{r_i(a) + \sum_j p_{ij}(a) w_j\} = x_i^1$.

Suppose that the inequality is valid for some t. Then, with Markov policy $R_* = (\pi^2, \pi^3, \dots)$,

$$\begin{aligned} v_i^{t+1}(R) + \textstyle\sum_j \mathbb{P}_{i,R}\{X_{t+2} = j\} w_j &= \textstyle\sum_a \pi_{ia}^1 \Big\{ r_i(a) + \sum_k p_{ik}(a) v_k^t(R_*) + \sum_{j,k} p_{ik}(a)\, \mathbb{P}_{k,R_*}\{X_{t+1} = j\} w_j \Big\} \\ &\leq max_a \Big\{ r_i(a) + \textstyle\sum_k p_{ik}(a) \big\{ v_k^t(R_*) + \sum_j \mathbb{P}_{k,R_*}\{X_{t+1} = j\} w_j \big\} \Big\} \\ &\leq max_a \{ r_i(a) + \textstyle\sum_k p_{ik}(a) x_k^t \} = x_i^{t+1}. \end{aligned}$$

Take any policy $R \in C^*$. Then, by the anne property of w, we have

$$liminf_{t\to\infty} \sum_j \mathbb{P}_{i,R}\{X_t = j\}w_j \geq liminf_{t\to\infty} \sum_j \mathbb{P}_{i,R}\{X_t = j\}(-w_j^-) = 0.$$

Hence, we obtain

$$v_i(R) = lim_{t\to\infty}\, v_i^t(R) \leq limsup_{t\to\infty} \{v_i^t(R) + \sum_j \mathbb{P}_{i,R}\{X_{t+1} = j\}w_j\} \leq x_i^t \leq w_i,\ i \in S,$$

and consequently, $v_i = sup_{R\in C^*}\, v_i^t(R) \leq w_i,\ i \in S$. □

We have seen in Section 4.1 that an optimal policy is conserving and that the reverse statement is not necessarily true. If the policy is also *equalizing*, i.e. $lim_{t\to\infty} \sum_j p_{ij}^t(f)v_j^+ = 0$ for all $i \in S$, then the policy is optimal as the next theorem shows.

Theorem 4.11

A policy $f^\infty \in C(D)$ is optimal if and only if f^∞ is conserving and equalizing.

Proof

$\Rightarrow$ Let f^∞ be an optimal policy, i.e. $v(f^\infty) = v$. Policy f^∞ is conserving, because

$$\begin{aligned} v &= v(f^\infty) = \sum_{t=1}^\infty P^{t-1}(f)r(f) = r(f) + P(f)\{\sum_{t=1}^\infty P^{t-1}(f)r(f)\} \\ &= r(f) + P(f)v(f^\infty) = r(f) + P(f)v. \end{aligned}$$

Iterating the above equation gives $v = \sum_{t=1}^n P^{t-1}(f)r(f) + P^n(f)v,\ n \in \mathbb{N}$. Since v is finite and $v = \sum_{t=1}^\infty P^{t-1}(f)r(f)$, we have $lim_{n\to\infty} P^n(f)v = 0$, i.e. $\sum_j p_{ij}^n(f)v_j = 0,\ i \in S$.

Since v has the property anne, $lim_{n\to\infty} \sum_j p_{ij}^n(f)v_j^- = 0,\ i \in S$, implying that also $lim_{n\to\infty} \sum_j p_{ij}^n(f)v_j^+ = 0,\ i \in S$, i.e. f^∞ is conserving and equalizing.

$\Leftarrow$ Since f^∞ is conserving, $v = r(f) + P(f)v$, implying $v = \sum_{t=1}^n P^{t-1}(f)r(f) + P^n(f)v,\ n \in \mathbb{N}$.

The equalizing property gives $limsup_{n\to\infty} \sum_j p_{ij}^n(f)v_j \leq lim_{n\to\infty} \sum_j p_{ij}^n(f)v_j^+ = 0,\ i \in S$.

Hence, we obtain

$v_i = lim_{n\to\infty}\{\sum_{t=1}^n \sum_j p_{ij}^{t-1}(f)r_j(f) + \sum_j p_{ij}^n(f)v_j\} \leq \sum_{t=1}^\infty \sum_j p_{ij}^{t-1}(f)r_j(f) = v_i(f^\infty)$,

i.e. f^∞ is optimal. □

Define by induction:

$$v_i^0 = 0 \text{ and } v_i^{n+1} = max\{r_i(a) + \sum_j p_{ij}(a)v_j^n\},\ i \in S. \tag{4.13}$$

An MDP is *stable* if $lim_{n\to\infty}\, v_i^n = v_i,\ i \in S$. Hence, in a stable MDP, the value vector can be approximated arbitrary close by value iteration. The next example shows that a convergent MDP is not necessarily stable.

Example 4.1 (continued)

$S = \{1, 2\};\ A(1) = \{1, 2\};\ A(1) = \{1\};\ p_{11}(1) = 1;\ p_{12}(2) = 1;\ p_{21}(1) = p_{22}(1) = 0;$
$r_1(1) = 0;\ r_1(2) = 2;\ r_2(1) = -1$. We have seen that $v = (1, -1)$.
It is easy to verify that $v^n = (2, -1),\ n \in \mathbb{N}$.

Theorem 4.12

Positive and negative MDPs are stable.

Proof

Firstly, assume that the MDP is positive. Then,

$$v_i = max_a\{r_i(a) + \sum_j p_{ij}(a)v_j\} \geq max_a\{r_i(a) + \sum_j p_{ij}(a)v_j^0\} = v_i^1 \geq v_i^0,\ i \in S.$$

Suppose that $v_i \geq v_i^k \geq v_i^{k-1},\ i \in S$ for some k. Then,

$$\begin{aligned} v_i &= max_a\{r_i(a) + \sum_j p_{ij}(a)v_j\} \geq max_a\{r_i(a) + \sum_j p_{ij}(a)v_j^k\} = v_i^{k+1} \\ &\geq max_a\{r_i(a) + \sum_j p_{ij}(a)v_j^{k-1}\} = v_i^k,\ i \in S. \end{aligned}$$

Hence, by induction, we have $v_i \geq v_i^{n+1} \geq v_i^n,\ i \in S,\ n \in \mathbb{N}$. Since for each i the sequence $\{v_i^n\}$ is monotone nondecreasing and bounded by $v_i < \infty$, $lim_{n\to\infty} v_i^n$ exists, say $v_i^\infty = lim_{n\to\infty} v_i^n,\ i \in S$, and $v^\infty \leq v$. By taking the limit for $n \to \infty$, it follows from $v_i^{n+1} = max\{r_i(a) + \sum_j p_{ij}(a)v_j^n\}$ that $v_i^\infty = max\{r_i(a) + \sum_j p_{ij}(a)v_j^\infty\},\ i \in S$. Hence, v^∞ is superharmonic and has the property anne, the last property because v^∞ is nonnegative. Therefore, $v^\infty \geq v$, and we have shown that $v_i = v_i^\infty = lim_{n\to\infty} v_i^n$, i.e. the positive MDPs are stable.

Next, we assume that the MDP is negative. Analogously to the positive case it can be shown that the sequence $\{v_i^n\}$ is monotone non-increasing in n, bounded below by v_i, with limit v_i^∞. Therefore, $v^\infty \geq v$ and satisfies $v_i^\infty = max\{r_i(a) + \sum_j p_{ij}(a)v_j^\infty\},\ i \in S$. Let $f^\infty \in C(D)$ be such that $v^\infty = r(f) + P(f)v^\infty$. Then, by induction on n,

$$v^\infty = \sum_{t=1}^n P^{t-1}(f)r(f) + P^n(f)v^\infty \leq \sum_{t=1}^n P^{t-1}(f)r(f),\ n \in \mathbb{N}.$$

Hence, $v \geq v(f^\infty) = lim_{n\to\infty} \sum_{t=1}^n P^{t-1}(f)r(f) \geq v^\infty$, implying that $v_i = v_i^\infty = lim_{n\to\infty} v_i^n$, i.e. the negative MDPs are stable. □

Remark

Since in negative MDPs every policy is equalizing, f^∞ is optimal if and only if f^∞ is conserving, i.e. $r(f) + P(f)v = v$. Hence, policy f_v^∞ is an optimal policy.

4.7 Special models

4.7.1 Red-black gambling

The red-black gambling model was introduced in section 1.3.1. The characteristics of this model are:

$$S = \{0, 1, \ldots, N\};\ A(0) = A(N) = \{0\},\ A(i) = \{1, 2, \ldots, min(i, N-i)\},\ 1 \leq i \leq N-1.$$

$$\text{For } 1 \leq i \leq N-1, a \in A(i):\ p_{ij}(a) = \begin{cases} p & , j = i+a \\ 1-p & , j = i-a \\ 0 & , j \neq i+a, i-a \end{cases} \quad \text{and } r_i(a) = 0.$$

$$p_{0j}(0) = p_{Nj}(0) = 0, j \in S;\ r_0(0) = 0, r_N(0) = 1.$$

The case $p = \frac{1}{2}$ was the subject of Exercise 1.4 in which the reader was asked to show that any $f^\infty \in C(D)$ is an optimal policy. This current section deals with the cases $p > \frac{1}{2}$ and $p < \frac{1}{2}$. In the red-black gambling model every policy is transient (see Exercise 4.8). Hence, we may use the results of section 4.3, e.g. that $v(f^\infty)$ is the unique solution of the linear system $x = r(f)+P(f)x$. In the red-black gambling model this system becomes

$$x_0 = 0;\ x_N = 1;\ x_i = px_{i+f(i)} + (1-p)x_{i-f(i)},\ 1 \le i \le N-1. \tag{4.14}$$

Let f_1^∞ be the *timid policy*, i.e. $f(i) = 1$ for all i. Then it is easy to verify that $v_i(f_1^\infty) = \frac{1-r^i}{1-r^N}$, $0 \le i \le N$ satisfies (4.14).

Case 1: $p > \frac{1}{2}$

In this case we will show that the timid policy f_1^∞ is optimal. For this purpose, it is sufficient to show that $v_i(f_1^\infty) \ge pv_{i+a}(f_1^\infty) + (1-p)v_{i-a}(f_1^\infty),\ (i,a) \in S \times A$.

Because $v_i(f_1^\infty) = \frac{1-r^i}{1-r^N},\ 0 \le i \le N$, we have to show

$1 - r^i \ge p(1-r^{i+a}) + q(1-r^{i-a})$, i.e. $-r^i \ge -pr^{i+a} - qr^{i-a}$, which is the same as

$1 \le pr^a + qr^{-a}$. For $F(a) = pr^a + qr^{-a}$, we have $F(1) = p\frac{q}{p} + q\frac{p}{q} = q + p = 1$.

It is sufficient to show that $F(a+1) \ge F(a)$ for all a.

$F(a+1) \ge F(a) \Leftrightarrow pr^{a+1} + qr^{-a-1} \ge pr^a + qr^{-a} \Leftrightarrow pr^{2a+2} + q \ge pr^{2a+1} + qr \Leftrightarrow$

$pr^{2a+1}(r-1) \ge q(r-1) \Leftrightarrow r^{2a+1} \le r \Leftrightarrow r \le 1 \Leftrightarrow p \ge \frac{1}{2}$.

Case 2: $p < \frac{1}{2}$

We will show that in this case *bold play*, i.e. stake $min(i, N-i)$ in state i, is optimal. Therefore we show that in value iteration with starting vector 0, i.e.

$v_i^0 = 0,\ i \in S;\ v_i^{n+1} = max_a\ \{pv_{i+a}^n + (1-p)v_{i-a}^n\},\ 1 \le i \le N-1;\ v_0^{n+1} = 0;\ v_N^{n+1} = 1$ for $n = 0, 1, \ldots$

the bold policy f_b^∞ satisfies $v^{n+1} = L_{f_b}v^n$. Since $v^n \to v$, this implies $v = L_{f_b}v$, i.e. f_b^∞ is an optimal policy.

Let $q = 1-p$ and let w_{ia}^n be the difference between the action $f_b(i)$ and a in the computation of v_i^{n+1}, i.e.

$$w_{ia}^n = \begin{cases} pv_{2i}^n - pv_{i+a}^n - qv_{i-a}^n & ,1 \le i \le N/2 \quad ,\ a \in A(i) \\ p + qv_{2i-N}^n - pv_{i+a}^n - qv_{i-a}^n & ,N/2 \le i \le N-1 \quad ,\ a \in A(i) \end{cases} \tag{4.15}$$

We have to show that $w_{ia}^n \ge 0$ for all i, a and n. To this end, we show by induction on n:

(1) $w_{ia}^n \ge 0$ for all i, a;

(2) $v_{i+a}^n \ge v_i^n + v_a^n$ for all i, a;

(3) $v_N^n + v_j^n \ge v_{N-k}^n + v_{j+k}^n$ for all j, k;

(4) $v_i^{n+1} = \begin{cases} pv_{2i}^n & ,1 \le i < N/2 \\ p + qv_{2i-N}^n & ,N/2 \le i \le N-1 \end{cases}$

For $n = 0$ it is easy to verify that the properties hold. Suppose that the properties are shown for n and consider $n+1$. Because w_{ia}^n has different expressions for the states below and above $N/2$, we have to distinguish between different intervals of the states.

Proof for (1):

For $i + a < N/2$ and $2i < N/2$:

$$w_{ia}^{n+1} = pv_{2i}^{n+1} - pv_{i+a}^{n+1} - qpv_{i-a}^{n+1} = p\{pv_{4i}^n - pv_{2(i+a)}^n - qv_{2(i-a)}^n\} = pw_{2i,2a}^n \geq 0.$$

For $i + a < N/2$ and $2i \geq N/2$:

$$\begin{aligned} w_{ia}^{n+1} &= pv_{2i}^{n+1} - pv_{i+a}^{n+1} - qpv_{i-a}^{n+1} \\ &= p\{p + qv_{4i-N}^n - pv_{2(i+a)}^n - qv_{2(i-a)}^n\} = pw_{2i,2a}^n \geq 0 \end{aligned}$$

For $i + a \geq N/2$ and $i < N/2$:

$$\begin{aligned} w_{ia}^{n+1} &= pv_{2i}^{n+1} - pv_{i+a}^{n+1} - qpv_{i-a}^{n+1} \\ &= p\{p + qv_{4i-N}^n - p - qv_{2(i+a)-N}^n - qv_{2(i-a)}^n\} \\ &= pq\{v_{4i-N}^n - v_{2(i+a)-N}^n - v_{2(i-a)}^n\} \geq 0 \end{aligned}$$

(the nonnegativity by property (2).

For $i + a \geq N/2,\ i \geq N/2,\ i - a < N/2$ and $2i - N < N/2$:

$$\begin{aligned} w_{ia}^{n+1} &= p + qv_{2i-N}^{n+1} - pv_{i+a}^{n+1} - qv_{i-a}^{n+1} \\ &= p + qpv_{4i-2N}^n - p\{p + qv_{2(i+a)}^n\} - qpv_{2(i-a)}^n \\ &= pq\{1 + v_{4i-2N}^n - v_{2(i+a)}^n - v_{2(i-a)}^n \\ &= pq\{v_N^n + v_{4i-2N}^n - v_{2(i+a)}^n - v_{2(i-a)}^n \geq 0 \end{aligned}$$

(the nonnegativity by property (3) with $j = 4i - 2N$ and $k = N - 2(i - a)$).

For $i + a \geq N/2,\ i \geq N/2,\ i - a < N/2$ and $2i - N \geq N/2$:

$$\begin{aligned} w_{ia}^{n+1} &= p + qv_{2i-N}^{n+1} - pv_{i+a}^{n+1} - qv_{i-a}^{n+1} \\ &= p + q\{p + qv_{4i-3N}^n\} - p\{p + qv_{2(i+a)-N}^n\} - qpv_{2(i-a)}^n \\ &= 2pq + q\{qv_{4i-3N}^n - pv_{2(i+a)-N}^n - pv_{2(i-a)}^n\} \\ &\geq pq\{2 + v_{4i-3N}^n - v_{2(i+a)-N}^n - v_{2(i-a)}^n\} \geq 0 \end{aligned}$$

(the nonnegativity because $v_{2(i+a)-N}^n + v_{2(i-a)}^n \leq 2$).

For $i + a \geq N/2,\ i \geq N/2,\ i - a \geq N/2$ and $2i - N < N/2$:

$$\begin{aligned} w_{ia}^{n+1} &= p + qv_{2i-N}^{n+1} - pv_{i+a}^{n+1} - qv_{i-a}^{n+1} \\ &= p + qpv_{4i-2N}^n - p\{p + qv_{2(i+a)-N}^n\} - q\{p + v_{2(i-a)-N}^n\} \\ &= q\{pv_{4i-2N}^n - pv_{2(i+a)-N}^n - qv_{2(i-a)-N}^n\} \\ &\geq qw_{2i-N,2a}^n \geq 0 \end{aligned}$$

For $i + a \geq N/2,\ i \geq N/2,\ i - a \geq N/2$ and $2i - N \geq N/2$:

$$\begin{aligned} w_{ia}^{n+1} &= p + qv_{2i-N}^{n+1} - pv_{i+a}^{n+1} - qv_{i-a}^{n+1} \\ &= p + q\{p + qv_{4i-3N}^n\} - p\{p + qv_{2(i+a)-N}^n\} - q\{p + v_{2(i-a)-N}^n\} \\ &= q\{p + v_{4i-3N}^n - pv_{2(i+a)-N}^n - qv_{2(i-a)-N}^n\} \\ &\geq qw_{2i-N,2a}^n \geq 0 \end{aligned}$$

Proof for (2):

For $i+a < N/2$:

$$v_{i+a}^{n+1} \;=\; pv_{2(i+a)}^{n} \;\geq\; p\{v_{2i}^{n} + v_{2a}^{n}\} \;=\; v_i^{n+1} + v_a^{n+1}.$$

For $i+a \geq N/2$ and $i < N/2$:

$$\begin{aligned} v_{i+a}^{n+1} &= p + qv_{2(i+a)-N}^{n} \;\geq \text{(because } w_{i+a,i-a}^{n} \geq 0) \\ &\geq pv_{2i}^{n} + qv_{2a}^{n} \;\geq\; pv_{2i}^{n} + pv_{2a}^{n} \;=\; v_i^{n+1} + v_a^{n+1}. \end{aligned}$$

For $i+a \geq N/2$ and $i \geq N/2$:

$$\begin{aligned} v_{i+a}^{n+1} &= p + qv_{2(i+a)-N}^{n} \;\geq\; p + q\{v_{2i-N}^{n} + v_{2a}^{n}\} \\ &\geq p + qv_{2i-N}^{n} + pv_{2a}^{n} \;=\; v_i^{n+1} + v_a^{n+1}. \end{aligned}$$

Proof for (3):

If $j \geq N-k$: $v_N^n + v_j^n \geq v_N^n + v_{N-k}^n \geq v_{j+k}^n + v_{N-k}^n$.

If $j \geq N-k$ and $j+k \leq N-k$:

For $N/2 \leq j \leq j+k \leq N-k$:

$v_N^{n+1} + v_j^{n+1} = 1 + p + qv_{2j-N}^n$ and $v_{j+k}^{n+1} + v_{N-k}^{n+1} = 2p + q\{v_{2(j+k)-N}^n + v_{2(N-k)-N}^n\}$.

Hence,

$v_N^{n+1} + v_j^{n+1} \geq v_{j+k}^n + v_{N-k}^n \;\leftrightarrow\; 1 + v_{2j-N}^n \geq v_{2(j+k)-N}^n + v_{2(N-k)-N}^n$,

which is true by property (3) (take in (3) $2j-N$ for j and $2k$ for k).

For $j < N/2 \leq j+k \leq N-k$:

$$\begin{aligned} v_{j+k}^{n+1} + v_{N-k}^{n+1} &= 2p + q\{v_{2(j+k)-N}^{n} + v_{2(N-k)-N}^{n}\} \text{ (by property (2))} \\ &\leq 2p + qv_{2j}^{n} \;=\; 2p + (1-p)v_{2j}^{n} \;=\; 2p + qv_{2j}^{n} \;=\; 1 + pv_{2j}^{n} - (1-2p)(1-v_{2j}^{n}) \\ &\leq 1 + pv_{2j}^{n} = v_N^{n+1} + v_j^{n+1}. \end{aligned}$$

For $j \leq j+k < N/2 \leq N-k$:

$$\begin{aligned} v_N^{n+1} + v_j^{n+1} &= 1 + pv_{2j}^{n} \geq q + p\{1 + v_{2j}^{n}\} \geq \text{ (take in (3) } 2j \text{ for } j \text{ and } 2k \text{ for } k) \\ &\geq q + p\{v_{2(j+k)}^{n} + v_{2(N-k)-N}^{n}\}. \\ v_{j+k}^{n+1} + v_{N-k}^{n+1} &= pv_{2(j+k)}^{n} + p + q\{v_{2(N-k)-N}^{n} \leq q + p\{v_{2(j+k)}^{n} + v_{2(N-k)-N}^{n}\}. \end{aligned}$$

Hence, $v_N^{n+1} + v_j^{n+1} \geq v_{N-k}^{n+1} + v_{j+k}^{n+1}$.

For $j \leq j+k \leq N-k < N/2$:

This case cannot occur, because $(j+k) + (N-k) = j + N \geq N$.

Proof for (4):

Since $w_{ia}^{n+1} \geq 0$ for all $(i,a) \in S \times A$, the bold actions maximize $pv_{i+a}^{n+1} + (1-p)v_{i-a}^{n+1},\ i \in S$.

Therefore, $v_i^{n+2} = \begin{cases} pv_{2i}^{n+1} & ,1 \leq i < N/2 \\ p + qv_{2i-N}^{n+1} & ,N/2 \leq i \leq N-1 \end{cases}$

4.7.2 Optimal stopping

The optimal stopping model was introduced in section 1.3.3. The characteristics of the model are:

$$S = \{1, 2, \ldots, N\};\ A(i) = \{1, 2\},\ i \in S;\ r_i(1) = r_i,\ i \in S;\ r_i(2) = -c_i,\ i \in S;$$
$$p_{ij}(1) = 0,\ i, j \in S;\ p_{ij}(2) = p_{ij},\ i, j \in S.$$

In this section we assume that $r_i \geq 0$ and $c_i \geq 0$ for all $i \in S$. Then, we have $0 \leq min_j\, r_j \leq v_i \leq max_j\, r_j < \infty,\ i \in S$. Furthermore, the model is convergent, because

$$\begin{aligned} v_i^+(R) &= \sum_{t=1}^{\infty} \sum_{(j,a)} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} \cdot r_j^+(a) \\ &= \sum_{t=1}^{\infty} \sum_j \mathbb{P}_{i,R}\{X_t = j, Y_t = 1\} \cdot r_j \leq max_j\, r_j < \infty. \end{aligned}$$

By Theorem 4.2, v satisfies the optimality equation

$$v_i = max\{r_i, -c_i + \sum_j p_{ij} v_j\},\ i \in S \tag{4.16}$$

Furthermore, since the value vector v is nonnegative, v has the property anne. Hence, by Theorem 4.10, v is the smallest superharmonic nonnegative vector, i.e. v is the unique optimal solution of the linear program

$$min \left\{ \sum_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i & ,i \in S \\ v_i \geq -c_i + \sum_j p_{ij} v_j & ,\ i \in S \end{array} \right\}. \tag{4.17}$$

Consider also the dual linear program

$$max \left\{ \sum_i r_i x_i - \sum_i c_i y_i \;\middle|\; \begin{array}{rcl} x_j + y_j - \sum_i p_{ij} y_i &=& 1,\ j \in S \\ x_i,\ y_i &\geq& 0,\ i \in S \end{array} \right\}. \tag{4.18}$$

Theorem 4.13

Let (x^*, y^*) be an extreme optimal solution of the dual program (4.18). Then, the policy f_*^∞ such that $f_*(i) = \begin{cases} 1 & \text{if } x_i^* > 0 \\ 2 & \text{if } x_i^* = 0 \end{cases}$ *is an optimal policy.*

Proof

If $x_j^* = 0$, then it follows from $x_j^* + y_j^* = 1 + \sum_i p_{ij} y_i^* \geq 1 > 0$ that $y_j^* > 0$. Since (x^*, y^*) is an extreme point, there are atmost N positive components. Hence, for each state $i \in S$, we have either $x_j^* > 0$ or $y_j^* > 0$. Furthermore, for $z^* = \begin{cases} x_i^* & \text{if } x_i^* > 0 \\ y_i^* & \text{if } x_i^* = 0 \end{cases}$ we obtain $z^* = e^T + (z^*)^T P(f_*)$.

Iterating this equality gives $z^* = \sum_{t=1}^n e^T P^{t-1}(f_*) + (z^*)^T P^n(f_*) \geq \sum_{t=1}^n e^T P^{n-1}(f_*)$ for all $n \in \mathbb{N}$. Hence, f_*^∞ is transient and $\{I - P(f_*)\}^{-1} = \sum_{t=1}^\infty P^{t-1}(f_*)$.

From the complementary slackness of linear programming it follows that $v = r(f_*) + P(f_*)v$. Iterating this equation gives $v = \{I - P(f_*)\}^{-1} r(f_*) = \sum_{t=1}^\infty e^T P^{t-1}(f_*) r(f_*) = v(f_*^\infty)$, i.e. f_*^∞ is an optimal policy. □

Algorithm 4.8 *Linear programming algorithm for optimal stopping*

1. Use the simplex method to compute optimal solutions v^* and (x^*, y^*) of the dual pair of linear programs:

$$min \left\{ \sum_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i & , \; i \in S \\ v_i \geq -c_i + \sum_j p_{ij} v_j & , \; i \in S \end{array} \right\}$$

and

$$max \left\{ \sum_i r_i x_i - \sum_i c_i y_i \;\middle|\; \begin{array}{rcl} x_j + y_j - \sum_i p_{ij} y_i & = & 1, \; j \in S \\ x_i, \; y_i & \geq & 0, \; i \in S \end{array} \right\}.$$

2. Take $f_*^\infty \in C(D)$ such that $f_*(i) = \begin{cases} 1 & \text{if } x_i^* > 0; \\ 2 & \text{if } x_i^* = 0. \end{cases}$

   v^* is the value vector and f_*^∞ an optimal policy (STOP).

Remark

An optimal stopping problem may be considered as a special case of the replacement problem that is discussed in section 8.1.1. In that section it is shown that an $\mathcal{O}(N^3)$ of Algorithm 4.8 is possible.

Example 4.4

$S = \{1, 2, 3, 4, 5\}$. The stopping rewards are: $r_1 = 0, \; r_2 = 2, \; r_3 = 2, \; r_4 = 3$ and $r_5 = 0$ and there are no costs for the continuing action ($c_i = 0, \; 1 \leq i \leq 5$). The states 1 and 5 are absorbing; in state i ($2 \leq i \leq 4$) there is a probability $\frac{1}{2}$ to go to state $i+1$ and a probability $\frac{1}{2}$ to go to state $i-1$. The dual LP program is:

$max \; 2x_2 + 2x_3 + 3x_4$

subject to:

$$\begin{array}{rrrrrrrrrrrrrrrrr}
x_1 & & & & & - & \frac{1}{2}y_2 & & & & & & & = & 1 \\
 & x_2 & & & & + & y_2 & - & \frac{1}{2}y_3 & & & & & = & 1 \\
 & & x_3 & & & - & \frac{1}{2}y_2 & + & y_3 & - & \frac{1}{2}y_4 & & & = & 1 \\
 & & & x_4 & & & & - & \frac{1}{2}y_3 & + & y_4 & & & = & 1 \\
 & & & & x_5 & & & & & - & \frac{1}{2}y_4 & & & = & 1 \\
 & & & & & x_1, & x_2, & x_3, & x_4, & x_5, & y_2, & y_3, & y_4 & \geq & 0
\end{array}$$

The optimal solution of the problem is:

$x_1 = 1, \; x_2 = \frac{3}{2}, \; x_3 = 0, \; x_4 = \frac{3}{2}; \; x_5 = 1; \; y_2 = 0; \; y_3 = 1$ and $y_4 = 0$.

Hence, the optimal policy is: continue in state 3 and stop in the other states.

The expected total reward is: $v_1 = 0, \; v_2 = 2, \; v_3 = 2\frac{1}{2}, \; v_4 = 3$ and $v_5 = 0$.

Let

$$S_0 = \{i \in S \mid r_i \geq -c_i + \sum_j p_{ij} r_j\},$$

i.e. S_0 is the set of states in which immediate stopping is not worse than continuing for one period and than choose to stop. The set S_0 follows directly from the data of the model. An optimal stopping problem is *monotone* if $p_{ij} = 0$ for all $i \in S_0,\ j \notin S_0$.

Theorem 4.14

In a monotone optimal stopping problem a one-step look ahead policy, i.e. a policy that stops in the states of S_0 and continues outside S_0, is an optimal policy.

Proof

Let v be the value vector of the optimal stopping problem. Define w by $w_i = \begin{cases} r_i, & i \in S_0; \\ v_i, & i \notin S_0. \end{cases}$

The value vector is the solution of the optimality equation: $v_i = max\{r_i, -c_i + \sum_j p_{ij} v_j\},\ i \in S$. Therefore, $w \leq v$. Furthermore, w is feasible for the LP problem, namely:

If $i \in S_0:\ w_i = r_i \geq -c_i + \sum_j p_{ij} r_j = -c_i + \sum_{j \in S_0} p_{ij} r_j = -c_i + \sum_{j \in S_0} p_{ij} w_j = -c_i + \sum_j p_{ij} w_j$.

If $i \notin S_0:\ w_i = v_i \geq -c_i + \sum_j p_{ij} v_j \geq -c_i + \sum_j p_{ij} w_j$.

It is obvious that $w_i \geq r_i,\ i \in S$. Because v is the smallest solution of the LP problem, we have $v = w$, i.e. $v_i = r_i,\ i \in S_0$, i.e. the stopping action is in S_0 optimal. If $i \notin S_0$, then we obtain, $r_i < -c_i + \sum_j p_{ij} r_j \leq -c_i + \sum_j p_{ij} v_j \leq v_i$: continue outside S_0 is optimal. □

Example 4.5

N different real numbers are drawn, one by one. The second number has a probability of $\frac{1}{2}$ to come on the right of the first number on the line of the real numbers (also a probability of $\frac{1}{2}$ to come on the left of the first number). The third number has a probability of $\frac{1}{3}$ to come in each of the three intervals on the right line, where the intervals are generated by the first two numbers. Etc. After each draw there are two possibilities: the last draw is the largest up to now or this is not the case. Only when the last number is the largest up to now we have the option to stop with as reward that largest number. If the last number is not the largest up to now or when we don't use the option to stop when the last number is the largest, we have to continue, unless all N numbers are drawn. Which policy maximizes the probability to stop with the largest of all N numbers?

We make the following model for this problem. Let $S = \{1, 2, \ldots, N\}$, where state i means that the i-th draw is the largest up to now. $A(i) = \{1, 2\},\ 1 \leq i \leq N-1;\ A(N) = \{1\};\ c_i = 0,\ i \in S$. As r_i we take the probability that, given that the i-th draw gives the largest number of the first i numbers, it is the largest number of all N numbers. The probability that the $(i+1)$-th number is the largest number of the first $i+1$ numbers is $\frac{1}{i+1}$; the probability that the $(i+2)$-th number is the largest of the first $i+2$ numbers is $\frac{1}{i+2}$, etc. Hence,

$$r_i = \Big(1 - \frac{1}{i+1}\Big)\Big(1 - \frac{1}{i+2}\Big) \cdots \Big(1 - \frac{1}{N}\Big) = \frac{i}{N}.$$

The transition probabilities are:

$$\begin{aligned} p_{ij} &= \text{the probability that the numbers } (i+1) \text{ up to and including number } (j-1) \text{ are} \\ &\quad \text{smaller than number } i \text{ and number } j \text{ is larger that number } i,\ j \geq i+1. \\ &= \left(1-\tfrac{1}{i+1}\right)\left(1-\tfrac{1}{i+2}\right)\cdots\left(1-\tfrac{1}{j-1}\right)\cdot\tfrac{1}{j} = \tfrac{i}{(j-1)j}. \\ S_0 &= \{i \in S \mid r_i \geq -c_i + \textstyle\sum_j p_{ij} r_j\} = \{i \in S \mid \tfrac{i}{N} \geq \sum_{j=i+1}^{N} \tfrac{i}{(j-1)j}\cdot\tfrac{j}{N}\} \\ &= \{i \in S \mid \tfrac{1}{i} + \tfrac{1}{i+1} + \cdots + \tfrac{1}{N-1} \leq 1\}. \end{aligned}$$

Because $\frac{1}{i} + \frac{1}{i+1} + \cdots + \frac{1}{N-1}$ is monotone decreasing in i, we have $S_0 = \{i \in S \mid i \geq i_*\}$, where $i_* = min\{i \mid \frac{1}{i} + \frac{1}{i+1} + \cdots + \frac{1}{N-1} \leq 1\}$. Because obviously S_0 is closed, the problem is monotone and therefore the optimal policy chooses the stopping action as soon as i_* drawn are made and that draw results in the largest number up to now. The value vector can be computed as follows.

If $i \geq i_*$: $v_i = r_i = \frac{i}{N}$.

If $i < i_*$: $v_i = -c_i + \sum_j p_{ij} v_j = \sum_{j=i+1}^{N} \frac{i}{(j-1)j}\cdot v_j = i \cdot \sum_{j=i+1}^{N} \frac{1}{(j-1)j}\cdot v_j$.

For $2 \leq i \leq i_* - 1$, we have:

$$\begin{aligned} v_{i-1} &= (i-1)\cdot\textstyle\sum_{j=i}^{N} \tfrac{1}{(j-1)j}\cdot v_j = (i-1)\cdot\left\{\tfrac{1}{i(i-1)} v_i + \sum_{j=i+1}^{N} \tfrac{1}{(j-1)j}\cdot v_j\right\} \\ &= \tfrac{1}{i} v_i + \tfrac{i-1}{i} v_i = v_i. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} v_1 &= v_2 = \cdots = v_{i_*-1} = (i_*-1)\cdot\textstyle\sum_{j=i_*}^{N} \tfrac{1}{(j-1)j}\cdot v_j \\ &= (i_*-1)\cdot\textstyle\sum_{j=i_*}^{N} \tfrac{1}{(j-1)j}\cdot\tfrac{j}{N} = \tfrac{i_*-1}{N}\cdot\sum_{j=i_*}^{N}\tfrac{1}{j-1}. \end{aligned}$$

Example 4.6

Problems like searching a target can often be modelled as an optimal stopping problem. Suppose that we are searching for an object with value r and that there is an apriori probability p that the object is in the search area. If we search in this area, for each search there are searching costs c and there is a probability β that we find the object, if it is in this area. A maximum of N searches is allowed. Of course, when the object is found, then we stop; but if the object is not found, will we do another search?

Let $S = \{0, 1, \ldots, N-1\}$, where state i means that i failured searches have been done.

In state i, we have the posteriori probability $p_i = \frac{p(1-\beta)^i}{p(1-\beta)^i + (1-p)}$ that the object is present.

Hence, we obtain $r_i = 0$ and $c_i = c - p_i \cdot \beta \cdot r,\ i \in S$, and $p_{ij} = \begin{cases} 1 - p_i\cdot\beta & ,\ i \in S,\ j = i+1; \\ 0 & ,\ i \in S,\ j \neq i+1. \end{cases}$

$S_0 = \{i \mid r_i \geq -c_i + \sum_j p_{ij} r_j\} = \{i \mid c_i \geq 0\} = \{i \mid p_i \leq \frac{c}{\beta\cdot r}\}$.

It is easy to verify that $p_0 \geq p_1 \geq \cdots \geq p_{N-1}$. Hence, $S_0 = \{i \mid i \geq i_*\}$, where $i_* = min\{ \mid p_i \leq \frac{c}{\beta\cdot r}\}$. It is obvious that S_0 is closed. Therefore, to stop in S_0 and continue outside S_0 is optimal. This result is intuitively clear: S_0 consists of the states where the expected netto costs (c_i) are nonnegative. A formula for the stopping states can also be given in the original data, namely

$$
\begin{aligned}
p_i \leq \tfrac{c}{\beta \cdot r} \quad &\Leftrightarrow \quad \tfrac{p(1-\beta)^i}{p(1-\beta)^i+(1-p)} \leq \tfrac{c}{\beta \cdot r} \quad \Leftrightarrow \quad p(1-\beta)^i \leq \tfrac{(1-p)c}{\beta \cdot r - c} \\
&\Leftrightarrow \quad p(1-\beta)^i \leq \tfrac{(1-p)c}{\beta \cdot r - c} \quad \Leftrightarrow \quad (1-\beta)^i \leq \tfrac{1-p}{p} \cdot \tfrac{c}{\beta \cdot r - c} \\
&\Leftrightarrow \quad i \geq \tfrac{log\left\{\frac{1-p}{p} \cdot \frac{c}{\beta \cdot r - c}\right\}}{log\,(1-\beta)} \quad = \quad i_*.
\end{aligned}
$$

4.8 Bibliographic notes

The study of Markov decision models with the expected total reward criterion originated with the book *How to gamble if you must* by Dubins and Savage ([61]), with appeared in 1965. Lemma 4.1 is due to [100]. Theorem 4.1 was shown by Blackwell in 1962 ([22]). Blackwell called such a policy *1-optimal*. Later, this property was called *Blackwell optimal* in honor to Blackwell. The proof of Theorem 4.2 follows the line of reasoning in the proof of Theorem 6.1 in [168]. The concepts 'conserving' and 'equalizing' are due to Dubins and Savage ([61]). Example 4.1 appeared in [206].

The name 'contracting' was introduced was introduced by Van Nunen and Wessels, who have studied this model systematically (cf. [211] and [208]). The equivalence between the first three statements in Theorem 4.3 appeared in [217]. The equivalence with statement (4) is due to Hordijk ([92]) and with (5) to Kallenberg ([108]). Related papers are [98] and [63].

Seminal papers on positive and negative MDPs are [23] and [193]. The sections 4.4 and 4.5 deal with linear programming and follow Kallenberg ([108], section 3.5 and 3.6). For value iteration we refer to Van der Wal ([203]) and for policy iteration to Puterman ([157]). References for convergent MDPs are Hordijk ([90],[91],[92]) and Van der Wal ([203]).

Gambling theory can be found in [61]. For the proof of the optimality of the timid policy (case $p > \frac{1}{2}$) we refer to [169]. The proof for the bold policy (case $p < \frac{1}{2}$) is based on unpublished work of Denardo ([48]).

4.9 Exercises

Exercise 4.1

Consider the following model.
$S = \{1,2\};\ A(1) = \{1,2\},\ A(2) = \{1\};\ p_{11}(1) = 1,\ p_{12}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1;$
$p_{21}(1) = 0,\ p_{22}(1) = 0.5;\ r_1(1) = r_1(2) = r_2(1) = 1.$
Construct a sequence of stationary Markov policies $\pi^\infty(n),\ n = 1,2,\ldots$ such that $v_1\big(\pi^\infty(n)\big) < \infty$ for $n = 1,2,\ldots$ and $sup_n\, v_1\big(\pi^\infty(n)\big) = +\infty$.

Exercise 4.2

A policy R is called *stopping* if $lim_{t\to\infty} \mathbb{P}_{i,R}\{X_t = j, Y_t = a\} = 0$ for all i, j and a.

a. Show that any transient policy is stopping.

b. Consider the model $S = \{1\};\ A(1) = \{1, 2\};\ p_{11}(1) = 1;\ p_{11}(2) = 0.5$ with policy R that takes action 2 at the time points $t = 2^n,\ n = 1, 2, 3, \ldots$. Show that R is stopping, but not transient.

c. Show that a stationary policy π^∞ is transient if and only if π^∞ is stopping.

Exercise 4.3

Prove case 3 of Lemma 4.1.

Exercise 4.4

Show that an MDP is contracting if and only if there exists a solution to the system

$$\begin{cases} \mu_i = max_a \{1 + \sum_j p_{ij}(a)\mu_j\} & ,\ i \in S \\ \mu_i \geq 0 & ,\ i \in S \end{cases}$$

Exercise 4.5

Consider a contracting MDP. An action $a \in A(i)$ is *suboptimal* if $r_i(a) + \sum_j p_{ij}(a)v_j < v_i$.
Consider the dual linear program in Algorithm 4.4 and let $f^\infty \in C(D)$ be the policy corresponding to some simplex tableau in which the x-variables have values $x_i^f(a),\ (i, a) \in S \times A$.
Let $d_i^f(a)$ be the value of the dual variable which corresponds to $x_i^f(a),\ (i, a) \in S \times A$.
Show the following properties:

a. If $d_i^f(a_i) > min_a\, d_i^f(a) + \sum_j p_{ij}(a_i)\{b_j - v_j(f^\infty)\}$, where b is an upper bound of the value vector v, then action a_i is an suboptimal action.

b. $b = v(f^\infty) - \frac{min_{(i,a)}\{d_i^f(a)/\mu_i\}}{1-\alpha} \cdot \mu$ is an upper bound of the value vector v, where α and μ are such that $\mu_i > 0,\ i \in S,\ \alpha \in [0, 1)$ and $\sum_j p_{ij}(a)\mu_j \leq \alpha \cdot \mu_i$ for all $(i, a) \in S \times A$.

Exercise 4.6

Consider the following model:
$S = \{1, 2, 3, 4, 5, 6, 7\};\ A(1) = A(2) = \{1, 2\},\ A(3) = \{1, 2, 3\},\ A(4) = \{1, 2\},\ A(5) = \{1, 2, 3\},$
$A(6) = \{1\},\ A(7) = \{1, 2\};\ p_{11}(1) = 1,\ p_{13}(2) = 1,\ p_{21}(1) = 1,\ p_{24}(2) = 1,\ p_{33}(1) = 0.5,$
$p_{31}(2) = 1,\ p_{37}(3) = 1,\ p_{43}(1) = 1,\ p_{42}(1) = 1,\ p_{54}(1) = 0.5,\ p_{53}(2) = 1,\ p_{56}(1) = 1,$
$p_{67}(1) = 0.5,\ p_{77}(1) = 0.5,\ p_{76}(2) = 1$ (the other transition probabilities are zero);
$r_1(1) = 0,\ r_1(2) = 0,\ r_2(1) = 0,\ r_2(2) = 2,\ r_3(1) = 1,\ r_3(2) = 1,\ r_3(3) = 1,\ r_4(1) = 1,$
$r_4(2) = 1,\ r_5(1) = 1,\ r_5(2) = 2,\ r_5(3) = 3,\ r_6(1) = 1,\ r_7(1) = 1,\ r_7(2) = 1.$
Use Algorithm 4.6 to determine an optimal policy. Take $\beta_i = \frac{1}{7},\ i \in S$.

Exercise 4.7

Consider the following model:

$S = \{1, 2, 3\};\ A(1) = \{1, 2\},\ A(2) = A(3) = \{1\};\ p_{11}(1) = p_{12}(2) = p_{23}(1) = p_{33}(1) = 1$

(the other transition probabilities are zero); $r_1(1) = 0,\ r_1(2) = 2,\ r_2(1) = -1,\ r_3(1) = 0.$

a. Is this model convergent?

b. Determine the value vector v and show that v has the property anne.

c. Compute the value vector and an optimal policy by linear programming.

d. What happens in value iteration, given by (4.13)?

e. Is the problem stable?

Exercise 4.8

Show that the red-black gambling model is contracting.

Exercise 4.9

Consider an optimal stopping problem with the data:

$S = \{1, 2, 3, 4\};\ r_1 = 0,\ r_2 = 1,\ r_3 = 2,\ r_4 = 2;\ c_i = 0,\ 1 \leq i \leq 4.$

$p_{11} = \frac{1}{2},\ p_{12} = \frac{1}{2},\ p_{13} = 0,\ p_{14} = 0;\ p_{21} = \frac{1}{8},\ p_{22} = \frac{1}{8},\ p_{23} = \frac{1}{2},\ p_{24} = \frac{1}{4};$

$p_{31} = \frac{1}{3},\ p_{32} = \frac{1}{3},\ p_{33} = \frac{1}{3},\ p_{34} = 0;\ p_{41} = \frac{1}{4},\ p_{42} = \frac{1}{8},\ p_{43} = \frac{1}{2},\ p_{44} = \frac{1}{8}.$

Determine an optimal policy for this problem.

Exercise 4.10

Every night a thief is going for robbery and with probability p_k he will capture an amount of k, $k = 0, 1, \ldots, n$. The probability to be catched is equal to p, and if he is catched, he will loose the total captures of all previous nights and he must stop.

At which captured amount, the thief will stop with robbery?

Show that the solution of this problem is: the thief stops aa soon as he has captured the amount $\frac{1-p}{p} \cdot \sum_{j=0}^{n} p_j \cdot j$ and gives an intuitive explanation of this result

Hint:

Use as state space $S = \{0, 1, \ldots\}$ and assume that the results of the optimal stopping problem are also true for this infinite state space.

Chapter 5

Average reward - general case

5.1 Introduction

When decisions are made frequently, so that the discount rate is very close to 1, or when performance criterion cannot easily be descibed in economic terms with discount factors, the decision maker may prefer to compare policies on the basis of their average expected reward instead of their expected total discounted reward. Consequently, the average reward criterion occupies a cornerstone of queueing control theory especially when applied to controlling computer systems and communication networks. In such systems, the controller makes frequent decisions and usually assesses system performance on the basis of throughput rate or the average time a job remains in the system. This optimality criterion may also be appropriate for inventory systems with frequent restocking decisions.

In the criterion of average reward the limiting behaviour of $\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t)$ is considered for $T \to \infty$. Since $lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r_{X_t}(Y_t)$ may not exist and interchanging limit and expectation is not allowed in general, there are four different evaluation measures which can be considered:

1. Lower limit of the average expected reward:
   $\phi_i(R) = liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i,R}\{r_{X_t}(Y_t)\},\ i \in S$, with value vector $\phi = sup_R\ \phi(R)$.
2. Upper limit of the average expected reward:
   $\overline{\phi}_i(R) = limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{i,R}\{r_{X_t}(Y_t)\},\ i \in S$, with value vector $\overline{\phi} = sup_R\ \overline{\phi}(R)$.
3. Expectation of the lower limit of the average reward:
   $\psi_i(R) = \mathbb{E}_{i,R}\{liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ r_{X_t}(Y_t)\},\ i \in S$, with value vector $\psi = sup_R\ \psi(R)$.
4. Expectation of the upper limit of the average reward:
   $\overline{\psi}_i(R) = \mathbb{E}_{i,R}\{limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ r_{X_t}(Y_t)\},\ i \in S$, with value vector $\overline{\psi} = sup_R\ \overline{\psi}(R)$.

As already mentioned in Section 1.2.2, these four criteria are equivalent in the sense that an optimal deterministic policy for one criterion is also optimal for the other criteria. We will use criterion 1, the lower limit of the average expected reward.

In this chapter we start with the classification of MDP models on the basis of the chain structure. Because the average reward criterion depends on the limiting behaviour of the underlying stochastic processes, this structure is of interest. In the subsequent section the stationary matrix, the fundamental matrix and the deviation matrix of a Markov chain is discussed. These matrices play an important role in the average reward criterion and also in more sensitive criteria. The most sensitive criteron is Blackwell optimality. Laurent series expansion relates the average reward to the total discounted reward. This is the subject of section 5.5. The last sections of this chapter deal with policy iteration, linear programming and value iteration.

5.2 Classification of MDPs

5.2.1 Definitions

There are several ways to classify MDPs. The first one distinguishes between *communicating* and *noncommunicating*. An MDP is communicating if for every $i, j \in S$ there exists a policy $f^\infty \in C(D)$, which may depend on i and j, such that in the Markov chain $P(f)$ state j is accessible from state i. An MDP is *weakly communicating* if $S = S_1 \cup S_2$, where $S_1 \cap S_2 = \emptyset$, S_1 is a closed communicating set under *some* policy $f^\infty \in C(D)$ and S_2 is a (possibly empty) set of states which are transient under *all* policies.
A second kind of classification concerns the ergodic structure. One distinguish between *irreducible, unichain* and *multichain* MDPs. An MDP is irreducible (also called *completely ergodic*) if the Markov chain $P(f)$ is irreducible for every $f^\infty \in C(D)$. An MDP is a unichain MDP if the Markov chain $P(f)$ is a unichain Markov chain (exactly one ergodic class plus a possibly empty set of transient states) for every $f^\infty \in C(D)$. An MDP is multichain if there exists a policy $f^\infty \in C(D)$ for which the Markov chain $P(f)$ has (at least) two ergodic classes.
The next result is obvious.

Lemma 5.1
An irreducible MDP is communicating and unichain.

5.2.2 Classification of Markov chains

For a single Markov chain it is easy to determine whether or not the Markov chain belangs to a certain class. Easy means polynomially solvable, i.e. the problem belongs in terms of the complexity theory to the class $\mathcal{P}$ of problems solvable in polynomial-time.
Consider a Markov chain with transition matrix P. The classification of the Markov chain can be executed in the *associated directed graph* $G(P) = \big(V(P), A(P)\big)$, where the nodes $V(P)$ are the states of the Markov chain and the arcs of $A(P)$ satisfy $A(P) = \{(i,j) \mid p_{ij} > 0\}$.
Since a strongly connected component of $G(P)$ is closed if and only if the corresponding states in the Markov chain are an ergodic class, the following algorithm determines the ergodic classes $E_1, E_2, \ldots, E_m$ and the set T of transient states.

Algorithm 5.1 *Ergodic classes and transient states of a Markov chain*

1. a. Determine the strongly connected components of $G(P)$, say $C_1, C_2, \ldots, C_n$.

   b. $m = 0$ and $T = \emptyset$.

2. For $i = 1$ to n do

   if C_i is closed: $m := m + 1$ and $E_m = C_i$;

   else: $T := T \cup C_i$.

The determination of the strongly connected components of a graph can be done in $\mathcal{O}(p) = \mathcal{O}(N^2)$, where p is the number of arcs of the graph (see [195]). For the examination whether the strongly connected components are closed or open, it is also sufficient to consider the arcs of the graph. Therefore, Algorithm 5.1 has complexity $\mathcal{O}(N^2)$.

5.2.3 Classification of Markov decision chains

An MDP has $\prod_{i \in S} |A(i)|$ different deterministic policies and each policy induces a Markov chain. Therefore, MDPs are also called *Markov decision chains.* The approach to analyse all Markov chains separately is prohibitive. The problem to determine whether or not an MDP belongs to a certain class is a combinatorial problem. It turns out that all classification problems are easy, i.e. polynomially solvable, except one. Checking the unichain condition is an $\mathcal{NP}$-hard problem.

For the analysis of the chain structure we use two directed graphs, G_1 and G_2, both with as node set the states of the MDP. In $G_1 = (S, A_1)$ the arc set $A_1 = \{(i, j) \mid p_{ij}(a) > 0 \text{ for every } a \in A(i)\}$. Hence, a path from i to j in G_1 means that state j is accessible from state i under *every* policy. In $G_2 = (S, A_2)$ the arc set $A_2 = \{(i, j) \mid p_{ij}(a) > 0 \text{ for some } a \in A(i)\}$, and a path from i to j in G_2 means that state j is accessible from state i under *some* policy. Let $M = \sum_{i \in S} |A(i)|$, then the construction of the graphs G_1 and G_2 has complexity $\mathcal{O}\big(\sum_{j \in S}\{\sum_{i \in S} |A(i)|\}\big) = \mathcal{O}(M \cdot N)$.

Communicating

The question whether or not an MDP is communication is solved by the following lemma.

Lemma 5.2

An MDP is communicating if and only if the graph G_2 is strongly connected.

Proof

$\Rightarrow$ Suppose that G_2 is not strongly connected, i.e. there are nodes i and j such that in G_2 is no path from i to j. This implies that for every $f^\infty \in C(D)$ in the Markov chain $P(f)$ state j is not accessible from state i. Consequently, the MDP is not communicating.

$\Leftarrow$ Suppose that G_2 is strongly connected and take any pair $i, j \in S$. Since G_2 is strongly connected there is a path from i to j. Hence, j is accessible from i under some policy. This implies the property communicating.

□

The above Lemma implies the following algorithm for checking the communicating property of an MDP. Since the construction of G_2 has complexity $\mathcal{O}(M \cdot N)$, and the determination of the strongly connected components is of order $N^2 \leq M \cdot N$, the total complexity is $\mathcal{O}(M \cdot N)$.

Algorithm 5.2 *Checking the communicating property of an MDP*

1. Construct the graph G_2.
2. Determine the strongly connected components of G_2, say $C_1, C_2, \ldots, C_n$.
3. If $n = 1$: the MDP is communicating (STOP).

   If $n \geq 2$: the MDP is noncommunicating.

If the outcome of Algorithm 5.2 is 'noncommunicating' ($n \geq 2$) one may ask whether the MDP is perhaps weakly communicating. If two or more of the strongly connected components are closed, then the MDP is not weakly communicating, since in that case there are two disjunct sets of states which both are ergodic under all policies.
If only one of the strongly connected components is closed, say C_1, one can try to find a state outside C_1, say state i, for which there is a positive transition probability to C_1 under all actions $a \in A(i)$. If such state does not exist, then there is a policy with the property that starting outside C_1 one never enters C_1. Hence, the MDP is not weakly communicating. Continuing in this way yields the following algorithm.

Algorithm 5.3 *Checking the communicating and weakly communicating property of an MDP*

1. Construct the graph G_2.
2. a. Determine the strongly connected components of G_2, say $C_1, C_2, \ldots, C_n$.

   b. $m = 0$ and $T = \emptyset$.
3. For $i = 1$ to n do

   if C_i is closed: $m := m + 1$ and $E_m = C_i$;

   else: $T := T \cup C_i$.
4. If $m \geq 2$: the MDP is not weakly communicating (STOP);

   else: if $T = \emptyset$: the MDP is communicating (STOP);

   else: go to step 5.
5. a. $S_1 = E_1,\ S_2 = \emptyset$

   b. Repeat

   $k = 0$

   For every $i \in T$ do:

   if $\sum_{j \in S_1 \cup S_2} p_{ij}(a) > 0$ for every $a \in A(i)$: $S_2 := S_2 \cup \{i\},\ T := T \backslash \{i\},\ k = 1$.

   Until $k = 0$

6. If $T = \emptyset$: the MDP is weakly communicating (STOP);

   else: the MDP is not weakly communicating.

For the complexity of Algorithm 5.3 we remark that the steps 1 until 4 are executed only once and have complexity $\mathcal{O}(M \cdot N)$. Step 5 is executed at most N times and each step has complexity of order $\sum_{i \in T} \sum_{a \in A(i)} |S_1 \cup S_2| \leq M \cdot N$. Hence, the complexity of step 5, and also the overall complexity of the algorithm is $\mathcal{O}(M \cdot N^2)$.

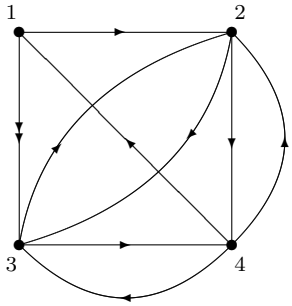
Irreducibility

For the irreducibility we use graph G_1. If G_1 is strongly connected, then the MDP is irreducible, because each pair of states communicates under every policy. If G_1 is not strongly connected we *condense* graph G_1 to graph G_1^c. The condensed graph G_1^c has a (compound) vertex for each strongly connected component of G_1. Let i and j be the compound vertices of G_1^c corresponding to the strongly connected components C_k and C_l, and let V_k and V_l be the vertex sets in G_1 of C_k and C_l, respectively. Then, (i, j) is an arc in G_1^c if every Markov chain in the MDP has a positive one-step transition from some state of V_k to some state of V_l, i.e.

$$max_{r \in V_k} \{ min_{a \in A(r)} \sum_{s \in V_l} p_{rs}(a) \} > 0.$$

Since states in the same strongly connected component communicate under every policy, an arc (i, j) in G_1^c means that any $s \in V_l$ is accessible from any $r \in V_k$ under every policy. It is easy to verify that the construction of the condensed graph G_1^c has complexity $\mathcal{O}(M \cdot N)$. The operation 'condensation' can be repeated until there are no changes in the graph. Let $\{G_1^c\}^*$ be the finally, after repeated condensations, obtained graph.

Example 5.1



Let $S = \{1, 2, 3, 4\}$; $A(1) = \{1, 2\}$, $A(2) = \{1\}$,
$A(3) = \{1\}$, $A(4) = \{1\}$.
$p_{12}(1) = 1$; $p_{13}(2) = 1$; $p_{23}(1) = p_{24}(1) = 0.5$;
$p_{32}(1) = p_{34}(1) = 0.5$; $p_{41}(1) = 0.5$, $p_{42}(1) = p_{43}(1) = 0.25$.
The graph at the right hand side presents the MDP model.
Graph G_1 is the same, but without the arcs $(1, 2)$ and $(1, 3)$.
The strongly connected components of G_1 are: $C_1 = \{1\}$ and $C_2 = \{2, 3, 4\}$.
$G_1^c = (V_1^c, A_1^c)$ with $V_1^c = \{1^*, 2^*\}$, where 1^* corresponds to state 1 and 2^* to the states $2, 3$ and 4, and $A_1^c = \{(1^*, 2^*), (2^*, 1^*)\}$.
When G_1^c is condensed we obtain $\{G_1^c\}^*$, consisting of a single vertex.

The next lemma shows that irreducibility is equivalent to the property that $\{G_1^c\}^*$ consists of a single vertex.

Lemma 5.3
An MDP is irreducible if and only if the ultimate condensation $\{G_1^c\}^*$ consists of a single vertex.

Proof
$\Rightarrow$ Suppose that $\{G_1^c\}^*$ has at least two vertices. Then, there is a (compound) vertex, say i, without an incoming arc (if each vertex has an incoming arc, there is a circuit and the graph can be condensed). Therefore, in each state of the (compound) vertices $j \neq i$ an action can be chosen with transition probability 0 to the states of i. The Markov chain under such policy is not irreducible.

$\Leftarrow$ Let $\{G_1^c\}^*$ consists of a single vertex. From the definition of condensation it follows that each two states communicate under any policy, i.e. the Markov chain is irreducible. □

Algorithm 5.4 *Checking the irreducibility property of an MDP*

1. Construct the graph G_1 and let $G = G_1$.
2. Determine the strongly connected components of G, say $C_1, C_2, \ldots, C_n$.
3. If all components consist of one vertex: go to step 4.
   else: construct the condensed graph of G, say G^c; $G := G^c$ and return to step 2.
4. If $n = 1$: the MDP is irreducible (STOP);
   else: the MPD is not irreducible.

The construction of G_1, the determination of the strongly connected components and the condensation operation have complexity of at most $\mathcal{O}(M \cdot N)$. In a new iteration the number of vertices of G decreases, so the number of iterations is at most N and the overall complexity of Algorithm 5.4 is $\mathcal{O}(M \cdot N^2)$.

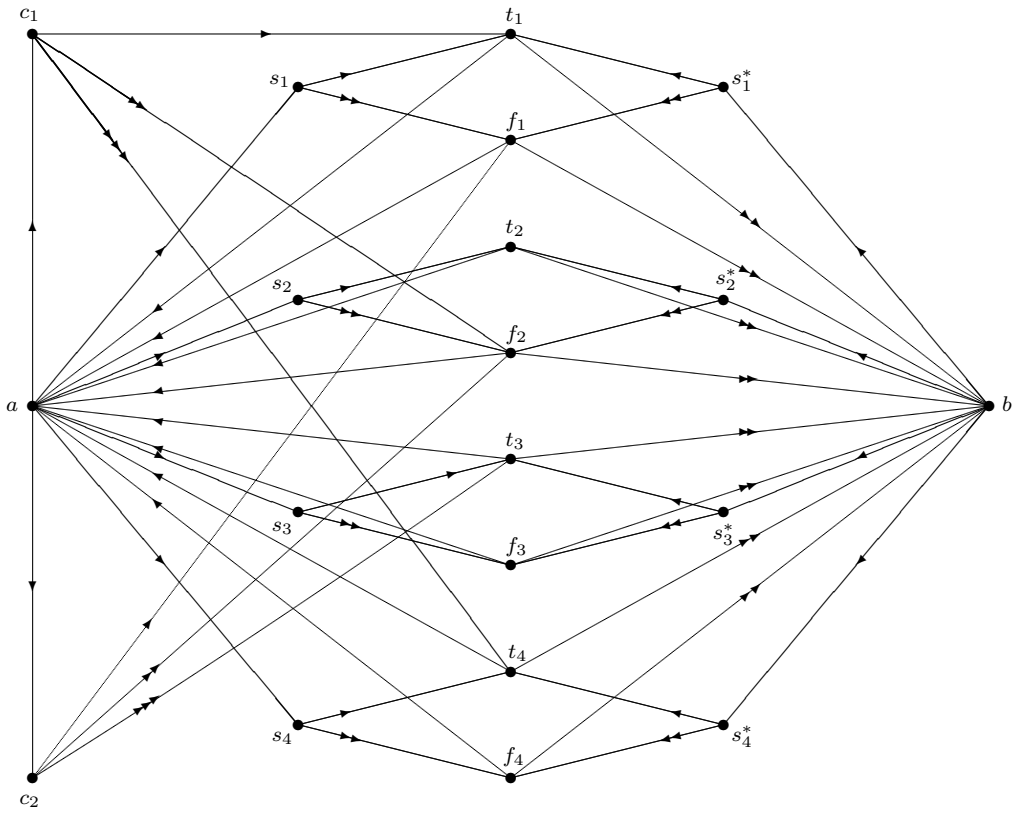
The last classification question concerns the distinction between unichain and multichain. It turns out that this decision problem is $\mathcal{NP}$-complete, so there is no hope of a polynomial algorithm.

Suppose that there exists a policy that results in multiple ergodic classes. Such a policy serves as a certificate that the answer is "yes". Since the determination of the ergodic classes of a Markov chain is polynomially (see Algorithm 5.1), the problem is in $\mathcal{NP}$.

To prove that the problem is $\mathcal{NP}$-complete we use a reduction to the 3-satisfiability problem ($3SAT$). An instance of $3SAT$ consists of n Boolean variables $x_1, x_2, \ldots, x_n$, and m clauses $C_1, C_2, \ldots, C_m$, with three literals per clause. Each clause is the disjunction of three literals, where a literal is either a variable x_i or its negative $\overline{x}_i$, for example $C = x_2 \cup \overline{x}_4 \cup x_5$. The question is whether there is an assigment of values ("true" or "false") to the variables such that all clauses are satisfied.

Suppose that we are given an instance of $3SAT$, with n variables and m clauses.
We construct an MDP as follows:
(a) two special states a and b;
(b) $4n$ states $s_i, s_i^*, t_i, f_i,\ i = 1, 2, \ldots, n$;
(c) m states $c_j,\ j = 1, 2, \ldots, m$.
For the actions and the transition probabilities, we have:
$A(a) = \{1\}$ and $p_{as_i}(1) = \frac{1}{n+m},\ 1 \le i \le n;\ p_{ac_j}(1) = \frac{1}{n+m},\ 1 \le j \le m$.
$A(b) = \{1\}$ and $p_{bs_i^*}(1) = \frac{1}{n},\ 1 \le i \le n$.
$A(s_i) = \{1, 2\}$ and $p_{s_it_i}(1) = 1,\ 1 \le i \le n;\ p_{s_if_i}(2) = 1,\ 1 \le i \le n$.
$A(s_i^*) = \{1, 2\}$ and $p_{s_i^*t_i}(1) = 1,\ 1 \le i \le n;\ p_{s_i^*f_i}(2) = 1,\ 1 \le i \le n$.
$A(t_i) = \{1, 2\}$ and $p_{t_ia}(1) = 1,\ 1 \le i \le n;\ p_{t_ib}(2) = 1,\ 1 \le i \le n$.
$A(f_i) = \{1, 2\}$ and $p_{f_ia}(1) = 1,\ 1 \le i \le n;\ p_{f_ib}(2) = 1,\ 1 \le i \le n$.
$A(c_j) = \{1, 2, 3\}$ and action a corresponds to the a-th literal of clause C_j. In particilar, if the a-th literal in clause C_j is of the form x_i, then $p_{c_jt_i}(a) = 1$; if the a-th literal in clause C_j is of the form $\overline{x}_i$, then $p_{c_jf_i}(a) = 1$.

Example 5.2
Suppose that $n = 4,\ m = 2$ and $C_1 = x_1 \cup \overline{x}_2 \cup x_4,\ C_2 = \overline{x}_1 \cup \overline{x}_2 \cup x_3$. Below we draw the corresponding MDP. The transition probabilities are 1, except from a (the probabilities are $\frac{1}{6}$) and from b (the probabilities are $\frac{1}{4}$).



We claim that we have a "yes" instance of $3SAT$ if and only if the correponding MDP is multichain. Suppose that we have a "yes" instance of $3SAT$. Consider an assignment of the variables such that all clauses are satisfied. We define the following policy:
(1) At every state c_j, consider a literal in the clause which is "true". If that literal is unnegated, say x_k, pick in state c_j the action that moves to state t_k; if that literal is negated, say $\overline{x}_k$, pick in state c_j the action that moves to state f_k.
(2) At every state s_i, let the next state be t_i if x_i is "true", and f_i if x_i is "false".
(3) At every state s_i^*, let the next state be f_i if x_i is "true", and t_i if x_i is "false".
(4) At every state t_i, let the next state be a if x_i is "true", and b if x_i is "false".
(5) At every state f_i, let the next state be b if x_i is "true", and a if x_i is "false".
(6) In the states a and b is only one action.

First, look at state a as starting state of the Markov chain. At the next time point the Markov chain is in some state s_i or in some state c_j.
If the next state is s_i, then the following happens:
- if x_i is "true" the next state is t_i and we return to state a;
- if x_i is "false" the next state is f_i and we return to state a.
If the next state is c_j, then the following happens:
- if the chosen action corresponds to an unnegated variable x_k: the next state is t_k and we return to state a;
- if the chosen action corresponds to a negated variable $\overline{x}_k$: the next state is f_k and we return to state a.
We conclude that a is a recurrent state and, starting from a, state b is never visited.

Next, look at state b as starting state of the Markov chain. At the next time point the Markov chain is in some state s_i^* and the following happens:
- if x_i is "true" the next state is f_i and we return to state b;
- if x_i is "false" the next state is t_i and we return to state b.
We conclude that b is a recurrent state and, starting from b, state a is never visited.
Therefore, the MDP is multichain.

For the converse, suppose that the MDP is multichain, and fix a policy that results in multiple ergodic classes. Given, the structure of the possible transitions, the state belongs to the set $\{a, b\}$ once every three transitions. Since we have multiple ergodic classes, it follows that a and b are both recurrent but do not belong to the same ergodic class. In particular, b is not accessible from a, and a is not accessible from b.

Consider the following assignment of the variables: if in state s_i action 1 is chosen, set x_i "true", and if in state s_i action 2 is chosen, set x_i "false". We need to show that with this assignment all clauses are satisfied.
Suppose that the transition out of s_i leads to t_i (i.e. $x_i = 1$). Since b is not accessible from a, it follows that b is not accessible from t_i, and therefore the action out of t_i leads back to a. Since

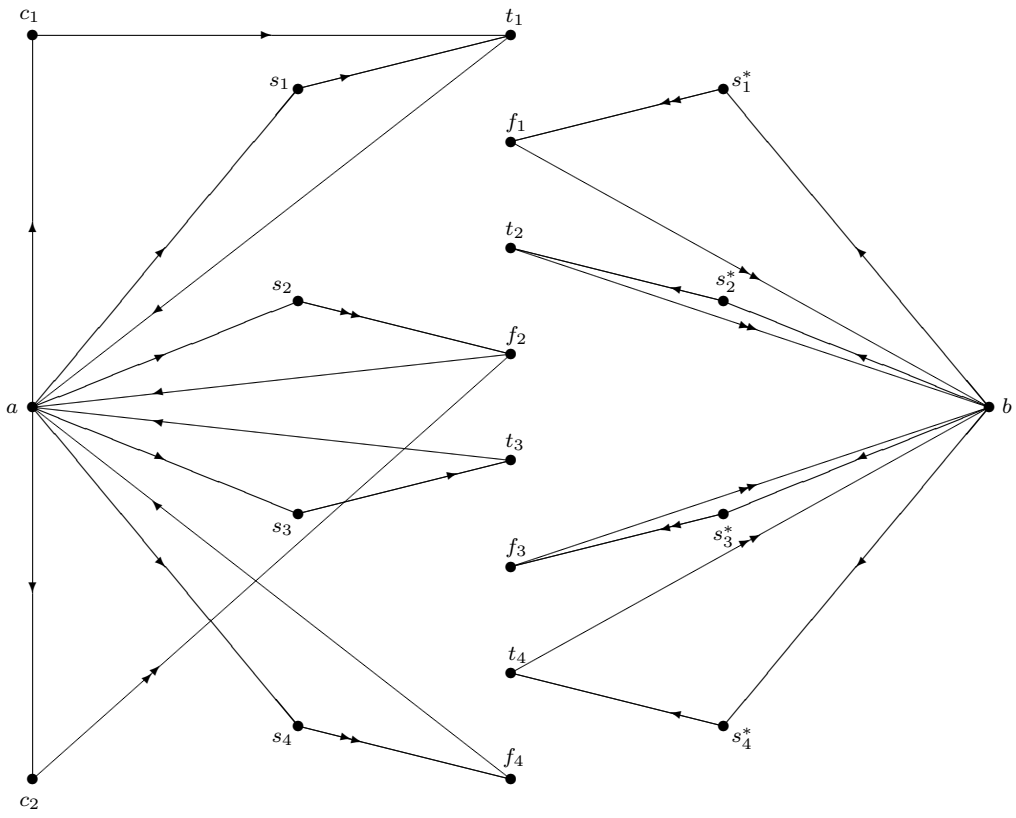
a is not accessible from b, the transaction out of s_i^* leads to f_i and then back to b. Similarly, suppose that the transition out of s_i leads to f_i (i.e. $x_i = 0$). Since b is not accessible from a, it follows that b is not accessible from f_i, and therefore the action out of f_i leads back to a. Since a is not accessible from b, the transaction out of s_i^* leads to t_i and then back to b.

Consider now a clause C_j and suppose that the transition in state c_j leads to t_i, i.e. x_i is part of clause C_j. Since b is not accessible from a, it follows that t_i leads back to a. Using the remarks above, it follows that the transition out of s_i leads to t_i, and therefore x_i is set to "true", and the clause is satisfied.
Suppose that the transition in state c_j leads to f_i, i.e. $\overline{x}_i$ is part of clause C_j. Since b is not accessible from a, it follows that f_i leads back to a. Using the earlier remarks, it follows that the transition out of s_i leads to f_i, and therefore x_i is set to "false", and the clause is satisfied.

By the above arguments, we have shown the following theorem.

Theorem 5.1
The determination problem whether or not an MDP is unichain or multichain is $\mathcal{NP}$-complete.

Example 5.2 (continued)
Consider the assignment $x_1 = 1,\ x_2 = 0,\ x_3 = 1,\ x_4 = 0$. As corresponding policy we take action 1 in state c_1 and action 2 in state c_2. The Markov chain of this policy is presented in the figure below. It is easy to see that this chain is has two ergodic classes.



5.3 Stationary, fundamental and deviation matrix

5.3.1 The stationary matrix

Consider a policy $f^\infty \in C(D)$. In average reward MDPs the limiting behaviour of $P^n(f)$ as n tends to infinity plays an important role. In general, $lim_{n\to\infty} P^n(f)$ does not exist (a counterexample is left to the reader). Therefore, we consider other types of convergence.

Let $\{b_n\}_{n=0}^\infty$ be a sequence. This sequence is called *Cesaro convergent* with Cesaro limit b if

$$lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{k=0}^{n-1} b_k \text{ exits and is equal to } b.$$

We denote this convergence by $lim_{n\to\infty} b_n =_c b$ or $b_n \to_c b$. The sequence is said to be *Abel convergent* with Abel limit b if

$$lim_{\alpha\uparrow 1}(1-\alpha) \textstyle\sum_{n=0}^\infty \alpha^n b_n \text{ exits and is equal to } b.$$

This convergence is denoted by $lim_{n\to\infty} b_n =_a b$ or $b_n \to_a b$. Ordinary convergence implies both Cesaro and Abel convergence, but the converse statements are not true in general (see Exercise 5.2). The next result is well known in the theory of the summability of series (e.g. Powell and Shah [155], p. 9).

Theorem 5.2
If the sequence $\{b_n\}_{n=0}^\infty$ is Cesaro convergent to b, then $\{b_n\}_{n=0}^\infty$ is also Abel convergent to b.

Remark
The converse statement of Theorem 5.2 is not true in general (see Exercise 5.3).

Theorem 5.3
Let P be any stochastic matrix, i.e. the matrix of a Markov chain. Then,
(1) $P^* := lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k$ exits, i.e. $P^n \to_c P^*$.
(2) $P^*P = PP^* = P^*P^* = P^*$.

Proof
Let $B^{(n)} = \frac{1}{n} \sum_{k=0}^{n-1} P^k$. Since P^k is stochastic for every k, $B^{(n)}$ is also a stochastic matrix. Hence, the series $\left\{B^{(n)}\right\}_{n=1}^\infty$ is bounded. Therefore, each infinite subsequence of $\left\{B^{(n)}\right\}_{n=1}^\infty$ has a point of accumulation. Furthermore, we have

$$B^{(n)} + \frac{1}{n}\{P^n - I\} = B^{(n)}P = PB^{(n)}, \ n \in \mathbb{N}. \tag{5.1}$$

Let $J = lim_{k\to\infty} B^{(n_k)}$, where $\left\{B^{(n_k)}\right\}_{k=1}^\infty$ is a convergent subsequence of $\left\{B^{(n)}\right\}_{n=1}^\infty$. From (5.1) we obtain

$$J = JP = PJ. \tag{5.2}$$

Let $\left\{B^{(m_k)}\right\}_{k=1}^{\infty}$ also be a convergent subsequence of $\left\{B^{(n)}\right\}_{n=1}^{\infty}$ with limit matrix K. From (5.1) it also follows that

$$K = KP = PK. \tag{5.3}$$

Hence, $J = P^n J = JP^n$ and $K = P^n K = KP^n$ for every n. Therefore, $J = B^{(n)}J = JB^{(n)}$ and $K = B^{(n)}K = KB^{(n)}$ for every n, implying that $J = KJ = JK$ and $K = JK = KJ$, i.e. $J = K$. The sequence $\left\{B^{(n)}\right\}_{n=1}^{\infty}$ has exactly one point of accumulation, i.e. $P^* := lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k$ exits and is the Cesaro limit of the sequence $\{P^n\}_{n=1}^{\infty}$. Furthermore, we have shown that $P^*P = PP^* = P^*P^* = P^*$. □

The matrix P^* is called the *stationary matrix* of the stochastic matrix P.

Corollary 5.1
$lim_{\alpha\uparrow 1} \sum_{n=0}^{\infty} \alpha^n (P^n - P^*) = 0$.

Proof
Since P^n is Cesaro convergent to P^*, $P^n - P^*$ is Cesaro convergent to 0, and consequently Abel convergent to 0, i.e. $lim_{\alpha\uparrow 1} \sum_{n=0}^{\infty} \alpha^n (P^n - P^*) = 0$. □

Let P be any stochastic matrix with ergodic classes $E_1, E_2, \ldots, E_m$ and transient states T. By renumbering of the states the matrix can be written in the following so-called *standard form*:

$$P = \begin{pmatrix} P_1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & P_2 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & & & \cdot & 0 \\ \cdot & & \cdot & \cdot & \cdot & & \cdot & 0 \\ \cdot & & & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & P_m & 0 \\ A_1 & A_2 & \cdot & \cdot & \cdot & \cdot & A_m & Q \end{pmatrix}, \tag{5.4}$$

where the matrix P_k corresponds to the ergodic class E_k, $1 \le k \le m$, and the matrix Q to the transient states. It is well known (e.g. Doob [60] p. 180) that $Q^n \to 0$ for $n \to \infty$. Since

$$(I - Q)(I + Q + \cdots + Q^{n-1}) = I - Q^n, \tag{5.5}$$

the right hand side of (5.5) tends to I, i.e. $I - Q$ is nonsingular and $(I - Q)^{-1} = \sum_{n=0}^{\infty} Q^n$.[1] From the theory of Markov chains it is also well known (see e.g. e.g. Chung [33] p. 33) that the stationary matrix of an ergodic class has strictly positive, identical rows, say π^k for P_k, and that π^k is the unique solution of the following system of linear equations

[1] A series $\sum_{n=0}^{\infty} A^n$ is a generalization of the geometric series and is often referred to as the *Neumann series*.

$$\begin{cases} \sum_{i\in Ek}(\delta_{ij}-p_{ij})x_i &= 0,\ j\in E_k \\ \sum_{i\in Ek} x_i &= 1 \end{cases} \tag{5.6}$$

Since (5.6) is a system of $|E_k|+1$ equations and $|E_k|$ variables, one of the equations, except the last normalization equation, can be deleted for the computation of π^k.

The following results are also well known (see e.g. Feller [68]).

Lemma 5.4
Let a_i^k be the probability that, starting from state $i\in T$, the Markov chain will be absorbed in ergodic class E_k, $1\le k\le m$. Then, a_i^k, $i\in T$, is the unique solution of the linear system $(I-Q)x=b^k$, where $b^k=A_ke$.

Theorem 5.4
Let P be any stochastic matrix written in the standard form (5.4). Then,

$$P^*=\begin{pmatrix} P_1^* & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & P_2^* & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & & & \cdot & \cdot \\ \cdot & & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & P_m^* & 0 \\ A_1^* & A_2^* & \cdot & \cdot & \cdot & \cdot & A_m^* & 0 \end{pmatrix}, \tag{5.7}$$

where P_k^* has identical rows π^k, which are the unique solution of (5.6) and $A_k^*=\{I-Q\}^{-1}\{A_ke\}\{\pi^k\}^T$, $1\le k\le m$.

Algorithm 5.5 *Determination of the stationary matrix P^**

1. Determine with Algorithm 5.1 the ergodic classes $E_1,E_2,\ldots,E_m$ and the transient states T and write P in standard form (5.4).

2. Determine for $k=1,2,\ldots,m$:

   a. the unique solution π_j^k, $j\in E_k$, of the linear system

   $$\begin{cases} \sum_{i\in E_k}(\delta_{ij}-p_{ij})x_i &= 0,\ j=2,3,\ldots,|E_k| \\ \sum_{i\in E_k} x_i = 1 \end{cases}$$

   b. the unique solution a_i^k, $i\in T$, of the linear sytem $\sum_{j\in T}(\delta_{ij}-p_{ij})x_j=\sum_{l\in E_k}p_{il}$, $i\in T$.

3. $p_{ij}^*=\begin{cases} x_j^k & i\in E_k,\ j\in E_k,\ k=1,2,\ldots,m \\ a_i^kx_j^k & i\in T,\ j\in E_k,\ k=1,2,\ldots,m \\ 0 & \text{else} \end{cases}$

Example 5.3

Consider the Markov chain with transition matrix $P = \begin{pmatrix} 0.5 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.4 & 0.6 \\ 0 & 0.2 & 0.4 & 0 & 0.4 \\ 0.7 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$.

Using Algorithm 5.1 we obtain the ergodic classes $E_1 = \{1, 4\},\ E_2 = \{5\},\ T = \{2, 3\}$.

The stand form of the matrix is: $P = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0.7 & 0.3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 0.4 & 0.2 & 0.4 \end{pmatrix}$.

$k = 1$: π^1 is the unique solution of $\begin{cases} 0.5x_1 - 0.7x_2 = 0 \\ x_1 + x_2 = 1 \end{cases} \rightarrow \pi_1^1 = \frac{7}{12},\ \pi_2^1 = \frac{5}{12}$.

a^1 is the unique solution of $\begin{cases} x_1 = 0.4 \\ -0.2x_1 + 0.6x_2 = 0 \end{cases} \rightarrow a_4^1 = \frac{2}{5},\ a_5^1 = \frac{2}{15}$.

$k = 2$: $\pi^2 = 1$ (state 3 is an absorbing state) and a^2 is the unique solution of

$$\begin{cases} x_1 = 0.6 \\ -0.2x_1 + 0.6x_2 = 0.4 \end{cases} \rightarrow a_4^2 = \frac{3}{5},\ a_5^2 = \frac{13}{15}.$$

The stationary matrix $P^* = \begin{pmatrix} \frac{7}{12} & \frac{5}{12} & 0 & 0 & 0 \\ \frac{7}{12} & \frac{5}{12} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{7}{30} & \frac{5}{30} & \frac{9}{15} & 0 & 0 \\ \frac{7}{90} & \frac{5}{90} & \frac{13}{15} & 0 & 0 \end{pmatrix}$.

5.3.2 The fundamental matrix and the deviation matrix

Theorem 5.5

Let P be an arbitrary stochastic matrix. Then, $I - P + P^*$ is nonsingular and $Z := (I - P + P^*)^{-1}$ satisfies $Z = lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=0}^{i-1} (P - P^*)^k$.

Proof

Since $P^*P = PP^* = P^*P^* = P^*$ (see Theorem 5.3) it follows, by induction on n, that $(P - P^*)^n = P^n - P^*,\ n \in \mathbb{N}$. Let $B := P - P^*$. Since

$$I - B^i = (I - B)(I + B + \cdots + B^{i-1}), \tag{5.8}$$

we have, by averaging (5.8),

$$I - \frac{1}{n} \sum_{i=1}^{n} B^i = (I - B) \cdot \frac{1}{n} \sum_{i=1}^{n} \sum_{k=0}^{i-1} B^k. \tag{5.9}$$

Since $\frac{1}{n}\sum_{i=1}^{n} B^i = \frac{1}{n}\sum_{i=1}^{n}(P^i - P^*) = \frac{1}{n}\sum_{i=1}^{n} P^i - P^*$, we obtain

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

i.e. $I - B = I - P + P^*$ is nonsingular and $Z = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P - P^*)^k$. □

The matrix $Z = (I - P + P^*)^{-1}$ is called the *fundamental matrix* of P.
The *deviation matrix* D is defined by $D := Z - P^* = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P - P^*)^k - P^*$.

Theorem 5.6
The deviation matrix D satisfies
(1) $D = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P^k - P^*)$.
(2) $P^*D = DP^* = (I - P)D + P^* - I = D(I - P) + P^* - I = 0$.

Proof
(1) Since $(P - P^*)^k = (P^k - P^*)$ for $k = 1, 2, \ldots$, we obtain

$\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P - P^*)^k = n \cdot I + \sum_{i=2}^{n}\sum_{k=1}^{i-1}(P - P^*)^k = n \cdot I + \sum_{i=2}^{n}\sum_{k=1}^{i-1}(P^k - P^*)$ and
$\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P^k - P^*) = n \cdot (I - P^*) + \sum_{i=2}^{n}\sum_{k=1}^{i-1}(P^k - P^*)$.

Therefore,

$$\begin{aligned} lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P^k - P^*) &= lim_{n\to\infty}\tfrac{1}{n}\{n \cdot (I - P^*) + \textstyle\sum_{i=2}^{n}\sum_{k=1}^{i-1}(P - P^*)^k \\ &= Z - P^*. \end{aligned}$$

(2) $P^*D = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1} P^*(P^k - P^*) = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P^* - P^*) = 0.$

$$\begin{aligned} (I - P)D &= lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{k=0}^{i-1}(I - P)(P^k - P^*) \\ &= lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{k=0}^{i-1}(P^k - P^{k+1}) \\ &= lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(I - P^k) = I - P^*. \end{aligned}$$

Similarly, it can be shown that $DP^* = 0$ and $D(I - P) = I - P^*$. □

The fundamental and the deviation matrix can be computed as follows. From (5.4) and (5.7) it follows that

$$I - P + P^* = \begin{pmatrix} C_1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & C_2 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & & & \cdot & \cdot \\ \cdot & & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & C_m & 0 \\ D_1 & D_2 & \cdot & \cdot & \cdot & \cdot & D_m & I - Q \end{pmatrix},$$

where $C_k = I - P_k + P_k^*$ and $D_k = -A_k + A_k^*$, $1 \leq k \leq m$. Hence,

$$Z = (I - P + P^*)^{-1} = \begin{pmatrix} C_1^{-1} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & C_2^{-1} & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & & & \cdot & \cdot \\ \cdot & & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 & C_m^{-1} & 0 \\ S_1 & S_2 & \cdot & \cdot & \cdot & \cdot & S_m & (I-Q)^{-1} \end{pmatrix},$$

where $S_k = -(I - Q)^{-1} D_k C_k^{-1}$, $1 \leq k \leq m$. The deviation matrix is $Z - P^*$.

Example 5.3 (continued)

$$I - P + P^* = \begin{pmatrix} \frac{13}{12} & -\frac{1}{12} & 0 & 0 & 0 \\ -\frac{7}{60} & \frac{67}{60} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{7}{30} & -\frac{7}{30} & 0 & 1 & 0 \\ \frac{7}{90} & \frac{5}{90} & \frac{7}{15} & -\frac{1}{5} & \frac{3}{15} \end{pmatrix} \rightarrow C_1 = \begin{pmatrix} \frac{13}{12} & -\frac{1}{12} \\ -\frac{7}{60} & \frac{67}{60} \end{pmatrix} \text{ and } C_2 = (1).$$

Hence, $C_1^{-1} = \begin{pmatrix} \frac{67}{72} & \frac{5}{72} \\ \frac{7}{72} & \frac{65}{72} \end{pmatrix}$ and $C_2^{-1} = (1)$. Since $I - Q = \begin{pmatrix} 1 & 0 \\ -\frac{1}{5} & \frac{3}{5} \end{pmatrix}$, we have $(I-Q)^{-1} = \begin{pmatrix} 1 & 0 \\ \frac{1}{3} & \frac{5}{3} \end{pmatrix}$.

Therefore, $S_1 = -(I-Q)^{-1} D_1 C_1^{-1} = -\begin{pmatrix} 1 & 0 \\ \frac{1}{3} & \frac{5}{3} \end{pmatrix} \begin{pmatrix} \frac{7}{30} & -\frac{7}{30} \\ \frac{7}{90} & \frac{5}{90} \end{pmatrix} \begin{pmatrix} \frac{67}{72} & \frac{5}{72} \\ \frac{7}{72} & \frac{65}{72} \end{pmatrix} = \begin{pmatrix} -\frac{7}{36} & \frac{7}{36} \\ -\frac{7}{36} & -\frac{1}{36} \end{pmatrix}$ and

$S_2 = -(I-Q)^{-1} D_2 C_2^{-1} = -\begin{pmatrix} 1 & 0 \\ \frac{1}{3} & \frac{5}{3} \end{pmatrix} \begin{pmatrix} 0 \\ \frac{7}{15} \end{pmatrix} (1) = \begin{pmatrix} 0 \\ -\frac{7}{9} \end{pmatrix}$.

It follows that $Z = \begin{pmatrix} \frac{67}{72} & \frac{5}{72} & 0 & 0 & 0 \\ \frac{7}{72} & \frac{65}{72} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -\frac{7}{36} & \frac{7}{36} & 0 & 1 & 0 \\ -\frac{7}{36} & -\frac{1}{36} & -\frac{7}{9} & \frac{1}{3} & \frac{5}{3} \end{pmatrix}$ and $D = \begin{pmatrix} \frac{25}{72} & -\frac{25}{72} & 0 & 0 & 0 \\ -\frac{35}{72} & \frac{35}{72} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{77}{180} & \frac{1}{36} & -\frac{3}{5} & 1 & 0 \\ -\frac{49}{180} & -\frac{1}{12} & -\frac{74}{45} & \frac{1}{3} & \frac{5}{3} \end{pmatrix}$.

In the theorems 5.5 and 5.6 the fundamental matrix Z and the deviation matrix D are expressed as Cesaro limits. These matrices can also be expressed in Abelian form as the next theorem shows.

Theorem 5.7

(1) $Z = lim_{\alpha \uparrow 1} \sum_{n=0}^{\infty} \alpha^n (P - P^*)^n$.

(2) $D = lim_{\alpha \uparrow 1} \sum_{n=0}^{\infty} \alpha^n (P^n - P^*)$.

Proof

(1) Similar as the proof that $(I-Q)^{-1} = \sum_{n=0}^{\infty} Q^n$, it can be shown that

$$H(\alpha) := \sum_{n=0}^{\infty} \{\alpha(P-P^*)\}^n = \{I-\alpha(P-P^*)\}^{-1}.$$

Hence, $I = H(\alpha)\{I-\alpha(P-P^*)\} = H(\alpha)(I-P+P^*) + (1-\alpha)H(\alpha)(P-P^*)$.

Since $P^n - P^*$ is Cesaro convergent to 0, $P^n - P^*$ is also Abel convergent to 0, i.e.

$lim_{a\uparrow 1}\,(1-\alpha)H(\alpha) = 0$. Therefore, $Z = (I-P+P^*)^{-1} = lim_{\alpha\uparrow 1} \sum_{n=0}^{\infty} \alpha^n (P-P^*)^n$.

(2) Because

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

we obtain

$$lim_{\alpha\uparrow 1} \sum_{n=0}^{\infty} \alpha^n (P^n - P^*) = lim_{\alpha\uparrow 1} \sum_{n=0}^{\infty} \alpha^n (P-P^*)^n - P^* = Z - P^* = D. \qquad \square$$

The following theorem gives the relation between average rewards, discounted rewards (over an infinite horizon) and total rewards over a finite horizon.

Theorem 5.8

Let f^∞ be a deterministic policy. Then,

(1) $\phi(f^\infty) = P^*(f)r(f)$.

(2) $\phi(f^\infty) = lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha(f^\infty)$.

(3) $v^T(f^\infty) = T\phi(f^\infty) + D(f)r(f) - P^T(f)D(f)r(f)$.

Proof

(1) $\phi(f^\infty) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P^t(f)r(f) = P^*(f)r(f)$.

(2) Since P^* is the Cesaro limit of P^t, it is also the Abel limit, i.e.

$\phi(f^\infty) = P^*(f)r(f) = lim_{\alpha\uparrow 1}\,(1-\alpha) \sum_{t=0}^{\infty} \{\alpha P(f)\}^t r(f) = v^\alpha(f^\infty)$.

(3) We apply induction on T.

$T=1$: $\phi(f^\infty) + D(f)r(f) - P(f)D(f)r(f) = \{P^*(f) + \{I-P(f)\}D(f)\}r(f) = r(f) = v^1(f)$,

using that $P^*(f) + \{I-P(f)\}D(f) = I$ (see Theorem 5.6, part (2)).

Suppose that the statement is true for T periods. Then, we can write

$(T+1)\phi(f^\infty) + D(f)r(f) - P^{T+1}(f)D(f)r(f) =$

$T\phi(f^\infty) + P^*(f)r(f) + D(f)r(r) - P^{T+1}(f)D(f)r(f) =$ (by the induction hypothesis)

$v^T(f^\infty) + P^T(f)D(f)r(f) + P^*(f)r(f) - P^{T+1}(f)D(f)r(f) =$

$v^T(f^\infty) + P^T(f)\{D(f) + P^*(f) - P(f)D(f)\}r(f) =$ (using Theorem 5.6, part (2))

$v^T(f^\infty) + P^T(f)r(f) = v^{T+1}(f^\infty)$. $\qquad \square$

The regular case

A Markov chain P is called a *regular* Markov chain if the chain is irreducible and aperiodic. In that case it can be shown [2] that $P^* = lim_{n\to\infty} P^n$. Since $(P - P^*)^n = P^n - P^*$ for $n = 1, 2, \ldots$, we have $(P - P^*)^n \to 0$ if $n \to \infty$. Therefore,

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

Because $D = Z - P^*$ and $Z = I + \sum_{n=1}^{\infty} (P - P^*)^n = I + \sum_{n=1}^{\infty} (P^n - P^*)$, we obtain

$$D = \sum_{n=0}^{\infty} (P^n - P^*),$$

i.e. D represents the total deviation with respect to the stationary matrix. This explains the name *deviation matrix*.

5.4 Extension of Blackwell's theorem

In Theorem 4.1 we proved the existence of a Blackwell optimal policy. The next theorem shows that the interval $[0, 1)$ can be partitioned in a finite number of subintervals such that in each subinterval there exists a deterministic policy which is optimal over the whole subinterval.

Theorem 5.9

There are numbers $\alpha_m, \alpha_{m-1}, \ldots, \alpha_0, \alpha_{-1}$ *and deterministic policies* $f_m^\infty, f_{m-1}^\infty, \ldots, f_0^\infty$ *such that*

(1) $0 = \alpha_m < \alpha_{m-1} < \cdots < \alpha_0 < \alpha_{-1} = 1$;

(2) $v^\alpha(f_j^\infty) = v^\alpha$ *for all* $\alpha \in [\alpha_j, \alpha_{j-1}),\ j = m, m-1, \ldots, 0.$

Proof

For any deterministic policy f^∞, $v^\alpha(f^\infty)$ is the unique solution of the linear system

$$\{I - \alpha P(f)\}x = r(f).$$

By Cramer's rule [3] $v_i^\alpha(f^\infty)$ is a rational functie in α for each component i.
Suppose that a deterministic Blackwell optimal policy does not exist. For any fixed α a deterministic α-discounted optimal policy exists. This implies a series $\{\alpha_k,\ k = 1, 2, \ldots\}$ and a series $\{f_k,\ k = 1, 2, \ldots\}$ such that

$$\alpha_1 \le \alpha_2 \le \cdots \text{ with } lim_{k\to\infty} \alpha_k = 1 \text{ and } v^\alpha = v^\alpha(f_k^\infty) > v^\alpha(f_{k-1}^\infty) \text{ for } \alpha = \alpha_k,\ k = 2, 3, \ldots.$$

Since there are only a finite number of deterministic policies, there must be a couple of policies, say f^∞ and g^∞, such that for some nondecreasing subsequence $\alpha_{k_n}, n = 1, 2, \ldots$ with $lim_{n\to\infty}\alpha_{k_n} = 1$,

$$\begin{cases} v^\alpha(f^\infty) > v^\alpha(g^\infty) & \text{for } \alpha = \alpha_{k_1}, \alpha_{k_3}, \ldots \\ v^\alpha(f^\infty) < v^\alpha(g^\infty) & \text{for } \alpha = \alpha_{k_2}, \alpha_{k_4}, \ldots \end{cases} \tag{5.10}$$

[2] For the proof see e.g. J. Kemeny and L. Snell: *Finite Markov chains*, Van Nostrand, 1960. p. 70.
[3] see e.g. J.B. Fraleigh and R.A. Beauregard: *Linear Algebra*, Addison Wesley, 1987, p. 214.

Let $h(\alpha) = v^\alpha(f^\infty) - v^\alpha(g^\infty)$, then $h_i(\alpha)$ is a continuous rational function in α on $[0,1)$ for each $i \in S$. From (5.10) it follows that $h_i(\alpha)$ has an infinite number of zeros, which is in contradiction with the rationality of $h_i(\alpha)$. Hence, there exists a deterministic Blackwell optimal policy, i.e. a policy f_0^∞ such that $v^\alpha(f_0^\infty) = v^\alpha$ for all $\alpha \in [\alpha_0, 1)$ for some $0 \leq \alpha_0 < 1$.
With similar arguments it can be shown that for each fixed $\alpha \in [0,1)$ there is a lower bound $L(\alpha) < \alpha$ and a deterministic policy $f_{L(\alpha)}^\infty$ such that $v^\alpha(f_{L(\alpha)}^\infty) = v^\alpha$ for all $\alpha \in \big(L(\alpha), \alpha\big)$. Similarly, for each fixed $\alpha \in (0,1]$ there is an upper bound $U(\alpha) > \alpha$ and a deterministic policy $f_{U(\alpha)}^\infty$ such that $v^\alpha(f_{U(\alpha)}^\infty) = v^\alpha$ for all $\alpha \in \big(\alpha, U(\alpha)\big)$.

The open intervals $\big(-1, U(0)\big)$, $\big\{\big(L(\alpha), U(\alpha)\big) \mid \alpha \in (0,1)\big\}$ and $\big(L(1), 2\big)$ are a covering of the compact set $[0,1]$. By the Heine-Borel-Lebesque covering theorem [4] the interval $[0,1]$ is covered by a finite number of intervals, say $\big(-1, U(0)\big)$, $\big\{\big(L(\alpha_j), U(\alpha_j)\big),\ j = m-1, m-2, \ldots, 1\big\}$ and $\big(L(1), 2\big)$. We may assume that

$$\alpha_m := 0 < \alpha_{m-1} < \cdots < \alpha_0 < \alpha_{-1} := 1,\ L(\alpha_{m-1}) < U(0),\ L(1) < U(\alpha_1)$$

and

$$L(\alpha_j) < L(\alpha_{j-1}) < U(\alpha_j) < U(\alpha_{j-1}),\ j = m-1, m-2, \ldots, 2.$$

Since the rational functions $v^\alpha(f_{L(\alpha_{j-1})}^\infty) = v^\alpha(f_{U(\alpha_j)}^\infty) = v^\alpha$ for all $\alpha \in \big(L(\alpha_{j-1}), U(\alpha_j)\big)$ we have

$$v^\alpha(f_{L(\alpha_{j-1})}^\infty) = v^\alpha(f_{U(\alpha_j)}^\infty),\ j = 0, 1, \ldots, m.$$

Let $f_j = f_{U(\alpha_j)},\ j = 0, 1, \ldots, m$. Then, $v^\alpha(f_j^\infty) = v^\alpha$ for all $\alpha \in (\alpha_j, \alpha_{j-1}),\ j = 0, 1, \ldots, m$. Since $v^\alpha(f^\infty)$ is continuous in α, also $v^\alpha(f_j^\infty) = v^\alpha$ for $\alpha = \alpha_j,\ j = 0, 1, \ldots, m$. □

5.5 The Laurent series expansion

Theorem 5.8 part (2) shows a relation between discounted and average reward when the discount factor tends to 1. This relation is based on the Laurent expansion of $v^\alpha(f^\infty)$ close to $\alpha = 1$ as expressed in the next theorem.

Theorem 5.10
Let $u^k(f),\ k = -1, 0, \ldots$ *be defined by* $u^{-1}(f) = P^*(f)r(f),\ u^0(f) = D(f)r(f)$ *and* $u^{k+1}(f) = -D(f)u^k(f),\ k \geq 0$. *Then,* $\alpha v^\alpha(f^\infty) = \sum_{k=-1}^\infty \rho^k u^k(f)$ *for* $\alpha_0(f) < \alpha < 1$, *where* $\rho = \frac{1-\alpha}{\alpha}$ *and* $\alpha_0(f) = \frac{\|D(f)\|}{1+\|D(f)\|}$.

Proof
Let $x(f) = \frac{1}{\alpha} \cdot \sum_{k=-1}^\infty \rho^k u^k(f) = \frac{\phi(f^\infty)}{1-\alpha} + \frac{1}{\alpha} \cdot \sum_{k=0}^\infty \rho^k u^k(f)$. Since $u^k(f) = D(f)\{-D(f)\}^k r(f)$ for $k \geq 0$, the series $\sum_{k=0}^\infty \rho^k u^k(f)$ is well defined if $\|\rho D(f)\| < 1$, i.e. $\alpha \geq \frac{\|D(f)\|}{1+\|D(f)\|}$.
Since $v^\alpha(f^\infty)$ is the unique solution of the linear system $\{I - \alpha P(f)\}x = r(f)$, it is sufficient to show that $\{I - \alpha P(f)\}x(f) = r(f)$, i.e. $y(f) := r(f) - \{I - \alpha P(f)\}x(f) = 0$.

[4] See e.g. A.C. Zaanen: *Integration*, North Holland, 1967.

$$\begin{aligned}
y(f) &= r(f) - \{I - \alpha P(f)\}\tfrac{P^*(f)r(f)}{1-\alpha} - \{I - \alpha P(f)\}\tfrac{D(f)}{\alpha}\textstyle\sum_{k=0}^{\infty}\{-\rho D(f)\}^k r(f) \\
&= r(f) - P^*(f)r(f) - \big\{\alpha\big(I - P(f)\big) + (1-\alpha)I\big\}\tfrac{D(f)}{\alpha}\textstyle\sum_{k=0}^{\infty}\{-\rho D(f)\}^k r(f) \\
&= \{I - P^*(f)\}r(f) - \{I - P(f)\}D(f)\textstyle\sum_{k=0}^{\infty}\{-\rho D(f)\}^k r(f) \\
&\qquad\qquad -\tfrac{1-\alpha}{\alpha}D(f)\textstyle\sum_{k=0}^{\infty}\{-\rho D(f)\}^k r(f) \\
&= \{I - P^*(f)\}r(f) - \{I - P^*(f)\}\textstyle\sum_{k=0}^{\infty}\{-\rho D(f)\}^k r(f) + \sum_{k=0}^{\infty}\{-\rho D(f)\}^{k+1} r(f) \\
&= \{I - P^*(f)\}r(f) - \textstyle\sum_{k=0}^{\infty}\{-\rho D(f)\}^k r(f) + P^*(f)r(f) + \sum_{k=1}^{\infty}\{-\rho D(f)\}^k r(f) \\
&= \{I - P^*(f)\}r(f) - r(f) - \textstyle\sum_{k=1}^{\infty}\{-\rho D(f)\}^k r(f) + P^*(f)r(f) + \sum_{k=1}^{\infty}\{-\rho D(f)\}^k r(f) \\
&= 0. \qquad \square
\end{aligned}$$

Corollary 5.2

$v^\alpha(f^\infty) = \frac{\phi(f^\infty)}{1-\alpha} + u^0(f) + \varepsilon(\alpha)$, *where* $\varepsilon(\alpha)$ *satisfies* $lim_{\alpha\uparrow 1}\ \varepsilon(\alpha) = 0$.

Proof

From Theorem 5.10 it follows that $v^\alpha(f^\infty) = \frac{\phi(f)}{1-\alpha} + \frac{u^0(f)}{\alpha} + \sum_{k=1}^{\infty}\ \frac{(1-\alpha)^k}{\alpha^{k+1}}u^k(f)$.

Since $\frac{1}{\alpha} = \frac{1}{1-(1-\alpha)} = 1 + (1-\alpha) + (1-\alpha)^2 + \cdots$, we may write

$v^\alpha(f^\infty) = \frac{\phi(f)}{1-\alpha} + u^0(f) + \varepsilon(\alpha)$, where $lim_{\alpha\uparrow 1}\,\varepsilon(\alpha) = 0$. $\square$

5.6 The optimality equation

In the discounted case, the value is the unique solution of an optimality equation. For the average reward criterion a similar result holds, but the equation is more complicated.

Theorem 5.11

Consider the system

$$\left\{\begin{array}{lcll} x_i & = & max_{a\in A(i)}\sum_j p_{ij}(a)x_j & ,\ i \in S \\ x_i + y_i & = & max_{a\in A(i,x)}\{r_i(a) + \sum_j p_{ij}(a)y_j\} & ,\ i \in S \end{array}\right. \tag{5.11}$$

where $A(i,x) = \{a \in A(i) \mid x_i = \sum_j p_{ij}(a)x_j\},\ i \in S$.

This system has the following properties:

(1) $x = u^{-1}(f_0),\ y = u^0(f_0)$, *where* f_0^∞ *is a Blackwell optimal policy, satisfies (5.11).*

(2) If (x,y) *is a solution of (5.11), then* $x = \phi$, *the value vector.*

Proof

Since f_0^∞ is a Blackwell optimal policy, for α sufficiently close to 1, say $\alpha \in [\alpha_0, 1)$, one can write

$$v_i^\alpha(f_0^\infty) = v_i^\alpha = max_{a\in A(i)}\{r_i(a) + \alpha\textstyle\sum_j p_{ij}(a)v_j^\alpha\} \geq r_i(a) + \alpha\sum_j p_{ij}(a)v_j^\alpha,\ (i,a) \in S \times A.$$

Combining this result with Corollary 5.2 gives for all $\alpha \in [\alpha_0, 1)$:

$$\begin{aligned}\tfrac{\phi_i(f_0^\infty)}{1-\alpha} + u_i^0(f_0) + \varepsilon_i(\alpha) &\geq r_i(a) + \{1-(1-\alpha)\}\textstyle\sum_j p_{ij}(a)\Big\{\tfrac{\phi_j(f_0^\infty)}{1-\alpha} + u_j^0(f_0) + \varepsilon_j(\alpha)\Big\} \\ &= r_i(a) + \textstyle\sum_j p_{ij}(a)\Big\{\tfrac{\phi_j(f_0^\infty)}{1-\alpha} + u_j^0(f_0) + \varepsilon_j(\alpha)\Big\} - \\ &\qquad (1-\alpha)\textstyle\sum_j p_{ij}(a)\Big\{\tfrac{\phi_j(f_0^\infty)}{1-\alpha} + u_j^0(f_0) + \varepsilon_j(\alpha)\Big\},\ (i,a) \in S \times A,\end{aligned}$$

i.e.

$$\tfrac{1}{1-\alpha}\Big\{\phi_i(f_0^\infty) - \textstyle\sum_j p_{ij}(a)\phi_j(f_0^\infty)\Big\} + \Big\{u_i^0(f_0) - r_i(a) - \textstyle\sum_j p_{ij}(a)u_j^0(f_0) - \sum_j p_{ij}(a)\phi_j(f_0^\infty)\Big\} + \varepsilon(\alpha) \geq 0.$$

Since this result holds for all $\alpha \in [\alpha_0, 1)$, the term multiplied by $\frac{1}{1-\alpha}$ has to be nonnegative, i.e.

$$\phi_i(f_0^\infty) \geq \sum_j p_{ij}(a)\phi_j(f_0^\infty) \text{ for all } i \in S \text{ and } a \in A(i). \tag{5.12}$$

Furthermore, when $\phi_i(f_0^\infty) = \sum_j p_{ij}(a)\phi_j(f_0^\infty)$, the next term has to be nonnegative, i.e.

$$u_i^0(f_0) \geq r_i(a) + \sum_j p_{ij}(a)u_j^0(f_0) - \sum_j p_{ij}(a)\phi_j(f_0^\infty) = r_i(a) + \sum_j p_{ij}(a)u_j^0(f_0) - \phi_i(f_0^\infty). \tag{5.13}$$

For $a = f_0(i),\ i \in S$, the inequalities in (5.12) and (5.13) are equalities, because:

$$\phi(f_0^\infty) = P^*(f_0)r(f_0) = P(f_0)P^*(f_0)r(f_0) = P(f_0)\phi(f_0^\infty)$$

and

$$u^0(f_0) = D(f_0)r(f_0) = \{I - P^*(f_0) + P(f_0)D(f_0)\}r(f_0) = r(f_0) - \phi(f_0^\infty) + P(f_0)u^0(f_0).$$

By these results, part (1) is shown. For part (2), let (x, y) be a solution of (5.11). Then, for any $f^\infty \in C(D)$, $x \geq P(f)x$, implying that $x \geq P^n(f)x$ for all $n \in \mathbb{N}$, and consequently, $x \geq P^*(f)x$. Furthermore, since $0 = P^*(f)\{x - P(f)\}$ and all elements of $P^*(f)$ and $x - P(f)$ are nonnegative, $p_{ij}^*(f)\{x - P(f)x\}_j = 0$ for all $i, j \in S$, implying that $p_{ii}^*(f)\{x - P(f)x\}_i = 0$ for all $i, \in S$.
For an ergodic state i, $p_{ii}^*(f) > 0$, and consequently $x_i - \sum_j p_{ij}(f)x_j = 0$, i.e. $f(i) \in A(i, x)$, and therefore, by (5.11) $x_i + y_i \geq r_i(f) + \sum_j p_{ij}(f)y_j$.
The columns of $P^*(f)$ corresponding to the transient states are zero, implying that $P^*(f)(x + y) \geq P^*(f)\{r(f) + P(f)y\} = \phi(f^\infty) + P^*(f)y$, i.e.

$$\phi(f^\infty) \leq P^*(f)x \leq x. \tag{5.14}$$

On the other hand, any solution of system (5.11) gives a policy g^∞ which satisfies $x = P(g)x$ and $x + y = r(g) + P(g)y$. Hence, $x = P^*(g)x$ and therefore,

$$\phi(g^\infty) = P^*(g)r(g) = P^*(g)\{x + y - P(g)y\} = x + P^*(g)\{y - P(g)y\} = x. \tag{5.15}$$

From (5.14) and (5.15) it follows that $x_i = max_{a \in A(i)} \sum_j p_{ij}(a)x_j = \phi_i,\ i \in S$. □

Remarks

1. Since the x-vector in (5.11) is unique, namely $x = \phi$, the set $A(i,x)$ is also unique for all $i \in S$.
2. If policy f^∞ satisfies $\phi = P(f)\phi$ and $\phi + y = r(f) + P(f)y$ for some vector y, then the policy is average optimal, namely $\phi = P^*(f)\phi = P^*(f)\{r(f) + P(f)y - y\} = \phi(f^\infty)$.
3. The proof suggests that a Blackwell optimal policy f_0^∞ is also average optimal, i.e. $\phi(f_0^\infty) \geq \phi(R)$ for every policy R. This result is shown below (Corollary 5.3).
4. If ϕ has identical components (e.g. if there is a unichain average optimal policy), then the first equation of (5.11) is superfluous and (5.11) can be replced by the single optimality equation

$$x + y_i = max_{a \in A(i)}\{r_i(a) + \sum_j p_{ij}(a)y_j\}, \ i \in S. \tag{5.16}$$

Theorem 5.12

$lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha(R) \geq \phi(R)$ for all policies R.

Proof

For $f^\infty \in C(D)$ we have shown in Theorem 5.8 part (2) that $lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha(f^\infty) = \phi(f^\infty)$.
For an arbitrary policy R the derivation is as follows.
Let $i \in S$ be any starting state and let $x_t = \sum_{(j,a)} \mathbb{P}_{i,R}\{X_t = j,\ Y_t = a\} \cdot r_j(a),\ t = 1,2,\ldots$.
Since the sequence $\{x_t \mid t = 1,2,\ldots\}$ is bounded, we may write

$$(1-\alpha)^{-1}v_i^\alpha(R) = \Big\{\sum_{t=1}^\infty \alpha^{t-1}\Big\} \cdot \{\sum_{t=1}^\infty \alpha^{t-1}x_t\} = \sum_{t=1}^\infty \Big\{\sum_{s=1}^t x_s\Big\} \cdot \alpha^{t-1}.$$

$(1-\alpha)^{-2} = \sum_{t=1}^\infty t\alpha^{t-1}$ for $\alpha \in (0,1)$, and therefore, $\phi_i(R) = \Big\{\sum_{t=1}^\infty t\alpha^{t-1}\Big\} \cdot (1-\alpha)^2 \cdot \phi_i(R)$.
Hence, $(1-\alpha)v_i^\alpha(R) - \phi_i(R) = (1-\alpha)^2 \cdot \sum_{t=1}^\infty \Big\{\frac{1}{t}\sum_{s=1}^t x_s - \phi_i(R)\Big\} \cdot t\alpha^{t-1}$.
Choose an arbitrary $\varepsilon > 0$. Since $\phi_i(R) = liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^T x_t$, there exists T_ε such that $\phi_i(R) < \frac{1}{T}\sum_{t=1}^T x_t + \varepsilon$ for all $T > T_\varepsilon$. This gives

$$(1-\alpha)^2 \sum_{t>T_\varepsilon} \Big\{\frac{1}{t}\sum_{s=1}^t x_s - \phi_i(R)\Big\}t\alpha^{t-1} > -\varepsilon(1-\alpha)^2 \sum_{t>T_\varepsilon} t\alpha^{t-1} \geq -\varepsilon(1-\alpha)^2 \sum_{t=1}^\infty t\alpha^{t-1} = -\varepsilon.$$

We also have,

$$(1-\alpha)^2 \sum_{t\leq T_\varepsilon} \Big\{\frac{1}{t}\sum_{s=1}^t x_s - \phi_i(R)\Big\}t\alpha^{t-1} \geq (1-\alpha)^2 min_{1\leq t\leq T_\varepsilon}\Big\{\frac{1}{t}\sum_{s=1}^t x_s - \phi_i(R)\Big\} \sum_{t\leq T_\varepsilon} t\alpha^{t-1} > -\varepsilon$$

for α sufficiently close to 1. Hence, $(1-\alpha)v_i^\alpha(R) - \phi_i(R) \geq -2\varepsilon$ for α sufficiently close to 1, i.e. $lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha(R) \geq \phi(R)$. □

Corollary 5.3

A Blackwell optimal policy f_0^∞ is also average optimal and consequently there exists a deterministic optimal policy.

Proof

Let f_0^∞ be a Blackwell optimal policy and R an arbitrary policy. Then,

$$\phi(f_0^\infty) = lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha(f_0^\infty) = lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha \geq lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha(R) \geq \phi(R).$$ □

5.7 Policy iteration

In policy iteration a sequence of policies $f_1^\infty, f_2^\infty, \ldots$ is constructed such that $\phi(f_{k+1}^\infty) \geq \phi(f_k^\infty)$ and $v^\alpha(f_{k+1}^\infty) > v^\alpha(f_k^\infty)$ for all $\alpha \in (\alpha_k, 1)$. Hence, each new policy in the sequence differs from the others. Since $C(D)$ is finite, the policy iteration method terminates after a finite number of iteration.

Theorem 5.13

Consider the following system

$$\left\{\begin{array}{rcl} \{I - P(f)\}x & = & 0 \\ x + \{I - P(f)\}y & = & r(f) \\ y + \{I - P(f)\}z & = & 0 \end{array}\right. \tag{5.17}$$

Then, for every $f^\infty \in C(D)$, the system (5.17) has a solution $\big(x(f), y(f), z(f)\big)$, where $x(f)$ and $y(f)$ are unique with $x(f) = u^{-1}(f) = \phi(f^\infty)$ and $y(f) = u^0(f)$.

Proof

First, we will show that $x(f) = u^{-1}(f)$, $y(f) = u^0(f)$ and $z(f) = u^1(f)$ is a solution of (5.17). We use the properties $P^*(f)D(f) = 0$ and $(I - P(f))D(f) = I - P^*(f)$ (see Theorem 5.6).

$$\begin{array}{rcl} \{I - P(f)\}x(f) & = & \{I - P(f)\}u^{-1}(f) = \{I - P(f)\}P^*(f)r(f) = 0 \\ x(f) + \{I - P(f)\}y(f) & = & P^*(f)r(f) + \{I - P(f)\}D(f)r(f) \\ & = & \{P^*(f) + \big(I - P(f)\big)D(f)\}r(f) = r(f) \\ y(f) + \{I - P(f)\}z(f) & = & D(f)r(f) - \{I - P(f)\}D^2(f)r(f) \\ & = & \{I - \big(I - P(f)\big)D(f)\}D(f)r(f) = P^*(f)D(f)r(f) = 0 \end{array}$$

Next, we show the second part of the theorem. Let (x, y, z) be any solution of (5.17). Then, $x = P(f)x$ implies $x = P^*(f)x = P^*(f)\{r(f) - \big(I - P(f)\big)y\} = P^*(f)r(f) = u^{-1}(f) = \phi(f^\infty)$. Since $y + \{I - P(f)\}z = 0$, we have $P^*(f)y = 0$, and consequently,

$$\{I - P(f) + P^*(f)\}y = \{I - P(f)\}y = r(f) - P^*(f)r(f),$$

i.e.

$$\begin{array}{rcl} y & = & \{I - P(f) + P^*(f)\}^{-1}\{I - P^*(f)\}r(f) \\ & = & Z(f)\{I - P^*(f)\}r(f) = \{D(f) + P^*(f)\}\{I - P^*(f)\}r(f) = D(f)r(f) = u^0(f). \end{array}$$ □

For every $i \in S$ and $f^\infty \in C(D)$, the action set $B(i, f)$ is defined by

$$B(i,f) = \left\{ a \in A(i) \;\middle|\; \begin{array}{l} \sum_j p_{ij}(a)\phi_j(f^\infty) > \phi_i(f^\infty) \text{ or} \\ \sum_j p_{ij}(a)\phi_j(f^\infty) = \phi_i(f^\infty) \text{ and } r_i(a) + \sum_j p_{ij}(a)u_j^0(f) > \phi_i(f^\infty) + u_i^0(f) \end{array} \right\}. \tag{5.18}$$

Theorem 5.14

(1) If $B(i,f)=\emptyset$ for every $i\in S$, then f^∞ is an average optimal policy.

(2) If $B(i,f)\neq\emptyset$ for at least one $i\in S$ and the policy $g^\infty\neq f^\infty$ satisfies $g(i)\in B(i,f)$ if $g(i)\neq f(i)$, then $\phi(g^\infty)\geq\phi(f^\infty)$ and $v^\alpha(g^\infty)>v^\alpha(f^\infty)$ for α sufficiently close to 1.

Proof

(1) Since $B(i,f)=\emptyset$ for every $i\in S$, for any $h^\infty\in C(D)$, we have $\sum_j p_{ij}(h)\phi_j(f^\infty)\leq\phi_i(f^\infty)$ and $r_i(h)+\sum_j p_{ij}(h)u_j^0(f)\leq\phi_i(f^\infty)+u_i^0(f)$ if $\sum_j p_{ij}(h)\phi_j(f^\infty)=\phi_i(f^\infty)$.

Let $R=(h,f,f,\dots)$. Then, $v^\alpha(R)=r(h)+\alpha P(h)v^\alpha(f^\infty)$ and, by Theorem 5.10,

$$\begin{aligned}\alpha v^\alpha(f^\infty) &= \tfrac{\alpha}{1-\alpha}\phi(f^\infty)+u^0(f)+\varepsilon_1(\alpha)=\{1-(1-\alpha)\}\tfrac{\phi(f^\infty)}{1-\alpha}+u^0(f)+\varepsilon_1(\alpha)\cdot e\\ &= \tfrac{\phi(f^\infty)}{1-\alpha}+u^0(f)-\phi(f^\infty)+\varepsilon_1(\alpha)\cdot e,\end{aligned}$$

(in this proof $\varepsilon_k(\alpha)$ satisfies $lim_{\alpha\uparrow 1}\varepsilon_k(\alpha)=0$) implying

$$\begin{aligned}v^\alpha(R) &= r(h)+P(h)\Big\{\tfrac{\phi(f^\infty)}{1-\alpha}+u^0(f)-\phi(f^\infty)+\varepsilon_1(\alpha)\cdot e\Big\}\\ &= \tfrac{P(h)\phi(f^\infty)}{1-\alpha}+r(h)+P(h)u^0(f)-P(h)\phi(f^\infty)+\varepsilon_1(\alpha)\cdot e.\end{aligned}$$

Since $v^\alpha(f^\infty)=\frac{\phi(f^\infty)}{1-\alpha}+u^0(f)+\varepsilon_2(\alpha)\cdot e$, we have

$$v^\alpha(f^\infty)-v^\alpha(R)=\frac{1}{1-\alpha}\{\phi(f^\infty)-P(h)\phi(f^\infty)\}+\{u^0(f)-r(h)-vP(h)u^0(f)+P(h)\phi(f^\infty)\}+\varepsilon_3(\alpha)\cdot e. \tag{5.19}$$

Since $\phi(f^\infty)-P(h)\phi(f^\infty)\geq 0$ and, if $\{\phi(f^\infty)-P(h)\phi(f^\infty)\}_i=0$,

$\{u^0(f)-r(h)-P(h)u^0(f)+P(h)\phi(f^\infty)\}_i=\{u^0(f)-r(h)-P(h)u^0(f)+\phi(f^\infty)\}_i\geq 0$,

we obtain $v^\alpha(f^\infty)-v^\alpha(R)\geq\varepsilon_3(\alpha)\cdot e$ for α sufficiently close to 1, i.e.

$$v^\alpha(f^\infty)\geq v^\alpha(R)+\varepsilon_3(\alpha)\cdot e=r(h)+\alpha P(h)v^\alpha(f^\infty)+\varepsilon_3(\alpha)\cdot e.$$

Hence, $\{I-\alpha P(h)\}v^\alpha(f^\infty)\geq r(h)+\alpha P(h)v^\alpha(f^\infty)+\varepsilon_3(\alpha)\cdot e$.

Therefore,

$$v^\alpha(f^\infty)\geq\{I-\alpha P(h)\}^{-1}\{r(h)+\varepsilon_3(\alpha)\cdot e\}=v^\alpha(h^\infty)+\frac{\varepsilon_3(\alpha)}{1-\alpha}\cdot e.$$

From the Laurent expansion follows $\phi(f^\infty)\geq\phi(h^\infty)$, i.e. f^∞ is an average optimal policy.

(2) Let $R=(g,f,f,\dots)$. Then,

if $g(i)=f(i)$, then the ith rows of $P(f)$ and $P(g)$ are identical and $r_i(f)=r_i(g)$, i.e.

$$v_i^\alpha(R)=\{r(g)+\alpha P(g)v^\alpha(f^\infty)\}_i=\{r(f)+\alpha P(f)v^\alpha(f^\infty)\}_i=v_i^\alpha(f^\infty).$$

if $g(i)\neq f(i)$, then $g(i)\in B(i,f)$, and because (5.19) holds for $h=g$, we have

$v_i^\alpha(f^\infty)-v_i^\alpha(R)=\frac{1}{1-\alpha}\{\phi(f^\infty)-P(h)\phi(f^\infty)\}_i+\{u^0(f)-r(g)-P(g)u^0(f)+P(g)\phi(f^\infty)\}_i$
$+\,\varepsilon_3(\alpha)\cdot e$ for α sufficiently close to 1.

Hence, for α sufficiently close to 1, $v^\alpha(R)=r(g)+\alpha P(g)v^\alpha(f^\infty)>v^\alpha(f^\infty)$, i.e.

$$\{I-\alpha P(g)\}v^\alpha(f^\infty)>r(g)\ \rightarrow\ v^\alpha(f^\infty)>\{I-\alpha P(g)\}^{-1}r(g)=v^\alpha(f^\infty).$$

Again, by the Laurent expansion, it follows that $\phi(g^\infty)\geq\phi(f^\infty)$. □

Algorithm 5.6 *Determination of an average optimal policy by policy iteration*

1. Take an arbitrary $f^\infty \in C(D)$.
2. Determine $\phi(f^\infty)$ and $u^0(f)$ as unique (x, y)-part in a solution of the system

$$\left\{\begin{array}{rcl} \{I - P(f)\}x & = & 0 \\ x + \{I - P(f)\}y & = & r(f) \\ y + \{I - P(f)\}z & = & 0 \end{array}\right.$$

3. Determine for every $i \in S$

$$B(i,f) = \left\{ a \in A(i) \;\middle|\; \begin{array}{l} \sum_j p_{ij}(a)\phi_j(f^\infty) > \phi_i(f^\infty) \text{ or} \\ \sum_j p_{ij}(a)\phi_j(f^\infty) = \phi_i(f^\infty) \text{ and } r_i(a) + \sum_j p_{ij}(a)u_j^0(f) > \phi_i(f^\infty) + u_i^0(f) \end{array} \right\}.$$

4. If $B(i,f) = \emptyset$ for every $i \in S$, then f^∞ is an average optimal policy (STOP).

   Otherwise: (a) take g such that $g \neq f$ and $g(i) \in B(i,f)$ if $g(i) \neq f(i)$;

   (b) $f := g$ and return to step 2.

Example 5.4

Consider the MDP of Example 3.1.

Start with the policy f^∞, where $f(1) = 3,\ f(2) = 2$ and $f(3) = 1$.

Iteration 1:

The solution of the linear system gives: $\phi(f^\infty) = \left(\frac{11}{2}, 4, \frac{11}{2}\right),\ u^0(f) = \left(-\frac{5}{4}, 0, \frac{5}{4}\right)$.

$B(1,f) = \emptyset,\ B(2,f) = \{1,3\};\ B(3,f) = \{3\}.\ g(1) = 3,\ g(2) = 3,\ g(3) = 3.$

$f(1) = 3,\ f(2) = 3,\ f(3) = 3.$

Iteration 2:

The solution of the linear system gives: $\phi(f^\infty) = (7,7,7),\ u^0(f) = (-4,-2,0)$.

$B(1,f) = \emptyset,\ B(2,f) = \emptyset,\ B(3,f) = \emptyset.$

f^∞ is an optimal policy.

5.8 Linear programming

To obtain the value vector and an average optimal policy by linear programming, we need a property for which the value vector is an extreme element. Such property, called superharmonicity, can be derived from the optimality equation. In the context of average reward, a vector $v \in \mathbb{R}^N$ is superharmonic if there exists a vector $u \in \mathbb{R}^N$ such that the pair (u, v) satisfies the following system of inequalities

$$\left\{\begin{array}{lcll} v_i & \geq & \sum_j p_{ij}(a)v_j & \text{for every } (i,a) \in S \times A \\ v_i + u_i & \geq & r_i(a) + \sum_j p_{ij}(a)u_j & \text{for every } (i,a) \in S \times A \end{array}\right. \tag{5.20}$$

Theorem 5.15

The value vector ϕ is the (componentswise) smallest superharmonic vector.

Proof

Let f_0^∞ be a Blackwell optimal policy. From Theorem 5.11 it follows that

$$\begin{cases} \phi_i & \geq & \sum_j p_{ij}(a)\phi_j & \text{for every } i \in S,\ a \in A(i) \\ \phi_i + u_i^0(f_0) & \geq & r_i(a) + \sum_j p_{ij}(a)u_j^0(f_0) & \text{for every } i \in S,\ a \in A(i,\phi) \end{cases} \tag{5.21}$$

where $A(i,\phi) = \{a \in A(i) \mid \phi_i = \sum_j p_{ij}(a)\phi_j\},\ i \in S$.

Let $A^*(i) = \{a \in A(i) \mid \phi_i + u_i^0(f_0) < r_i(a) + \sum_j p_{ij}(a)u_j^0(f_0)\},\ i \in S$.

Define

$$s_i(a) := \phi_i - \sum_j p_{ij}(a)\phi_j,\ t_i(a) = \phi_i + u_i^0(f_0) - r_i(a) - \sum_j p_{ij}(a)u_j^0(f_0),\ (i,a) \in S \times A,$$

$$M := \begin{cases} min\left\{\frac{s_i(a)}{t_i(a)} \;\middle|\; a \in A^*(i),\ i \in S\right\} & \text{if } \cup_{i\in S}\ A^*(i) \neq \emptyset \\ 0 & \text{if } \cup_{i\in S}\ A^*(i) = \emptyset \end{cases} \quad \text{and } u := u^0(f_0) - M \cdot \phi.$$

For $a \in A(i,\phi)$, we have

$$\phi_i = \sum_j p_{ij}(a)\phi_j$$

and

$$\phi_i + u_i = \phi_i + u_i^0(f_0) - M \cdot \phi_i \geq r_i(a) + \sum_j p_{ij}(a)\{u_j^0(f_0) - M \cdot \phi_j\} = r_i(a) + \sum_j p_{ij}(a)u_j.$$

For $a \in A^*(i)$, we have

$$\phi_i > \sum_j p_{ij}(a)\phi_j$$

and

$$\begin{aligned} \phi_i + u_i &= \phi_i + u_i^0(f_0) - M \cdot \{s_i(a) + \textstyle\sum_j p_{ij}(a)\phi_j\} \\ &= t_i(a) + r_i(a) + \textstyle\sum_j p_{ij}(a)u_j^0(f_0) - M \cdot s_i(a) \geq r_i(a) + \sum_j p_{ij}(a)u_j. \end{aligned}$$

For $a \notin \{a \in A(i,\phi) \cup A^*(i)\}$, we have

$$\phi_i > \sum_j p_{ij}(a)\phi_j$$

and

$$\begin{aligned} \phi_i + u_i &= \phi_i + u_i^0(f_0) - M \cdot \phi_i \geq t_i(a) + r_i(a) + \textstyle\sum_j p_{ij}(a)\{u_j^0(f_0) - M \cdot \phi_j\} \\ &= t_i(a) + r_i(a) + \textstyle\sum_j p_{ij}(a)u_j \geq r_i(a) + \sum_j p_{ij}(a)u_j. \end{aligned}$$

Hence, the value vector ϕ is superharmonic.

Suppose that y is also superharmonic with corresponding vector x, Then, $y \geq P(f_0)y$, implying that $y \geq P^*(f_0)y \geq P^*(f_0)\{r(f_0) + \big(P(f_0) - I\big)x\} = P^*(f_0)r(f_0) = \phi(f_0^\infty) = \phi$, i.e. ϕ is the smallest superharmonic vector. □

Corollary 5.4

From the proof of Theorem 5.15 it follows that there exists a solution of the modified optimality equation

$$\begin{cases} x_i & = & max_{a\in A(i)} \sum_j p_{ij}(a)x_j & ,\ i \in S \\ x_i + y_i & = & max_{a\in A(i)}\{r_i(a) + \sum_j p_{ij}(a)y_j\} & ,\ i \in S \end{cases} \tag{5.22}$$

with $x = \phi$ as unique x-vector in this solution.

Example 5.5

$S = \{1,2,3\};\ A(1) = \{1,2\},\ A(2) = \{1\},\ A(2) = \{1\};\ r_1(1) = 3,\ r_1(2) = 1,\ r_2(1) = 0,\ r_3(1) = 2.$
$p_{11}(1) = 1,\ p_{12}(1) = p_{13}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1,\ p_{13}(2) = 0;$
$p_{21}(1) = 0,\ p_{22}(1) = 1,\ p_{23}(1) = 0;\ p_{31}(1) = p_{32}(1) = 0,\ p_{33}(1) = 1.$
The modified optimality equation for this model is:

$x_1 = max\{x_1, x_2\};\ x_2 = max\{x_2, x_3\};\ x_3 = max\{x_3\}.$

$x_1 + y_1 = max\{3 + y_1, 1 + y_2\};\ x_2 + y_2 = max\{0 + y_2, 1 + y_3\};\ x_3 + y_3 = max\{2 + y_3\}.$

This equation has as solution $x = (3,2,2)$ and $y = (a, b-1, b)$ for any a and b with $3 + a \geq b$. The original optimality equation is considerably more complex, because the equations in the second part depend on the values of x_1, x_2 and x_3.

Corollary 5.5

The value vector ϕ is the unique v-part of an optimal solution (u,v) of the linear program

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{rcll} \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j & \geq & 0 & \text{for every } (i,a) \in S \times A \\ v_i + \sum_j \big(\delta_{ij} - p_{ij}(a)\big) u_j & \geq & r_i(a) & \text{for every } (i,a) \in S \times A \end{array} \right\}, \tag{5.23}$$

where $\beta_j > 0,\ j \in S$, is arbitrarily chosen.

The dual linear program of (5.23) is

$$max \left\{ \sum_{(i,a)} r_i(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_a x_j(a) \ + \ \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} y_i(a) & = & \beta_j,\ j \in S \\ x_i(a), y_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}. \tag{5.24}$$

Theorem 5.16

Let (x,y) be an extreme optimal solution of (5.24). Then, any $f^\infty \in C(D)$, where $x_i\big(f(i)\big) > 0$ if $\sum_a x_i(a) > 0$ and $y_i\big(f(i)\big) > 0$ if $\sum_a x_i(a) = 0$ is an average optimal policy.

Proof

First, notice that f^∞ is well defined, because for every $j \in S$,

$$\sum_a x_j(a) + \sum_a y_j(a) = \sum_{(i,a)} p_{ij}(a) y_i(a) + \beta_j > 0,\ j \in S.$$

Let $S_x = \big\{i \in S \mid \sum_a x_i(a) > 0\big\}$. Since $x_i\big(f(i)\big) > 0,\ i \in S_x$ and $y_i\big(f(i)\big) > 0,\ i \notin S_x$, it follows from the complementary slackness property of linear programming that

$$\phi_i + \sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\} u_j = r_i\big(f(i)\big),\ i \in S_x \tag{5.25}$$

and

$$\sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\} \phi_j = 0,\ i \notin S_x. \tag{5.26}$$

The primal program (5.23) implies $\sum_j\{\delta_{ij}-p_{ij}(a)\}\phi_j \geq 0,\ (i,a)\in S\times A$. Suppose that $\sum_j\{\delta_{kj}-p_{kj}(f(k))\}\phi_j > 0$ for some $k\in S_x$. Since $x_k\big(f(k)\big)>0$, this implies that $\sum_j\{\delta_{kj}-p_{kj}(f(k))\}\phi_j\cdot x_k\big(f(k)\big)>0$. Furthermore, $\sum_j\{\delta_{ij}-p_{ij}(a)\}\phi_j\cdot x_i(a)\geq 0,\ (i,a)\in S\times A$. Hence,

$$\sum_{(i,a)}\sum_j\{\delta_{ij}-p_{ij}(a)\}\phi_j\cdot x_i(a)>0.$$

On the other hand, this result is contradictory to the constraints of the dual program (5.24) from which follows that

$$\sum_{(i,a)}\sum_j\{\delta_{ij}-p_{ij}(a)\}\phi_j\cdot x_i(a)=\sum_j\Big\{\sum_{(i,a)}\big(\delta_{ij}-p_{ij}(a)\big)x_i(a)\Big\}\phi_j=0.$$

This contradiction implies that

$$\sum_j\{\delta_{ij}-p_{ij}\big(f(i)\big)\}\phi_j=0,\ i\in S_x. \tag{5.27}$$

From (5.26) and (5.27) it follows that

$$\sum_j\{\delta_{ij}-p_{ij}\big(f(i)\big)\}\phi_j=0,\ i\in S. \tag{5.28}$$

Next, we show that S_x is closed under $P(f)$, i.e. $p_{ij}\big(f(i)\big)=0,\ i\in S_x,\ j\notin S_x$. Suppose that $p_{kl}\big(f(k)\big)>0$ for some $k\in S_x,\ l\notin S_x$. From the constraints of dual program (5.24) it follows that

$$0=\sum_a x_l(a)=\sum_{(i,a)}p_{il}(a)x_i(a)\geq p_{kl}\big(f(k)\big)x_k\big(f(k)\big)>0, \tag{5.29}$$

implying a contradiction.

We now show that the states of $S\backslash S_x$ are transient in the Markov chain induced by $P(f)$. Suppose that $S\backslash S_x$ has an ergodic state. Since S_x is closed, the set $S\backslash S_x$ contains an ergodic class, say $J=\{j_1,j_2,\ldots,j_m\}$. Since (x,y) is an extreme solution and $y_j\big(f(j)\big)>0,\ j\in J$, the corresponding columns in (5.24) are linearly independent. Because these columns have zeros in the first N rows, the second parts of these vectors are also independent vectors. Since for $j\in J$ and $k\notin J$, we have $\delta_{jk}-p_{jk}\big(f(j)\big)=0-0=0$, the vectors $b^i,\ 1\leq i\leq m$, where b^i has components $\delta_{j_ik}-p_{j_ik}\big(f(j_i)\big),\ k\in J$, are also linear independent.

However, $\sum_{k=1}^m b_k^i=\sum_{k=1}^m\{\delta_{j_ij_k}-p_{j_ij_k}\big(f(j_i)\big)\}=1-1=0,\ i=1,2,\ldots,m$, which contradicts the independency of $b^1,b^2,\ldots,b^m$.

We finish the proof as follows. From (5.27) it follows that $\phi=P(f)\phi$, and consequently we have $\phi=P^*(f)\phi$. Since that states of $S\backslash S_x$ are transient in the Markov chain induced by $P(f)$, the columns of $P^*(f)$ corresponding to $S\backslash S_x$ are zero-vectors. Hence, by (5.25),

$$\phi(f^\infty)=P^*(f)r(f)=P^*(f)\Big\{\phi+\{I-P(f)\}u\Big\}=P^*(f)\phi=\phi,$$

i.e. f^∞ is an average optimal policy. □

Algorithm 5.7 *Determination of an average optimal policy by linear programming*

1. Take any vector β, where $\beta_j > 0,\ j \in S$.

2. Use the simplex method to compute optimal solutions (u, v) and (x, y) of the programs

$$min\left\{\sum_j \beta_j v_j \;\middle|\; \begin{array}{rcll} \sum_j\{\delta_{ij} - p_{ij}(a)\}v_j & \geq & 0 & \text{for every } (i,j) \in S \times A \\ v_i + \sum_j \big(\delta_{ij} - p_{ij}(a)\big)u_j & \geq & r_i(a) & \text{for every } (i,j) \in S \times A \end{array}\right\}$$

and

$$max\left\{\sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rl} \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}x_i(a) & = 0,\ j \in S \\ \sum_a x_j(a) \ + \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}y_i(a) & = \beta_j,\ j \in S \\ x_i(a), y_i(a) & \geq 0,\ (i,a) \in S \times A \end{array}\right\}.$$

3. Take $f^\infty \in C(D)$ such that $x_i\big(f(i)\big) > 0$ if $\sum_a x_i(a) > 0$ and $y_i\big(f(i)\big) > 0$ if $\sum_a x_i(a) = 0$.

   Then, f^∞ is an average optimal policy and ϕ is the value vector.

The next example shows an optimal solution (x, y) of the dual program (5.24) which has in some state i more than one positive $x_i(a)$ or $y_i(a)$ variable.

Example 5.6

Consider the MDP of Example 3.1. The dual linear program is:

$$max\{x_1(1) + 2x_1(2) + 3x_1(3) + 6x_2(1) + 4x_2(2) + 5x_2(3) + 8x_3(1) + 9x_3(2) + 7x_3(3)\}$$

subject to the constraints

$$\begin{aligned}
&x_1(2) + x_1(3) - x_2(1) - x_3(1) = 0 \\
&x_2(1) + x_2(3) - x_1(2) - x_3(2) = 0 \\
&x_3(1) + x_3(2) - x_1(3) - x_2(3) = 0 \\
&x_1(1) + x_1(2) + x_1(3) + y_1(2) + y_1(3) - y_2(1) - y_3(1) = \tfrac{1}{3} \\
&x_2(1) + x_2(2) + x_2(3) + y_2(1) + y_2(3) - y_1(2) - y_3(2) = \tfrac{1}{3} \\
&x_3(1) + x_3(2) + x_3(3) + y_3(1) + y_3(2) - y_1(3) - y_2(3) = \tfrac{1}{3} \\
&x_1(1), x_1(2), x_1(3), x_2(1), x_2(2), x_2(3), x_3(1), x_3(2), x_3(3) \geq 0
\end{aligned}$$

If we solve this linear program with an LP-solver, we obtain:
$x_1(1) = x_1(2) = x_1(3) = x_2(1) = x_2(2) = 0,\ x_2(3) = \frac{1}{2},\ x_3(1) = 0,\ x_3(2) = \frac{1}{2},\ x_3(3) = 0.$
$y_1(1) = 0,\ y_1(2) = \frac{1}{6},\ y_1(3) = \frac{1}{6},\ y_2(1) = y_2(2) = y_2(3) = y_3(1) = y_3(2) = y_3(3) = 0.$
Hence, the optimal policy is: $f(1) = 2$ (or $f(1) = 3$), $f(2),\ 3,\ f(3) = 2$.
The value vector $\phi = (7, 7, 7,)$.

In the average reward case there is in general no one-to-one correspondence between the feasible solutions of the dual program (5.24) and the set of stationary policies. The natural formula for mapping feasible solutions (x, y) to the set of stationary policies is:

$$\pi_{ia}^{x,y} = \begin{cases} \frac{x_i(a)}{\sum_a x_i(a)},\ a \in A(i),\ i \in S_x; \\ \frac{y_i(a)}{\sum_a y_i(a)},\ a \in A(i),\ i \in S \backslash S_x. \end{cases}$$

In the next example [5] two different solutions are mapped on the same deterministic policy.

Example 5.7

Consider the MDP with $S = \{1,2,3,4\};\ A(1) = \{1\},\ A(2) = A(3) = \{1,2\},\ A(4)\{1\};$
$r_1(1) = r_2(1) = r_2(2) = r_3(1) = r_3(2) = r_4(1) = 1;$
$p_{11}(1) = 0,\ p_{12}(1) = 1,\ p_{13}(1) = p_{14}(1) = 0;\ p_{21}(1) = p_{22}(1) = 0,\ p_{23}(1) = 1,\ p_{24}(1) = 0;$
$p_{21}(2) = p_{22}(2) = p_{23}(2) = 0,\ p_{24}(2) = 1;\ p_{31}(1) = p_{32}(1) = 0,\ p_{33}(1) = 1,\ p_{34}(1) = 0;$
$p_{31}(2) = 1,\ p_{32}(2) = p_{33}(2) = p_{34}(2) = 0;\ p_{41}(1) = p_{42}(1) = p_{43}(1) = 0,\ p_{44}(1) = 1.$
The dual linear program becomes (take $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \frac{1}{4}$).

$max\{x_1(1) + x_2(1) + x_2(2) + x_3(1) + x_3(2) + x_4(1)\}$

subject to the constraints

$$\begin{array}{rrrrrrrrrrrrrcl}
x_1(1) & & & & -x_3(2) & & & & & & & & = & 0 \\
-x_1(1) & +x_2(1) & +x_2(2) & & & & & & & & & & = & 0 \\
 & -x_2(1) & & & +x_3(2) & & & & & & & & = & 0 \\
 & & -x_2(2) & & & & & & & & & & = & 0 \\
x_1(1) & & & & & & +y_1(1) & & & & -y_3(2) & = & \frac{1}{4} \\
 & x_2(1) & +x_2(2) & & & & -y_1(1) & +y_2(1) & +y_2(2) & & & = & \frac{1}{4} \\
 & & & x_3(1) & +x_3(2) & & & -y_2(1) & & +y_3(2) & = & \frac{1}{4} \\
 & & & & & x_4(1) & & & -y_2(2) & & = & \frac{1}{4}
\end{array}$$

$x_1(1), x_2(1), x_2(2), x_3(1), x_3(2), x_4(1), y_1(1), y_2(1), y_2(2), y_3(2) \geq 0.$

The following two feasible solutions (x^1, y^1) and (x^2, y^2) are mapped on the same deterministic policy f^∞, where $f(1) = f(2) = 1,\ f(3) = 2$ and $f(4) = 1$:
$x_1^1(1) = x_2^1(1) = \frac{1}{4},\ x_2^1(2) = x_3^1(1) = 0,\ x_3^1(2) = x_4^1(1) = \frac{1}{4};\ y_1^1(1) = y_2^1(1) = y_2^1(2) = y_3^1(2) = 0$
and $x_1^2(1) = x_2^2(1) = \frac{1}{6},\ x_2^2(2) = x_3^2(1) = 0,\ x_3^2(2) = \frac{1}{6},\ x_4^2(1) = \frac{1}{2};\ y_1^2(1) = \frac{1}{6},\ y_2^2(1) = y_2^2(2) = \frac{1}{4},$
$y_3^2(2) = \frac{1}{12}.$

For any $\pi^\infty \in C(S)$ we can define a feasibe solution (x^π, y^π) of the dual program as follows. Consider the Markov chain induced by $P(\pi)$ and suppose that this Markov chain has m recurrent sets, say $S_1, S_2, \ldots, S_m$, and let T be the set of transient states. Define (x^π, y^π) by

$$x_i^\pi(a) = \{\beta^T P^*(\pi)\}_i \cdot \pi_{ia},\ (i,a) \in S \times A \tag{5.30}$$

$$y_i^\pi(a) = \{\beta^T D(\pi) + \gamma^T P^*(\pi)\}_i \cdot \pi_{ia},\ (i,a) \in S \times A, \tag{5.31}$$

where $\gamma_i = \begin{cases} 0 & i \in T; \\ max_{l \in S_j} \left\{ -\frac{\sum_{k \in S} \beta_k d_{kl}(\pi)}{\sum_{k \in S_j} p_{kl}^*(\pi)} \right\} & i \in S_j,\ 1 \leq j \leq m. \end{cases}$

Notice that γ is constant on every ergodic set.

[5] More examples to illustrate irregularities in the linear programming method for multichain MDPs can be found in: L.C.M. Kallenberg: *Remarks on old results*, Technical Report, University of Leiden, 2005.

Theorem 5.17

(x^π, y^π), defined by (5.30) and (5.31), is a feasible solution of the dual program (5.24).

Proof

$\sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}x_i^\pi(a) = \sum_j x_j^\pi(a) - \sum_{(i,a)} p_{ij}(a)x_i^\pi(a) = \{\beta^T P^*(\pi)\}_j - \{\beta^T P^*(\pi)P(\pi)\}_j = 0.$

$\sum_j x_j^\pi(a) + \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}y_i^\pi(a)$

$$\begin{aligned} &= \{\beta^T P^*(\pi)\}_j + \{\beta^T D(\pi) + \gamma^T P^*(\pi)\}_j - \{\beta^T D(\pi)P(\pi) + \gamma^T P^*(\pi)P(\pi)\}_j \\ &= \left\{\beta^T\{P^*(\pi) + D(\pi)(I - P(\pi)\}\right\}_j + \{\gamma^T P^*(\pi)(I - P(\pi))\}_j \\ &= \left\{\beta^T\{P^*(\pi) + I - P^*(\pi)\}\right\}_j = \beta_j. \end{aligned}$$

The nonnegativity of $x_i^\pi(a)$ is obvious. For the nonnegativity of $y_i^\pi(a)$ we distinguish between $i \in T$ and $i \in S_j$ for some $1 \le j \le m$. Notice that $y_i^\pi(a) = \{\sum_k \beta_k d_{ki}(\pi) + \sum_k \gamma_k p_{ki}^*(\pi)\} \cdot \pi_{ia}$.

If $i \in T$:

$p_{ki}^*(\pi) = 0$ for all k and, by Theorem 5.7, $d_{ki}(\pi) = \sum_{t=0}^\infty \{P^t(\pi)\}_{ki}$.

Therefore, $y_i^\pi(a) = \{\sum_k \beta_k \cdot (\sum_{t=0}^\infty \{P^t(\pi)\}_{ki}) \cdot \pi_{ia} \ge 0$.

If $i \in S_j$:

$p_{ki}^*(\pi) = 0$ for all $k \notin (S_j \cup T)$. Hence,

$$\begin{aligned} y_i^\pi(a) &= \{\textstyle\sum_{k\in S} \beta_k d_{ki}(\pi) + \sum_{k\in S_j} \gamma_k p_{ki}^*(\pi)\} \cdot \pi_{ia} \\ &\ge \{\textstyle\sum_{k\in S} \beta_k d_{ki}(\pi) + \sum_{k\in S_j} \left\{ - \frac{\sum_{k\in S} \beta_k d_{ki}(\pi)}{\sum_{k\in S_j} p_{ki}^*(\pi)}\right\} p_{ki}^*(\pi)\Big\} \cdot \pi_{ia} \\ &= \textstyle\sum_{k\in S} \beta_k d_{ki}(\pi) - \sum_{k\in S} \beta_k d_{ki}(\pi) = 0. \end{aligned}$$

□

Theorem 5.18

The correspondence between the stationary policies and the feasible solutions of program (5.24) preserves the optimality property, i.e.

(1) If π^∞ is an average optimal policy, then (x^π, y^π) is an optimal solution of (5.24).

(2) If (x, y) is an optimal solution of (5.24), then the stationary policy $\pi^{x,y}$ is an average optimal policy.

Proof

(1) Since (x^π, y^π) is feasible for (5.24) it is sufficient to show that $\sum_{(i,a)} r_i(a)x_i^\pi(a) = \sum \beta_j\phi_j$.

$\sum_{(i,a)} r_i(a)x_i^\pi(a) = \sum_{(i,a)} r_i(a)\{\beta^T P^*(\pi)\}_i \cdot \pi_{ia} = \{\beta^T P^*(\pi)\}_i r_i(\pi) = \beta^T\phi(\pi^\infty) = \beta^T\phi.$

(2) The proof of this part has the same structure as the proof of Theorem 5.16.

Suppose that $(v = \phi, u)$ is an optimal solution of the primal program (5.23).

Let $S_x = \{i \in S \mid \sum_a x_i(a) > 0\}$ and $A^+(i) = \{a \in A(i) \mid \pi_{ia}^{x,y} > 0\}$, $i \in S$.

Since $x_i(a) > 0$, $i \in S_x$, $a \in A^+(i)$ and $y_i(a) > 0$, $i \notin S_x$, $a \in A^+(i)$, it follows from the complementary slackness property of linear programming that

$$\phi_i + \sum_j \{\delta_{ij} - p_{ij}(a)\}u_j = r_i(a), \ i \in S_x, \ a \in A^+(i) \tag{5.32}$$

and

$$\sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j = 0,\ i \notin S_x,\ a \in A^+(i). \tag{5.33}$$

The primal program (5.23) implies $\sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j \geq 0,\ (i,a) \in S \times A$.

Suppose that $\sum_j \{\delta_{kj} - p_{kj}(a_k)\}\phi_j > 0$ for some $k \in S_x$ and some $a_k \in A^+(k)$.

Since $\pi^{x,y}_{ka_k} > 0$, we also have $x_{ka_k} > 0$, and $\sum_j \{\delta_{kj} - p_{kj}(a_k)\}\phi_j \cdot x_k(a_k) > 0$.

Furthermore, $\sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j \cdot x_i(a) \geq 0,\ (i,a) \in S \times A$. Hence,

$$\sum_{(i,a)} \sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j \cdot x_i(a) > 0.$$

On the other hand, this result is contradictory to the constraints of the dual program (5.24) from which follows that

$$\sum_{(i,a)} \sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j \cdot x_i(a) = \sum_j \Big\{ \sum_{(i,a)} \big(\delta_{ij} - p_{ij}(a)\big) x_i(a) \Big\} \phi_j = 0.$$

This contradiction implies that

$$\sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j = 0,\ i \in S_x,\ a \in A^+(i). \tag{5.34}$$

From (5.33) and (5.34) it follows that

$$\sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j = 0,\ i \in S,\ a \in A^+(i). \tag{5.35}$$

Next, we show that S_x is closed under $P(\pi^{x,y})$, i.e. $p_{ij}(\pi^{x,y}) = 0,\ i \in S_x,\ j \notin S_x$.

Suppose that $p_{kl}(\pi^{x,y}) > 0$ for some $k \in S_x,\ l \notin S_x$. Since $p_{kl}(\pi^{x,y}) = \sum_a p_{kl}(a)\pi^{x,y}_{ka}$, there exists an action a_k such that $p_{kl}(a_k) > 0$ and $\pi^{x,y}_{ka_k} > 0$. From the constraints of dual program (5.24) it follows that

$$0 = \sum_a x_l(a) = \sum_{(i,a)} p_{il}(a)x_i(a) \geq p_{kl}(a_k)x_k(a_k) > 0, \tag{5.36}$$

implying a contradiction.

Then, we show that the states of S_x are the recurrent states of the Markov chain induced by $P(\pi^{x,y})$. Let $x_i = \sum_a x_i(a),\ i \in S$. Since $x_i(a) = \pi^{x,y}_{ia} \cdot x_i$ for all (i,a), the constraints of (5.24) imply $x^T = x^T P(\pi^{x,y})$, and consequently, $x^T = x^T P^*(\pi^{x,y})$. Because, for $i \in T$, $x_i = \sum_j x_j p^*_{ji}(\pi^{x,y}) = 0$, we have $T \subseteq S \backslash S_x$. Suppose that $T \neq S \backslash S_x$. Since S_x is closed under $P(\pi^{x,y})$, there exists an ergodic set $S_1 \subseteq S \backslash S_x$. Hence, $0 = \sum_{j \notin S_1} \sum_{i \in S_1} p_{ij}(\pi^{x,y})$, implying $0 = \sum_{j \notin S_1} \sum_{i \in S_1} \sum_a p_{ij}(a) y_i(a)$. We also have, denoting $\sum_a y_i(a)$ by $y_i,\ i \in S$,

$$
\begin{aligned}
0 \ &< \ \textstyle\sum_{j \in S_1} \beta_j = \sum_{j \in S_1} y_j - \sum_{j \in S_1} \sum_{(i,a)} p_{ij}(a) y_i(a) \\
&= \ \textstyle\sum_{j \in S_1} y_j - \sum_{j \in S} \sum_{(i,a)} p_{ij}(a) y_i(a) + \sum_{j \notin S_1} \sum_{(i,a)} p_{ij}(a) y_i(a) \\
&= \ \textstyle\sum_{j \in S_1} y_j - \sum_{j \in S} \sum_{i \in S_1} \sum_a p_{ij}(a) y_i(a) - \sum_{j \in S} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) \\
&\qquad\qquad + \textstyle\sum_{j \notin S_1} \sum_{i \in S_1} \sum_a p_{ij}(a) y_i(a) + \sum_{j \notin S_1} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) \\
&= \ \textstyle\sum_{j \in S_1} y_j - \sum_{j \in S} \sum_{i \in S_1} \sum_a p_{ij}(a) y_i(a) - \sum_{j \in S} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) \\
&\qquad\qquad + \textstyle\sum_{j \notin S_1} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) \\
&= \ \textstyle\sum_{j \in S_1} y_j - \sum_{i \in S_1} y_i - \sum_{j \in S} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) + \sum_{j \notin S_1} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) \\
&= \ \textstyle- \sum_{j \in S} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) + \sum_{j \notin S_1} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) \\
&= \ \textstyle- \sum_{j \in S_1} \sum_{i \notin S_1} \sum_a p_{ij}(a) y_i(a) \leq 0,
\end{aligned}
$$

implying a contraction. So, S_x is the set of the recurrent states in the Markov chain $P(\pi^{x,y})$. We finish the proof as follows. From (5.35) it follows that

$$
\begin{aligned}
\phi_i \ &= \ \textstyle\sum_j p_{ij}(a) \phi_j,\ i \in S,\ a \in A^+(i) \\
\phi_i \ &= \ \textstyle\sum_j \sum_a p_{ij}(a) \pi^{x,y}_{ia} \phi_j,\ i \in S \\
\phi_i \ &= \ \textstyle\sum_j p_{ij}(\pi^{x,y}) \phi_j,\ i \in S,
\end{aligned}
$$

or, in vector notation, $\phi = P(\pi^{x,y})\phi$, implying $\phi = P^*(\pi^{x,y})\phi$. Since $S \backslash S_x$ is the set of transient states, we have $p^*_{ij}(\pi^{x,y}) = 0,\ j \in S \backslash S_x$. Therefore, we can write using (5.32),

$$
\begin{aligned}
\phi(\pi^{x,y}) \ &= \ P^*(\pi^{x,y}) r(\pi^{x,y}) \\
\phi(\pi^{x,y}) \ &= \ P^*(\pi^{x,y}) \{\phi + \big(I - P(\pi^{x,y})\big) u\} = P^*(\pi^{x,y}) \phi = \phi,
\end{aligned}
$$

implying that policy $\pi^{x,y}$ is an average optimal policy. □

5.9 Value iteration

For the method of value iteration the following scheme is used:

$$
\begin{cases}
v_i^{n+1} = max_a \{r_i(a) + \sum_j p_{ij}(a) v_j^n\},\ i \in S,\ n = 0, 1, \ldots \\
v_i^0 \text{ arbitrarily chosen},\ i \in S
\end{cases}
\tag{5.37}
$$

Let $f^\infty_{n+1} \in C(D)$ be such that

$$
v^{n+1} = r(f_{n+1}) + P(f_{n+1}) v^n,\ n = 0, 1, \ldots . \tag{5.38}
$$

However, in general, neither the sequence $\{v^n\}_{n=0}^{\infty}$ nor the sequence $\{v^{n+1}-v^n\}_{n=0}^{\infty}$ is convergent as the next example shows.

Example 5.8
Let $S=\{1,2\},\ A(1)=A(2)=\{,\ p_{11}(1)=0,\ p_{12}(1)=1,\ p_{21}(1)=1,\ p_{22}(1)=0,$
$r_1(1)=2,\ r_2(1)=0$, and let $v^0=(0,0)$.
Then, $v^{2n}=(2n,2n)$ and $v^{2n+1}=(2n+2,2n),\ n=0,1,\dots$.
Hence, no convergence for the sequence $\{v^n\}_{n=0}^{\infty}$ nor for $\{v^{n+1}-v^n\}_{n=0}^{\infty}$.

Remark
We will show that

$$\phi = lim_{n\to\infty}\frac{1}{n}v^n.$$

However, this is numerically an instable computation scheme if v^n tends to infinity. Fortunately, we may also use the property

$$\phi = lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\{v^{k+1}-v^k\}.$$

These properties can be shown by using the sequence $\{e^n\}_{n=0}^{\infty}$, where

$$e^n = v^n - n\cdot\phi - u,$$

with u defined as in Theorem 5.15, i.e.

$$u = u^0(f_0) - M\cdot\phi \text{ for some } M$$

(see the proof of Theorem 5.15) and f_0^{∞} a Blackwell optimal policy. In case the Markov chains $P(f)$ are aperiodic for all $f^{\infty}\in C(D)$, which we may assume without loss of the generality (see Lemma 5.7), we can show that

$$\phi = lim_{n\to\infty}\{v^{n+1}-v^n\}.$$

Let f_0^{∞} be a Blackwell optimal policy. From the proof of Theorem 5.15 it follows that (ϕ,u) with $u=u^0(f_0)-M\cdot\phi$, satisfies for any $f^{\infty}\in C(D)$

$$\begin{cases} \phi & \geq & P(f)\phi \\ \phi+u & \geq & r(f)+P(f)u \end{cases} \tag{5.39}$$

Furthermore, we define $F_0=\{f \mid \phi=P(f)\phi;\ \phi+u=r(f)+P(f)u\}$. Notice that $f_0\in F_0$ and that $f\in F_0$ implies that f^{∞} is an average optimal policy, since

$$\phi(f^{\infty}) = P^*(f)r(f) = P^*(f)\{\phi+u-P(f)u\} = P^*(f)\phi = \phi.$$

Lemma 5.5

(1) *If $f\in F_0$, then $P(f)e^n\leq e^{n+1}\leq P(f_{n+1})e^n,\ n=0,1,\dots.$*

(2) *$\{e^n\}_{n=0}^{\infty}$ is bounded.*

(3) *$\phi = lim_{n\to\infty}\frac{1}{n}v^n.$*

(4) *$\phi = lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\{v^k-v^{k-1}\}.$*

Proof

(1) Let $n \in \mathbb{N}_0$ and $f \in F_0$ be arbitrarily chosen. Then,

$$\begin{aligned} P(f)e^n &= P(f)\{v^n - n\cdot\phi - u\} = P(f)v^n - n\cdot P(f)\phi - P(f)u \\ &= \{P(f)v^n + r(f)\} - n\cdot P(f)\phi - \{P(f)u + r(f)\} \\ &\leq v^{n+1} - (n+1)\cdot\phi - u = e^{n+1}. \\ P(f_{n+1})e^n &= P(f_{n+1})v^n - n\cdot P(f_{n+1})\phi - P(f_{n+1})u \\ &\geq P(f_{n+1})v^n - n\cdot\phi - \{u + \phi - r(f_{n+1})\} = v^{n+1} - (n+1)\cdot\phi - u = e^{n+1}. \end{aligned}$$

(2) From part (1), we obtain

$$\begin{aligned} P^n(f_0)(v^0 - u) &= P^n(f_0)e^0 \leq P^{n-1}(f_0)e^1 \leq \cdots \leq P^0(f_0)e^n = e^n \leq P(f_n)e^{n-1} \\ &\leq P(f_n)P(f_{n-1})e^{n-2} \leq \cdots \leq P(f_n)P(f_{n-1})\cdots P(f_1)e^0 \\ &= P(f_n)P(f_{n-1})\cdots P(f_1)(v^0 - u), \end{aligned}$$

implying that $min_i\,(v_i^0 - u_i)\cdot e \leq e^n \leq max_i\,(v_i^0 - u_i)$.

(3) Since $\phi = \frac{1}{n}\{v^n - e^n - u\}$ and $\{e^n\}_{n=0}^{\infty}$ is bounded, we have $\phi = lim_{n\to\infty}\frac{1}{n}v^n$.

(4) From $\frac{1}{n}\sum_{k=1}^{n}\{v^k - v^{k-1}\} = \frac{1}{n}(v^n - v^0)$ and part (3), we obtain
$\phi = lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\{v^k - v^{k-1}\}$. □

Lemma 5.6

Let, for all $i \in S$, $A_n(i) = \left\{a \in A(i) \;\middle|\; max_b\{r_i(b) + \sum_j p_{ij}(b)v_j^{n-1}\} = r_i(a) + \sum_j p_{ij}(a)v_j^{n-1}\right\}$ *and* $A_*(i) = \{a \in A(i) \mid \phi_i = \sum_j p_{ij}(a)\phi_j\}$. *Then, for* n *sufficiently large,* $A_n(i) \subseteq A_*(i),\ i \in S$.

Proof

Suppose the contrary. Then, there exists a pair $(i,a) \in S \times A$ and a sequence $\{n_k\},\ k = 1,2,\ldots$ such that $a \in A_{n_k}(i),\ k = 1,2,\ldots$ and $a \notin A_*(i)$. Since $\frac{1}{n_k}v_i^{n_k} = \frac{1}{n_k}\{r_i(a) + \sum_j p_{ij}(a)v_i^{n_k-1}\}$, and by part (3) of Lemma 5.5, we obtain $\phi_i = \sum_j p_{ij}(a)\phi_j$, i.e. $a \in A_*(i)$: contradiction. □

Next, we show that we may assume that for every $f^\infty \in C(D)$ the Markov chain $P(f)$ is aperiodic. In that case we have $P^*(f) = lim_{n\to\infty}P^n(f)$.[6]
Consider for an arbitrary $\lambda \in (0,1)$ the transition probabilities

$$p_{ij}(a)(\lambda) = \lambda\delta_{ij} + (1-\lambda)p_{ij}(a),\ (i,a) \in S \times A,\ j \in S. \tag{5.40}$$

Since $p_{ii}(a)(\lambda) \geq \lambda > 0,\ i \in S$, the transition matrix is aperiodic. Let $\phi_\lambda(f^\infty)$ be the average reward of policy f^∞ with respect to the transitions $p_{ij}(a)(\lambda)$. The following lemma shows that $\phi_\lambda(f^\infty) = \phi_(f^\infty),\ f^\infty \in C(D)$.

[6]See e.g. H.M. Taylor and S. Karlin: *An introduction to stochastic modeling*, 3rd edition, 1998, chapter 4.

Lemma 5.7

$\phi_\lambda(f^\infty) = \phi(f^\infty),\ f^\infty \in C(D)$.

Proof

$P_\lambda(f)\phi(f^\infty) = \{\lambda I + (1-\lambda)P(f)\}\phi(f^\infty) = \lambda\phi(f^\infty) + (1-\lambda)\phi(f^\infty) = \phi(f^\infty)$, and consequently, $P^*_\lambda(f)\phi(f^\infty) = \phi(f^\infty)$. We also have

$$\begin{aligned} r(f) + P_\lambda(f)D(f)r(f) - D(f)r(f) &= r(f) + \{\lambda I + (1-\lambda)P(f)\}D(f)r(f) - D(f)r(f) \\ &= r(f) + (\lambda-1)\{I - P(f)\}D(f)r(f) \\ &= r(f) + (\lambda-1)\{I - P^*(f)\}r(f) \\ &= \lambda r(f) + (1-\lambda)\phi(f^\infty). \end{aligned}$$

Hence, $(1-\lambda)r(f) + \{P_\lambda(f) - I\}D(f)r(f) = (1-\lambda)\phi(f^\infty)$. Multiplying this equality by $P^*_\lambda(f)$ gives $(1-\lambda)P^*_\lambda(f)r(f) = (1-\lambda)P^*_\lambda(f)\phi(f^\infty)\} = (1-\lambda)\phi(f^\infty)$, i.e. $\phi_\lambda(f^\infty) = \phi_(f^\infty)$. □

Theorem 5.19

Let $s_i(a) = r_i(a) - \phi_i + \sum_j p_{ij}(a)u_j - u_i,\ (i,a) \in S \times A$, *and* $m_i = liminf_{n\to\infty}\, e^n_i$,

$M_i = limsup_{n\to\infty}\, e^n_i$, *and* $A_*(i) = \{a \in A(i) \mid \phi_i = \sum_j p_{ij}(a)\phi_j\},\ i \in S$.

Then, $max_{a\in A_*(i)}\{s_i(a) + \sum_j p_{ij}(a)m_j\} \le m_i \le M_i \le max_{a\in A_*(i)}\{s_i(a) + \sum_j p_{ij}(a)M_j\},\ i \in S$.

Proof

For n sufficiently large, we obtain by Lemma 5.6

$$\begin{aligned} max_{a\in A_*(i)}\{s_i(a) + \textstyle\sum_j p_{ij}(a)e^n_j\} &= max_{a\in A_*(i)}\{r_i(a) - \phi_i + \textstyle\sum_j p_{ij}(a)u_j - u_i + \sum_j p_{ij}(a)e^n_j\} \\ &= max_{a\in A_*(i)}\{r_i(a) - \phi_i + \textstyle\sum_j p_{ij}(a)(u_j + e^n_j) - u_i\} \\ &= max_{a\in A_*(i)}\{r_i(a) - \phi_i + \textstyle\sum_j p_{ij}(a)(v^n_j - n\cdot\phi_j) - u_i\} \\ &= max_{a\in A_*(i)}\{r_i(a) - (n+1)\phi_i + \textstyle\sum_j p_{ij}(a)v^n_j - u_i\} \\ &= max_{a\in A_*(i)}\{r_i(a) + \textstyle\sum_j p_{ij}(a)v^n_j\} - (n+1)\phi_i - u_i \\ &= v^{n+1}_i - (n+1)\phi_i - u_i = e^{n+1}_i. \end{aligned}$$

Hence,

$$\begin{aligned} m_i &= liminf_{n\to\infty}\, e^{n+1}_i = liminf_{n\to\infty}\, max_{a\in A_*(i)}\{s_i(a) + \textstyle\sum_j p_{ij}(a)e^n_j\} \\ &\ge max_{a\in A_*(i)}\{s_i(a) + \textstyle\sum_j p_{ij}(a)\big(liminf_{n\to\infty}\, e^n_j\big)\} = max_{a\in A_*(i)}\{s_i(a) + \textstyle\sum_j p_{ij}(a)m_j\} \end{aligned}$$

and

$$\begin{aligned} M_i &= limsup_{n\to\infty}\, e^{n+1}_i = limsup_{n\to\infty}\, max_{a\in A_*(i)}\{s_i(a) + \textstyle\sum_j p_{ij}(a)e^n_j\} \\ &\le max_{a\in A_*(i)}\{s_i(a) + \textstyle\sum_j p_{ij}(a)\big(limsup_{n\to\infty}\, e^n_j\big)\} = max_{a\in A_*(i)}\{s_i(a) + \textstyle\sum_j p_{ij}(a)M_j\}. \end{aligned}$$

□

Theorem 5.20

Under the aperiodicity assumption, the sequence $\{e^n\}_{n=0}^{\infty}$ *is convergent.*

Proof

Suppose that $m_j = lim_{k\to\infty} e_j^{p_k}$ and $M_j = lim_{k\to\infty} e_j^{q_k},\ j \in S$, for some subsequences $\{p_k\}$ and $\{q_k\}$ of $\{0, 1, 2, \dots\}$. Choose for every $k \in \mathbb{N}$ an integer $h(k)$ such that $r_k := q_{h(k)} - p_k \geq k$.

From Lemma 5.5 part (1), we obtain $e^{q_{h(k)}} = e^{r_k + p_k} \geq \{P(f)\}^{r_k} \cdot e^{p_k}$ for any $f \in F_0$ and $k \in \mathbb{N}$.

Hence, for any $f \in F_0$,

$$\begin{aligned} M &= lim_{k\to\infty} e^{q_k} = lim_{k\to\infty} e^{q_{h(k)}} \geq lim_{k\to\infty} \{P(f)\}^{r_k} \cdot e^{p_k} \\ &= lim_{k\to\infty} \{P(f)\}^k \cdot e^{p_k} = \left\{lim_{k\to\infty} \{P(f)\}^k\right\} \cdot \left\{lim_{k\to\infty} e^{p_k}\right\} = P^*(f)m. \end{aligned}$$

Similarly, we obtain $m \geq P^*(f)M$.

Since $m \geq P^*(f)M \geq P^*(f)P^*(f)m = P^*(f)m$ and $P^*(f)\{m - P^*(f)m\} = 0$, we have $m_j = \{P^*(f)m\}_j$, and similarly $M_j = \{P^*(f)M\}_j$, for every state recurrent under $P(f)$.

Therefore, $m_j \geq \{P^*(f)M\}_j = M_j$ for every recurrent state, i.e. $m_j = M_j$ for every state which is recurrent under $P(f)$ for some $f \in F_0$.

Let f_* satisfy $f_*(i) \in A_*(i)$ and $\{s(f_*) + P(f_*)M\}_i = max_{a\in A_*(i)} \{s_i(a) + \sum_j p_{ij}(a)M_j\},\ i \in S$.

By Theorem 5.19, $s(f_*)+P(f_*)M \geq M$, implying $P^*(f_*)s(f_*) \geq 0$. Since (ϕ, u) is superharmonic, $s(f_*) \leq 0$, i.e. $P^*(f_*)s(f_*) = 0$ and therefore, $s_j(f_*) = 0$ for j recurrent under f_*.

Take policy f^∞ equal to f_*^∞ in the states which are recurrent under $P(f_*)$, and equal to a policy f_0^∞, where $f_0 \in F_0$, in the transient states of $P(f_*)$. Then, $f^\infty \in F_0$ and the states which are recurrent under $P(f_*)$ are a subset of the states which are recurrent under $P(f)$.

Hence, $m_j = M_j$ for the states j recurrent under $P(f_*)$. By Theorem 5.19, we obtain $s(f_*)+P(f_*)m \leq m \leq M \leq s(f_*)+P(f_*)M$, i.e. $P(f_*)(M-m) \geq M-m \geq 0$, and consequently,

$$P^*(f_*)(M - m) \geq M - m \geq 0. \tag{5.41}$$

Since $m_j = M_j$ for the states which are recurrent under $P(f_*)$, we have $P(f_*)(M - m) = 0$, and by (5.41), $M = m$. □

Corollary 5.6

$\phi = lim_{n\to\infty} (v^{n+1} - v^n)$.

Proof

$\phi = (v^{n+1} - v^n) - (e^{n+1} - e^n)$. Since the sequence $\{e^n\}_{n=0}^{\infty}$ converges, $lim_{n\to\infty} (e^{n+1} - e^n) = 0$, and consequently, $\phi = lim_{n\to\infty} (v^{n+1} - v^n)$. □

We will close this section by an algorithm to compute an ε-optimal policy under the following assumption.

Assumption 5.1
Every Markov chain $P(f)$ is aperiodic and the value vector is constant, i.e. $\phi = \phi_0 \cdot e$.

Theorem 5.21
Let $l_n = min_i\,(v_i^n - v_i^{n-1})$ and $u_n = max_i\,(v_i^n - v_i^{n-1})$. Then,

(1) $l_n \uparrow \phi_0$ *and* $u_n \downarrow \phi_0$.

(2) $l_n \cdot e \leq \phi(f_n^\infty) \leq \phi_0 \cdot e \leq u_n \cdot e$ *for every* $n \geq 1$.

Proof

(1) $v^{n+1} - v^n \geq \{r(f_n) + P(f_n)v^n\} - \{r(f_n) + P(f_n)v^{n-1}\} = P(f_n)\{v^n - v^{n-1}\}$
$\geq P(f_n) \cdot min_i\,\{v^n - v^{n-1}\}_i \cdot e = l_n \cdot e$, implying $l_{n+1} \geq l_n$.

Similarly, it can be shown that $u_{n+1} \leq u_n$.

Hence, with Corollary 5.6, we obtain $l_n \uparrow \phi_0$ and $u_n \downarrow \phi_0$.

(2) For any $n \geq 1$, we have $u_n \geq u_{n+1} \geq \cdots \geq lim_{k\to\infty}\, u_{n+k} = \phi_0$ and

$\phi(f_n^\infty) = P^*(f_n)r(f_n) = P^*(f_n)\{v^n - P(f_n)v^{n-1}\} = P^*(f_n)\{v^n - v^{n-1}\}$
$\geq P^*(f_n) \cdot min_i\,\{v^n - v^{n-1}\}_i \cdot e = min_i\,\{v^n - v^{n-1}\}_i \cdot e = l_n \cdot e.$ □

From the above theorem we can derive an algorithm. However, since $\phi = lim_{n\to\infty}\frac{1}{n}v^n$ (see Lemma 5.5 part (3)), v^n grows linearly in n, which may cause numerical difficulties. To overcome these difficulties, we use the following transformation. Let

$$w_i^n = v_i^n - v_N^n,\ i \in S,\ n \geq 0 \text{ and } g^n = v_N^n - v_N^{n-1},\ n \geq 1.$$

Then, we have $w_i^n = \{e_i^n + n\phi_0 + u_i\} - \{e_N^n + n\phi_0 + u_N\} = \{e_i^n - e_N^n\} + \{u_i - u_N\}$, which is a bounded sequence, and $g^n = \{e_N^n + n\phi_0 + u_N\} - \{e_N^{n-1} + (n-1)\phi_0 + u_N\} = \{e_N^n - e_N^{n-1}\} + \phi_0$, which is also a bounded sequence. Furthermore, the recurrence relations become

$$\begin{aligned} g^{n+1} &= v_N^{n+1} - v_N^n = max_{a\in A(N)}\,\{r_N(a) + \textstyle\sum_j\, p_{Nj}(a)(v_j^n - v_N^n)\} \\ &= max_{a\in A(N)}\,\{r_N(a) + \textstyle\sum_j\, p_{Nj}(a)w_j^n\}, \end{aligned}$$

and

$$\begin{aligned} w_i^{n+1} &= v_i^{n+1} - v_N^{n+1} = max_{a\in A(i)}\,\{r_i(a) + \textstyle\sum_j\, p_{ij}(a)(v_j^n - v_N^n)\} + (v_N^n - v_N^{n+1}) \\ &= max_{a\in A(i)}\,\{r_i(a) + \textstyle\sum_j\, p_{ij}(a)w_j^n\} - g^{n+1},\ i \in S. \end{aligned}$$

For the bounds l_n and u_n, we obtain

$$l_n = min_i\,(v_i^n - v_i^{n-1}) = min_i\,\{(w_i^n + v_N^n) - (w_i^{n-1} + v_N^{n-1})\} = min_i\,(w_i^n - w_i^{n-1}) + g^n$$

and

$$u_n = max_i\,(v_i^n - v_i^{n-1}) = max_i\,\{(w_i^n + v_N^n) - (w_i^{n-1} + v_N^{n-1})\} = max_i\,(w_i^n - w_i^{n-1}) + g^n.$$

In step 2 of algorithm 5.8 (see below) we use v for w^n, w for w^{n+1}, g for g^{n+1}, u for $u_{n+1} - g^{n+1}$ and l for $l_{n+1} - g^{n+1}$.

Algorithm 5.8 *Value iteration (aperiodicity and constant value vector case)*

1. Choose $\varepsilon > 0$ and take $v \in \mathbb{R}^N$ arbitrary with $v_N = 0$.
2. a. Compute $y_i(a) = r_i(a) + \sum p_{ij}(a)v_j,\ i \in S,\ a \in A(i)$.
   b. $g = max_{a \in A(N)}\, y_N(a)$.
   c. $w_i = max_{a \in A(i)}\, y_i(a) - g,\ i \in S$, and take f such that $w = r(f) + P(f)v - g \cdot e$.
   d. $u = max_i\,(w_i - v_i),\ l = min_i\,(w_i - v_i)$.
3. If $u - l \leq \varepsilon$: f^∞ is an ε-optimal policy and $\frac{1}{2}(u+l) + g$ is an $\frac{1}{2}\varepsilon$-approximation of ϕ_0

   (STOP);

   Otherwise: $v := w$ and return to step 2.

Example 5.9
Consider the MDP of Example 3.1. Notice that the value vector is constant ($\phi_0 = 7$). Although the requirement of aperiodicity is not fulfilled, the algorithm works, as one can see below. Take $\varepsilon = 0.1$ and $v^0 = (0, 0, 0)$.
Iteration 1:
$y_1(1) = 1,\ y_1(2) = 2,\ y_1(3) = 3;\ y_2(1) = 6,\ y_2(2) = 4,\ y_2(3) = 5;\ y_3(1) = 8,\ y_3(2) = 9,\ y_3(3) = 7.$
$g = 9;\ w = (-6, -3, 0);\ f(1) = 3,\ f(2) = 1,\ f(3) = 2;\ u = 0,\ l = -6;\ v = (-6, -3, 0).$
Iteration 2:
$y_1(1) = -5,\ y_1(2) = -1,\ y_1(3) = 0;\ y_2(1) = 0,\ y_2(2) = 1,\ y_2(3) = 5;\ y_3(1) = 2,$
$y_3(2) = 6,\ y_3(3) = 7.$
$g = 7;\ w = (-7, -2, 0);\ f(1) = 3,\ f(2) = 3,\ f(3) = 3;\ u = 0,\ l = -1;\ v = (-7, -2, 0).$
Iteration 3:
$y_1(1) = -6,\ y_1(2) = 0,\ y_1(3) = 3;\ y_2(1) = -1,\ y_2(2) = 2,\ y_2(3) = 5;\ y_3(1) = 1,$
$y_3(2) = 7,\ y_3(3) = 7.$
$g = 7;\ w = (-4, -2, 0);\ f(1) = 3,\ f(2) = 3,\ f(3) = 3;\ u = 0,\ l = 0$:
f^∞ with $f(1) = 3,\ f(2) = 3,\ f(3) = 3$ is an optimal policy and $\phi_0 = 7$.

5.10 Bibliographic notes

The concept 'communicating' was introduced by Bather [7]. Platzman [150] introduced the notion 'weakly communicating' under the name *simply connected*. In Kallenberg [111] the classification of MDPs is discussed and the question whether checking the unichain condition can be done in

polynomial time was raised in that paper. Tsitsiklis [199] has solved this problem by proving Theorem 5.1. From the paper McCuaig [136] it follows that for *deterministic MDPs* (each transition probability in $\{0,1\}$) this problem is solvable in polynomial time. Feinberg and Yang have shown ([67]) that other special cases (the so-called *recurrent* and *absorbing* cases) are also polynomially solvable.

Cesaro published in 1890 his idea concerning the convergence of averages ([30]). Theorem 5.3 can be found in many textbooks on Markov chains, e.g. Kemeny and Snell [120]. The proof of this theorem and also the theorems 5.5 and 5.6 follows Veinott [218]. Theorem 5.7 is due to Blackwell [22].

Blackwell [22] provided a theoretical framework for analyzing multichain models. His observation that the average reward model may be viewed as a limit of expected discounted reard models, in which the discount rate approaches 1, stimulated extensive research on average reward models. He introduced the concept of the so-called *1-optimality*, which later was renamed to Blackwell optimality. He also showed that the partial Laurent series expansion, given in Corollary 5.2, provided a link between these two models. The complete Laurent expansion, as presented in Theorem 5.10, is due to Miller and Veinott [138].

The average reward optimality equation (5.11) appears implicitly in Blackwell [22]; an explicit statement appears in Derman's book [55]. This optimality equation is extensively investigated by Schweitzer and Federgruen [177]. Theorem 5.12 is a Tauberian result which can be found in Hordijk [89].

Howard [101] presented the policy iteration algorithm. However, he did not show that the algoritm terminates in finitely many steps. Veinott [214] completed this analysis by establishing that the algorithm cannot cycle.

The linear programming approach for the average reward criterion was independently introduced by De Ghellinck [39] and Manne [134] for the completely ergodic case. The first analysis for the multichain case has been presented in Denardo and Fox [50] who proved Theorem 5.15. Denardo [46] and Derman [55] improved these results slightly. Hordijk and Kallenberg [95] have solved the remaining problems and proved Theorem 5.16. Kallenberg [108] provides a comprehensive analysis of all aspects of linear programming for MDP models.

The value iteration scheme (5.37) was proposed by Bellman [11] and Howard [101]. Lemma 5.5 is due to Brown [27]. The data transformation (5.40) to assure aperiodicity was proposed by Schweitzer [175]. Bounds on the value vector, as given in Theorem 5.21, can be found in Hastings [85]. Denardo [47] proved the convergence of the sequence $\{e^n\}_{n=0}^{\infty}$ under the unichain and aperiodicity assumption. This result was generalized to the multichain case, as shown in Theorem 5.20, by Schweitzer and Federgruen ([176], [178]). Algorithm 5.8, called the *relative value iteration*, is due to White [230].

5.11 Exercises

Exercise 5.1

Consider the following model:
$S = \{1, 2\};\ A(1) = A(2) = \{1, 2\};\ p_{11}(1) = 1,\ p_{12}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1;$
$p_{21}(1) = 0,\ p_{22}(1) = 1;\ p_{21}(2) = 1,\ p_{22}(2) = 0;\ r_1(1) = 2,\ r_1(2) = 2;\ r_2(1) = -2,\ r_2(1) = -2.$
In each state one can choose to stay in that state (action 1) or to move to the other state (action 2). Consider the nonstationary policy R which, on starting in state 1, remains in state 1 for one period, proceeds to state 2 and remains there for three periods, returns to state 1 and remains there for $3^2 = 9$ periods, proceeds to state 2 and remains there for $3^3 = 27$ periods, etc.
Compute for this policy

$$\phi_1(R) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\}$$

and

$$\overline{\phi}_1(R) = limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{1,R}\{r_{X_t}(Y_t)\}.$$

Exercise 5.2

a. Show that ordinary convergence implies Cesaro convergence.
b. Show, without making use of Theorem 5.2, that ordinary convergence implies Abel convergence.
c. Give a counterexample that Cesaro convergence does not imply ordinary convergence.
d. Give a counterexample that Abel convergence does not imply ordinary convergence.

Exercise 5.3

Give a counterexample that Abel convergence does not imply Cesaro convergence.

Exercise 5.4

Consider the stochastic matrix

$$P = \begin{pmatrix} 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0.25 & 0.25 & 0 & 0.25 & 0 & 0 & 0 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0.5 \end{pmatrix}.$$

a. Determine the ergodic sets and the transient states; write the matrix in standard form.
b. Determine the stationary matrix P^*.

Exercise 5.5

Consider the stochastic matrix

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0.25 & 0 & 0 & 0.25 & 0.5 \end{pmatrix}.$$

Determine the stationary matrix, the fundamental matrix and the deviation matrix.

Exercise 5.6

Show that the deviation matrix D satisfies $De = 0$, i.e. the rows sum up to 0.

Exercise 5.7

Let P be an irreducible double stochastic matrix, i.e. an irreducible stochastic matrix with $\sum_{i=1}^{n} p_{ij} = 1,\ j = 1, 2, \ldots, n$: both the rows and the columns sum up to 1.
Determine the stationary matrix P^*.

Exercise 5.8

Consider the following MDP:

$S = \{1, 2\};\ A(1) = \{1, 2, 3\},\ A(2) = \{1\};\ r_1(1) = 1,\ r_1(2) = \frac{3}{4},\ r_1(3) = \frac{1}{2};\ r_2(1) = 0;$

$p_{11}(1) = 0,\ p_{12}(1) = 1;\ p_{11}(2) = \frac{1}{2},\ p_{12}(2) = \frac{1}{2};\ p_{11}(3) = 1,\ p_{12}(3) = 0;,\ p_{21}(1) = 0,\ p_{22}(1) = 1.$

Determine for the deterministic policy f^∞ with $f(1) = 2,\ f(2) = 1$:

a. On which subinterval of $[0, 1)$ is f^∞ an α-discounted optimal policy.

b. $u^k(f)$ for $k = -1, 0, 1, \ldots$.

c. $v^T(f^\infty)$ for all $T = 1, 2, \ldots$.

d. The Laurent expansion for $v^\alpha(f^\infty)$ and $\alpha_0(f) = \frac{\|D(f)\|}{1+\|D(f)\|}$.

Exercise 5.9

Suppose that the MDP is irreducible. Then the value vector has identical components, say ϕ.
Show the following properties:

(1) $(x, y) = (\phi, u^0(f_0))$ is a solution of the equation

$x + y_i = max_{a \in A(i)}\{r_i(a) + \sum_j p_{ij}(a) y_j\}, i \in S$, where f_0^∞ is a Blackwell optimal policy.

(2) If (x, y) is a solution of the above equation, then $x = \phi$ and $y = u^0(f_0) + c \cdot e$ for some $c \in \mathbb{R}$.

(3) Consider the following MDP, which is obviously not an irreducible model, but multichain.

$S = \{1, 2, 3\},\ A(1) = A(2) = \{1, 2\},\ A(3) = 1.$

$r_1(1) = 3,\ r_1(2) = 1,\ r_2(1) = 0,\ r_2(2) = 1,\ r_3(1) = 2.$

$p_{11}(1) = 1,\ p_{12}(1) = p_{13}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1,\ p_{13}(2) = 0;$

$p_{21}(1) = 0,\ p_{22}(1) = 1,\ p_{23}(1) = 0;\ p_{21}(2) = p_{22}(2) = 0,\ p_{23}(2) = 1;$

$p_{31}(1) = p_{32}(1) = 0,\ p_{33}(1) = 1.$

Show that the optimality equation of part (1) doesn't has a solution for this multichain model.

Exercise 5.10

Suppose that the MDP is unichained. Then, for every f^∞, the average reward vector $\phi(f^\infty)$ has identical components, also denoted by $\phi(f^\infty)$.

Show the following properties:

(1) The linear system $\begin{cases} x \cdot e + \{I - P(f)\}y &= r(f) \\ y_1 &= 0 \end{cases}$ has a unique solution $x = \phi(f^\infty)$ and $y = u^0(f) - u_1^0(f) \cdot e$.

(2) Show that the set $B(i, f)$, defined in (5.13), in the unichain case can be simplified to

$$B(i, f) = \{a \in A(i) \mid r_i(a) + \textstyle\sum_j p_{ij}(a) y_j > x + y_i\},$$

where x and y are the solution of the system of part (1) of this exercise.

(3) Formulate the policy iteration algorithm for the unichain case.

(4) Consider the following MDP:

$S = \{1, 2\};\ A(1) = A(2) = \{1, 2\};\ r_1(1) = 4,\ r_1(2) = 2,\ r_2(1) = 3,\ r_2(2) = 1.$

$p_{11}(1) = \frac{1}{3}, p_{12}(1) = \frac{2}{3}; p_{11}(2) = \frac{2}{3}, p_{12}(2) = \frac{1}{3}; p_{21}(1) = \frac{1}{2}, p_{22}(1) = \frac{1}{2}; p_{21}(2) = \frac{1}{2}, p_{22}(2) = \frac{1}{2}.$

Show that the model is a unichain MDP and compute an average optimal policy by the algorithm of part (3), starting with the policy f^∞, where $f(1) = f(2) = 2$.

Exercise 5.11

Consider an MDP with $\rho := min_{i,j,a}\, p_{ij}(a) > 0$. Show the following properties.

(1) $P(f_n)y^n \le y^{n+1} \le P(f_{n+1})y^n,\ n \in \mathbb{N}$, where $y^n = v^n - v^{n-1}$.

(2) $span\, y^{n+1} \le (1 - N\rho) \cdot span\, y^n,\ n \in \mathbb{N}$.

(3) Algorithm 5.8 terminates in at most T iterations with $T = \frac{log\left\{\frac{\varepsilon}{u_0 - l_0}\right\}}{log(1 - N\rho)}$.

Exercise 5.12

Consider the operators L_f and U defined on $\mathbb{R}^N$ by:

$$L_f\, x = r(f) + P(f)x \text{ and } (Ux)_i = max_{a \in A(i)} \{r_i(a) + \textstyle\sum_j p_{ij}(a) x_j\},\ i \in S.$$

Show that for any $x \in \mathbb{R}^N$ and $f^\infty \in C(D)$ (without any assumption about the chain structure):

(1) If $L_f\, x \le y$, then $\phi(f) \le max_i\, (y - x)_i \cdot e$.

(2) If $L_f\, x \ge y$, then $\phi(f) \ge min_i\, (y - x)_i \cdot e$.

(3) $min_i\, (Ux - x)_i \cdot e \le \phi(f_x^\infty) \le \phi \le max_i\, (Ux - x)_i \cdot e$, where f_x satisfies $Ux = L_{f_x}\, x$.

Chapter 6

Average reward - special cases

6.1 The irreducible case

In this section we assume

Assumption 6.1
The Markov chain $P(f)$ is irreducible for every $f^\infty \in C(D)$.

We have seen in section 5.2.3 that checking the irreducibility property can be done in polynomial time (polynomial in $M = \sum_{i \in S} |A(i)|$).

In this case, for every policy f^∞ the stationary matrix has identical and strictly positive rows, and consequently the vector $\phi(f)$ has identical components. Therefore we may $\phi(f)$, and also the value vector ϕ, consider as a scalar. The irreducible case looks like the discounted case. Most results are similar and can be obtainted by the properties that the stationary matrix has identical rows with strictly positive elements.

6.1.1 Optimality equation

Theorem 6.1
Consider the optimality equation

$$x + y_i = max_{a \in A(i)} \Big\{ r_i(a) + \sum_j p_{ij}(a) y_j \Big\}, \ i \in S. \tag{6.1}$$

(1) $(x, y) = \big(\phi, u^0(f_0)\big)$, *where ϕ is the value and f_0^∞ a Blackwell optimal policy is a solution of the optimality equation.*

(2) *If (x, y) is a solution of the optimality equation, then $x = \phi$ and $y = u^0(f_0) + c \cdot e$ for some $c \in \mathbb{R}$.*

Proof

(1) We have seen in Theorem 5.11 that $\big(\phi, u^0(f_0)\big)$ is a solution of (5.11). Since the vector ϕ has identical components, $A(i, \phi) = A(i)$ for all i. Hence, $\big(\phi, u^0(f_0)\big)$ is a solution of (6.1).

(2) It follows also from Theorem 5.11 that if (x, y) is a solution of (5.11), then $x = \phi$.
From the optimality equation, we obtain $\phi + y \geq r(f_0) + P(f_0)y$. Furthermore, the property $\{I - P(f_0)\}D(f_0) = I - P^*(f_0)$ implies, $\phi + u^0(f_0) = r(f_0) + P(f_0)u^0(f_0)$.
Let $z = y - u^0(f_0)$, then $z \geq P(f_0)z$, i.e. $z - P(f_0)z \geq 0$. Since $P^*(f_0)\{z - P(f_0)z\} = 0$ and $P^*(f_0)$ has stricty positive elements, we have $z = P(f_0)z$. Consequently, $z = P^*(f_0)z = c \cdot e$ for some $c \in \mathbb{R}$, (the last equality because $P^*(f_0)$ has identical rows): $y = u^0(f_0) + c \cdot e$. □

Example 6.1

The following model does not satisfy the irreducibility assumption. We show that in that case the optimality equation (6.1) cannot be used.

$S = \{1, 2, 3\},\ A(1) = A(2) = \{1, 2\},\ A(3) = 1.$

$r_1(1) = 3,\ r_1(2) = 1,\ r_2(1) = 0,\ r_2(2) = 1,\ r_3(1) = 2.$

$p_{11}(1) = 1,\ p_{12}(1) = p_{13}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1,\ p_{13}(2) = 0;\ p_{21}(1) = 0,\ p_{22}(1) = 1,$
$p_{23}(1) = 0;\ p_{21}(2) = p_{22}(2) = 0,\ p_{23}(2) = 1;\ p_{31}(1) = p_{32}(1) = 0,\ p_{33}(1) = 1.$

The optimality equation (6.1) becomes:

$$x + y_1 = max\{3 + y_1, 1 + y_2\};\ x + y_2 = max\{0 + y_2, 1 + y_3\};\ x + y_3 = 2 + y_3.$$

The third equation gives $x = 2$. If we use this value in the first equation, we obtain:
$2 + y_1 = max\{3 + y_1, 1 + y_2\} \geq 3 + y_1$, implying that the system is infeasible.

Remark

We give a heuristic derivation that this optimality equation can be derived from the optimality equation for the discounted reward when the discount factor α tends to 1. First, we write the the optimality equation for the discounted reward as

$$0 = max_{a \in A(i)}\{r_i(a) + \alpha \sum_j p_{ij}(a)v_j^\alpha - v_i^\alpha\},\ i \in S.$$

Then, we use the first terms of the Laurent expansion: $v^\alpha = \frac{\phi \cdot e}{1-\alpha} + u^0 + \varepsilon(\alpha)$. We obtain

$$\begin{aligned} 0 &= max_{a \in A(i)}\Big\{r_i(a) + \alpha \textstyle\sum_j p_{ij}(a)\{\frac{\phi}{1-\alpha} + u_j^0\} - \{\frac{\phi}{1-\alpha} + u_i^0\} + \varepsilon(\alpha)\Big\},\ i \in S. \\ 0 &= max_{a \in A(i)}\{r_i(a) + \alpha \textstyle\sum_j p_{ij}(a)u_j^0 - \phi - u_i^0 + \varepsilon(\alpha)\},\ i \in S. \end{aligned}$$

If α tends to 1, denote ϕ as x and u^0 as y, we establish that

$$0 = max_{a \in A(i)}\{r_i(a) + \sum_j p_{ij}(a)y_j - x - y_i\},\ i \in S,$$

or

$$x + y_i = max_{a \in A(i)}\Big\{r_i(a) + \sum_j p_{ij}(a)y_j\Big\},\ i \in S.$$

6.1.2 Policy iteration

Theorem 6.2

The linear system $\left\{ \begin{array}{rcl} x \cdot e + \{I - P(f)\}y & = & r(f) \\ y_1 & = & 0 \end{array} \right.$ *has a unique solution* $x = \phi(f^\infty)$ *and* $y = u^0(f) - u_1^0(f) \cdot e$.

Proof

Multiply the first equality by $P^*(f)$:

$x \cdot P^*(f)e + P^*(f)\{I - P(f)\}y = P^*(f)r(f) = \phi(f^\infty) \cdot e \ \rightarrow \ x \cdot e = \phi(f^\infty) \cdot e$, i.e. $x = \phi(f^\infty)$.

Since $x = \phi(f^\infty) = P^*(f)r(f)$, the first equation can be written as:

$\{I - P(f) + P^*(f)\}y = r(f) - P^*(f)r(f) + P^*(f)y$, implying

$$\begin{aligned} y &= \{I - P(f) + P^*(f)\}^{-1}\big\{\big(I - P^*(f)\big)r(f) + P^*(f)y\big\} \\ &= \{D(f) + P^*(f)\}\big\{\big(I - P^*(f)\big)r(f) + P^*(f)y\big\} = D(f)r(f) + P^*(f)y = u^0(f) + c \cdot e, \end{aligned}$$

($P^*(f)y = c \cdot e$ because $P^*(f)$ has identical rows). Because $y_1 = 0$, we have, $c = -u_1^0(f)$. Hence, $y = u^0(f) - u_1^0(f) \cdot e$. □

For every $i \in S$ and $f^\infty \in C(D)$, we define the action set $B(i, f)$ by

$$B(i, f) = \big\{a \in A(i) \mid r_i(a) + \textstyle\sum_j p_{ij}(a)u_j^0(f) > \phi(f^\infty) + u_i^0(f)\big\}.$$

Since $u^0(f)$ and the solution y of the system in Theorem 6.1 differ a constant vector, we also have

$$B(i, f) = \big\{a \in A(i) \mid r_i(a) + \sum_j p_{ij}(a)y_j > \phi(f^\infty) + y_i\big\}. \tag{6.2}$$

Theorem 6.3

(1) If $B(i, f) = \emptyset$ *for every* $i \in S$*, then* f^∞ *is an average optimal policy.*

(2) If $B(i, f) \neq \emptyset$ *for at least one* $i \in S$ *and the policy* $g^\infty g \neq f^\infty$ *satisfies* $g(i) \in B(i, f)$ *if* $g(i) \neq f(i)$*, then* $\phi(g^\infty) > \phi(f^\infty)$.

Proof

(1) If $B(i, f) = \emptyset$ for every $i \in S$, then $r(g) + P(g)u^0(f) \leq \phi(f^\infty) \cdot e + u^0(f)$ for all $g^\infty \in C(D)$. Hence, $P^*(g)r(g) + P^*(g)P(g)u^0(f) \leq \phi(f^\infty) \cdot P^*(g)e + P^*(g)u^0(f)$ for all $g^\infty \in C(D)$, i.e. $\phi(g^\infty) \cdot e + P^*(g)u^0(f) \leq \phi(f^\infty) \cdot e + P^*(g)u^0(f)$. Therefore, $\phi(g^\infty) \leq \phi(f^\infty)$ for all $g^\infty \in C(D)$: f^∞ is average optimal.

(2) If $g(i) = f(i)$, then row i of $P(f)$ is identical to row i of $P(g)$, and also $r_i(f) = r_i(g)$. Hence, $\{r(g) + P(g)u^0(f)\}_i = \{r(f) + P(f)u^0(f)\}_i = \{P^*(f)r(f) + u^0(f\}_i = \phi(f^\infty) + u_i^0(f)$, the last equality but one because $I + P(f)D(f) = P^*(f) + D(f)$.

If $g(i) \neq f(i)$, then $g(i) \in B(i, f)$ and $\{r(g) + P(g)u^0(f)\}_i > \phi(f^\infty) + u_i^0(f)$.

Therefore, $r(g) + P(g)u^0(f) > \phi(f^\infty) \cdot e + u^0(f)$. Since the elements of $P^*(g)$ are strictly positive, we obtain $P^*(g)r(g) + P^*(g)u^0(f) > \phi(f^\infty) \cdot e + P^*(g)u^0(f)$: $\phi(g^\infty) > \phi(f^\infty)$. □

Algorithm 6.1 *Determination of an average optimal policy by policy iteration (irreducible case)*

1. Take an arbitrary $f^\infty \in C(D)$.

2. Determine the unique solution $(x = \phi(f^\infty), y)$ of the system

$$\begin{cases} x \cdot e + \{I - P(f)\}y &= r(f) \\ y_1 &= 0 \end{cases}$$

3. Determine for every $i \in S$

$B(i,f) = \{a \in A(i) \mid r_i(a) + \sum_j p_{ij}(a)y_j > \phi(f^\infty) + y_i\}$.

4. If $B(i,f) = \emptyset$ for every $i \in S$, then f^∞ is an average optimal policy (STOP).

Otherwise: (a) take g such that $g \neq f$ and $g(i) \in B(i,f)$ if $g(i) \neq f(i)$;

(b) $f := g$ and return to step 2.

Example 6.2

Apply Algorithm 6.1 to the following model (easy to check that the model is irreducible).

$S = \{1,2\};\ A(1) = A(2) = \{1,2\};\ r_1(1) = 4,\ r_1(2) = 2,\ r_2(1) = 3,\ r_2(2) = 1.$

$p_{11}(1) = \frac{1}{3}, p_{12}(1) = \frac{2}{3};\ p_{11}(2) = \frac{2}{3}, p_{12}(2) = \frac{1}{3};\ p_{21}(1) = \frac{1}{2}, p_{22}(1) = \frac{1}{2};\ p_{21}(2) = \frac{1}{2}, p_{22}(2) = \frac{1}{2}.$

Start with f^∞, where $f(1) = 2,\ f(2) = 1$.

Iteration 1:

The system is:
$$\begin{cases} x + \frac{1}{3}y_1 - \frac{1}{3}y_2 = 2 \\ x - \frac{1}{2}y_1 + \frac{1}{2}y_2 = 1 \\ y_1 = 0 \end{cases}$$

with solution $x = \frac{8}{5},\ y_1 = 0,\ y_2 = -\frac{6}{5}$.

$B(1,f) = B(2,f) = \{1\}$. Take $g(1) = g(2) = 1$. Then, $f(1) = f(2) = 1$

Iteration 2:

The system is:
$$\begin{cases} x + \frac{2}{3}y_1 - \frac{2}{3}y_2 = 4 \\ x - \frac{1}{2}y_1 + \frac{1}{2}y_2 = 3 \\ y_1 = 0 \end{cases}$$

with solution $x = \frac{24}{7},\ y_1 = 0,\ y_2 = -\frac{6}{7}$.

$B(1,f) = B(2,f) = \emptyset$: f^∞, where $f(1) = f(2) = 1$, is an optimal policy and the value is $\frac{24}{7}$.

6.1.3 Linear programming

Since the value vector is the smallest superharmonic vector (cf. Theorem 5.15), in the case where ϕ is a constant, ϕ is the unique x-solution of the linear program

$$min \left\{ x \;\middle|\; x + \sum_j \{\delta_{ij} - p_{ij}(a)\} y_j \geq r_i(a),\ i \in S,\ a \in A(i) \right\}. \tag{6.3}$$

The dual of (6.3) is:

$$max \left\{ \sum_{i,a} r_i(a) x_i(a) \;\middle|\; \begin{array}{lcl} \sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_{i,a} x_i(a) & = & 1 \\ x_i(a) \geq 0,\ i \in S,\ a \in A(i) & & \end{array} \right\}. \tag{6.4}$$

Remark

We show that the dual linear program (6.4) can be considered as the dual linear program (3.32) for the discounted reward in which the discount factor tends to 1. First, we remark that if we summing up the constraints of (3.32), we obtain $\sum_{i,a} (1-\alpha) x_i(a) = \sum_j \beta_j$. If we take $\beta_j = \frac{1}{N}$, $j \in S$, add this (redundant) constraint $\sum_{i,a} (1-\alpha) x_i(a) = 1$ and multiply both the objective function as the constraints of (3.32) with $(1-\alpha)$, the program becomes:

$$max \left\{ \sum_{i,a} r_i(a)(1-\alpha) x_i(a) \;\middle|\; \begin{array}{lcl} \sum_{i,a} \{\delta_{ij} - \alpha p_{ij}(a)\}(1-\alpha) x_i(a) & = & (1-\alpha)\beta_j,\ j \in S \\ \sum_{i,a} (1-\alpha) x_i(a) & = & 1 \\ (1-\alpha) x_i(a) \geq 0,\ i \in S,\ a \in A(i) & & \end{array} \right\}. \tag{6.5}$$

Let $\alpha \uparrow 1$ and denote $lim_{\alpha \uparrow 1} (1-\alpha) x_i(a)$ again by $x_i(a)$, $i \in S, a \in A(i)$, then we get (6.4).

Theorem 6.4

Let (ϕ, y^*) and x^* be optimal solutions of (6.3) and (6.4), respectively. Let f_*^∞ be such that $x_i^*\big(f_*(i)\big) > 0$, $i \in S$. Then, f_*^∞ is well-defined and an optimal policy.

Proof

Let x a feasible solution of (6.4) ((6.4) is feasible, because (6.3) has a finite optimal solution) and let $x_i = \sum_a x_i(a)$, $i \in S$. Let $\pi^\infty \in C(S)$ defined by $\pi_i(a) = \begin{cases} \frac{x_i(a)}{x_i} & \text{if } x_i > 0,\ i \in S; \\ \text{arbitrary} & \text{if } x_i = 0,\ i \in S. \end{cases}$

Hence, $x_i(a) = \pi_i(a) \cdot x_i$, $(i,a) \in S \times A$ and $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} \pi_i(a) \cdot x_i = 0$, $j \in S$. Therefore, $x^T \{I - P(\pi)\} = 0$, where $\{P(\pi\}_{ij} = \sum_a p_{ij}(a) \pi_i(a)$ for all $i, j \in S$, i.e. x is a stationary distribution of the Markov chain $P(\pi)$. Since the chain is irreducible, we have $x_i > 0$, $i \in S$.

Therefore, $x_i^* = \sum_a x_i^*(a) > 0$, $i \in S$, i.e. f_*^∞ is a well-defined policy.

From the orthogonality property of linear programming it follows that:

$$x_i^*(a) \cdot \Big\{ \phi + \textstyle\sum_j \{\delta_{ij} - p_{ij}(a)\} y_j^* - r_i(a) \Big\} = 0,\ i \in S,\ a \in A(i).$$

Hence, $\phi \cdot e + \{I - P(f_*)\} y^* = r(f_*)$. Multiply with $P^*(f_*)$: $\phi \cdot e = P^*(f_*) r(f_*) = \phi(f_*^\infty) \cdot e$, implying that f_*^∞ is an optimal policy. □

Algorithm 6.2

Determination of an average optimal policy by linear programming (irreducible case)

1. Determine an optimal solution x^* of the linear program (6.4).
2. Take $f_*^\infty \in C(D)$ such that $x_i^*\big(f_*(i)\big) > 0$ for every $i \in S$.

   The value vector ϕ is the optimum of the program and f_*^∞ is an optimal policy (STOP).

Example 6.3

Apply Algorithm 6.2 to the following model (easy to check that the model is irreducible).

$S = \{1,2\};\ A(1) = A(2) = \{1,2\};\ r_1(1) = 1,\ r_1(2) = 0,\ r_2(1) = 2,\ r_2(2) = 5.$

$p_{11}(1) = \frac{1}{2}, p_{12}(1) = \frac{1}{2};\ p_{11}(2) = \frac{1}{4}, p_{12}(2) = \frac{3}{4};\ p_{21}(1) = \frac{2}{3}, p_{22}(1) = \frac{1}{3};\ p_{21}(2) = \frac{1}{3}, p_{22}(2) = \frac{2}{3}.$

The linear program (6.4) is:

$$max\{1 \cdot x_1(1) + 0 \cdot x_1(2) + 2 \cdot x_2(1) + 5 \cdot x_2(2)\}$$

subject to

$$x_1(1) + x_1(2) = \tfrac{1}{2}x_1(1) + \tfrac{1}{4}x_1(2) + \tfrac{2}{3}x_2(1) + \tfrac{1}{3}x_2(2);$$
$$x_2(1) + x_2(2) = \tfrac{1}{2}x_1(1) + \tfrac{3}{4}x_1(2) + \tfrac{1}{3}x_2(1) + \tfrac{2}{3}x_2(2);$$
$$x_1(1) + x_1(2) + x_2(1) + x_2(2) = 1;$$
$$x_1(1),\ x_1(2),\ x_2(1),\ x_2(2) \geq 0.$$

The optimal optimal solution is:

$x_1^*(1) = 0,\ x_1^*(2) = \frac{4}{13},\ x_2^*(1) = 0,\ x_2^*(2) = \frac{9}{13};\ optimum = \frac{45}{13}.$

Therefore, the optimal policy is: $f_*(1) = 2,\ f_*(2) = 2$ and the value $\phi = \frac{45}{13}$.

As in the discounted case, there is a bijection between the feasible solutions of the dual program (6.5) and the set $C(S)$ of stationary policies. Let π^∞ be a stationary policy and let $x(\pi)$ be the stationary distribution of $P(\pi)$. Define x^π by

$$x_i^\pi(a) = x_i(\pi) \cdot \pi_{ia},\ (i,a) \in S \times A. \tag{6.6}$$

Reversely, let x be a feasible solution of (6.5). Define π^x by

$$\pi_{ia}^x = \frac{x_i(a)}{\sum_a x_i(a)},\ (i,a) \in S \times A. \tag{6.7}$$

Theorem 6.5

The mapping (6.6) is a bijection between the feasible solutions of the dual program (6.5) and the set $C(S)$ with (6.7) as the reverse mapping. Furthermore, the extreme solutions of (6.5) correspond to the set $C(D)$ of deterministic policies.

Proof

Let π^∞ be any stationary policy. Then, x^π, defined by (6.6), satisfies

$$\begin{cases} \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}x_i^\pi(a) = \sum_i\{\delta_{ij} - p_{ij}(\pi)\}x_i(\pi) = \left\{\{x(\pi)\}^T\{I - P(\pi)\}\right\}_j = 0, \ j \in S \\ \sum_{i,a} x_i^\pi(a) = \sum_i x_i(\pi) = 1 \text{ and } x_i^\pi(a) \geq 0 \text{ for all } (i,a) \in S \times A \end{cases}$$

Hence, x^π is a feasible solution of (6.5).

Conversely, let x be a feasible solution of (6.5). From the proof of Theorem 6.4 it follows that $\sum_a x_i(a) > 0, \ i \in S$, so the policy π^x is well-defined. Since $\pi_{ia}^{x^\pi} = \pi_{ia}, \ (i,a) \in S \times A$, (6.6) is a bijection with (6.7) as the reverse mapping.

Let $f^\infty \in C(D)$ and suppose that x^f is not an extreme point, i.e. $x^f = \lambda x^1 + (1-\lambda)x^2$, where $\lambda \in (0,1), \ x^1 \neq x^2$ and x^1, x^2 are feasible solutions of (6.5). Since $x_i^f(a) = 0$ for $a \neq f(i)$, also $x_i^1(a) = x_i^2(a) = 0, \ i \neq f(i), \ i \in S$. Therefore, both x^1 and x^2 are solutions of the same linear system $x^T\{(I - P(f)\} = 0, \ x^T e = 1$, which has a unique solution: $x^1 = x^2$, implying a contradiction, showing that x^f is an extreme point.

Finally, let x be an extreme point of (6.5). Since the sum of the first N components is zero in every column, the rank of the whole system ($N+1$ equations) is at most N. Therefore, any extreme solution has at most N positive components. Since, $\sum_a x_i(a) > 0, \ i \in S$, x has in each state exacty one positive component. Hence, the corresponding policy is deterministic. □

Next, we show the equivalence between linear programming and policy iteration. Consider a deterministic policy f^∞. We have seen that x^f is an extreme point of (6.5) and that $x_i^f(f(i)) > 0$ for every $i \in S$. In the simplex tableau corresponding to x^f, the column of a nonbasic variable $x_i(a)$ has as reduced costs (the transformed objective function value)

$$d_i(a) = x + \sum_j \{\delta_{ij} - p_{ij}(a)\}y_j - r_i(a).$$

Since $x_i^f(f(i)) > 0, \ i \in S$, it follows from the complementary slackness property of linear programming that $d_i(f(i)) = 0, \ i \in S$, implying $x \cdot e + \{I - P(f)\}y = r(f)$, and consequently $x \cdot e = P^*(f)r(f) = \phi(f^\infty) \cdot e$. Therefore, we have $x = \phi(f^\infty)$ and $\phi(f^\infty) \cdot e + \{I - P(f)\}y = r(f)$. Since, we also have $\phi(f^\infty) \cdot e + \{I - P(f)\}u^0(f) = r(f)$, we obtain $\{I - P(f)\}\{y - u^0(f)\} = 0$, i.e. $y - u^0(f) = P(f)\{y - u^0(f)\}$. Hence, $y - u^0(f) = P^*(f)\{y - u^0(f)\} = c \cdot e$ for some scalar c.

This implies that the reduced costs satisfy $d_i(a) = \phi(f^\infty) + \sum_j \{\delta_{ij} - p_{ij}(a)\}u^0(f) - r_i(a)$. Since $a \in B(i,f)$ if and only if $d_i(a) < 0$, it follows that the set of actions from which $g(i)$ may be chosen in policy iteration corresponds to the possible choices for the pivot column in the simplex method, which yieds the following theorem.

Theorem 6.6

(1) Any policy iteration algorithm is equivalent to a block-pivoting simplex algorithm.

(2) Any simplex algorithm is equivalent to a particular policy iteration algorithm.

6.1.4 Value iteration

In section 5.9 we presented an algorithm for value iteration under the assumption that the value vector is constant and the Markov chains $P(f),\ f^\infty \in C(D)$, are aperiodic. The last part of this assumption is not a serious restriction: by a data tranformation the original model can be transformed into a model in which every Markov chain $P(f),\ f^\infty \in C(D)$ is aperiodic and has the same average reward as the original Markov chain. In case of irreducibility no better algorithm than Algorithm 5.8 is known.

6.1.5 Modified policy iteration

In average reward models, value iteration may converge very slowly, and policy iteration may be inefficient in models with many states because of the need to solve large linear systems of equations. As in discounted models, modified policy iteration provides a compromis between these two algorithms. It avoids many value iterations and it avoids solving the linear system. Let the operators T and T_f, for $f\infty \in C(D)$ be defined by

$$(Tx)_i = max_a\{r_i(a) + \sum_j p_{ij}(a)x_j\},\ i \in S;\ T_f\, x = r(f) + P(f)x. \tag{6.8}$$

Notice that for $k \in \mathbb{N}$, $T_f^k\, x = r(f) + P(f)r(f) + \cdots + P^{k-1}(f)r(f) + P^k(f)x = v^k(f^\infty) + P^k(f)x$. The modified policy iteration algorithm is as follows.

Algorithm 6.3 *Modified value iteration (irreducible case)*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.

2. a. Choose $k \in \mathbb{N}$ arbitrary;

   b. Determine f such that $T_f\, x = T\, x$;

   c. Let $l = min_i\,(Tx - x)$ and $u = max_i\,(Tx - x)$.

3. If $u - l \leq \varepsilon$: f^∞ is an ε-optimal policy and $\frac{1}{2}(u + l)$ is a $\frac{1}{2}\varepsilon$-approximation of the value ϕ (STOP);

   Otherwise: $x := T_f^k\, x$ and return to step 2.

We work in the remaining part of this subsection under the following *strong aperiodicity* assumption.

Assumption 6.2
$p_{ii}(a) \geq \lambda > 0,\ i \in S$ *for some* $\lambda \in (0, 1)$.

We have seen in section 5.9 that the data transformation (5.40) gives strong aperiodicity without changing the average reward.

If $k = 1$, the method becomes the standard value iteration method. We will also argue that policy iteration corresponds to $k = \infty$. Let $\{x^n\}$, $\{f_n^\infty\}$ and $\{k_n\}$ the values of x, f and k in iteration $n+1$ of the algorithm: $T_{f_n} x^n = T x^n$ and $x^{n+1} = T_{f_n}^{k_n} x^n$. Since, by Theorem 5.8, $v^k(f^\infty) = k \cdot \phi(f^\infty) + u^0(f) - P^k(f)u^0(f)$ for every policy f^∞, we obtain

$$x^{n+1} = T_{f_n}^{k_n} x^n = v^{k_n}(f_n^\infty) + P^{k_n}(f_n)x^n = k_n \cdot \phi(f_n^\infty) + u^0(f_n) - P^{k_n}(f_n)\{u^0(f_n) - x^n\}.$$

If $k_n \to \infty$, $P^{k_n}(f_n) \to P^*(f_n)$ and $P^{k_n}(f_n)\{u^0(f_n) - x^n\}$ converges to a constant vector. Hence, x^{n+1} and $u^0(f_n)$ differ a constant vector for a large k_n. In policy iteration with best improving actions, a new policy in state i is obtained by maximizing $r_i(a) + \sum_j p_{ij}(a)u_j^0(f_n)$, wich gives the same policy as maximizing $r_i(a) + \sum_j p_{ij}(a)x_j^{n+1} = \left(Tx^{n+1}\right)_i$, which is the determination of the policy in step 2a of Algorithm 6.3. Let

$$g^n = Tx^n - x^n,\ l_n = min_i\, g_i^n, \text{ and } u_n = max_i\, g_i^n,\ i \in S,\ n = 0, 1, \ldots.$$

Lemma 6.1

$l_n \le \phi(f_n^\infty) \le \phi \le u_n$ for all $f^\infty \in C(D)$ and all $n \in \mathbb{N}$.

Proof

For all $f^\infty \in C(D)$, we have $P^*(f)\{r(f) + P(f)x^n - x^n\} = \phi(f^\infty) \cdot e$. So, with $f = f_n$,

$$\phi(f_n^\infty) \cdot e \quad = \quad P^*(f_n)\{r(f_n) + P(f_n)x^n - x^n\} = P^*(f_n)\{Tx^n - x^n\} \ge P^*(f_n)l_n \cdot e = l_n \cdot e.$$

Clearly, $\phi(f_n^\infty) \le \phi$, and with $f = f_*$, where f_*^∞ is an average optimal policy,

$$\begin{aligned} \phi \cdot e &= \phi(f_*^\infty) \cdot e = P^*(f_*)\{r(f_*) + P(f_*)x^n - x^n\} \\ &\le P^*(f_*)\{Tx^n - x^n\} \ge P^*(f_*)u_n \cdot e = u_n \cdot e. \end{aligned}$$

□

Lemma 6.2

The sequence $\{l_n,\ n = 0, 1, \ldots\}$ is monotonically nondecresing.

Proof

$$\begin{aligned} T\,x^{n+1} - x^{n+1} &\ge T_{f_n}\,x^{n+1} - x^{n+1} = T_{f_n}^{k_n+1}\,x^n - T_{f_n}^{k_n}\,x^n \\ &= \{r(f_n) + P(f_n)r(f_n) + \cdots + P^{k_n}(f_n)r(f_n)) + P^{k_n+1}(f_n)x^n\} - \\ &\qquad \{r(f_n) + P(f_n)r(f_n) + \cdots + P^{k_n-1}(f_n)r(f_n) + P^{k_n}(f_n)x^n\} \\ &= P^{k_n}(f_n)\{T_{f_n}\,x^n - x^n\} = P^{k_n}(f_n)\{T\,x^n - x^n\} \ge l_n \cdot P^{k_n}(f_n)e = l_n \cdot e. \end{aligned}$$

Hence, $min_i\,(Tx^{n+1} - x^{n+1})_i = l_{n+1} \ge l_n$. □

In the special case $k = 1$ (value iteration), also the sequence $\{u_n,\ n = 0, 1, \ldots\}$ is monotone, actually nonincreasing (see Theorem 5.21). However, this is not the case if $k \ge 2$, as the next example shows.

Example 6.4

$S = \{1, 2\}$; $A(1) = \{1\}$, $A(2) = \{1, 2\}$; $r_1(1) = 100$, $r_2(1) = 0$, $r_2(2) = 10$.

$p_{11}(1) = 1, p_{12}(1) = 0$; $p_{21}(1) = 0.9$, $p_{22}(1) = 0.1$; $p_{21}(2) = 0.1$, $p_{22}(2) = 0.9$.

Start with $x^0 = (0,0)$ and take $k = 2$.

$(T\,x^0)_1 = 100 + 1 \times 0 = 100;\ (T\,x^0)_2 = max\{0 + 0.9 \times 0 + 0.1 \times 0, 10 + 0.1 \times 0 + 0.9 \times 0\} = 10.$

$f_0(1) = 1,\ f_0(2) = 2;\ l_0 = 10,\ u_0 = 100.$

$x^1 = T^2_{f_0}\,x^0 = T_{f_0}\{T\,x^0\} = (100 + 100, 10 + 0.1 \times 0, 10 + 0.1 \times 100 + 0.9 \times 10) = (200, 29).$

$(T\,x^1)_1 = 100 + 1 \times 200 = 300;$

$(T\,x^1)_2 = max\{0 + 0.9 \times 200 + 0.1 \times 29, 10 + 0.1 \times 200 + 0.9 \times 29\} = 182.9.$

$f_1(1) = 1,\ f_1(2) = 1;\ l_1 = 100,\ u_1 = 153.9.$ Hence, $u_1 > u_0$.

In the next lemma we show that the aperiodicity and irreducibility assumptions together imply that $\gamma > 0$, where

$$\gamma = min_{i,j \in S}\, min_{h_1,h_2,\ldots,h_{N-1}} \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij}. \tag{6.9}$$

Lemma 6.3

$\gamma > 0$, where γ is defined in (6.9).

Proof

It is sufficient to show that $\{P(h_1)P(h_2)\cdot P(h_{N-1})\}_{ij} > 0$ for all $h_1, h_2, \ldots, h_{N-1}$ and $i, j \in S$. Let $h_1, h_2, \ldots, h_{N-1}$ be arbitrary decision rules. We define for $n = 0, 1, \ldots, N-1$ and $i \in S$:

$$S(i,0) = \{i\};\ S(i,n) = \{j \in S \mid \{P(h_1)P(h_2)\cdot P(h_n)\}_{ij} > 0\},\ n = 1, 2, \ldots, N-1.$$

Then, it has to be shown that $S(i, N-1) = S$ for all $i \in S$. Clearly, $S(i,n) \subseteq S(i,n+1)$, namely: if $j \in S(i,n)$, i.e. $\{P(h_1)P(h_2)\cdot P(h_n)\}_{ij} > 0$ and the strong aperiodicity holds, then

$$\begin{aligned} \{P(h_1)P(h_2)\cdots P(h_{n+1})\}_{ij} &= \textstyle\sum_k \{P(h_1)P(h_2)\cdots P(h_n)\}_{ik} P(h_{n+1})_{kj} \\ &\geq \{P(h_1)P(h_2)\cdot P(h_n)\}_{ij} P(h_{n+1})_{jj} > 0. \end{aligned}$$

Hence, it remains to show that the sets $S(i,n)$ are strictly increasing in n as long as $S(i,n) \neq S$. Suppose $S(i,n+1) = S(i,n) \neq S$. Then, we have for all $j \in S(i,n)$ and all $k \notin S(i,n)$ that $P(h_{n+1})_{jk} = 0$, otherwise $k \in S(i,n+1)$. Therefore, $S(i,n)$ is closed under $P(h_{n+1})$, which contradicts the irreducibility of the Markov chain $P(h_{n+1})$. □

The following lemma implies that the sequence $\{l_n,\ n = 0, 1, \ldots\}$ converges to the value ϕ exponentially fast.

Lemma 6.4

If $n, m \in \mathbb{N}$ satisfy $\sum_{i=0}^{m-1} k_{n+i} \geq N-1$, then $\phi - l_{n+m} \leq (1-\gamma)(\phi - l_n)$.

Proof

It follows from the proof of Lemma 6.2 that $g^{n+1} \geq P^{k_n}(f_n)g^n$. Consequently, for all $m = 1, 2, \ldots,$

$$g^{n+m} \geq P^{k_{n+m-1}}(f_{n+m-1})P^{k_{n+m-2}}(f_{n+m-2})\cdots P^{k_n}(f_n)g^n. \tag{6.10}$$

Let j_0 such that $u_n = g^n_{j_0}$. Then, for all $i \in S$ and all h_1, h_2, h_{N-1},

$$\begin{aligned}
\{P(h_1)P(h_2)\cdots P(h_{N-1})g^n\}_i &= \textstyle\sum_j \{P(h_1)P(h_2)\cdots P(h_{N-})\}_{ij} g^n_j \\
&= \textstyle\sum_{j\neq j_0} \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij} g^n_j + \\
&\qquad \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij_0} g^n\}_{j_0} \\
&= \textstyle\sum_{j\neq j_0} \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij} g^n_j + \\
&\qquad \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij_0} u_n \\
&\geq \textstyle\sum_{j\neq j_0} \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij} l_n + \\
&\qquad \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij_0} u_n \\
&= \big\{1 - \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij_0}\big\} l_n + \\
&\qquad \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij_0} u_n \\
&= l_n + \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ij_0}(u_n - l_n) \\
&\geq l_n + \gamma(u_n - l_n) = (1-\gamma)l_n + \gamma u_n \geq (1-\gamma)l_n + \gamma\phi,
\end{aligned}$$

the last inequality by Lemma 6.1. So, $P(h_1)P(h_2)\cdots P(h_{N-1})g^n \geq \{(1-\gamma)l_n + \gamma\phi\}\cdot e$.

Then, also for $k > N-1$ and all $h_1, h_2, \ldots, h_k$,

$$\begin{aligned}
P(h_1)P(h_2)\cdots P(h_k)g^n &\geq \{P(h_1)P(h_2)\cdots P(h_{k-N+1})\}\{P(h_{k-N+2})P(h_{k-N+3})\cdots P(h_k)g^n\} \\
&\geq \{P(h_1)P(h_2)\cdots P(h_{k-N+1})\}\{(1-\gamma)l_n + \gamma\phi\}\cdot e \\
&= \{(1-\gamma)l_n + \gamma\phi\}\cdot e.
\end{aligned}$$

Hence, with (6.10), for all n, m such that $\sum_{i=0}^{m-1} k_{n+i} \geq N-1$,

$$g^{n+m} \geq P^{k_{n+m-1}}(f_{n+m-1})P^{k_{n+m-2}}(f_{n+m-2})\cdots P^{k_n}(f_n)g^n \geq \{(1-\gamma)l_n + \gamma\phi\}\cdot e.$$

Thus, $l_{n+m} \geq (1-\gamma)l_n + \gamma\phi$, i.e. $\phi - l_{n+m} \leq (1-\gamma)(\phi - l_n)$. □

We now know that l_n converges to ϕ exponentially fast. Thus, f^∞_n will be ε-optimal for n sufficiently large. The problem, however, is to recognize this. Therefore, we want that also u_n converges to ϕ. Define δ by

$$\delta = min_{i,j\in S}\, min_{f^\infty\in C(D)} \{P^*(f)\}_{ij} > 0. \tag{6.11}$$

Lemma 6.5
$u_n - l_n \leq \frac{1}{\delta}(\phi - l_n)$ for all $n = 0, 1, \ldots$.

Proof
$\phi\cdot e \geq \phi(f^\infty_n)\cdot e = P^*(f_n)\{rf_n) + P(f_n)x^n - x^n\} = P^*(f_n)\{Tx^n - x^n\} = P^*(f_n)g^n$.

Let j_0 such that $u_n = g^n_{j_0}$. Then, for all $i \in S$,

$$\begin{aligned}
\{P^*(f_n)g^n\}_i &= \textstyle\sum_{j\neq j_0} p^*_{ij} g^n_j + p^*_{ij_0} g^n_{j_0} \geq \sum_{j\neq j_0} p^*_{ij} l_n + p^*_{ij_0} u_n \\
&= (1 - p^*_{ij_0})l_n + p^*_{ij_0} u_n = l_n + p^*_{ij_0}(u_n - l_n) \\
&\geq l_n + \delta(u_n - l_n) = (1-\delta)l_n + \delta u_n.
\end{aligned}$$

Hence, $P^*(f_n)g^n \geq \{(1-\delta)l_n + \delta u_n\}\cdot e$, and so, $\phi \geq (1-\delta)l_n + \delta u_n$, i.e. $u_n - l_n \leq \frac{1}{\delta}(\phi - l_n)$. □

Corollary 6.1
$u_n - l_n \to 0$ *for* $n \to \infty$.

Theorem 6.7
Algorithm 6.3 is correct.

Proof
Since $u_n - l_n$ converges to 0, the algorithm terminates. By Lemma 6.1, $l_n \leq \phi(f_n^\infty) \leq \phi \leq u_n$. Hence, if $u_n - l_n < \varepsilon$, f_n^∞ is an ε-optimal policy. Furthermore, $|\phi - \frac{1}{2}(u_n + l_n)| < \frac{1}{2}\varepsilon$, i.e. $\frac{1}{2}(u_n + l_n)$ is a $\frac{1}{2}\varepsilon$-approximation of ϕ. □

6.2 Unichain case

In this section we assume

Assumption 6.3
For every $f^\infty \in C(D)$ *the Markov chain* $P(f)$ *has exactly one ergodic class plus a possibly empty set of transient states.*

We have seen in section 5.2.3 that checking the unichain property is, in general, $\mathcal{NP}$-complete.

Also in this case, for every policy f^∞ the stationary matrix has identical rows, and consequently the vector $\phi(f)$ has identical components and we may $\phi(f)$ and the value vector ϕ consider as a scalar.

6.2.1 Optimality equation

We first argue that the result of Theorem 6.1 also holds in the unichain case. Following the proof of Theorem 6.1, we obtain part (1) and part (2) until $z - P(f_0)z \geq 0$ and $P^*(f_0)\{z - P(f_0)z\} = 0$. We have to give another proof that $z = P^*(f_0)z$, because the property that $P^*(f_0)$ has strictly positive elements doesn't hold in the unichain case. The columns of $P^*(f_0)$ are zero for the transient states and therefore the vector $P^*(f_0)z$ doesn't depend on the values of z_i for transient states i. Hence, we have to show that $z_j = \{P^*(f_0)z\}_j$ for the ergodic states j. But the states in this (only) ergodic class generate an irreducible Markov chain and the proof is similar as in Theorem 6.1.

Example 6.5
$S = \{1, 2\};\ A(1) = \{1, 2\},\ A(2) = \{1\};\ r_1(1) = 5,\ r_1(2) = 10,\ r_2(1) = -1.$
$p_{11}(1) = 0.5,\ p_{12}(1) = 0.5;\ p_{11}(2) = 0,\ p_{12}(2) = 1;\ p_{21}(1) = 0,\ p_{22}(1) = 1.$
It is obvious that this MDP is unichain and not irreducible (state 2 is absorbing and state 1 transient under all policies).

The optimality equation is: $\begin{cases} x + y_1 = max\{5 + 0.5y_1 + 0.5y_2, 10 + y_2\} \\ x + y_2 = -1 + y_2 \end{cases}$

From the second equation we obtain $x = -1$. If we take $y_2 = 0$, the first equation becomes $-1 + y_1 = max\{5 + 0.5y_1, 10\}$. This equation has the solution $y_1 = 12$.
This model has two deterministic stationary policies: f_1^∞ and f_2^∞ with $f_1(1) = 1$ and $f_2(1) = 2$. Both policies are average optimal ($\phi_i(f_1^\infty) = \phi_i(f_2^\infty) = -1,\ i = 1, 2$) and $u_1^0(f_1) = 12,\ u_1^0(f_2) = 11,\ u_2^0(f_1) = u_2^0(f_2) = 0$ (easy to verify). Hence, $\big(\phi(f_1^\infty), u^0(f_1)\big)$ satisfies the optimality equation and $\big(\phi(f_2^\infty), u^0(f_2)\big)$ does not.

For $y \in \mathbb{R}^N$ a decision rule g is *y-improving* if $r(g) + P(g)y = max_f \{r(f) + P(f)y\}$.

Lemma 6.6
Let (ϕ, y) be a solution of the optimality equation (6.1) and let the decision rule g be y-improving. Then, g^∞ is an optimal policy.

Proof
From the optimality equation and the y-improving property of g it follows that

$$r(g) + P(g)y = max_f \{r(f) + P(f)y\} = \phi \cdot e + y.$$

By multiplying this equality by $P^*(g)$, we obtain $\phi(g^\infty) \cdot e = \phi \cdot e$: g^∞ is an optimal policy. □

From Example 6.4 it follows that the reverse statement (if g^∞ is an optimal policy, then g is y-improving) need not hold. The policy f_2^∞ is average optimal, but not y-maximizing:
$y = (12, 0),\ r_1(f_2) + \sum_j p_{1j}(f_2)y = 10$ and $max_f \{r_1(f) + \sum_j p_{1j}(f)y\} = 11$.

6.2.2 Policy iteration

The policy iteration method is the same as in the irreducible case, but the proof of finiteness is different. In the irreducible case for subsequent policies the average reward increases strictly. This is not true in the unichain case, where we have increasing in the following lexicographic sense: either the average reward increases strictly or there is no decrease in the average reward, but there is a strict increase in the bias term u^0. We will discuss the version of Algorithm 6.1 in which the 'best' improving actions are taken.

Algorithm 6.4 *Determination of an average optimal policy by policy iteration (unichain case)*

1. Take an arbitrary $f^\infty \in C(D)$.

2. Determine the unique solution $(x = \phi(f^\infty), y)$ of the system

$$\begin{cases} x \cdot e + \{I - P(f)\}y &= r(f) \\ y_1 &= 0 \end{cases}$$

3. Determine for every $i \in S$

   $B(i,f) = \{a \in A(i) \mid r_i(a) + \sum_j p_{ij}(a)y_j > \phi(f^\infty) + y_i\}$.

4. If $B(i,f) = \emptyset$ for every $i \in S$, then f^∞ is an average optimal policy (STOP).

   Otherwise: (a) take g such that $r_i(g) + \sum_j p_{ij}(g)y_j = max_a \{r_i(a) + \sum_j p_{ij}(a)y_j\},\ i \in S$;

   (b) $f := g$ and return to step 2.

Theorem 6.8

(1) If $B(i,f) = \emptyset$ for every $i \in S$, then f^∞ is an average optimal policy.

(2) If $B(i,f) \neq \emptyset$ for at least one $i \in S$, then

(a) $r(g) + P(g)y > \phi(f^\infty) \cdot e + y$;

(b) if $\{r(g) + P(g)y\}_i > \phi(f^\infty) + y_i$ for some state i which is recurrent under $P(g)$, then $\phi(g^\infty) > \phi(f^\infty)$.

(c) if $\{r(g) + P(g)y\}_i = \phi(f^\infty) + y_i$ for all states which are recurrent under $P(g)$, $\phi(g^\infty) = \phi(f^\infty)$ and $u^0(g) > u^0(f)$.

(3) Algorithm 6.3 terminates with an average optimal policy.

Proof

The proofs of (1) and (2) part (a) are similar to the proof of the same results in Theorem 6.3.

Part (2b):

Since $P^*(g)$ has identical rows which elements are strictly positive for recurrent states, we obtain $P^*(g)\{r(g) + P(g)y\} > P^*(g)\{\phi(f^\infty) + y\}$, i.e. $\phi(g^\infty) \cdot e > \phi(f^\infty) \cdot e$.

Part (2c):

Since $P^*(g)$ has identical rows which elements are zero for transient states, we obtain $P^*(g)\{r(g) + P(g)y\} = P^*(g)\{\phi(f^\infty) + y\}$, i.e. $\phi(g^\infty) \cdot e = \phi(f^\infty) \cdot e$.

Since $g(i) = f(i),\ i \in R(g)$, the set of states which are recurrent under $P(g)$, $R(g)$ is also a recurrent class under $P(f)$. Because the Markov chains have only one recurrent set, $R(g) = R(f)$ and $T(g) = T(f)$ (the last sets are the sets of transient states). Hence, also $P^*(g) = P^*(f)$.

We have to use the structure of $P(g)$ as a unichain Markov chain: $P(g) = \begin{pmatrix} P_{RR} & O \\ P_{TR} & P_{TT} \end{pmatrix}$, where $(I - P_{TT})$ is a nonsingular matrx and $(I - P_{TT})^{-1} = \sum_{t=0}^\infty P_{TT}^t$. $P^*(g) = \begin{pmatrix} P^*_{RR} & O \\ P^*_{TR} & O \end{pmatrix}$, and consequently $\big(I - P(g) + P^*(g)\big) = \begin{pmatrix} A & O \\ B & C \end{pmatrix}$, and $Z(f) = \big(I - P(f) + P^*(f)\big)^{-1} = \begin{pmatrix} W & O \\ X & Y \end{pmatrix}$, where $W = A^{-1},\ Y = C^{-1} = (I - P_{TT})^{-1} = \sum_{t=0}^\infty P_{TT}^t \geq I_T$ and $X = -C^{-1}BA^{-1}$.

$D(g) = Z(g) - P^*(g) = \begin{pmatrix} D_{RR} & O \\ D_{TR} & Y \end{pmatrix}$.

Let $z = r(g) + P(g)u^0(f) - \phi(f^\infty) \cdot e - u^0(f) = r(g) + P(g)y - \phi(f^\infty) \cdot e - y$.

Then, $z_i = 0,\ i \in R(g),\ z_i \geq 0,\ i \in T(g)$ and $z_i > 0$ for at least one $i \in T(g)$.

Hence, $D(g)z = \begin{pmatrix} D_{RR} & O \\ D_{TR} & Y \end{pmatrix} \begin{pmatrix} 0 \\ z_T \end{pmatrix} = \begin{pmatrix} 0 \\ Yz_T \end{pmatrix} \geq \begin{pmatrix} 0 \\ z_T \end{pmatrix} > 0$.

Therefore,

$$\begin{aligned} D(g)z &= D(g)\{r(g) + P(g)u^0(f) - \phi(f^\infty)\cdot e - u^0(f)\} \\ &= u^0(g) + D(g)\{P(g) - I\}u^0(f) - \phi(f^\infty)\cdot D(g)e \\ &= u^0(g) + \{P^*(g) - I\}u^0(f) = u^0(g) + \{P^*(f) - I\}u^0(f) = u^0(g) - u^0(f). \end{aligned}$$

Since $D(g)z > 0$, we obtain $u^0(g) > u^0(f)$.

Part (3):

From part (2) it follows that the vectors $\big(\phi(f^\infty), u^0(f)\big)$ of the iterates are lexicographically strictly increasing. Hence, each policy occurs at most once and after a finite amount of iterations $B(i,f) = \emptyset$ for every $i \in S$, i.e. the algorithm terminates with an average optimal policy. □

Example 6.4 (continued)

We apply Algorithm 6.4 to this model starting with $f(1) = 2,\ f(2) = 1$.

Iteration 1

Consider the system $\begin{cases} x + y_1 - y_2 = 10 \\ x + y_2 - y_2 = -1 \\ y_1 = 0 \end{cases} \rightarrow\ x = \phi(f^\infty) = -1,\ y_1 = 0,\ y_2 = -11.$

$B(1,f) = \{1\},\ B(2,f\} = \emptyset.\ g(1) = 1,\ g(2) = 1;\ f(1) = 1,\ f(2) = 1.$

Iteration 2

Consider the system $\begin{cases} x + y_1 - 0.5y_1 - 0.5y_2 = 5 \\ x + y_2 - y_2 = -1 \\ y_1 = 0 \end{cases} \rightarrow\ x = \phi(f^\infty) = -1,\ y_1 = 0,\ y_2 = -12.$

$B(1,f) = B(2,f\} = \emptyset : f^\infty$ is an average optimal policy.

6.2.3 Linear programming

In the unichain case the same linear program can be used as in the irreducible case, but the result is slightly different. The value ϕ is again the unique x-part of program (6.3), but we loose the property that every feasible solution x of the dual program (6.4) satisfies $\sum_a x_i(a) > 0,\ i \in S$. It turns out that this property can only be shown for recurrent states. However, since there is only one recurrent set, the actions in transient states doesn't influence the average reward for that states. The following theorem shows the result.

Theorem 6.9

Let (ϕ, y^*) and x^* be optimal solutions of the linear programs (6.3) and (6.4), respectively. Define f_*^∞ such that for every $i \in S$ $x_i^*\big(f_*(i)\big) > 0$ if $\sum_a x_i^*(a) > 0$ and $f_*(i)$ is an arbitrary action if $\sum_a x_i^*(a) = 0$. Then, f_*^∞ is an average optimal policy.

Proof

Suppose that $\sum_a x_j^*(a) = 0$. The constraints of (6.4) imply $0 = \sum_a x_j^*(a) = \sum_{i,a} p_{ij}(a)x_i^*(a)$. Hence, $p_{ij}(a)x_i^*(a) = 0,\ (i,a) \in S \times A$. Therefore, in states i with $\sum_a x_i^*(a) > 0$ we have

$p_{ij}(f_*) = 0$, i.e. the set $S_{x^*} = \{i \mid \sum_a x_i^*(a) > 0\}$ is closed in the Markov chain $P(f_*)$.
Since this Markov chain has only one ergodic set, the states $S \backslash S_{x^*}$ are transient under $P(f_*)$.
From the orthogonality property of linear programming it follows that:

$$x_i^*(a) \cdot \Big\{\phi + \sum_j \{\delta_{ij} - p_{ij}(a)\} y_j^* - r_i(a)\Big\} = 0, \ i \in S, \ a \in A(i).$$

Hence, $\phi + \{\big(I - P(f_*)\big)y^*\}_i - r_i(f_*) = 0, \ i \in S_{x^*}$. Multiply $\phi \cdot e + \{I - P(f_*)y^*\} - r(f_*)$ by $P^*(f_*)$ and notice that the columns of $P^*(f_*)$ are zeros for the states in $S \backslash S_{x^*}$, because these states are transient. Therefore, we obtain $0 = \phi \cdot e - P^*(f_*)r(f_*) = \phi \cdot e - \phi(f_*^\infty) \cdot e$, implying that f_*^∞ is an optimal policy. □

Algorithm 6.5

Determination of an average optimal policy by linear programming (unichain case)

1. Determine an optimal solution x^* of the linear program (6.4).
2. Let $S_{x^*} = \{i \mid \sum_a x_i^*(a) > 0\}$.
3. Take any $f_*^\infty \in C(D)$ such that $x_i^*\big(f_*(i)\big) > 0$ for every $i \in S_{x^*}$.

   The value vector ϕ is the optimum of the program and f_*^∞ is an optimal policy (STOP).

Example 6.4 (continued)

The linear program (6.4) for this model becomes

$max\{5x_1(1) + 10x_1(2) - x_2(1)\}$

subject to

$x_1(1) + x_1(2) = \frac{1}{2}x_1(1); \ x_2(1) = \frac{1}{2}x_1(1) + x_1(2) + x_2(1);$

$x_1(1) + x_1(2) + x_2(1) = 1; \ x_1(1), x_1(2), x_2(1) \geq 0.$

The optimal optimal solution is $x_1^*(1) = x_1^*(2) = 0, \ x_2^*(1) = 1; \ optimum = -1$.
Therefore, in state 1 any action can be chosen and the two deterministic policies are both optimal.
From this examples it also follows that in the unichain case there is no one-to-one correspondence between policy iteration and linear prgramming. Furthermore, that there is no one-to-one correspondence between the feasible solutions of the dual program and stationary policies: the optimal solution x^* corresponds to the two deterministic policies.

6.2.4 Value iteration

In section 5.9 we presented an algorithm for value iteration under the assumption that the value vector is constant and the Markov chains $P(f), \ f^\infty \in C(D)$, are aperiodic. The last part of this assumption is not a serious restriction: by a data tranformation the original model can be transformed into a model in which every Markov chain $P(f), \ f^\infty \in C(D)$ is aperiodic and has the same average reward as the original Markov chain. In case of unichain models no better algorithm than Algorithm 5.8 is known.

6.2.5 Modified policy iteration

We discuss the modified policy iteration for unichain MDPs under the same strong aperiodicity Assumption 6.2 as in the irreducible case. We also use the same algorithm (Algorithm 6.3), but for notation convenience we take in each iteration the same k. However, the proof of its correctness is more complicated. For the unichain case Lemma 6.3 no longer holds and the constant δ, defined in (6.11), may be zero, so Lemma 6.5 can no longer be used. Notice that Lemma 6.1, Lemma 6.2 and relation (6.10) hold also in the unichain case.

First, we will derive a similar lemma as Lemma 6.3 (Lemma 6.7), which enables us to show that $span(x^n) = max\, x_i^n - min\, x_i^n$ is bounded. Next, it is shown that the boundedness of $span(x^n)$ implies that l_n converges to ϕ. Finally we show that there exists a subsequence of $\{u_n\}$ which converges to ϕ. Define

$$\gamma = min_{i,j\in S}\, min_{h_1,h_2,\ldots,h_{N-1}} \sum_k min\Big\{\{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{ik}, \{P(h_1)P(h_2)\cdots P(h_{N-1})\}_{jk}\Big\}. \tag{6.12}$$

Then, the unichain condition and the strong aperiodicity assumption yield the following result, which states that any two states i and j have a common successor after $N-1$ transitions.

Lemma 6.7
$\eta > 0$.

Proof
Let $h_1, h_2, \ldots, h_{N-1}$ be an arbitrary sequence of policies and define $S(i,n)$ for $n = 0, 1, , \ldots$ and $i \in S$ as in the proof of Lemma 6.3. Clearly $S(i,n) \subseteq S(i,n+1)$, and if $S(i,n) = S(i,n+1)$, then $S(i,n)$ is closed under $P(h_{n+1})$. Now it has to be shown that $S(i,N-1) \cap S(j,N-1) \neq \emptyset$ for all pairs $i, j \in S$.

Suppose $S(i,N-1) \cap S(j,N-1) = \emptyset$ for some $i, j \in S$. Then $S(i,N-1)$ and $S(j,N-1)$ are both proper subsets of S, so there exists $0 \leq m, n \leq N-2$ such that $S(i,m) = S(i,m+1)$ and $S(j,n) = S(j,n+1)$. But this implies that $S(i,m)$ is closed under $P(h_{m+1})$ and that $S(j,n)$ is closed under $P(h_{n+1})$, and since $S(i,N-1) \cap S(j,N-1) = \emptyset$, $S(i,m) \cap S(j,n)$ is also empty. Let f^∞ be the policy with $f(s) = h_{m+1},\ s \in S(i,m)$ and $f(s) = h_{n+1},\ s \in S(j,n)$ (outside $S(i,m) \cup S(j,n)$ choose $f(s)$ arbitrary). Then, $P(f)$ has two disjunct, nonempty closed subsets, namely $S(i,m)$ and $S(j,n)$, which contradict the unichain condition. □

Lemma 6.8
For all $x \in \mathbb{R}^N$ and all decision rules $h_1, h_2, \ldots, h_{N-1}$, we have $span(Qx) \leq (1-\eta)span(x)$, with $Q = P(h_1)P(h_2)\cdots P(h_{N-1})$.

Proof

Let i and j such that $span(Qx) = (Qx)_i - (Qx)_j$. Then,

$$\begin{aligned} span(Qx) &= \textstyle\sum_k \{Q_{ik} - Q_{jk}\}x_k \\ &= \textstyle\sum_k \big\{Q_{ik} - min\{Q_{ik}, Q_{jk}\}\big\}x_k - \sum_k \big\{Q_{jk} - min\{Q_{ik}, Q_{jk}\}\big\}x_k \\ &\leq \textstyle\sum_k \big\{Q_{ik} - min\{Q_{ik}, Q_{jk}\}\big\}max_k\, x_k - \sum_k \big\{Q_{jk} - min\{Q_{ik}, Q_{jk}\}\big\}min_k\, x_k \\ &= span(x) - \textstyle\sum_k\, min\{Q_{ik}, Q_{jk}\}\big\}\{max_k\, x_k - min_k\, x_k\} \leq (1-\eta)span(x). \end{aligned}$$ □

Define K by $K = max_{i,a}\, r_i(a) - min_{i,a}\, r_i(a)$. Then, $span\big(r(f)\big) \leq K$ for all $f^\infty \in C(D)$.

Lemma 6.9

$span(T_{h_1}T_{h_2}\cdots T_{h_{N-1}}x) \leq (N-1)K + (1-\eta)span(x)$ for all $x \in \mathbb{R}^N$ and all decision rules $h_1, h_2, \ldots, h_{N-1}$.

Proof

Since $span(y+z) \leq span(y) + span(z)$ for all y, z and $span\{P(f)y\} \leq span(y)$ for all decision rules f and all y, we obtain

$$\begin{aligned} &span(T_{h_1}T_{h_2}\cdots T_{h_{N-1}}x) = span\Big\{r(h_1) + P(h_1)r(h_2) + \cdots + \\ &\qquad\qquad P(h_1)P(h_2)\cdots P(h_{N-2})r(h_{N-1}) + P(h_1)P(h_2)\cdots P(h_{N-1})x\Big\} \\ &\leq span\{r(h_1)\} + span\{r(h_2)\} + \cdots + span\{r(h_{N-1})\} + span\{P(h_1)P(h_2)\cdots P(h_{N-1})x\} \\ &\leq (N-1)K + span\{P(h_1)P(h_2)\cdots P(h_{N-1})x\} \leq (N-1)K + (1-\eta)span(x), \end{aligned}$$

the last inequality by Lemma 6.8. □

In order to prove that $span(x^n)$ is bounded, we introduce the following notation:

$$w^{nk+p} = T^p_{f_n}\, x^n,\ n = 0, 1, \ldots \text{ and } p = 0, 1, \ldots, k-1.$$

Then, $w^{nk} = T^0_{f_n}\, x^n = x^n$, and consequently, $w^{nk+p} = T^p_{f_n}\, w^{nk},\ n = 0, 1, \ldots;\ p = 0, 1, \ldots, k-1$.

Theorem 6.10

$span(x^n) \leq \frac{1}{\eta}(N-1)K + span(x^0)$.

Proof

It follows from Lemma 6.9 that for all $l = 0, 1, \ldots$ and all $q = 0, 1, \ldots, N-2$, we have

$$\begin{aligned} span\{w^{l(N-1)+q}\} &\leq (N-1)K + (1-\eta)span\{w^{(l-1)(N-1)+q}\} \\ &\leq (N-1)K + (1-\eta)(N-1)K + (1-\eta)span\{w^{(l-2)(N-1)+q}\} \leq \cdots \\ &\leq (N-1)K + (1-\eta)(N-1)K + \cdots + (1-\eta)^{l-1}(N-1)K + (1-\eta)^l span\{w^q\}. \end{aligned}$$

Furthermore, it follows from the proof of Lemma 6.9 that

$$span\{w^q\} = T^q_{f_0}\, x^0 \leq qK + span(x^0) \leq (N-1)K + span(x^0) \text{ for } q = 0, 1, \ldots, N-2.$$

Hence,

$$span\{w^{l(N-1)+q}\} \leq \textstyle\sum_{j=0}^{l} (1-\eta)^j(N-1)K + (1-\eta)^l span(x^0) \leq \frac{1}{\eta}(N-1)K + span(x^0),$$

implying that $span(w^n) \leq \frac{1}{\eta}(N-1)K + span(x^0)$ for all $n = 0, 1, \ldots$. Since $w^{nk} = x^n$ for all $n \geq 0$, the theorem is proven. □

Before we can prove that $lim_{n\to\infty} l_n = \phi$, we first have to derive some other results.

$$\begin{aligned} x^{n+1} - x^n &= L^k_{f_n} x^n - x^n \\ &= r(f_n) + P(f_n)r(f_n) + \cdots + P^{k-1}(f_n)r(f_n) + P^k(f_n)x^n - x^n \\ &= \{I + P(f_n) + \cdots + P^{k-1}(f_n)\}\{r(f_n) + P(f_n)x^n - x^n\} \\ &= \{I + P(f_n) + \cdots + P^{k-1}(f_n)\}\{T_{f_n}x^n - x^n\} \\ &= \{I + P(f_n) + \cdots + P^{k-1}(f_n)\}\{Tx^n - x^n\} \\ &= \{I + P(f_n) + \cdots + P^{k-1}(f_n)\}g^n. \end{aligned}$$

So, we obtain for $n = 0, 1, \ldots$ and $m = 1, 2, \ldots$

$$x^{n+m} - x^n = \sum_{l=n}^{n+m-1} \{I + P(f_l) + \cdots + P^{k-1}(f_l)\}g^l. \tag{6.13}$$

Consider the iterates $x^n, x^{n+1}, \ldots, x^{n+m-1}$. Since $u_l \geq \phi$ for all l, there has to be a state $j_0 \in S$ with $g^l_{j_o} \geq \phi$ for at least $\frac{m}{N}$ of the m indices $l = n, n+1, \ldots, n+m-1$. Using (6.13), where $x^{n+m} - x^n$ is expressed as a sum of km terms and using the property that $g^l \geq l_l \cdot e \geq l_n \cdot e$ for $l = n, n+1, \ldots, n+m-1$, we obtain

$$x^{n+m} - x^n \geq \Big\{\frac{m}{N}\phi + \big\{km - \frac{m}{N}\big\}l_n\Big\} \cdot e = \big\{km \cdot l_n + \frac{m}{N}(\phi - l_n)\big\} \cdot e. \tag{6.14}$$

From (6.10) it follows that $g^{n+m} \geq P^{k_{n+m-1}}(f_{n+m-1})P^{k_{n+m-2}}(f_{n+m-2}) \cdots P^{k_n}(f_n)g^n$. Hence, by the strong aperiodicity condition and the proof technique of Lemma 6.4, we have

$$g^{n+m}_i \geq \alpha^{km} g^n_i + (1 - \alpha^{km})l_n.$$

For $q = 0, 1, \ldots, k-1$, we obtain

$$\begin{aligned} g^{n+m}_i &\geq \alpha^{km-q}\{P^q(f_n)g^n\}_i + (1 - \alpha^{km-q})\, min_j \{P^q(f_n)\}_j \\ &\geq \alpha^{km-q}\{P^q(f_n)g^n\}_i + (1 - \alpha^{km-q})l_n \;\geq\; \alpha^{km}\{P^q(f_n)g^n\}_i + (1 - \alpha^{km})l_n, \;\; i \in S. \end{aligned}$$

Let $l_* = lim_{n\to\infty} l_n$ and let $i_0 \in S$ satisfy $g^{n+m}_{i_0} = l_{n+m}$. Then, for $q = 0, 1, \ldots, k-1$,

$$\begin{aligned} \{P^q(f_n)g^n\}_{i_0} &\leq \alpha^{-km}\{g^{n+m}_{i_0} - (1 - \alpha^{km})l_n\} \;=\; l_n + \alpha^{-km}\{g^{n+m}_{i_0} - l_n\} \\ &= l_n + \alpha^{-km}\{l_{n+m} - l_n\} \;\leq\; l_n + \alpha^{-km}\{l_* - l_n\}. \end{aligned}$$

Hence, $\big\{\{I + P(f_n) + \cdots + P^{k-1}(f_n)\}g^n\big\}_{i_0} \leq k \cdot l_n + k \cdot \alpha^{-km}\{l_* - l_n\}$ for all n.

So, because the sequence l_n is nondecreasing, we have for $l = n, n+1, \ldots, n+m+1$

$$\begin{aligned} \big\{\{I + P(f_l) + \cdots + P^{k-1}(f_l)\}g^l\big\}_{i_0} &\leq k \cdot l_l + k \cdot \alpha^{-km}\{l_* - l_l\} \\ &= k \cdot \alpha^{-km} l_* + k \cdot l_l(1 - \alpha^{-km}) \\ &\leq k \cdot \alpha^{-km} l_* + k \cdot l_n(1 - \alpha^{-km}), \end{aligned}$$

the last inequality because $\alpha < 1$ implies $1 - \alpha^{-km} < 0$. We obtain

$$(x^{n+m} - x^n)_{i_0} = \Big\{\sum_{l=n}^{n+m-1} \{I + P(f_l) + \cdots + P^{k-1}(f_l)\}g^l\Big\}_{i_0} \leq km \cdot l_n + km \cdot \alpha^{-km}(l_* - l_n). \tag{6.15}$$

It follows from (6.13) and (6.15) that

$$\begin{aligned} span(x^{n+m} - x^n) &\geq \{km \cdot l_n + \tfrac{m}{N}(\phi - l_n)\} - \{km \cdot l_n + km \cdot \alpha^{-km}(l_* - l_n)\} \\ &= \tfrac{m}{N}(\phi - l_n) - km \cdot \alpha^{-km}(l_* - l_n). \end{aligned} \tag{6.16}$$

Theorem 6.11
$l_n \uparrow \phi$.

Proof
From Lemma 6.2 and Lemma 6.1 it follows that $l_n \uparrow$ and $l_n \leq \phi$ for all n. So, $l_* = lim_{n\to\infty} l_n \leq \phi$. Suppose that $l_* < \phi$. Since, by Theorem 6.10, $span(x^n)$ is bounded, there exists a K_1 such that $span(x^n) \leq K_1$ for all n. Choose m such that $\frac{m}{N}(\phi - l_n) \geq 2K_1 + K_2$, where K_2 is some positive constant. Next, choose n such that $km \cdot \alpha^{-km}(l_* - l_n) < K_2$. Then, it follows from (6.16) that $span(x^{n+m} - x^n) > (2K_1 + K_2) - K_2 = 2K_1$. Since $span(x) \geq span(x - y) - span(y)$ for every x and y (see Exercise 6.1), we obtain

$$K_1 \geq span(x^{n+m}) \geq span(x^{n+m} - x^n) - span(x^n) > 2K_1 - K_1 = K_1,$$

implying a contradiction. □

We now know that l_n converges to ϕ and, by Lemma 6.1, that f_n^∞ is ε-optimal for n sufficiently large. In order to be able to recognize that n is sufficiently large one needs the following result.

Theorem 6.12
ϕ is a limit point of the sequence $\{u_n\}$.

Proof
We know from Lemma 6.1 that $u_n \geq \phi$ for all n. Furthermore, it follows from the boundedness of $span(x^n)$ that also $span(Tx^n - x^n) = span(g^n) = u_n - l_n$ is bounded. Hence, also $\{u_n\}$ is bounded. Let u_* be the smallest limit point of $\{u_n\}$ and suppose that $u_* > \phi$. Then one may construct, similar as in the proof of Theorem 6.11 where we supposed that $l_* < \phi$ (we also need a similar expression as (6.16)), a contradiction. Hence, ϕ is the smallest limit point of the sequence $\{u_n\}$. □

Finally, we show that $span(g^n)$ converges to zero geometrically fast. Since $span(x^n)$ is bounded, also $x^n - x_N^n \cdot e$ is bounded. Furthermore, $\phi \cdot e$ is the limit of $\{g^n\}$. Because there are only a finite number of policies, there exists a subsequence of $\{x^n\}$ and $\{g^n\}$ with $g^{n_m} \to \phi \cdot e$, $f_{n_m}^\infty = f$ and $x^{n_m} - x_N^{n_m} \cdot e \to x$ for some $f^\infty \in C(D)$ and some $x \in \mathbb{R}^N$.

Then, for all m, $max_{g^\infty \in C(D)} T_g\, x^{n_m} - x^{n_m} = T_{f_{n_m}}\, x^{n_m} - x^{n_m} = T_f\, x^{n_m} - x^{n_m} = g^{n_m}$. Letting m tends to infinity yields

$$max_{g^\infty \in C(D)} T_g\, x - x = T_f\, x - x = \phi \cdot e,$$

where it is used that $T_g\, x^{n_m} - x^{n_m} = T_g\, \{x^{n_m} - x_N^{n_m} \cdot e\} - \{x^{n_m} - x_N^{n_m} \cdot e\}$.

Lemma 6.10

If $span(x^n - x) \leq \varepsilon$ *and* $T_{f_n}\, x = x + \phi \cdot e$, *then* $span(x^{n+1} - x) \leq \varepsilon$ *and* $T_{f_{n+1}}\, x \geq x + \phi \cdot e - \varepsilon \cdot e$.

Proof

$x^{n+1} = T_{f_n}\, x^n = T_{f_n}\, x + P^k(f_n)(x^n - x) = x + k \cdot \phi \cdot e + P^k(f_n)(x^n - x)$, implying that

$$span(x^{n+1} - x) = span(P^k(f_n)(x^n - x) \leq span(x^n - x) \leq \varepsilon.$$

Furthermore,

$$\begin{aligned} T_{f_{n+1}}\, x - x &= T_{f_{n+1}}\, x^{n+1} + P(f_{n+1})(x - x^{n+1}) - x \\ &\geq T_f\, x^{n+1} + P(f_{n+1})(x - x^{n+1}) - x \\ &= T_f\, x - x + P(f)(x^{n+1} - x) - P(f_{n+1})(x^{n+1} - x) \\ &\geq \phi \cdot e + min_i\, (x^{n+1} - x)_i \cdot e - max_i\, (x^{n+1} - x)_i \cdot e \\ &= \phi \cdot e - span(x^{n+1} - x) \cdot e \geq \phi \cdot e - \varepsilon \cdot e. \end{aligned}$$

□

Remark

Since $C(D)$ has a finite number of policies g^∞, there is also only a finite number of vectors $T_g\, x - x$ and at least one of them, namely $T_f\, x - x$, equals $\phi \cdot e$. Hence, there exists an $\varepsilon > 0$ such that $T_g\, x - x \geq \phi \cdot e - \varepsilon \cdot e$ implies $T_g\, x - x = \phi \cdot e$. If $\varepsilon > 0$ is taken in this way, Lemma 6.10 gives the following result.

Corollary 6.2

If $span(x^n - x) \leq \varepsilon$ *and* $T_{f_n}\, x = x + \phi \cdot e$, *then* $span(x^{n+1} - x) \leq \varepsilon$ *and* $T_{f_{n+1}}\, x = x + \phi \cdot e - \varepsilon \cdot e$, *where* $\varepsilon > 0$ *is taken as in the above remark.*

$x^{n_m} - x_N^{n_m} \cdot e - x \to 0$ if $m \to \infty$. Therefore, also $span(x^{n_m} - x) = span(x^{n_m} - x_N^{n_m} \cdot e - x) \to 0$ if $m \to \infty$. Furthermore, since $f_{n_m}^\infty = f$ for all m, $T_{f_{n_m}}\, x - x = \phi \cdot e$ for all m. Hence, there exists a number n such that $span(x^n - x) \leq \varepsilon$ and $T_{f_n}\, x - x = \phi \cdot e$. Then, by Corollary 6.2,

$$span(x^{n+m} - x) \leq \varepsilon \text{ and } T_{f_{n+m}}\, x - x = \phi \cdot e \text{ for all } m = 0, 1, \ldots.$$

Furthermore,

$$\begin{aligned} x^{n+m} &= T_{f_{n+m-1}}^k T_{f_{n+m-2}}^k \cdots T_{f_n}^k\, x^n \\ &= T_{f_{n+m-1}}^k T_{f_{n+m-2}}^k \cdots T_{f_n}^k\, x + P^k(f_{n+m-1}) P^k(f_{n+m-2}) \cdots P^k(f_n)(x^n - x) \\ &= x + mk\, \phi \cdot e + P^k(f_{n+m-1}) P^k(f_{n+m-2}) \cdots P^k(f_n)(x^n - x). \end{aligned}$$

Hence, $span(x^{n+m} - x) = span\{P^k(f_{n+m-1}) P^k(f_{n+m-2}) \cdots P^k(f_n)(x^n - x)\}$, and by Lemma 6.8 $span(x^{n+m} - x)$ decreases exponentially fast to zero as $m \to \infty$. Then, also g^{n+m} converges to $\phi \cdot e$ exponentially fast, since

$$\begin{aligned} g^{n+m} &= T_{f_{n+m}}\, x^{n+m} - x^{n+m} = T_{f_{n+m}}\, x^{n+m} - x^{n+m} - T_{f_{n+m}}\, x + x + \phi \cdot e \\ &= \{P(f_{n+m}) - I\}(x^{n+m} - x) + \phi \cdot e. \end{aligned}$$

Therefore, the convergence of the modified policy iteration method is exponentially fast.

6.3 Communicating case

In this section we make the following assumption.

Assumption 6.4
For every $i, j \in S$ there exists a policy $f^\infty \in C(D)$, which may depend on i and j, such that in the Markov chain $P(f)$ state j is accessible from state i.

Clearly, this assumption is equivalent to the property that every *completely mixed stationary policy* π^∞, i.e. $\pi_{ia} > 0$ for every $(i, a) \in S \times A$, is irreducible. We have seen in section 5.2.3 that checking the communicating property can be done in polynomial time (polynomial in $M = \sum_{i \in S} |A(i)|$). In this case, policies with two or more recurrent sets are possible, but there is an optimal policy which has only one recurrent set. Hence, the value vector ϕ has identical components and there exists a unichain optimal policy.

6.3.1 Optimality equation

In the communicating case the value vector ϕ is a constant vector. Hence, by Theorem 5.11, ϕ is the unique x-part in the optimality equation (6.1). The next example shows that in the communicating case the property that the y-vector is unique up to a constant does not hold.

Example 6.6
$S = \{1, 2\},\ A(1) = A(2) = \{1, 2\}.\ r_1(1) = 1,\ r_1(2) = 0;\ r_2(1) = 1,\ r_2(2) = 0.$
$p_{11}(1) = 1,\ p_{12}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1;\ p_{21}(1) = 0,\ p_{22}(1) = 1;\ p_{21}(2) = 1,\ p_{22}(2) = 0.$
This is a multichain, but communicating model. The optimality equation (6.1) becomes:

$$x + y_1 = max\{1 + y_1, 0 + y_2\};\ x + y_2 = max\{1 + y_2, 0 + y_1\}.$$

Two different solutions are: $x = 1,\ y_1 = 0,\ y_2 = 1$ and $x = 1,\ y_1 = 1,\ y_2 = 0$.
The difference between the y-vectors is the non-constant vector $(-1, 1)$.

6.3.2 Policy iteration

Since there exists a unichain optimal policy one might conjecture that we can solve such a problem using Algorithm 6.4. The following example shows that this need not happen, even when we start with a unichain policy.

Example 6.7
$S = \{1, 2, 3\},\ A(1) = \{1, 2\},\ A(2) = \{1, 2, 3\},\ A(3) = \{1, 2\}.$
$r_1(1) = 0,\ r_1(2) = 2;\ r_2(1) = 1,\ r_2(2) = 1,\ r_2(3) = 3;\ r_3(1) = 2;\ r_3(2) = 4.$
$p_{12}(1) = p_{11}(2) = p_{23}(1) = p_{21}(2) = p_{22}(3) = p_{32}(1) = p_{33}(2) = 1$ (other transitions are 0).
This is a multichain and communicating model.
Algorithm 6.4 with starting policy $f(1) = 2,\ f(2) = 2,\ f(3) = 1$ gives:

Iteration 1:

Consider the system $\begin{cases} x + y_1 - y_1 = 2 \\ x + y_2 - y_1 = 1 \\ x + y_3 - y_2 = 2 \\ \qquad\quad y_1 = 0 \end{cases} \rightarrow x = \phi(f^\infty) = 2,\ y_1 = 0,\ y_2 = y_3 = -1.$

$B(1,f) = \emptyset,\ B(2,f) = \{3\},\ B(3,f) = \{2\}.$

$g(1) = 2,\ g(2) = 3,\ g(3) = 2.\ f(1) = 2,\ f(2) = 3,\ f(3) = 2.$

Iteration 2:

Consider the system $\begin{cases} x + y_1 - y_1 = 2 \\ x + y_2 - y_2 = 3 \\ x + y_3 - y_3 = 4 \\ \qquad\quad y_1 = 0 \end{cases} \rightarrow$ inconsistent system (multichain policy).

Below we state the following modification of the multichain policy iteration algorithm (Algorithm 5.6), which exploits the communication structure by finding a 'unichain improvement' which indicates whether or not the current policy is known to be unichain.

Algorithm 6.6

Determination of an average optimal policy by policy iteration (communicating case)

1. a. Select $f^\infty \in C(D)$

   b. If $P(f^\infty)$ is unichain, set unichain = 1;

   otherwise, set unichain = 0.

2. a. If *unichain* = 0, then go to step 2b;

   otherwise, go to step 2c.

   b. (i) Determine $\phi(f^\infty)$ and $y = u^0(f)$ as unique (x,y)-part in a solution of the system

$$\begin{cases} \{I - P(f)\}x & & & = & 0 \\ x + \{I - P(f)\}y & & = & r(f) \\ y + \{I - P(f)\}z & = & 0 \end{cases}$$

   (ii) Go to step 3.

   c. (i) Determine $\phi(f^\infty)$ and $y = u^0(f)$ as unique (x,y)-part in a solution of the system

$$\begin{cases} x \cdot e + \{I - P(f)\}y & = & r(f) \\ y + \{I - P(f)\}z & = & 0 \end{cases}$$

   (ii) Go to step 3.

3. a. If $\phi(f^\infty)$ is constant, then go to step 3c;

   otherwise, go to step 3b.

   b. (i) Let $S_0 = \{i \in S \mid \phi_i(f^\infty) = max_k\, \phi_k(f^\infty)\}$; $g(i) = f(i),\ i \in S$; $T = S \backslash S_0$; $W = S_0$.

   (ii) If $T = \emptyset$, then go to step 3b (v).

   (iii) Obtain $j \in T$ and $a_j \in A(j)$ for which $\sum_{k \in W}\, p_{jk}(a_j) > 0$.

   (iv) Set $T := T \backslash \{j\},\ W := W \cup \{j\}$ and $g(j) = a_j$; return to step 3b (ii).

   (v) $unichain = 1,\ f := g$ and return to step 2.

   c. (i) $B(i,f) = \{a \in A(i) \mid r_i(a) + \sum_j\, p_{ij}(a) u_j^0(f) > \phi(f^\infty) + u_i^0(f)\}$.

   (ii) If $B(i,f) = \emptyset$ for every $i \in S$, then f^∞ is an average optimal policy (STOP).

   Otherwise: (a) take g such that

$$r_i(g) + \sum_j p_{ij}(g) u_j^0(f) = max_a \{r_i(a) + \sum_j p_{ij}(a) u_j^0(f)\},\ i \in S;$$

   (b) $f := g,\ unichain = 0$ and return to step 2.

Example 6.7 (continued)

We apply Algorithm 6.6 to the model of Example 6.7, starting with $f(1) = f(2) = 2,\ f(3) = 1$.

Iteration 1:

step 1b: $unichain = 1$.

step 2c: $\phi(f^\infty) = 2;\ u^0(f) = (0, -1, -1)$.

step 3c: $B(1,f) = \emptyset,\ B(2,f) = \{3\},\ B(3,f) = \{2\}$. $g(1) = 2,\ g(2) = 3,\ g(3) = 2$.

$f(1) = 2,\ f(2) = 3,\ f(3) = 2;\ unichain = 0$.

Iteration 2:

step 2b: $\phi(f^\infty) = (2, 3, 4);\ u^0(f) = (0, 0, 0)$.

step 3b: $S_0 = \{3\},\ g(1) = 2,\ g(2) = 3,\ g(3) = 2;\ T = \{1, 2\},\ W = \{3\}$.

$j = 2,\ a_j = 1;\ T = \{1\},\ W = \{2, 3\},\ g(2) = 1$.

$j = 1,\ a_j = 1;\ T = \emptyset,\ W = \{1, 2, 3\},\ g(1) = 1$.

$unichain = 1;\ f(1) = f(2) = 1,\ f(3) = 2$.

Iteration 3:

step 2c: $\phi(f^\infty) = 4;\ u^0(f) = (-7, -3, 0)$.

step 3c: $B(1,f) = B(2,f) = B(3,f) = \emptyset : f(1) = f(2) = 1,\ f(3) = 2$ is an optimal policy.

Theorem 6.13

Algorithm 6.6 terminates in a finite number of iterations with an optimal policy.

Proof

Case 1: $\phi(f^\infty)$ is not constant: step 3b is executed.

In this step a unichain policy g^∞ is found with $\phi(g^\infty) > \phi(f^\infty)$, namely:

During step 2b we have $S = T \cup W$ and $T \cap W = \emptyset$. At the start of this step $T \neq \emptyset$ (otherwise $\phi(f^\infty)$ is constant). The communicating assumption guarantees that there exists at least one

pair of states $k \in W$ and $j \in T$ with $a_j \in A(j)$ where $p_{jk}(a_j) > 0$. Hence, after $|T|$ subiterations of step 3b the set T is empty and for the policy g^∞ the average reward $\phi(g^\infty)$ is constant. So, $\phi_i(g^\infty) > \phi_i(f^\infty),\ i \notin S_0$ and this case can occur in only a finite number of iterations. Note that g^∞ is also unichain except in the case where S_0 consists of more than one recurrent class (this last situation allows *unichain* $= 1$ because we only use in case *unichain* $= 1$ that the average reward is constant).

Case 2: $\phi(f^\infty)$ is constant: step 3c is executed.

In this case one step of Algorithm 6.6 is the same one step in the multichain case (Algorithm 5.6), because for a constant $\phi(f^\infty)$ the action set $B(i,f)$ of (5.18) becomes

$$B(i,f) = \{a \in A(i) \mid r_i(a) + \textstyle\sum_j p_{ij}(a)u_j^0(f) > \phi(f^\infty) + u_i^0(f)\}.$$

From Theorem 5.14 it follows that if $B(i,f) = \emptyset,\ i \in S$, then f^∞ is an average optimal policy, and if $B(i,f) \neq \emptyset$ for at least one $i \in S$, then g^∞ is 'better' than f^∞.

Above we have shown that all policies are different, so the algorithm terminates, and that at termination the last policy is optimal. □

6.3.3 Linear programming

Since the value vector ϕ is constant in communicating models, we would suspect some simplification in the linear programming approach. The property that ϕ is the smallest superharmonic vector implies in this case that ϕ is the unique v-part of an optimal solution (u,v) of the linear program

$$min \left\{ x \;\middle|\; x + \sum_j \{\delta_{ij} - p_{ij}(a)\} y_j \geq r_i(a),\ i \in S,\ a \in A(i) \right\}. \tag{6.17}$$

The dual of (6.17) is:

$$max \left\{ \sum_{i,a} r_i(a) x_i(a) \;\middle|\; \begin{array}{lcl} \sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_{i,a} x_i(a) & = & 1 \\ x_i(a) \geq 0,\ i \in S,\ a \in A(i) & & \end{array} \right\}. \tag{6.18}$$

The next example shows that - in contrast with the irreducible and the unichain case - in the communicating case the optimal solution of the dual program doesn't provide an optimal policy, in general.

Example 6.7 (continued)

The dual linear program (6.18) of this model is (without the nonnegativity of the variables)

$$maximize\ 2x_1(2) + x_2(1) + x_2(2) + 3x_2(3) + 2x_3(1) + 4x_3(2)$$

subject to

$$\begin{array}{rcrcrcrcrcrcl}
 & x_1(1) & & & & - & x_2(2) & & & & & = & 0 \\
- & x_1(1) & & & x_2(1) & + & x_2(2) & & - & x_3(1) & & = & 0 \\
 & & & - & x_2(1) & & & & + & x_3(1) & & = & 0 \\
 & x_1(1) & + & x_1(2) & + & x_2(1) & + & x_2(2) & + & x_2(3) & + & x_3(1) & + & x_3(2) & = & 1
\end{array}$$

The optimal solution is: $x_1(1) = x_1(2) = x_2(1) = x_2(2) = x_2(3) = x_3(1) = 0;\ x_3(2) = 1$. The objective function value equals 4. Proceeding as if this were a unichain model, we choose arbitrary actions in the states 1 and 2. Clearly, this approach could generate a nonoptimal policy, e.g. $f(1) = 2,\ f(2) = 3$. We do have the following results.

Theorem 6.14

Let x^* be an extreme optimal solution of (6.18) and let $S_* = \{i \mid \sum_a x_i^*(a) > 0\}$. Choose any policy f_*^∞ such that $x_i^*\big(f_*(i)\big) > 0,\ i \in S_*$. Then, $\phi_j(f_*^\infty) = \phi,\ j \in S_*$.

Proof

First, we show that S_* is closed under $P(f_*)$. Suppose $p_{kl}(f_*) > 0$ for some $k \in S_*$ and $l \notin S_*$. From (6.18) it follows that $0 = \sum_a x_l^*(a) = \sum_{i,a} p_{il}(a) x_i^*(a) \geq p_{kl}(f_*) x_k^*\big(f_*(k)\big)) > 0$, implying a contradiction.

Consider the Cartesian product $S_* \times A_* = \{(i,a) \mid x_i^*(a) > 0\}$. Then,

$$\begin{cases} 0 = \sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} x_i^*(a) = \sum_{(i,a) \in S_* \times A_*} \{\delta_{ij} - p_{ij}(a)\} x_i^*(a),\ j \in S \\ 1 = \sum_{i,a} x_i^*(a) = \sum_{(i,a) \in S_* \times A_*} x_i^*(a) \end{cases} \tag{6.19}$$

Since S_* is closed under $P(f_*)$, S_* contains at least one ergodic set $S_1 \subseteq S_*$. Let z^* be the unique stationary distribution of $P(f_*)$, restricted to the states of S_1, i.e.

$$\begin{cases} 0 = \sum_{i \in S_1} \{\delta_{ij} - p_{ij}(f_*)\} z_i^*,\ j \in S \\ 1 = \sum_{i \in S_1} z_i^* \end{cases} \tag{6.20}$$

Let $S_1 \times A_1$ be the Cartesian product $\{(i,a) \mid i \in S_1,\ a = f_*(i)\}$. Subtract (6.20) from (6.19), which yields

$$\begin{cases} 0 = \sum_{(i,a) \in (S_* \times A_*) \setminus (S_1 \times A_1)} \{\delta_{ij} - p_{ij}(a)\} x_i^*(a) + \sum_{(i,a) \in S_1 \times A_1} \{\delta_{ij} - p_{ij}(a)\} \{x_i^*(a) - z_i^*\},\ j \in S \\ 0 = \sum_{(i,a) \in (S_* \times A_*) \setminus (S_1 \times A_1)} x_i^*(a) + \sum_{(i,a) \in S_1 \times A_1} \{x_i^*(a) - z_i^*\} \end{cases} \tag{6.21}$$

Since x^* is an extreme solution, the columns corresponding to the positive variables, i.e. the columns $\{\delta_{ij} - p_{ij}(a)\},\ j \in S$ for $(i,a) \in S_* \times A_*$ are linear independent. Hence, the coefficients in (6.21) are zero. So, $(S_* \times A_*) \setminus (S_1 \times A_1) = \emptyset$ and $x_i^*(a) = z_i^*,\ (i,a) \in S_1 \times A_1$.

Because the optima of (6.17) and (6.14) are equal, we obtain

$$\begin{aligned} \phi &= \sum_{i,a} r_i(a) x_i(a) = \sum_{(i,a) \in S_* \times A_*} r_i(a) x_i(a) = \sum_{(i,a) \in S_1 \times A_1} r_i(a) x_i(a) \\ &= \sum_{i \in S_1} r_i(f_*) x_i\big(f_*(i)\big) = \sum_{i \in S_*} r_i(f_*) z_i^* = \sum_{i \in S_*} p_{ji}^* r_i(f_*) = \sum_i p_{ji}^* r_i(f_*) = \phi_j(f_*^\infty). \end{aligned}$$ □

Theorem 6.15

An MDP is communicating if and only if for every $b \in \mathbb{R}^N$ such that $\sum_i b_i = 0$ there exists a $y \in \mathbb{R}^{|S \times A|}$ such that $y_i(a) \geq 0$ for $(i,a) \in S \times A$ and $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i(a) = b_j,\ j \in S$.

Proof

$\Rightarrow$ Consider a communicating MDP and let π^∞ be a completely mixed stationary policy. Then, $P(\pi)$ is an irreducible Markov chain. Let x be the (strictly positive) stationary distribution of $P(\pi)$ and let $Z(\pi) = \{I - P(\pi) + P^*(\pi)\}$ be the fundamental matrix of $P(\pi)$. Choose any row vector $b \in \mathbb{R}^N$ such that $\sum_i b_i = 0$. Define $d \in \mathbb{R}^N$ by $d = bZ(\pi) + c \cdot x$ with $c \geq 0$ sufficiently large to assure $d \geq 0$. Take $y_i(a) = d_i \pi_i(a),\ (i,a) \in S \times A$. Then, $y_i(a) \geq 0,\ (i,a) \in S \times A$. Notice that

$$\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i(a) = b_j,\ j \in S \ \Leftrightarrow\ \sum_i \{\delta_{ij} - p_{ij}(\pi)\} d_i = b_j,\ j \in S \ \Leftrightarrow$$

$$d\{I - P(\pi)\} = b \ \Leftrightarrow\ \{bZ(\pi) + c \cdot x\}\{I - P(\pi)\} = b \ \Leftrightarrow\ bZ(\pi)\{I - P(\pi)\} = b \ \Leftrightarrow$$

$$b\{I - P^*(\pi)\} = b \ \Leftrightarrow\ bP^*(\pi) = 0.$$

Since $P^*(\pi)$ has identical rows, we obtain $\{b^*(\pi)\}_j = \sum_i b_i p^*_{ij}(\pi) = p^*_{jj} \sum_i b_i = 0$.

$\Leftarrow$ Assume that for every $b \in \mathbb{R}^N$ such that $\sum_i b_i = 0$ there exists a $y \in \mathbb{R}^{|S \times A|}$ such that $y_i(a) \geq 0$ for $(i,a) \in S \times A$ and $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i(a) = b_j,\ j \in S$. Suppose that the MDP is not communicating. Then, there exists a pair of states (k,l) such that $\{P^t(f)\}_{kl} = 0$ for all $f^\infty \in C(D)$ and all $t \geq 1$.

Define $S_l = \{i \in S \mid \{P^t(f)\}_{il} > 0$ for some $f^\infty \in C(D)$ and some $t \geq 1\}$. Suppose that $S_l = \emptyset$. Then, for any b such that $b_l < 0,\ \sum_i b_i = 0$ with corresponding y such that $y_i(a) \geq 0$ for $(i,a) \in S \times A$ and $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i(a) = b_j,\ j \in S$, we have $0 > b_l = \sum_{i,a} \{\delta_{il} - p_{il}(a)\} y_i(a) = \sum_{i,a} \delta_{il} y_i(a) = \sum_a y_l(a)$: contradiction. Hence, $S_l \neq \emptyset$. Take any b such that $b_j < 0,\ j \in S_l,\ b_j > 0,\ j \notin S_l$ and $\sum_i b_i = 0$ with correponding y, i.e. $y_i(a) \geq 0$ for $(i,a) \in S \times A$ and $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i(a) = b_j,\ j \in S$. Define $y \in \mathbb{R}^N$ by $y_i = \sum_a y_i(a),\ i \in S$ and a stationary policy π^∞ by $\pi_{ia} = \begin{cases} \frac{y_i(a)}{y_i} & \text{if } y_i > 0 \\ \text{arabitrary} & \text{if } y_i = 0 \end{cases}$

Then, $y_i(a) = y_i \pi_{ia},\ (i,a) \in S \times A$ and $y_j - \sum_i p_{ij}(\pi) y_i = b_j,\ j \in S$. Note that for $j \in S_l$, $p_{ij}(\pi) = 0$ if $i \notin S_l$. Hence, we obtain $y_j - \sum_{i \in S_l} p_{ij}(\pi) y_i = b_j,\ j \in S_l$. Write this equation as $y_j = b_j + \sum_{i \in S_l} p_{ij}(\pi) y_i,\ j \in S_l$. Summing up over $j \in S_l$, we get

$$\sum_{j \in S_l} y_j = \sum_{j \in S_l} b_j + \sum_{i \in S_l} \{\sum_{j \in S_l} p_{ij}(\pi)\} y_i \leq \sum_{j \in S_l} b_j + \sum_{i \in S_l} y_i.$$

Hence, $\sum_{j \in S_l} b_j \geq 0$, which gives the desired contradiction. □

In a unichain model, we can choose arbitrary actions in transient states because under any action the system eventually reaches the single recurrent class and achieves the maximal average reward. In a communicating model, such an approach can result in nonoptimal policies because it could keep the system outside of S_* indefinitely. Either one of the following approches solves this.

1. Search procedure

Obtain an optimal solution x^* of program (6.18). For $i \in S_* = \{i \mid \sum_a x_i^*(a) > 0\}$, take for $f_*(i)$ the action which satisfies $x_i^*\big(f_*(i)\big) > 0$. For the remaining states use the following search procedure.

While $S_* \neq S$

a. select $i \notin S_*,\ j \in S_*,\ f_*(i) \in A(i)$ satisfying $p_{ij}\big(f_*(i)\big) > 0$;

b. $S_* := S_* \cup \{i\}$.

By the communicating property this search procedure will find in each state of $S \backslash S_*$ an action which drives the system to S_* with positive probability.

2. Determination y variables

Obtain an optimal solution x^* of program (6.18). Choose $\beta \in \mathbb{R}^N$ satisfying $\beta_j > 0,\ j \in S$ and $\sum_j \beta_j = 1$. Let $b_j = \beta_j - \sum_a x_j^*(a),\ j \in S$. Then, $\sum_j b_j = \sum_j \beta_j - \sum_{j,a} x_j^*(a) = 1 - 1 = 0$. Because the model is communicating, by Theorem 6.15 there exists a y^* such that $y_i^*(a) \geq 0$ for $(i,a) \in S \times A$ and $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i^*(a) = b_j,\ j \in S$. Notice that (x^*, y^*) is an optimal solution of the dual linear program (5.24) (the feasible solution (x^*, y^*) is optimal, since in the proof of Theorem 6.14 is shown that the value of the objective function is ϕ). If $i \in S_*$, take for $f_*(i)$ the action which satisfies $x_i^*\big(f_*(i)\big) > 0$; if $i \notin S_*$, take for $f_*(i)$ an action which satisfies $y_i^*\big(f_*(i)\big) > 0$. In Theorem 5.16 is shown that f_* is an optimal policy.

Algorithm 6.7

Determination of an average optimal policy by linear programming (communicating case)

1. Determine an extreme optimal solution x^* of the linear program

$$max \left\{ \sum_{i,a} r_i(a) x_i(a) \;\middle|\; \begin{array}{lll} \sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_{i,a} x_i(a) & = & 1 \\ x_i(a) \geq 0,\ i \in S,\ a \in A(i) & & \end{array} \right\}.$$

2. Choose $f_*(i)$ such that $x_i^*\big(f_*(i)\big) > 0,\ i \in S_* := \{i \mid \sum_a x_i^*(a) > 0\}$.

3. Either go to step 4 (search procedure) or go to step 5 (determination y variables).

4. a. While $S_* \neq S$:

   (1) select $i \notin S_*,\ j \in S_*,\ f_*(i) \in A(i)$ such that $p_{ij}\big(f_*(i)\big) > 0$; (2) $S_* := S_* \cup \{i\}$.

   b. Go to step 6.

5. a. Choose $\beta \in \mathbb{R}^N$ such that $\beta_j > 0,\ j \in S$ and $\sum_j \beta_j = 1$ and let $b_j = \beta_j - \sum_a x_j^*(a),\ j \in S$.

   b. Determine y^* such that $y_i^*(a) \geq 0,\ (i,a) \in S \times A$ and $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i^*(a) = b_j,\ j \in S$.

   c. Choose $f_*(i)$ such that $y_i^*\big(f_*(i)\big) > 0,\ i \in S \backslash S_*$.

   d. Go to step 6.

6. f_*^∞ is an average optimal policy and $\phi = \sum_{i,a} r_i(a) x_i^*(a)$. (STOP).

Example 6.7 (continued)

We apply Algorithm 6.7 to the model of Example 6.7 with $\beta_j = \frac{1}{3},\ j = 1, 2, 3$.

We execute both step 4 (search procedure) as step 5 (determination y variables).

Step 1:

We have already seen that the linear program (6.18) has as optimal solution x^* satisfying:

$x_1^*(1) = x_1^*(2) = x_2^*(1) = x_2^*(2) = x_2^*(3) = x_3^*(1) = 0;\ x_3^*(2) = 1$ with objective function value 4.

Step 2:

$S_* = \{3\};\ f_*(3) = 2.$

Step 4 (search procedure):

$i = 2;\ j = 3;\ f_*(2) = 1;\ S_* = \{2, 3\}.$

$i = 1;\ j = 2;\ f_*(1) = 1;\ S_* = \{1, 2, 3\}.$

Step 5 (determination y variables):

$b_1 = \frac{1}{3},\ b_2 = \frac{1}{3},\ b_3 = -\frac{21}{3}$. The system becomes (without the nonnegativity of the variables):

$$\begin{array}{rcrcrcrcl} & y_1(1) & & & - & y_2(2) & & = & \frac{1}{3} \\ - & y_1(1) & + & y_2(1) & + & y_2(2) & - & y_3(1) & = & \frac{1}{3} \\ & & - & y_2(1) & & & + & y_3(1) & = & -\frac{2}{3} \end{array}$$

with feasible solution $y_1^*(1) = \frac{1}{3},\ y_2^*(1) = \frac{2}{3},\ y_2^*(2) = y_3^*(1) = 0$. Hence, $f_*(1) = f_*(2) = 1$.

Step 6:

The optimal policy is $f_*(1) = f_*(2) = 1$ and $f_*(3) = 2$; the value $\phi = 4$.

Remark

It turns out that Algorithm 6.7 with the search procedure can also be used in the so-called *weak unichain case*, i.e. for each optimal stationary policy f^∞ the associated Markov chain $P(f)$ is unichain (cf. Exercise 6.8).

6.3.4 Value iteration

In section 5.9 we presented an algorithm for value iteration under the assumption that the value vector is constant and the Markov chains $P(f),\ f^\infty \in C(D)$, are aperiodic. The last part of this assumption is not a serious restriction: by a data tranformation the original model can be transformed into a model in which every Markov chain $P(f),\ f^\infty \in C(D)$ is aperiodic and has the same average reward as the original Markov chain. In case of unichain models no better algorithm than Algorithm 5.8 is known.

6.3.5 Modified value iteration

Algorithm 6.3 is again used as the modified value iteration method for communicating MDPs. In Section 6.2.5 the convergence proof for the unichain case has been given in two stages. First, the unichain assumption and the strong aperiodicity assumption were used to prove that $span(x^n)$ is bounded (Theorem 6.10). In the second stage we used the boundedness of $span(x^n)$ and the property $u_n \geq \phi$ for all n to prove that $l_n \uparrow \phi$ and that ϕ is a limit point of the sequence $\{u_n\}$

(Theorems 6.11 and 6.12). From these proofs it is clear that the modified policy iteration method will convergence whenever $span(x^n)$ is bounded and the value vector ϕ is independent of the initial state (if the strong aperiodicity assumption holds), which is the case for communicating MDPs. Therefore, we have to show that the sequence $\{span(x^n)\}$ is also bounded in the communicating case.

Define: $M = max_{i,a} |r_i(a)|$; $L_n = min_i\, x_i^n$; $U_n = max_i\, x_i^n$; $\theta = min_{i,j,a} \{p_{ij}(a) \mid p_{ij}(a) > 0\}$.

Lemma 6.11

For all $n = 0, 1, \ldots,$ we have $L_{n+1} \geq L_n - k \cdot M$ and $U_{n+1} \leq U_n + k \cdot M$.

Proof

$$\begin{aligned} x^{n+1} &= T_{f_n}^k\, x^n = r(f_n) + P(f_n)r(f_n) + \cdots + P^{k-1}(f_n)r(f_n) + P^k(f_n)x^n \\ &\geq -Me - Me - \cdots - Me + P^k(f_n)x^n \geq -k \cdot Me + L_n e. \end{aligned}$$

Hence, $L_{n+1} \geq L_n - k \cdot M$. Similarly it can be shown that $U_{n+1} \leq U_n + k \cdot M$. □

Lemma 6.12

If $span(x^{n+m-1}) \geq span(x^n)$, then for all l with $n \leq l \leq n + m - 2$, $L_{l+1} - L_l \leq (2m - 3)kM$.

Proof

From Lemma 6.11 we obtain

$$\begin{aligned} span(x^{n+m-1}) &= U_{n+m-1} - L_{n+m-1} \\ &= \textstyle\sum_{j=n}^{n+m-2} \{(U_{j+1} - U_j) - (L_{j+1} - L_j)\} + U_n - L_n \\ &= \textstyle\sum_{j=n}^{n+m-2} (U_{j+1} - U_j) - \sum_{j=n,\ j\neq l}^{n+m-2} (L_{j+1} - L_j) - (L_{l+1} - L_l) + span(x^n) \\ &\leq (m-1)kM + (m-2)kM - (L_{l+1} - L_l) + span(x^{n+m-1}). \end{aligned}$$

Hence, $L_{l+1} - L_l \leq (2m - 3)kM$. □

Lemma 6.13

If $span(x^{n+m-1}) \geq span(x^n)$ and $x_i^{l+1} \leq c + L_{l+1}$ for some $i \in S$ and some $n \leq l \leq n + m - 2$, then $x_j^l \leq L_l + \lambda^{1-k}\theta^{-1}\{c + 2kM(m-1)\}$ for all $j \in S$ for which an action $a \in A(i)$ with $p_{ij}(a) > 0$ exists ($\lambda \in (0,1)$ is the constant in the strong aperiodicity assumption (5.40)).

Proof

Since

$$(Ux^l)_i = max_a \{r_i(a) + \textstyle\sum_j\, p_{ij}(a)x_j^l\} \geq -M + max_a \{\sum_j\, p_{ij}(a)x_j^l\},\ i \in S$$

and

$$x^{l+1} = T_{f_l}^k x^l = T_{f_l}^{k-1}(Tf_l x^l) = T_{f_l}^{k-1}(Ux^l),$$

we have

$$x^{l+1} \geq T_{f_l}^{k-1}\{-Me + max_f\, P(f)x^l\} \geq -kMe + P_{f_l}^{k-1}\{max_f\, P(f)x^l\}.$$

Notice that

$$\begin{aligned} P_{fl}^{k-1}\{max_f\, P(f)x^l\} &= P_{fl}^{k-1}\{max_f\,\big(P(f)x^l + L_l e - L_l e\big)\} \\ &= L_l e + P_{fl}^{k-1}\{max_f\,\big(P(f)x^l - L_l e\big)\} \\ &\geq L_l e + \lambda^{k-1}\{max_f\,\big(P(f)x^l - L_l e\big)\} \\ &= (1-\lambda^{k-1})L_l e + \lambda^{k-1}\{max_f\, P(f)x^l\}. \end{aligned}$$

Hence,

$$c + L_{l+1} \geq x_i^{l+1} \geq -kM + (1-\lambda^{k-1})L_l + \lambda^{k-1}\cdot max_a \textstyle\sum_j p_{ij}(a)x_j^l.$$

Then, by Lemma 6.12,

$$c + L_l + (2m-3)kM \geq c + L_{l+1} \geq -kM + (1-\lambda^{k-1})L_l + \lambda^{k-1}\cdot max_a \textstyle\sum_j p_{ij}(a)x_j^l,$$

or

$$max_a \textstyle\sum_j p_{ij}(a)(x_j^l - L_l) \leq \lambda^{1-k}\{c + 2(m-1)kM\}.$$

Take any $j \in S$ for which an action $a \in A(i)$ with $p_{ij}(a) > 0$ exists. Then, $p_{ij}(a) \geq \theta$ and

$$\theta(x_j^l - L_l) \leq p_{ij}(a)(x_j^l - L_l) \leq \lambda^{1-k}\{c + 2(m-1)kM\},$$

implying

$$x_j^l \leq L_l + \lambda^{1-k}\theta^{-1}\{c + 2kM(m-1)\}.$$

□

Define $c_0 = 0$ and $c_n = \lambda^{1-k}\theta^{-1}\{c_{n-1} + 2kM(N-1)\},\ n = 1, 2, \ldots, N-1$.

Lemma 6.14

If $span(x^{n+N-1}) \geq span(x^n)$, *then* $span(x^n) \leq c_{N-1}$.

Proof

Let $i \in S$ such that $x_i^{n+N-1} = L_{n+N-1}$ and define the sets $S(t),\ t = 0, 1, \ldots, N-1$ by

$$\begin{aligned} S(0) &= \{i\} \\ S(t+1) &= \{j \in S \mid \exists\ k \in S(t) \text{ and } a \in A(k) \text{ such that } p_{kj}(a) > 0\},\ t = 0, 1, \ldots, N-2. \end{aligned}$$

From $p_{jj}(a) \geq \lambda > 0,\ (j,a) \in S \times A$, it follows that $S(t) \subseteq S(t+1)$. Furthermore, it follows from the communicatingness that $S(N-1) = S$. Then, Lemma 6.13 with $c = c_0 = 0,\ m = N$ and $l = n + N - 2$ implies that $x_j^{n+N-2} - L_{n+N-2} \leq c_1$ for all $j \in S(1)$. Next, again by Lemma 6.13 with $c = c_1,\ m = N$ and $l = n + N - 3$, we obtain $x_j^{n+N-3} - L_{n+N-3} \leq c_2$ for all $j \in S(2)$. Continuing in this way, we get $x_j^n - L_n \leq c_{N-1}$ for all $j \in S(N-1) = S$. Hence, $span(x^n) = max\, x_j^n - min\, x_j^n = max\, x_j^n - L_n \leq c_{N-1}$.

□

Finally, we prove in the next theorem that the sequence $\{span(x^n)\}$ is bounded.

Theorem 6.16

$span(x^n) \leq max\{span(x^0) + 2kM(N-2),\ c_{N-1} + 2kM(N-1)\},\ n = 0, 1, \ldots.$

Proof

By Lemma 6.11, we have

$$\begin{aligned} span(x^n) &\leq span(x^{n-1}) + 2kM \leq \cdots \leq span(x^0) + 2nkM \\ &\leq span(x^0) + 2(N-2)kM,\ n = 0, 1, \ldots, N-2, \end{aligned}$$

and consequently

$$span(x^n) \leq max\{span(x^0) + 2kM(N-2),\ c_{N-1} + 2kM(N-1)\},\ n = 0, 1, \ldots, N-2.$$

Furthermore, also by Lemma 6.11, we obtain

$$\begin{aligned} span(x^{n+N-1}) &= U_{n+N-1} - L_{n+N-1} \\ &\leq (U_{n+N-2} + kM) - (L_{n+N-2} - kM) = U_{n+N-2} - L_{n+N-2} + 2kM \\ &\leq (U_{n+N-3} + kM) - (L_{n+N-3} - kM) + 2kM = U_{n+N-3} - L_{n+N-3} + 4kM \\ &\leq \cdots \leq U_n - L_n + 2kM(N-1) = span(x^n) + 2kM(N-1). \end{aligned}$$

If $span(x^{n+N-1}) \geq span(x^n)$, then by Lemma 6.14 $span(x^n) \leq c_{N-1}$, and consequently $span(x^{n+N-1}) \leq c_{N-1} + 2kM(N-1)$, implying

$$span(x^{n+N-1}) \leq max\{span(x^n),\ c_{N-1} + 2kM(N-1)\},\ n = 0, 1, \ldots. \tag{6.22}$$

For any $n \geq N-1$, we write $n = p + q(N-1)$ for some $q \geq 1$ and some $0 \leq p \leq N-2$.

Then, by (6.22),

$$\begin{aligned} span(x^n) &\leq max\{span(x^p),\ c_{N-1} + 2kM(N-1)\} \\ &\leq max\{span(x^0) + 2kM(N-2),\ c_{N-1} + 2kM(N-1)\}. \end{aligned}$$

□

The proofs for the modified policy iteration method in the unichain and the communicating case depend heavily on the strong aperiodicity assumption. One might wonder whether only aperiodicity, as in the standard value iteration method, would not suffice. The following example demonstrates one of the problems one can encounter under the weaker assumption that all Markov chains $P(f),\ f^\infty \in C(D)$, are aperiodic and unichain.

Example 6.8

$S = \{1, 2, 3, 4, 5, 6, 7\};\ A(1) = \{1\}\ A(2) = \{1\};\ A(3) = \{1, 2\};\ A(4) = \{1, 2\};\ A(5) = \{1\};$
$A(6) = \{1, 2\};\ A(7) = \{1\}$. There are 8 different policies (only in the states 3, 4 and 6 there are two choices).

The transition probabilities are (we give only the strictly positive probabilities):

$p_{12}(1) = p_{13}(1) = \frac{1}{2};\ p_{23}(1) = 1;\ p_{33}(1) = 1;\ p_{34}(2) = 1;\ p_{41}(1) = 1;\ p_{45}(2) = p_{46}(2) = \frac{1}{2};$
$p_{56}(1) = 1;\ p_{67}(1) = 1;\ p_{63}(2) = 1;\ p_{73}(1) = 1.$

It can easily be verified that all policies are unichain and aperiodic.

For the rewards we choose: $r_1(1) = 2;\ r_2(1) = 4;\ r_3(1) = 4;\ r_3(2) = 6;\ r_4(1) = 4;$
$r_4(2) = 6;\ r_5(1) = 6;\ r_6(1) = 2;\ r_6(2) = 0;\ r_7(1) = 0.$

Let f_1^∞ be the policy that takes in state 3 action 1, in state 4 action 2 and in state 6 action 1; let f_2^∞ be the policy that takes in state 3 action 2, in state 4 action 1 and in state 6 action 2;

let f_*^∞ be the policy that takes in state 3 action 2, in state 4 action 2 and in state 6 action 2. Then, it can be verfied that $\phi(f_1^\infty) = \phi(f_2^\infty) = 4$ and $\phi(f_*^\infty) = \frac{30}{7}$ (this is the optimal policy). Choose $x^0 = (1, 4, 2, 0, 0, 0, 0)$ and take $k = 2$.
Iteration 1:
$Tx^0 = T_{f_1}x^0 = (5, 6, 6, 6, 6, 2, 2)$; $l = 2$; $u = 6$. $x^1 = T_{f_1}^2 x^0 = (8, 10, 10, 10, 8, 4, 6)$.
Iteration 2:
$Tx^1 = T_{f_2}x^1 = (12, 14, 16, 12, 10, 10, 10)$; $l = 2$; $u = 6$. $x^2 = T_{f_1}^2 x^1 = (17, 20, 18, 16, 16, 16, 16)$.
Since $x^2 = x^0 + 16e$ cycling will occur between the two nonoptimal policies f_1^∞ and f_2^∞.

6.4 Bibliographic notes

In the irreducible and unichain case the solution of the optimality equation (6.1) can be exhibited as the fixed-point of an N-step contraction (cf. Federgruen, Schweitzer and Tijms ([66]).
The policy iteration algorithm 6.1 was introduced by Howard ([101]), where he demonstrated finite convergence under the irreducibility assumption. Various treatments of policy iteration in special cases can also be found in Schweitzer ([174]), Denardo ([47]) and Lassere [126]. Haviv and Puterman ([86]) discuss the communicating policy algorithm 6.6.
The pioneering work in solving undiscounted MDPs was made by De Ghellinkck ([39]) and Manne ([134]), who independently formulated the linear programs (6.3) and (6.4) for the irreducible case. The relation between the discounted and undiscounted linear programs is described in Nazareth and Kulkarni ([141]. For the unichain case we refer to Denardo and Fox ([50], Denardo ([46]), Derman ([55]) and Kallenberg ([108]). In the irreducible and unichain case also a suboptimality test can be implemented (cf. Hastings ([85]) and Lasserre ([125]).
Kallenberg ([108]) and Filar and Schultz ([75]) discuss the linear programming approach for communicating models.
Van der Wal ([202] and [203]) analyzed the modified policy iteration method for the irreducible, unichain and communicating case.

6.5 Exercises

Exercise 6.1
Show that $span(x) \geq span(x - y) - span(y)$ for every x and y.

Exercise 6.2
Let P be a unichain Markov chain with stationary distribution π and let $x \geq Px$ or $x \leq Px$. Prove that $x_i = \pi^T x$ for every recurrent state i.

Exercise 6.3

Consider the following model:

$S = \{1, 2, 3\},\ A(1) = A(2) = A(3) = \{1, 2\}.$

$r_1(1) = 1,\ r_1(2) = 0,\ r_2(1) = 2,\ r_2(2) = 1,\ r_3(1) = 1,\ r_3(2) = 2.$

$p_{11}(1) = \frac{1}{2},\ p_{12}(1) = \frac{1}{4},\ p_{13}(1) = \frac{1}{4};\ p_{11}(2) = \frac{1}{4},\ p_{12}(2) = \frac{1}{4},\ p_{13}(2) = \frac{1}{2};$

$p_{21}(1) = \frac{1}{2},\ p_{22}(1) = \frac{1}{2},\ p_{23}(1) = 0;\ p_{21}(2) = 0,\ p_{22}(2) = \frac{1}{2},\ p_{23}(2) = \frac{1}{2};$

$p_{31}(1) = \frac{1}{2},\ p_{32}(1) = 0,\ p_{33}(1) = \frac{1}{2};\ p_{31}(2) = 0,\ p_{32}(2) = \frac{1}{2},\ p_{33}(2) = \frac{1}{2}.$

a. Show that this model is unichain, communicating, but not irreducible.

b. Formulate the optimality equation (6.1).

c. Determine an optimal policy by the Policy Iteration Algorithm 6.4, starting with $f(1) = f(2) = f(3) = 1$.

Exercise 6.4

Consider the model of Exercise 6.2. Formulate the primal and dual linear program to solve this unichain model. Apply Algorithm 6.5 to determine the value and an optimal policy.

Exercise 6.5

Consider the model of Exercise 6.2. Execute three iterations of the modified policy iteration algorithm 6.3 with $k = 2$ and $x^0 = (1, 1, 1)$.

Exercise 6.6

For the standard method of value iteration (see section 5.9) the series $\{x^n - n \cdot \phi\}$ is bounded (see Lemma 5.5), even if all policies are periodic. One might wonder whether in the modified policy iteration the series $\{x^n - nk \cdot \phi\}$ is bounded.

Let $S = \{1, 2\};\ A(1) = \{1, 2\},\ A(2) = \{1\};\ r_1(1) = 4,\ r_1(2) = 3,\ r_2(1) = 0;$

$p_{11}(1) = 0,\ p_{12}(1) = 1;\ p_{11}(2) = 1,\ p_{12}(2) = 0;\ p_{21}(1) = 1,\ p_{22}(2) = 0.$

Consider the modified policy iteration algorithm with $k = 2$ and $x^0 = (0, 0)$.

Show that $x^n = (4n, 4n)$ and that $\{x^n - nk \cdot \phi\}$ is unbounded.

Exercise 6.7

Assume there is a state, say state 0, and a scalar $\alpha \in (0, 1)$ such that $p_{i0} \geq 1 - \alpha$ for all $(i, a) \in S \times A$. So, the model is unichain. Consider a new decision process with identical state and action spaces and identical rewards but with transition probabilities given by

$$\overline{p}_{ij}(a) = \begin{cases} \frac{1}{\alpha}\, p_{ij}(a) & j \neq 0; \\ \frac{1}{\alpha}\{p_{i0}(a) - (1 - \alpha)\} & j = 0. \end{cases}$$

Show that the optimality equation of the α-discounted new process is equivalent to the optimality equation (6.1).

Exercise 6.8

Consider an MDP with the *weak unichain assumption*, i.e. for each optimal stationary policy f^∞ the associated Markov chain $P(f)$ is unichain. For this model the following algorithm is proposed.

1. Determine an optimal solution x^* of the dual linear program

$$max \left\{ \sum_{i,a} r_i(a) x_i(a) \;\middle|\; \begin{array}{lcl} \sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_{i,a}\ x_i(a) & = & 1 \\ x_i(a) \geq 0,\ i \in S,\ a \in A(i) & & \end{array} \right\}.$$

2. Choose $f_*(i)$ such that $x_i^*\big(f_*(i)\big) > 0,\ i \in S_* := \{i \mid \sum_a\ x_i^*(a) > 0\}$.

3. While $S_* \neq S$:

   (1) select $i \notin S_*,\ j \in S_*,\ f_*(i) \in A(i)$ such that $p_{ij}\big(f_*(i)\big) > 0$; (2) $S_* := S_* \cup \{i\}$.

4. f_*^∞ is an average optimal policy and $\phi = \sum_{i,a} r_i(a) x_i^*(a)$. (STOP).

Prove the correctness of this algorithm.

Chapter 7

More sensitive optimality criteria

7.1 Introduction

In the two previous chapters we have considered the long-run average reward criterion. This criterion ignores transient rewards. The following examples shows why this is (sometimes) undesirable.

Example 7.1
$S = \{1, 2\};\ A(1) = \{1, 2\},\ A(2) = \{1\};\ r_1(1) = 1000,\ r_1(2) = 0,\ r_2(1) = 0.$
$p_{11}(1) = 0,\ p_{12}(1) = 1;\ p_{11}(2) = 0,\ p_{12}(1) = 1;\ p_{21}(1) = 0,\ p_{22}(1) = 1.$
This model has two deterministic policies and both policies are average optimal with average reward 0. However, the policy which chooses in state 1 the first action has a total reward of 1000 and would be prefered. The average reward criterion ignores this distinction.

We address this deficiency of the average reward criterion by more sensitive optimality criteria, the so-called n-discount optimality and the n-average optimality. We may restrict ourselves in this chapter to policies $f^\infty \in C(D)$ by the following arguments:

1. We have shown the existence of a deterministic Blackwell optimal policy.
2. We will show in this chapter (see Lemma 7.2) that a Blackwell optimal policy is n-discount optimal for any $n \geq -1$.
3. It can also be shown that n-discount optimality is equivalent to n-average optimality for all $n \geq -1$.

In section 1.2.2, the concept of n-discount optimality, for $n = -1, 0, 1, \ldots$, is defined as

$$lim_{\alpha \uparrow 1}\ (1 - \alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha\} = 0. \tag{7.1}$$

From this definition it is easy to see the following lemma.

Lemma 7.1
If a policy is n-discount optimal, then it is m-discount optimal for $m = -1, 0, \ldots, n$.

Lemma 7.2

Suppose that f_0^∞ is a Blackwell optimal policy. Then, f_0^∞ is n-discount optimal for any $n \geq -1$.

Proof

Take any $n \geq -1$. Since f_0^∞ is a Blackwell optimal policy, we have for some $0 < \alpha_0 < 1$, $v^\alpha(f_0^\infty) = v^\alpha$ for all $\alpha \in [\alpha_0, 1)$. Hence, $(1-\alpha)^{-n}\{v^\alpha(f_0^\infty) - v^\alpha\} = 0$ for all $\alpha \in [\alpha_0, 1)$. Therefore, $lim_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_0^\infty) - v^\alpha\} = 0$, i.e. f_0^∞ is n-discount optimal. □

Also in section 1.2.2, the concept of an n-average optimal policy R_*, for $n = -1, 0, 1, \ldots$, is defined as

$$liminf_{T\to\infty} \frac{1}{T}\{v^{n,T}(R_*) - v^{n,T}(R)\} \geq 0 \text{ for every policy } R, \tag{7.2}$$

where the vector $v^{n,T}(R)$ is defined by

$$v^{n,T}(R) = \begin{cases} v^T(R) & \text{for } n = -1 \\ \sum_{t=1}^T v^{n-1,t}(R) & \text{for } n = 0, 1, \ldots \end{cases} \tag{7.3}$$

So, (-1)-average optimality is the same as average optimality.

Lemma 7.3

If a policy is n-average optimal, then it is m-average optimal for $m = -1, 0, \ldots, n$.

Proof

The proof is left to the reader (see Exercise 7.1).

7.2 Equivalence between n-discount and n-average optimality

Sladky ([186]) has shown that a policy R_* is n-average optimal policy if and only if R_* is n-discount optimal. We show in this section this result only for $n = -1$ and $n = 0$ and restrict ourselves to deterministic policies. In the case of arbitrary policies the notation will be more complicated; for $n \geq 1$ the analysis is much more sophisticated.

For $n = -1$, the criteria (-1)-discount optimality and (-1)-average optimality become

$$lim_{\alpha\uparrow 1}\,(1-\alpha)\{v^\alpha(f_*^\infty) - v^\alpha\} = 0 \text{ and } \phi(f_*^\infty) \geq \phi(f^\infty) \text{ for every } f^\infty \in C(D),$$

respectively. The following theorem shows that average optimality is equivalent to (-1)-discount optimality.

Theorem 7.1

Average optimality is equivalent to (-1)-discount optimality.

Proof

In Theorem 5.8 (2) is shown that $\phi(f^\infty) = lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha(f^\infty),\ f^\infty \in C(D)$. For f_0^∞ a Blackwell optimal policy, we obtain $\phi = lim_{\alpha\uparrow 1}(1-\alpha)v^\alpha$. Let f_*^∞ be (-1)-discount optimal, then $0 = lim_{\alpha\uparrow 1}(1-\alpha)\{v^\alpha(f_*^\infty) - v^\alpha\} = \phi(f_*^\infty) - \phi$, i.e. $\phi(f_*^\infty)$ is average optimal.
Conversely, let f_*^∞ be an average optimal policy, and let f_0^∞ be Blackwell optimal. Then, we can write

$$\begin{aligned} 0 &\geq lim_{\alpha\uparrow 1}(1-\alpha)\{v^\alpha(f_*^\infty) - v^\alpha\} = lim_{\alpha\uparrow 1}(1-\alpha)\{v^\alpha(f_*^\infty) - v^\alpha(f_0^\infty)\} \\ &= \phi(f_*^\infty) - \phi(f_0^\infty) \geq 0, \end{aligned}$$

i.e. $lim_{\alpha\uparrow 1}(1-\alpha)\{v^\alpha(f_*^\infty) - v^\alpha\} = 0$: f_*^∞ is (-1)-discount optimal. □

For $n = 0$, the criteria 0-discount optimality and 0-average optimality become

$$lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha\} = 0 \text{ and } liminf_{T\to\infty}\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\{v^t(f_*) - v^t(f)\} \geq 0,\ f^\infty \in C(D),$$

respectively. The following theorem shows that 0-average optimality is equivalent to 0-discount optimality. This criterion is also called *bias optimality*.

Theorem 7.2

0-average optimality is equivalent to 0-discount optimality.

Proof

Suppose that f_*^∞ is 0-average optimal, and let f_0^∞ be a Blackwell optimal policy. Then, both f_*^∞ and f_0^∞ are average optimal policies. In Theorem 5.8 part (3) we have shown

$$v^t(f^\infty) = \textstyle\sum_{s=1}^{t} P^{s-1}(f)r(f) = t\phi(f^\infty) + u^0(f) - P^t(f)D(f)r(f),\ f^\infty \in C(D).$$

Hence,

$$\begin{aligned} 0 &\leq liminf_{T\to\infty}\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\{v^t(f_*) - v^t(f)\} \\ &= liminf_{T\to\infty}\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\big\{\{t\phi(f_*^\infty) + u^0(f_*) - P^t(f_*)D(f_*)r(f_*)\} - \\ &\qquad\qquad \{t\phi(f_0^\infty) + u^0(f_0) - P^t(f_0)D(f_0)r(f_0)\}\big\} \\ &= liminf_{T\to\infty}\big\{\tfrac{1}{2}(T+1)\{\phi(f_*^\infty) - \phi(f_0^\infty)\} + \{u^0(f_*) - u^0(f_0)\} - \\ &\qquad\qquad \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\big\{P^t(f_*)D(f_*)r(f_*) - P^t(f_0)D(f_0)r(f_0)\big\}\big\} \\ &= \{u^0(f_*) - u^0(f_0)\} + liminf_{T\to\infty}\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\big\{P^t(f_0)D(f_0)r(f_0) - P^t(f_*)D(f_*)r(f_*)\big\} \end{aligned}$$

Since $lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} P^t(f)D(f) = P^*(f)D(f) = 0$ for every $f^\infty \in C(D)$, we obtain

$0 \leq u^0(f_*) - u^0(f_0)$. Then, by the Laurent expansion, we can write

$$\begin{aligned} 0 &\geq lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha\} = lim_{\alpha\uparrow 1})\{v^\alpha(f_*^\infty) - v^\alpha(f_0^\infty)\} \\ &= lim_{\alpha\uparrow 1}(1-\alpha)^{-1}\{\phi(f_*^\infty) - \phi(f_0^\infty)\} + \{u^0(f_*) - u^0(f_0)\} = u^0(f_*) - u^0(f_0) \geq 0. \end{aligned}$$

Hence, $lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha\} = 0$, i.e. f_*^∞ is 0-discount optimal.

Conversely, suppose that $lim_{\alpha\uparrow 1}\{v^\alpha(f_*^\infty) - v^\alpha\} = 0$. Take any $f^\infty \in C(D)$. Then, by the Laurent expansion, $\phi(f_*^\infty) \geq \phi(f^\infty)$ and if $\phi_i(f_*^\infty) = \phi_i(f^\infty)$ for some $i \in S$, then $u_i^0(f_*) \geq u_i^0(f)$. Hence, we can write

$$liminf_{T\to\infty} \tfrac{1}{T}\sum_{t=1}^{T}\{v^t(f_*^\infty) - v^t(f^\infty)\} =$$

$$liminf_{T\to\infty}\Big\{\tfrac{1}{2}(T+1)\{\phi(f_*^\infty) - \phi(f^\infty)\} + \{u^0(f_*) - u^0(f_0)\} -$$

$$\tfrac{1}{T}\sum_{t=1}^{T}\left\{P^t(f_*)D(f_*)r(f_*) - P^t(f_0)D(f_0)r(f_0)\right\}\Big\} =$$

$$liminf_{T\to\infty}\Big\{\tfrac{1}{2}(T+1)\{\phi(f_*^\infty) - \phi(f^\infty)\} + \{u^0(f_*) - u^0(f_0)\}\Big\} \geq 0,$$

i.e. f_*^∞ is 0-average optimal. □

Example 7.2

$S = \{1,2,3\};\ A(1) = \{1,2\},\ A(2) = \{1\},\ A(3) = \{1\};\ r_1(1) = 1,\ r_1(2) = 2,\ r_2(1) = 1,\ r_3(1) = 0.$
$p_{11}(1) = 0,\ p_{12}(1) = 1,\ p_{13}(1) = 0;\ p_{11}(2) = p_{12}(2) = 0,\ p_{13}(2) = 1;$
$p_{21}(1) = p_{22}(1) = 0,\ p_{23}(1) = 1;\ p_{31}(1) = p_{32}(1) = 0,\ p_{33}(1) = 1.$
This model has two deterministic policies. If we look at the discounted reward, for f_1^∞ with $f_1(1) = 1$, we have $v_1^\alpha(f_1^\infty) = 1 + \alpha,\ v_2^\alpha(f_1^\infty) = 1$, and $v_3^\alpha(f_1^\infty) = 0$; for f_2^∞ with $f_2(1) = 2$, we obtain $v_1^\alpha(f_2^\infty) = 2,\ v_2^\alpha(f_1^\infty) = 1$, and $v_3^\alpha(f_1^\infty) = 0$.
Hence, $v^\alpha = (2,1,0)$ and f_2^∞ is α-discounted optimal for all discount factors $\alpha \in [0,1)$.
For f_1^∞, we have $lim_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_1^\infty) - v^\alpha\} = lim_{\alpha\uparrow 1}\ (1-\alpha)^{-n}(\alpha - 1, 0, 0)$.
For $n = -1$ and $n = 0$ this is equal to $(0,0,0)$, for $n = 1$ the limit is $(-1,0,0)$ and for $n \geq 1$, this limit is the vector $(-\infty, 0, 0)$.
Therefore, f_2^∞ is n-discount optimal for all $n = -1, 0, 1, \ldots$, and f_1^∞ is only for $n = 0$ and $n = 1$ n-discount optimal.

7.3 Stationary optimal policies and optimality equations

In this section we use the Laurent series expansion to interpret the n-discount optimality within the class of stationary policies. Furthermore, we provide a system of optimality equations which characterize a stationary n-discount optimal policy. First, we present a lemma which shows that n-discount optimality as defined in (7.1) is equivalent to

$$liminf_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha(f^\infty\} \geq 0 \text{ for all } f^\infty \in C(D). \tag{7.4}$$

Lemma 7.4

The definitions (7.1) and (7.4) for n-discount optimality are equivalent.

Proof

Assume that f_*^∞ is n-discount optimal in the sense of (7.1). Take an arbitrary policy $f^\infty \in C(D)$. Since $v^\alpha(f_*^\infty) \leq v^\alpha$, we have $(1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha(f^\infty\} \geq (1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha\},\ \alpha \in (0,1)$. Hence,

$$liminf_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha(f^\infty\} \geq liminf_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha\} = 0.$$

Conversely, suppose that f_*^∞ satisfies (7.4). Let f_0^∞ be a Blackwell optimal policy. Then, we can write

$$0 \leq liminf_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha(f_0^\infty\} = liminf_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha\} \leq 0.$$

Therefore,

$$lim_{\alpha\uparrow 1}\ (1-\alpha)^{-n}\{v^\alpha(f_*^\infty) - v^\alpha\} = 0.$$

□

Instead of the discount factor α we can also use the interest rate ρ, where the relation between the two are given by $\alpha = \frac{1}{1+\rho}$ or $\rho = \frac{1-\alpha}{\alpha}$.

Since $lim_{\alpha\uparrow 1}\,(1-\alpha)^{-n}\,v^\alpha(f^\infty) = lim_{\alpha\uparrow 1}\,(\frac{1-\alpha}{\alpha})^{-n}\,v^\alpha(f^\infty)$ and $\alpha \uparrow 1$ if and only if $\rho \downarrow 0$, the concept of n-discount optimality is equivalent to

$$liminf_{\rho\downarrow 0}\ \rho^{-n}\{v^\rho(f_*^\infty) - v^\rho(f^\infty)\} \geq 0 \text{ for all } f^\infty \in C(D).$$

In Theorem 5.10 we have shown that for $0 < \rho \leq \|D(f)\|^{-1}$, $\alpha v^\alpha(f^\infty) = \sum_{k=-1}^\infty \rho^k u^k(f)$. Hence, as function of the interst rate ρ, the total expected discounted reward is written as

$$v^\rho(f^\infty) = (1+\rho)\cdot \sum_{k=-1}^{\infty} \rho^k u^k(f). \tag{7.5}$$

Let $F_\infty = \{\{f_0^\infty \mid f_0^\infty \text{ is Blackwell optimal}\}$ and $F_n = \{f_*^\infty \mid f_*^\infty \text{ is } n\text{-discount-optimal}\}$ for $n \geq -1$. We have already seen that $F_{n+1} \subseteq F_n$ for $n = -1, 0, 1, \ldots$ and that $f_*^\infty \in F_{-1}$ if and only if $u^-(f_*) \geq u^{-1}(f),\ f^\infty \in C(D)$. In the next theorem is shown that $F_\infty = \cap_{n=-1}^\infty F_n$ and that $F_n = \{f_*^\infty \in F_{n-1} \mid u^n(f_*) \geq u^n(f),\ f^\infty \in F_{n-1}\}$ for all $n \geq 0$.

Theorem 7.3

$F_n = \{f_*^\infty \in F_{n-1} \mid u^n(f_*) \geq u^n(f)$ for all $f^\infty \in F_{n-1}\},\ n \geq 0$ and $F_\infty = \cap_{n=-1}^\infty F_n$.

Proof

We use induction on n. Let $f_*^\infty \in F_0$, i.e. $liminf_{\rho\downarrow 0}\,\{v^\rho(f_*^\infty) - v^\rho(f^\infty)\} \geq 0,\ f^\infty \in C(D)$. Take an arbitrary policy $f^\infty \in F_{-1}$. Then, f_*^∞ and f^∞ are both average optimal policies, i.e. $u^{-1}(f_*) = u^{-1}(f)$. From (7.5) it follows that

$$v^\rho(f_*^\infty) - v^\rho(f^\infty) = (1+\rho)\Big\{\{u^0(f_*) - u^0(f)\} + \sum_{k=1}^{\infty} \rho^k\{u^k(f_*) - u^k(f)\}\Big\}. \tag{7.6}$$

Hence, $0 \leq liminf_{\rho\downarrow 0}\,\{v^\rho(f_*^\infty) - v^\rho(f^\infty)\} = u^0(f_*) - u^0(f)$. Consequently, $u^0(f_*) \geq u^0(f)$.

Conversely, suppose that $f_*^\infty \in F_{-1}$ and that $u^0(f_*) \geq u^0(f)$ for all $f^\infty \in F_{-1}$. Notice that $liminf_{\rho \downarrow 0} \{v^\rho(f_*^\infty) - v^\rho(f^\infty)\} = liminf_{\rho \downarrow 0} \Big\{ \frac{1}{\rho}\{u^{-1}(f_*) - u^{-1}(f)\} + \{u^0(f_*) - u^0(f)\} \Big\}$.
Since $u^{-1}(f_*) \geq u^{-1}(f),\ f^\infty \in C(D)$ and $u^0(f_*) \geq u^0(f),\ f^\infty \in F_{-1}$, we have
$liminf_{\rho \downarrow 0} \{v^\rho(f_*^\infty) - v^\rho(f^\infty)\} \geq 0,\ f^\infty \in C(D)$, i.e. f_*^∞ is 0-discount optimal.
The proof of the induction step follows by similar arguments and is left to the reader.
Suppose that f_*^∞ is Blackwell optimal. Then, by Lemma 7.2, f_*^∞ is n-discount optimal for $n = -1, 0, 1, \ldots$. So, $f_*^\infty \in \cap_{n=-1}^\infty F_n$.
Finally, suppose that $f_*^\infty \in \cap_{n=-1}^\infty F_n$. Take an arbitrary $f^\infty \in C(D)$. If $f^\infty \in \cap_{n=-1}^\infty F_n$, then $u^n(f_*) = u^n(f)$ for $n = -1, 0, 1, \ldots$, and consequently $v^\rho(f_*^\infty) = v^\rho(f^\infty)$ for all $\rho > 0$.
Suppose that $f^\infty \notin \cap_{n=-1}^\infty F_n$ and let $f^\infty \notin F_n$ for some n and n minimal for which this holds. Then, $v^\rho(f_*^\infty) - v^\rho(f^\infty) = (1+\rho)\rho^n \Big\{ \{u^n(f_*) - u^n(f)\} + \sum_{k=n+1}^\infty \rho^k \{u^k(f_*) - u^k(f)\} \Big\}$ with $u^n(f_*) > u^n(f)$. Hence, we can find a $\rho(f)$ such that $v^\rho(f_*^\infty) \geq v^\rho(f^\infty)$ for $0 < \rho \leq \rho(f)$. Since $C(D)$ is finite, there exists a ρ_* sucht that $v^\rho(f_*^\infty) \geq v^\rho(f^\infty)$ for $0 < \rho \leq \rho_*$ for all $f^\infty \in C(D)$, i.e. f_*^∞ is Blackwell optimal. □

Remarks

1. $f_*^\infty \in F_n$ if and only if $\big(u^{-1}(f), u^0(f), \ldots, u^n(f)\big)$ is lexicographically the largest vector over the set vectors $\big(u^{-1}(g), u^0(g), \ldots, u^n(g)\big)\ g^\infty \in C(D)$.
2. Suppose that, for some $n \geq -1$, F_n contains a single policy f_*^∞. Then, f_*^∞ is a Blackwell optimal policy.

Next, we derive the optimality equations by using the optimality properties of a Blackwell optimal policy. Suppose that f_0^∞ is a Blackwell optimal policy. Then, by definition, f_0^∞ is discounted optimal for each $\rho,\ 0 < \rho \leq \rho_*$, so it satisfies

$$max_{f^\infty \in C(D)} \Big\{ r(f) + \{\frac{1}{1+\rho} P(f) - I\} v^\alpha(f_0^\infty) \Big\} = 0.$$

Noting that $\frac{1}{1+\rho} P(f) - I = \frac{1}{1+\rho}\{(P(f) - I) - \rho I\}$ and that $v^\alpha(f_0^\infty)$ has the Laurent series expansion for ρ near to 0, we obtain the equation

$$max_{f^\infty \in C(D)} \Big\{ r(f) + \{P(f) - I - \rho I\} \{ \sum_{k=-1}^{\infty} \rho^k u^k(f_0) \} \Big\} = 0.$$

Rearranging terms yields

$$max_{f^\infty \in C(D)} \Big\{ \{P(f) - I\} \tfrac{u^{-1}(f_0)}{\rho} + \{r(f) - u^{-1}(f_0) + (P(f) - I) u^0(f_0) + \sum_{k=1}^\infty \rho^k \{-u^{k-1}(f_0) + (P(f) - I) u^k(f_0)\} \Big\} = 0.$$

For the above equality to hold for all ρ near 0 it requires that:

1. $P(f)u^{-1}(f_0) - u^{-1}(f_0) \leq 0$ for all policies $f^\infty \in C(D)$.
2. For those f for which $\{P(f)u^{-1}(f_0)\}_i = u_i^{-1}(f_0)$ for some $i \in S$, we have the requirement $r_i(f) - u_i^{-1}(f_0) + \{P(f)u^0(f_0)\}_i - u_i^0(f_0) \leq 0$.
3. For those f for which $\{P(f)u^{-1}(f_0)\}_i = u_i^{-1}(f_0)$ and $r_i(f) - u_i^{-1}(f_0) + \{P(f)u^0(f_0)\}_i - u_i^0(f_0) = 0$ for some $i \in S$, we have $-u_i^0(f_0) + \{P(f)u^1(f_0)\}_i - u_i^1(f_0) \leq 0$.

In this way one can formulate the requirements.

This observation shows that the following system of inductively defined equations characterizes the coefficients of the Laurent series expansion of a Blackwell optimal policy:

$$max_{a\in A(i)} \Big\{ \sum_j p_{ij}(a)x_j^{-1} - x_i^{-1} \Big\} = 0. \tag{7.7}$$

$$max_{a\in A^{(-1)}(i,x^{-1})} \Big\{ r_i(a) + \sum_j p_{ij}(a)x_j^0 - x_i^0 - x_i^{-1} \Big\} = 0 \tag{7.8}$$

$$max_{a\in A^{(k-1)}(i,x^{-1},x^0,\ldots,x^{k-1})} \Big\{ \sum_j p_{ij}(a)x_j^k - x_i^k - x_i^{k-1} \Big\} = 0,\ k = 1,2,\ldots \tag{7.9}$$

where

$A^{(-1)}(i,x^{-1}) = argmax_{a\in A(i)} \Big\{ \sum_j p_{ij}(a)x_j^{-1} - x_i^{-1} \Big\}$;

$A^{(0)}\big(i,x^{-1},x^0\big) = argmax_{a\in A^{(-1)}(i,x^{-1})} \big\{ r_i(a) + \sum_j p_{ij}(a)x_j^0 - x_i^0 - x_i^{-1} \big\}$;

$A^{(k-1)}\big(i,x^{-1},x^0,\ldots,x^{k-1}\big) = argmax_{a\in A^{(k-2)}(i,x^{-1},x^0,\ldots,x^{k-2})} \big\{ \sum_j p_{ij}(a)x_j^{k-1} - x_i^{k-1} - x_i^{k-2} \big\},\ k \geq 2.$

We refer to this system as the *sensitive discount optimality equations* and to the individual equations as the (-1)th equation, 0th equation, 1th equation, etc. The sets of maximizing decision rules depend on the sequence of the x^k and consequently the system of equations is highly non-linear. Observe that the (-1)th and the 0th equation are the multichain average reward optimality equations. From the results of Chaper 5 it follows that if x^{-1} and x^0 satisfy these two equations and $f(i) \in A^{(-1)}(i,x^{-1}),\ i \in S$, then $x^{-1} = \phi$ and f^∞ is an average optimal policy. We generalize this observation to n-discount optimality below.

Theorem 7.4

If the vector $(x^{-1},x^0,\ldots,x^n)$ satisfies the following linear system

$$\begin{cases} \{I - P(f)\}x^{-1} & & & & & = & 0 \\ x^{-1} & + & \{I - P(f)\}x^0 & & & = & r(f) \\ & & x^{k-1} & + & \{I - P(f)\}x^k & = & 0,\ 1 \leq k \leq n \end{cases}$$

then $x^k = u^k(f),\ k = -1,0,1,\ldots,n-1$ and, if in addition either $-x^n + \{I - P(f)\}x^{n+1} = 0$ or $P^*(f)x^n = 0$, then $x^n = u^n(f)$.

Proof

Notice that $-x^n + \{I - P(f)\}x^{n+1} = 0$ implies $P^*(f)x^n = 0$. So, it is sufficient to consider the case with $P^*(f)x^n = 0$. It is straightforward to see that $\big(u^{-1}(f), u^0(f), \ldots, u^n(f)\big)$ is a solution. Consider an arbitrary solution $\big(x^{-1}, x^0, \ldots, x^n\big)$. Then, $\{I - P(f)\}x^{-1} = 0$ implies $x^{-1} = P^*(f)x^{-1}$, and consequently (from the second equation of the system), we get $x^{-1} = P^*(f)r(f) = u^{-1}(f)$. From the third equation, for $k = 1$, it follows that $P^*(f)x^0 = 0$, and from the second equation, $\{I - P(f) + P^*(f)\}x^0 = \{I - P^*(f)\}r(f)$. Therefore, we have $x^0 = \{I - P(f) + P^*(f)\}^{-1}\{I - P^*(f)\}r(f) = \{D(f) + P^*(f)\}\{I - P^*(f)\}r(f) = D(f)r(f) = u^0(f)$. By induction on k, we will show that $x^k = u^k(f),\ k \geq 1$ (for $k \leq 0$ this is shown above). Assume that $x^{-1} = u^{-1}(f), x^0 = u^0(f), \ldots, x^k = u^k(f)$. Since $x^k + \{I - P(f)\}x^{k+1} = 0$ and $P^*(f)x^{k+1} = 0$, we obtain

$$\begin{aligned} x^{k+1} &= -\{I - P(f) + P^*(f)\}^{-1}u^k(f) = -\{D(f) + P^*(f)\}(-1)^k\{D(f)\}^{k+1}r(f) \\ &= (-1)^{k+1}\{D(f)\}^{k+2}r(f) = u^{k+1}(f). \end{aligned}$$

□

7.4 Lexicographic ordering of Laurent series

Let, for $0 < \rho \leq \|D(f)\|^{-1}$, the matrix $H^\rho(f)$ be defined by

$$H^\rho(f) = (1 + \rho) \cdot \Big\{P^*(f) + \sum_{k=0}^{\infty} (-1)^k \rho^{k+1} D^{k+1}(f)\Big\}. \tag{7.10}$$

Theorem 7.5

(1) $H^\rho(f) = \rho \cdot \big\{I - \frac{1}{1+\rho}P(f)\big\}^{-1}$.

(2) $H^\rho(f)r(f) = \rho \cdot v^\rho(f^\infty)$.

(3) $\rho \cdot \{v^\rho(f^\infty) - x\} = H^\rho(f)\big\{r(f) + \frac{1}{1+\rho}P(f)x - x\big\}$ *for every* $x \in \mathbb{R}^N$.

Proof

$$\begin{aligned} (1)\quad \big\{I - \tfrac{1}{1+\rho}P(f)\big\}H^\rho(f) &= \big\{(1+\rho)I - P(f)\big\}\tfrac{1}{1+\rho}H^\rho(f) \\ &= \{(1+\rho)I - P(f)\}\Big\{P^*(f) + \textstyle\sum_{k=0}^{\infty}(-1)^k\rho^{k+1}D^{k+1}(f)\Big\} \\ &= \rho\big\{P^*(f) + \textstyle\sum_{k=0}^{\infty}(-1)^k\rho^{k+1}D^{k+1}(f)\big\} + \\ &\qquad \{I - P(f)\}\big\{P^*(f) + \textstyle\sum_{k=0}^{\infty}(-1)^k\rho^{k+1}D^{k+1}(f)\big\} \\ &= \rho\big\{P^*(f) + \textstyle\sum_{k=0}^{\infty}(-1)^k\rho^{k+1}D^{k+1}(f)\big\} + \\ &\qquad \{I - P(f)\}D(f)\textstyle\sum_{k=0}^{\infty}(-1)^k\rho^{k+1}D^k(f) \\ &= \rho \cdot P^*(f) + \textstyle\sum_{k=0}^{\infty}(-1)^k\rho^{k+2}D^{k+1}(f)\big\} + \\ &\qquad \{I - P^*(f)\}\textstyle\sum_{k=0}^{\infty}(-1)^k\rho^{k+1}D^k(f) \\ &= \rho \cdot P^*(f) + \textstyle\sum_{k=1}^{\infty}(-1)^{k-1}\rho^{k+1}D^k(f)\big\} + \\ &\qquad \rho \cdot \{I - P^*(f)\} + \textstyle\sum_{k=1}^{\infty}(-1)^k\rho^{k+1}D^k(f) \\ &= \rho \cdot I. \end{aligned}$$

(2) $\rho \cdot v^\rho(f^\infty) = \rho \cdot \{I - \frac{1}{1+\rho}P(f)\}^{-1} r(f) = H^\rho(f) r(f).$

(3) $$\begin{aligned} H^\rho(f)\{r(f) + \tfrac{1}{1+\rho}P(f)x - x\} &= \rho \cdot v^\rho(f^\infty) - H^\rho(f)\{I - \tfrac{1}{1+\rho}P(f)\}x \\ &= \rho \cdot v^\rho(f^\infty) - \rho \cdot x = \rho \cdot \{v^\rho(f^\infty) - x\}. \end{aligned}$$ □

Define the sets LS_1 and LS_2 of Laurent series by

$$LS_1 = \{u(\rho) \mid u(\rho) = \textstyle\sum_{k=-1}^{\infty} \rho^k u^k;\ u^k \in \mathbb{R}^N,\ k \geq -1;\ limsup_{k\to\infty} \|u^k\|^{1/k} < \infty\}.$$

$$LS_2 = \{x(\rho) \mid x(\rho) = (1+\rho)\textstyle\sum_{k=-1}^{\infty} \rho^k x^k;\ x^k \in \mathbb{R}^N,\ k \geq -1;\ limsup_{k\to\infty} \|x^k\|^{1/k} < \infty\}.$$

Lemma 7.5
$LS_1 = LS_2$.

Proof
Take any $u(\rho) \in LS_1$. Then, since $(1+\rho)^{-1} = 1 - \rho + \rho^2 - \rho^3 + \cdots$, we have

$$u(\rho) = (1+\rho)\{\textstyle\sum_{j=0}^{\infty}(-\rho)^j\}\sum_{k=-1}^{\infty}\rho^k u^k = (1+\rho)\sum_{k=-1}^{\infty}\sum_{j=0}^{\infty}(-1)^j\rho^{k+j}u^k.$$

Let $i = k + j$, then $i = -1, 0, 1, \ldots$ and $k \leq i$. Therefore, we may write

$$u(\rho) = (1+\rho)\textstyle\sum_{i=-1}^{\infty}\rho^i\{\sum_{k=-1}^{i}(-1)^{i-k}u^k\} = (1+\rho)\sum_{i=-1}^{\infty}\rho^i x^i,$$

where $x^i = \sum_{k=-1}^{i}(-1)^{i-k}u^k$. Because $\|x^i\| \leq \sum_{k=-1}^{i}\|u^k\| \leq (i+2)\cdot max_{-1\leq k\leq i}\|u^k\|$, we obtain $\|x^i\|^{1/i} \leq (i+2)^{1/i}\cdot\{max_{-1\leq k\leq i}\|u^k\|\}^{1/i}$, and consequently
$limsup_{i\to\infty}\|x^i\|^{1/i} \leq \{limsup_{i\to\infty}(i+2)^{1/i}\}\cdot\{limsup_{i\to\infty}\|u^i\|^{1/i}\} = limsup_{i\to\infty}\|u^i\|^{1/i} < \infty$,
i.e. $u(\rho) \in LS_2$.

Conversely, let $x(\rho) \in LS_2$. Then,

$$x(\rho) = (1+\rho)\textstyle\sum_{k=-1}^{\infty}\rho^k x^k = \rho^{-1}x^{-1} + \sum_{k=0}^{\infty}\rho^k\{x^k + x^{k-1}\} = \sum_{k=-1}^{\infty}\rho^k u^k,$$

where $u^{-1} = x^{-1}$ and $u^k = x^k + x^{k-1},\ k \geq 0$.
Since $limsup_{k\to\infty}\|u^k\|^{1/k} \leq 2\cdot limsup_{k\to\infty}\|k^k\|^{1/k} < \infty$, we have $x(\rho) \in LS_1$. □

Because the sets LS_1 and LS_2 are identical, we denote this set as LS. Notice that LS is a linear vector space. We define a *lexicographic ordering* on LS: $u(\rho)$ is *nonnegative (nonpositive)* if the first nonzero vector of $(u^{-1}, u^0, u^1, \ldots)$ is nonnegative (nonpositive), i.e.

$$\begin{cases} u(\rho) \geq_l 0 & \text{if } limin f_{\rho\downarrow 0}\ \rho^{-k}u(\rho) \geq 0 \text{ for } k = -1, 0, 1, \ldots. \\ u(\rho) >_l 0 & \text{if } u(\rho) \geq_l 0 \text{ and } u(\rho) \neq 0. \end{cases}$$

For $f^\infty \in C(D)$, let $L_f^\rho : LS \to LS$ be defined by

$$L_f^\rho\, x(\rho) = r(f) + (1+\rho)^{-1}P(f)x(\rho).$$

The Laurent expansion of $L_f^\rho\, x(\rho) - x(\rho)$ becomes:

$$\begin{aligned} L_f^\rho\, x(\rho) - x(\rho) &= r(f) + (1+\rho)^{-1} P(f)\big\{(1+\rho)\textstyle\sum_{k=-1}^{\infty} \rho^k x^k\big\} - (1+\rho)\sum_{k=-1}^{\infty} \rho^k x^k \\ &= r(f) + \textstyle\sum_{k=-1}^{\infty} \rho^k P(f) x^k - \sum_{k=-1}^{\infty} \rho^k x^k - \sum_{k=-1}^{\infty} \rho^{k+1} x^k \\ &= r(f) + \textstyle\sum_{k=-1}^{\infty} \rho^k P(f) x^k - \sum_{k=-1}^{\infty} \rho^k x^k - \sum_{k=0}^{\infty} \rho^k x^{k-1} \\ &= \rho^{-1}\{P(f)x^{-1} - x^{-1}\} + \{r(f) + P(f)x^0 - x^0 - x^{-1}\} \\ &\qquad + \textstyle\sum_{k=1}^{\infty} \rho^k \{P(f)x^k - x^k - x^{k-1}\}. \end{aligned}$$

The equation above implies that $L_f^\rho\, x(\rho) \in LS$. The next theorem shows that $v^\rho(f^\infty)$ is a fixed-point of L_f^ρ.

Lemma 7.6
$L_f^\rho\, v^\rho(f^\infty) - v^\rho(f^\infty) = 0.$

Proof
For $x = v^\rho(f^\infty) = (1+\rho)\sum_{k=-1}^{\infty} \rho^k u^k(f)$, we have $x^k = u^k(f),\ k = -1, 0, 1, \ldots$. Hence,

$$\begin{aligned} L_f^\rho\, x(\rho) - x(\rho) = \rho^{-1}\{P(f)u^{-1}(f) - u^{-1}(f)\} + \{r(f) + P(f)u^0(f) - u^0(f) - u^{-1}(f)\} \\ + \textstyle\sum_{k=1}^{\infty} \rho^k \{P(f)u^k(f) - u^k(f) - x^{k-1}(f)\}. \end{aligned}$$

To establish the fixed-point, note that

$$\begin{aligned} &P(f)u^{-1}(f) - u^{-1}(f) = \{P(f)P^*(f) - P^*(f)\}r(f) = 0. \\ &r(f) + P(f)u^0(f) - u^0(f) - u^{-1}(f) = \{I + P(f)D(f) - D(f) - P^*(f)\}r(f) = 0. \\ &P(f)u^k(f) - u^k(f) - u^{k-1}(f) = (-1)^{k-1}\{D(f)\}^k\{-P(f)D(f) + D(f) - I\}r(f) \\ &\qquad\qquad = (-1)^{k-1}\{D(f)\}^k\{-P^*(f)\}r(f) = 0,\ k \geq 1. \end{aligned}$$

□

Consider the mapping $B : LS \to LS$, where for $x(\rho) = (1+\rho)\sum_{k=-1}^{\infty} \rho^k x^k,\ Bx(\rho)$ is defined by

$$Bx(\rho) = \sum_{k=-1}^{\infty} \rho^k B^{(k)}\big(x^{-1}, x^0, \ldots, x^k\big)$$

with

$$\begin{aligned} &\big\{B^{(-1)}\big(x^{-1}\big)\big\}_i = max_{a\in A(i)}\big\{\textstyle\sum_j p_{ij}(a)x_j^{-1} - x_i^{-1}\big\} \text{ and} \\ &A^{(-1)}\big(i, x^{-1}\big) = argmax_{a\in A(i)}\big\{\textstyle\sum_j p_{ij}(a)x_j^{-1} - x_i^{-1}\big\},\ i \in S, \\ &\big\{B^{(0)}\big(x^{-1}, x^0\big)\big\}_i = max_{a\in A^{(-1)}(i,x^{-1})}\big\{r_i(a) + \textstyle\sum_j p_{ij}(a)x_j^0 - x_i^0 - x_i^{-1}\big\} \text{ and} \\ &A^{(0)}\big(i, x^{-1}, x^0\big) = argmax_{a\in A^{(-1)}(i,x^{-1})}\big\{r_i(a) + \textstyle\sum_j p_{ij}(a)x_j^0 - x_i^0 - x_i^{-1}\big\},\ i \in S, \end{aligned}$$

and for $k \geq 1$

$$\begin{aligned} &\big\{B^{(k)}\big(x^{-1}, x^0, \ldots, x^k\big)\big\}_i = max_{a\in A^{(k-1)}(i,x^{-1},x^0,\ldots,x^{k-1})}\big\{\textstyle\sum_j p_{ij}(a)x_j^k - x_i^k - x_i^{k-1}\big\} \text{ and} \\ &A^{(k)}\big(i, x^{-1}, x^0, \ldots, x^k\big) = argmax_{a\in A^{(k-1)}(i,x^{-1},x^0,\ldots,x^{k-1})}\big\{\textstyle\sum_j p_{ij}(a)x_j^k - x_i^k - x_i^{k-1}\big\},\ i \in S. \end{aligned}$$

Since we have derived that

$$\rho^{-1}\{P(f)x^{-1} - x^{-1}\} + \{r(f) + P(f)x^0 - x^0 - x^{-1}\} + \sum_{k=1}^{\infty} \rho^k\{P(f)x^k - x^k - x^{k-1}\} =$$

$$L_f^\rho\, x(\rho) - x(\rho) = r(f) + (1+\rho)^{-1}P(f)x(\rho) - x(\rho) \in LS \text{ for all } f^\infty \in C(D),$$

$Bx(\rho)$ is an element of LS which is the result of lexicographically maximizing the elements $r(f) + (1+\rho)^{-1}P(f)x(\rho) - x(\rho)$ over the set $C(D)$, i.e.

$$Bx(\rho) = lexmax_{f^\infty \in C(D)}\{r(f) + (1+\rho)^{-1}P(f)x(\rho) - x(\rho)\} = lexmax_{f^\infty \in C(D)}\{L_f^\rho\, x(\rho) - x(\rho)\}. \tag{7.11}$$

Because, by Lemma 7.6, $L_g^\rho\, v^\rho(g^\infty) - v^\rho(g^\infty) = 0$ for all $g^\infty \in C(D)$, we obtain for all $g^\infty \in C(D)$,

$$Bv^\rho(g^\infty) = lexmax_{f^\infty \in C(D)}\{L_f^\rho\, v^\rho(g^\infty) - v^\rho(g^\infty)\} \geq_l L_g^\rho\, v^\rho(g^\infty) - v^\rho(g^\infty) = 0, \tag{7.12}$$

i.e. $Bv^\rho(g^\infty)$ is lexicographically nonnegative for all $g^\infty \in C(D)$.

Lemma 7.7

$H^\rho(f)\{r(f) + (1+\rho)^{-1}P(f)v^\rho(g^\infty) - v^\rho(g^\infty)\} = \rho \cdot \{v^\rho(f^\infty) - v^\rho(g^\infty)\}$.

Proof

By Theorem 7.5 part (3), we obtain

$H^\rho(f)\{r(f) + (1+\rho)^{-1}P(f)v^\rho(g^\infty) - v^\rho(g^\infty)\} = \rho \cdot \{v^\rho(f^\infty) - v^\rho(g^\infty)\}$. □

Next, we will show that $H^\rho(f)$ is a *positive operator* for every $f^\infty \in C(D)$, i.e.

$$H^\rho(f)\{u(\rho)\} \geq_l 0 \text{ if } u(\rho) \geq_l 0 \text{ and } H^\rho(f)\{u(\rho)\} >_l 0 \text{ if } u(\rho) >_l 0.$$

Theorem 7.6

$H^\rho(f)$ is a positive operator for all $f^\infty \in C(D)$.

Proof

$$\begin{aligned} H^\rho(f)u(\rho) &= \rho \cdot \left\{I - (1+\rho)^{-1}P(f)\right\}^{-1}u(\rho) = \rho \cdot \sum_{t=0}^{\infty}\{(1+\rho)^{-1}P(f)\}^t u(\rho) \\ &= \rho \cdot u(\rho) + \rho \cdot \sum_{t=1}^{\infty}\{(1+\rho)^{-1}P(f)\}^t u(\rho). \end{aligned}$$

Hence, it is sufficient to show that $P^t(f)u(\rho) \geq_l 0$ for all $t \geq 1$ and all $u(\rho) \geq_l 0$.

Take any $t \geq 1,\ i \in S$, and let $T(i) = \{j \in S \mid p_{ij}^t(f) > 0\}$.

Suppose that k is such that $\{P^t(f)u^m\}_i = 0,\ -1 \leq m \leq k-1$ and $\{P^t(f)u^k\}_i \neq 0$.

Because $0 = \{P^t(f)u^m\}_i = \sum_{j \in T(i)} p_{ij}^t(f)u_j^m$, we have $u_j^m = 0,\ j \in T(i),\ -1 \leq m \leq k-1$.

Since $u(\rho) \geq_l 0$, we have $u_j^k \geq 0,\ j \in T(i)$ and consequently, $\{P^t(f)u^k\}_i = \sum_{j \in T(i)} p_{ij}^t(f)u_j^k \geq 0$.

Because $\{P^t(f)u^k\}_i \neq 0$, we get $\{P^t(f)u^k\}_i > 0$, i.e. $P^t(f)u(\rho) \geq_l 0$. □

Theorem 7.7

(1) The equation $Bx = 0,\ x \in LS$, has a unique solution $x = v^\rho(f_0^\infty)$, where f_0^∞ is a Blackwell optimal policy.

(2) If $f^\infty \in C(D)$ satisfies $Bx = r(f) + (1+\rho)^{-1}P(f)x - x = 0$, then f^∞ is Blackwell optimal.

Proof

Let f_0^∞ be a Blackwell optimal policy, i.e. $v^\rho(f_0^\infty) \geq_l v^\rho(f^\infty)$ for all $f^\infty \in C(D)$.

We first show that $Bv^\rho(f_0^\infty) = 0$ and $r(f) + (1+\rho)^{-1}P(f)v^\rho(f_0^\infty) - v^\rho(f_0^\infty) \leq_l 0,\ f^\infty \in C(D)$.

In (7.12) it is shown that $Bv^\rho(f^\infty) \geq_l 0$ for all $f^\infty \in C(D)$. Hence, $Bv^\rho(f_0^\infty) \geq_l 0$.

Suppose that $Bv^\rho(f_0^\infty) >_l 0$. Hence, from (7.11), it follows that there is a policy f^∞ satisfying $r(f) + (1+\rho)^{-1}P(f)v^\rho(f_0^\infty) - v^\rho(f_0^\infty) >_l 0$. Then, by Theorem 7.6 and Lemma 7.7, we obtain $H^\rho(f)\{r(f) + (1+\rho)^{-1}P(f)v^\rho(f_0^\infty) - v^\rho(f_0^\infty)\} = \rho \cdot \{v^\rho(f^\infty) - v^\rho(f_0^\infty)\} >_l 0$, contradicting the Blackwell optimality of f_0^∞. Hence, we have shown that

$$Bv^\rho(f_0^\infty) = 0 \text{ and } r(f) + (1+\rho)^{-1}P(f)v^\rho(f_0^\infty) - v^\rho(f_0^\infty) \leq_l 0,\ f^\infty \in C(D). \tag{7.13}$$

Next, suppose that $Bx = 0$ for some $x \in LS$. Then, $r(f_0) + (1+\rho)^{-1}P(f_0)x - x \leq_l 0$.

Since $H^\rho(f_0)$ is a positive operator, we obtain by Theorem 7.5 part (3),

$$0 \geq_l H^\rho(f_0)\{r(f_0) + (1+\rho)^{-1}P(f_0)x - x\} = \rho \cdot \{v^\rho(f_0^\infty) - x\},\ \text{i.e. } v^\rho(f_0^\infty) \leq_l x. \tag{7.14}$$

Therefore, $v^\rho(f_0^\infty)$ is the lexicographically smallest solution of the functional equation $Bx = 0$.

Finally, suppose that $Bx = r(f) + (1+\rho)^{-1}P(f)x - x = 0$ for some $f^\infty \in C(D)$. Then, we obtain

$$0 = H^\rho(f)\{r(f) + (1+\rho)^{-1}P(f)x - x\} = \rho \cdot \{v^\rho(f^\infty) - x\},\ \text{i.e. } v^\rho(f^\infty) = x. \tag{7.15}$$

Combining (7.14) and (7.15) gives $v^\rho(f^\infty) \geq_l v^\rho(f_0^\infty)$. Since f_0^∞ is Blackwell optimal, we also have $v^\rho(f_0^\infty) \geq_l v^\rho(f^\infty)$, i.e. $v^\rho(f^\infty) = v^\rho(f_0^\infty)$, implying that f^∞ is also a Blackwell optimal policy and the functional equation $Bx = 0$ has a unique solution $x = v^\rho(f_0^\infty)$. □

7.5 Policy iteration for n-discount optimality

In this section we derive for any $n \in \mathbb{N}$ a policy iteration algorithm which computes a policy that lexicographically maximes the vector $\left(u^{-1}(f), u^0(f), \ldots, u^n(f)\right)$ over all $f^\infty \in C(D)$, i.e. an n-discount optimal policy. Furthermore, we will show that an n-discount optimal policy for $n \geq N-1$ is a Blackwell optimal policy.

Algorithm 7.1 *Determination of an n-discount optimal policy by policy iteration*

1. Take an arbitrary $f^\infty \in C(D)$.
2. Determine $\left(u^{-1}(f), u^0(f), \ldots, u^{n+1}(f)\right)$ as unique solution of the system

$$\begin{cases} \{I - P(f)\}x^{-1} & = 0 \\ x^{-1} + \{I - P(f)\}x^0 & = r(f) \\ x^{k-1} + \{I - P(f)\}x^k & = 0,\ 1 \leq k \leq n+1;\ P^*(f)x^{n+1} = 0 \end{cases}$$

3. (a) Determine for all $i \in S$:
   (1) $max_{a \in A(i)} \{\sum_j p_{ij}(a) u_j^{-1}(f) - u_i^{-1}(f)\}$.
   (2) $A^{-1}(i) = argmax_{a \in A(i)} \{\sum_j p_{ij}(a) u_j^{-1}(f) - u_i^{-1}(f)\}$.

   (b) If $max_{a \in A(i)} \{\sum_j p_{ij}(a) u_j^{-1}(f) - u_i^{-1}(f)\} = 0,\ i \in S$, then go to step 3c.
   Otherwise: Take g such that $g \in A^{-1}(i),\ i \in S$, and go to step 5.

   (c) Determine for all $i \in S$:
   (1) $max_{a \in A^{-1}(i)} \{r_i(a) + \sum_j p_{ij}(a) u_j^0(f) - u_i^0(f) - u_i^{-1}(f)\}$.
   (2) $A^0(i) = argmax_{a \in A^{-1}(i)} \{r_i(a) + \sum_j p_{ij}(a) u_j^0(f) - u_i^0(f) - u_i^{-1}(f)\}$.

   (d) If $max_{a \in A^{-1}(i)} \{r_i(a) + \sum_j p_{ij}(a) u_j^0(f) - u_i^0(f) - u_i^{-1}(f)\} = 0,\ i \in S$: go to step 3e.
   Otherwise: Take g such that $g \in A^0(i),\ i \in S$, and go to step 5.

   (e) For $k = 0$ until n do
   Determine for all $i \in S$:
   (1) $max_{a \in A^k(i)} \{\sum_j p_{ij}(a) u_j^{k+1}(f) - u_i^{k+1}(f) - u_i^k(f)\}$.
   (2) $A^{k+1}(i) = argmax_{a \in A^k(i)} \{\sum_j p_{ij}(a) u_j^{k+1}(f) - u_i^{k+1}(f) - u_i^k(f)\}$.
   If $max_{a \in A^{k+1}(i)} \{\sum_j p_{ij}(a) u_j^{k+1}(f) - u_i^{k+1}(f) - u_i^k(f)\} > 0$ for some $i \in S$:
   take g such that $g \in A^{k+1}(i),\ i \in S$ and go to step 5.

4. f^∞ is n-discount optimal.

5. $f(i) := g(i),\ i \in S$ and return to step 2.

Remarks:

1. In step 2 of the algorithm, we may instead of the last requirement $P^*(f)x^{n+1} = 0$ also solve the additional equation $x^{n+1} + \{I - P(f)\}x^{n+2} = 0$, which implies $P^*(f)x^{n+1} = 0$.
2. If $|A^k(i)| = 1$ for one or more states, then for that states i step 3 of the algorithm can be skipped, because $A^{k+1}(i)$ consists of the same single action as $A^k(i)$.

Example 7.2 (continued)

We compute a 1-discount optimal policy, starting with the policy $f(1) = f(2) = f(3) = 1$.

Iteration 1:

Step 2:

$$P(f) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \rightarrow \quad P^*(f) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \rightarrow \quad D(f) = \begin{pmatrix} 1 & 1 & -2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

$u^{-1}(f) = (0, 0, 0);\ u^0(f) = (2, 1, 0);\ u^1(f) = (-3, -1, 0);\ u^2(f) = (4, 1, 0)$.

Step 3: (only for $i = 1$, because $|A(2)| = |A(3)| = 1$.)

a. $a = 1$: $\sum_j p_{ij}(a) u_j^{-1}(f) - u_1^{-1}(f) = u_2^{-1}(f) - u_1^{-1}(f) = 0 - 0 = 0$.
   $a = 2$: $\sum_j p_{ij}(a) u_j^{-1}(f) - u_1^{-1}(f) = u_3^{-1}(f) - u_1^{-1}(f) = 0 - 0 = 0$.
   $A^{-1}(1) = \{1, 2\}$.

c. $a=1$: $r_1(a)+\sum_j p_{ij}(a)u_j^0(f)-u_1^0(f)-u_1^{-1}(f)=r_1(1)+u_2^0(f)-u_1^0(f)-u_1^{-1}(f)=1+1-2-0=0.$

$a=2: r_1(a)+\sum_j p_{ij}(a)u_j^0(f)-u_1^0(f)-u_1^{-1}(f)=r_1(2)+u_3^0(f)-u_1^0(f)-u_1^{-1}(f)=2+0-2-0=0.$

$A^0(1)=\{1,2\}$.

e. $k=0$: $a=1$: $\sum_j p_{ij}(a)u_j^1(f)-u_1^1(f)-u_1^0(f)=u_2^1(f)-u_1^1(f)-u_1^0(f)=-1+3-2=0.$

$a=2$: $\sum_j p_{ij}(a)u_j^1(f)-u_1^1(f)-u_1^0(f)=u_3^1(f)-u_1^1(f)-u_1^0(f)=0+3-2=1.$

$g(1)=2$.

Step 5:

$f(1)=2,\ f(2)=1,\ f(3)=1.$

Iteration 2:

Step 2:

$$P(f)=\begin{pmatrix}0&0&1\\0&0&1\\0&0&1\end{pmatrix}\ \rightarrow\ P^*(f)=\begin{pmatrix}0&0&1\\0&0&1\\0&0&1\end{pmatrix}\ \rightarrow\ D(f)=\begin{pmatrix}1&0&-1\\0&1&-1\\0&0&0\end{pmatrix}.$$

$u^{-1}(f)=(0,0,0);\ u^0(f)=(2,1,0);\ u^1(f)=(-2,-1,0);\ u^2(f)=(2,1,0).$

Step 3:

a. $a=1$: $\sum_j p_{ij}(a)u_j^{-1}(f)-u_1^{-1}(f)=u_2^{-1}(f)-u_1^{-1}(f)=0-0=0.$

$a=2$: $\sum_j p_{ij}(a)u_j^{-1}(f)-u_1^{-1}(f)=u_3^{-1}(f)-u_1^{-1}(f)=0-0=0.$

$A^{-1}(1)=\{1,2\}$.

c. $a=1$: $r_1(a)+\sum_j p_{ij}(a)u_j^0(f)-u_1^0(f)-u_1^{-1}(f)=r_1(1)+u_2^0(f)-u_1^0(f)-u_1^{-1}(f)=1+1-2-0=0.$

$a=2: r_1(a)+\sum_j p_{ij}(a)u_j^0(f)-u_1^0(f)-u_1^{-1}(f)=r_1(2)+u_3^0(f)-u_1^0(f)-u_1^{-1}(f)=2+0-2-0=0.$

$A^0(1)=\{1,2\}$.

e. $k=0$: $a=1$: $\sum_j p_{ij}(a)u_j^1(f)-u_1^1(f)-u_1^0(f)=u_2^1(f)-u_1^1(f)-u_1^0(f)=-1+2-2=-1.$

$a=2$: $\sum_j p_{ij}(a)u_j^1(f)-u_1^1(f)-u_1^0(f)=u_3^1(f)-u_1^1(f)-u_1^0(f)=0+2-2=0.$

$A^1(1)=\{2\}$.

Since $A^1(1)$ consists of one element the policy f^∞ with $f(1)=2,\ f(2)=f(3)=1$ is 1-optimal.

In order to show the correctness of Algorithm 7.1 we need some theorems which we present below. We introduce the following notation for two policies $f^\infty, g^\infty \in C(D)$,

$$\begin{aligned}\psi^{-1}(f,g) &= P(g)u^{-1}(f)-u^{-1}(f)\\ \psi^0(f,g) &= r(g)+P(g)u^0(f)-u^0(f)-u^{-1}(f)\\ \psi^k(f,g) &= P(g)u^k(f)-u^k(f)-u^{k-1}(f)\end{aligned}$$

Theorem 7.8

For every $f^\infty, g^\infty \in C(D)$ and every $m\in\mathbb{N}$, we have

$\alpha v^\alpha(g^\infty)=\sum_{k=-1}^{m-1}\rho^k\{u^k(f)+\sum_{t=1}^\infty \alpha^t P^{t-1}(g)\psi^k(f,g)\}-\rho^m\sum_{t=1}^\infty \alpha^t P^{t-1}(g)u^{m-1}(f).$

Proof

$$\begin{aligned}
\alpha v^{\alpha}(g^{\infty}) &= \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) r(g) \\
&= \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{r(g) + P(g)u^0(f) - u^0(f) - u^{-1}(f)\} + \\
&\qquad \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^{-1}(f) - \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^0(f) - u^0(f)\} \\
&= \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^0(f,g) + \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^{-1}(f) - \\
&\qquad \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^0(f) - u^0(f)\}.
\end{aligned}$$

$$\begin{aligned}
\textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^{-1}(f) &= \alpha u^{-1}(f) + \textstyle\sum_{t=2}^{\infty} \alpha^t P^{t-1}(g) u^{-1}(f) \\
&= \alpha u^{-1}(f) + \textstyle\sum_{t=2}^{\infty} \alpha^t \big\{ u^{-1}(f) + \sum_{s=1}^{t-1} \{P^s(g)u^{-1}(f) - P^{s-1}(g)u^{-1}(f)\}\big\} \\
&= \alpha(1-\alpha)^{-1}u^{-1}(f) + \textstyle\sum_{t=2}^{\infty} \alpha^t \sum_{s=1}^{t-1} P^{s-1}(g)\psi^{-1}(f,g) \\
&= \alpha(1-\alpha)^{-1}u^{-1}(f) + \textstyle\sum_{s=1}^{\infty} \big(\sum_{t=s+1}^{\infty} \alpha^t\big) P^{s-1}(g)\psi^{-1}(f,g) \\
&= \rho^{-1}u^{-1}(f) + \textstyle\sum_{s=1}^{\infty} \alpha^{s+1}(1-\alpha)^{-1} P^{s-1}(g)\psi^{-1}(f,g) \\
&= \rho^{-1}\{u^{-1}(f) + \textstyle\sum_{s=1}^{\infty} \alpha^s P^{s-1}(g)\psi^{-1}(f,g)\}.
\end{aligned}$$

For $k \geq 0$, we obtain

$$\begin{aligned}
\rho \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^k(f) &= \big(\tfrac{1}{\alpha} - 1\big) \textstyle\sum_{t=1}^{\infty} \alpha^s P^{t-1}(g) u^k(f) \\
&= \textstyle\sum_{t=1}^{\infty} \alpha^{t-1} P^{t-1}(g) u^k(f) - \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^k(f) \\
&= u^k(f) + \textstyle\sum_{t=2}^{\infty} \alpha^{t-1} P^{t-1}(g) u^k(f) - \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^k(f) \\
&= u^k(f) + \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^k(f) - u^k(f)\},
\end{aligned}$$

i.e. $\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^k(f) - u^k(f)\} = \rho \cdot \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^k(f) - u^k(f)$.

Since $u^k(f) = P(g)u^{k+1}(f) - u^{k+1}(f) - \psi^{k+1}(f,g)$, we can write

$$\begin{aligned}
&\textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^k(f) - u^k(f)\} = \\
&\rho \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^{k+1}(f) - u^{k+1}(f) - \psi^{k+1}(f,g)\} - u^k(f) = \\
&\rho \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^{k+1}(f) - u^{k+1}(f)\} - \rho \cdot \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^{k+1}(f,g)\} - u^k(f).
\end{aligned}$$

Hence, using this formula for $k = 0$ and then for $k = 1$, we obtain

$$\begin{aligned}
&\textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^0(f) - u^0(f)\} = \\
&\rho \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^1(f) - u^1(f)\} - \rho \cdot \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^1(f,g) - u^0(f) = \\
&\rho \cdot \Big\{\rho \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^2(f) - u^2(f)\} - \rho \cdot \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^2(f,g) - u^1(f)\Big\} \\
&\qquad\qquad -\rho \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^1(f,g) - u^0(f) = \\
&\rho^2 \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^2(f) - u^2(f) - \psi^2(f,g)\} - \rho \cdot u^1(f) - \rho \cdot \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^1(f,g) - u^0(f).
\end{aligned}$$

Similarly, by induction on m, it can be shown that

$$\begin{aligned}
&\textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^0(f) - u^0(f)\} = \\
&\rho^m \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\{P(g)u^m(f) - u^m(f) - \psi^m(f,g)\} \\
&\qquad\qquad - \textstyle\sum_{k=1}^{m-1} \rho^k \big\{u^k(f) + \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^k(f,g)\big\} - u^0(f) = \\
&\rho^m \cdot \textstyle\sum_{t=1}^{\infty} \alpha^t P^{t-1}(g) u^{m-1}(f) - \sum_{k=1}^{m-1} \rho^k \big\{u^k(f) + \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^k(f,g)\big\} - u^0(f).
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}\alpha v^{\alpha}(g^{\infty}) &= \sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^0(f,g)+\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)u^{-1}(f)-\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\{P(g)u^0(f)-u^0(f)\}\\ &= \sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^0(f,g)+\rho^{-1}\{u^{-1}(f)+\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^{-1}(f,g)\}\\ &\quad -\rho^m\cdot\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)u^{m-1}(f)+\sum_{k=1}^{m-1}\rho^k\big\{u^k(f)+\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)\big\}+u^0(f)\\ &= \sum_{k=-1}^{m-1}\rho^k\big\{u^k(f)+\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)\big\}-\rho^m\cdot\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)u^{m-1}(f).\end{aligned}$$

□

Theorem 7.9

If f^{∞} and g^{∞} are subsequent policies in Algorithm 7.1, then $v^{\rho}(g^{\infty}) > v^{\rho}(f^{\infty})$ for ρ sufficiently small.

Proof

Since $\psi^k(f,f)=0$ for $k=-1,0,1\ldots$, we obtain from Theorem 7.8 with $m=n+2$,

$$v^{\rho}(g^{\infty})-v^{\rho}(f^{\infty}) = (1+\rho)\Big\{\sum_{k=-1}^{n+1}\rho^k\big\{\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)\big\}+\rho^{n+2}\cdot\sum_{t=1}^{\infty}\alpha^t\big\{P^{t-1}(f)-P^{t-1}(g)\big\}u^{n+1}(f)\Big\}.$$

Since

$$\begin{aligned}\|\rho^{n+2}\cdot\sum_{t=1}^{\infty}\alpha^t\big\{P^{t-1}(f)-P^{t-1}(g)\big\}u^{n+1}(f)\| &\leq \rho^{n+2}(1-\alpha)^{-1}\|P^{t-1}(f)-P^{t-1}(g)\|\cdot\|u^{n+1}(f)\|\\ &\leq 2\rho^{n+2}(1+\rho)\cdot\|u^{n+1}(f)\|,\end{aligned}$$

$\rho^{n+2}\cdot\sum_{t=1}^{\infty}\alpha^t\big\{P^{t-1}(f)-P^{t-1}(g)\big\}u^{n+1}(f)$ is arbitrary close to 0 for ρ sufficiently small.

Since f^{∞} and g^{∞} are subsequent policies in Algorithm 7.1, we have for some $-1\leq m\leq n+1$, $\psi^k(f,g)=0,\ -1\leq k\leq m-1$ and $\psi^m(f,g)>0$.

If we define $\psi^k(f,g)=0$ for $k\geq n+2$, then $\sum_{k=-1}^{\infty}\rho^k\psi^k(f,g)\in LS$, and $\sum_{k=-1}^{\infty}\rho^k\psi^k(f,g)>_l 0$.

Since, by Theorem 7.5, part (1), $H^{\rho}(g)=\rho\cdot\big\{I-\frac{1}{1+\rho}P(g)\big\}^{-1}$ and, by Theorem 7.6, $H^{\rho}(g)$ is a positive operator, $\big\{I-\frac{1}{1+\rho}P(g)\big\}^{-1}$ is also a positive operator, implying that $\big\{I-\frac{1}{1+\rho}P(g)\big\}^{-1}\big\{\sum_{k=-1}^{\infty}\rho^k\psi^k(f,g)\big\}>_l 0$.

We can also write

$$\big\{I-\tfrac{1}{1+\rho}P(g)\big\}^{-1}\big\{\sum_{k=-1}^{\infty}\rho^k\psi^k(f,g)\big\}=\sum_{t=1}^{\infty}\alpha^{t-1}P^{t-1}(g)\big\{\sum_{k=-1}^{\infty}\rho^k\psi^k(f,g)\big\}=$$

$$\sum_{t=1}^{\infty}\alpha^{t-1}P^{t-1}(g)\big\{\sum_{k=-1}^{n+1}\rho^k\psi^k(f,g)\big\}=\tfrac{1}{\alpha}\Big\{\sum_{k=-1}^{n+1}\rho^k\big\{\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)\big\}\Big\}=$$

$$(1+\rho)\Big\{\sum_{k=-1}^{n+1}\rho^k\big\{\sum_{t=1}^{\infty}\alpha^t P^{t-1}(g)\psi^k(f,g)\big\}\Big\}.$$

Since we have shown that $v^{\rho}(g^{\infty})-v^{\rho}(f^{\infty})$ consists of two terms, where the first term is lexicographically positive and the second term is arbitrary close to 0 for ρ sufficiently small, we have shown that $v^{\rho}(g^{\infty})>v^{\rho}(f^{\infty})$ for ρ sufficiently small.

□

Theorem 7.10

Algorithm 7.1 is correct, i.e. it terminates with an n-discount optimal policy.

Proof

From Theorem 7.9 it follows that each subsequent policy $f^\infty \in C(D)$ is different from all previous. Since $C(D)$ is a finite set, the algorithm terminates, say with policy f^∞. In Theorem 7.9 is shown that for any policy g^∞,

$$v^\rho(g^\infty) - v^\rho(f^\infty) = \\ (1+\rho)\Big\{ \textstyle\sum_{k=-1}^{n+1} \rho^k \big\{ \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^k(f,g)\big\} + \rho^{n+2} \cdot \sum_{t=1}^{\infty} \alpha^t \big\{P^{t-1}(f) - P^{t-1}(g)\big\} u^{n+1}(f)\Big\}.$$

Since the algorithm terminates, we have $\sum_{k=-1}^{n+2} \rho^k \psi^k(f,g) \leq_l 0$. Analogously as in the proof of Theorem 7.9 this implies that $(1+\rho)\Big\{ \sum_{k=-1}^{n+1} \rho^k \big\{ \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^k(f,g)\big\}\Big\} \leq_l 0$.
Since the second term of the above expression for $v^\rho(g^\infty) - v^\rho(f^\infty)$ is again arbitrary close to 0 for ρ sufficiently small, we have shown that $lim_{\rho \downarrow 0}\, \rho^{-n}\{v^\rho(g^\infty) - v^\rho(f^\infty)\} \geq 0$ for every $g^\infty \in C(D)$, i.e. f^∞ is an n-discount optimal policy. □

We close this section with the proof that for $n \geq N-1$ an n-discount optimal policy is also Blackwell optimal. Therefore we need the following lemma.

Lemma 7.8

If $\psi^k(f,g) = 0$ for $k = 1, 2, \ldots, N$, then $\psi^k(f,g) = 0$ for $k \geq N+1$.

Proof

Let $L = \{x \mid \{P(f) - P(g)\}x = 0\}$. For $k \geq 1$, we have

$$\begin{aligned} \psi^k(f,g) &= P(g)u^k(f) - u^k(f) - u^{k-1}(f) = P(g)u^k(f) - (-1)^k\{D(f) - I\}D^k(f)r(f) \\ &= P(g)u^k(f) - (-1)^k\{P(f)D(f) - P^*(f)\}D^k(f)r(f) \\ &= P(g)u^k(f) - (-1)^k P(f)D^{k+1}(f)r(f) = P(g)u^k(f) - P(f)u^k(f), \end{aligned}$$

i.e. $u^k(f) \in L$ for $k = 1, 2, \ldots$.
Since L is a linear vector space in $\mathbb{R}^N$, the $N+1$ vectors $u^k(f),\ 1 \leq k \leq N+1$ are linear dependent. Because $u^k(f) = B^{k-1}x_0$ for $k \geq 1$, where $x_0 = u^1(f) = D(f)r(f)$ and $B = -D(f)$,
the $N+1$ vectors $x_0, Bx_0, B^2x_0, \ldots, B^Nx_0$ are linear dependent, i.e. for some $1 \leq k \leq N$, we have $B^kx_0 = \sum_{j=0}^{k-1} \lambda_j B^j x_0$ for some scalars $\lambda_0, \lambda_1, \ldots, \lambda_{k-1}$. Hence, $B^kx_0 \in L$. Since $B^{k+1}x_0 = \sum_{j=0}^{k-1} \lambda_j B^{j+1}x_0$, the vector $B^{k+1}x_0$ is a linear combination of the elements $Bx_0, B^2x_0, \ldots, B^kx_0$, which all belong to L, so $B^{k+1}x_0 \in L$. Similarly, by induction, it can be shown that $u^k(f) = B^{k-1}x_0 \in L,\ k \geq 1$, i.e. implying that $\psi^k(f,g) = 0$ for $k \geq 1$. □

Theorem 7.11

If Algorithm 7.1 is used to determine an $(N-1)$-discount optimal policy f^∞, then f^∞ is a Blackwell optimal policy.

Proof

If the algorithm terminates with policy f^∞, we have $\sum_{k=-1}^{N} \rho^k \psi^k(f,g) \leq_l 0$ for every policy g^∞, i.e. either $\sum_{k=-1}^{N} \rho^k \psi^k(f,g) <_l 0$ or $\sum_{k=-1}^{N} \rho^k \psi^k(f,g) = 0$.

In the first case, we obtain analogously to Theorem 7.9 that $v^\rho(g^\infty) < v^\rho(f^\infty)$ for ρ sufficiently small. In the second case, we have $\psi^k(f,g) = 0,\ 1 \leq k \leq N$. From Lemma 7.8 it follows that $\psi^k(f,g) = 0$ also for $k \geq N+1$. Hence, $\psi^k(f,g) = 0,\ k \geq 1$.

If we let $m \to \infty$ in Theorem 7.8, then we obtain

$$\alpha v^\alpha(g^\infty) = \sum_{k=-1}^{\infty} \rho^k \{u^k(f) + \sum_{t=1}^{\infty} \alpha^t P^{t-1}(g)\psi^k(f,g)\} = \sum_{k=-1}^{\infty} \rho^k u^k(f) = \alpha v^\alpha(f^\infty)$$

for every $\alpha \in [0,1)$, Hence, f^∞ is a Blackwell optimal policy. □

7.6 Linear programming and n-discount optimality (irreducible case)

Without any assumption about the chain structure, only for the criteria average and bias optimality there exist a satisfactory treatment by linear programming. In this section we show that, under the assumption of irreducibility, a nice treatment for n-discount optimality, based on *nested linear programs*. Throughout this section we have the following assumption.

Assumption 7.1

For any policy $f^\infty \in C(D)$, the Markov chain $P(f)$ is irreducible.

7.6.1 Average optimality

The special linear programming approach for average rewards in the irreducible case was treated in section 6.1.3. There, we have shown that the value vector ϕ and an optimal policy can be found by the linear programs

$$min\Big\{v \;\Big|\; v + \sum_j \{\delta_{ij} - p_{ij}(a)\}u_j \geq r_i(a),\ (i,a) \in S \times A\Big\} \tag{7.16}$$

and

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & = & 0,\ j \in S \\ \sum_{(i,a)} x_i(a) & = & 1 \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}, \tag{7.17}$$

respectively. Furthermore, we have shown (Theorem 6.5) that there is a bijection between the feasible solutions of the dual program (7.17) and the set $C(S)$ of stationary policies such that extreme solutions correspond to the set $C(D)$ of deterministic policies. This bijection is given by

$$x_i^{\pi}(a) = x_i(\pi) \cdot \pi_{ia}, \ (i,a) \in S \times A \text{ and } \pi_{ia}^{x} = \frac{x_i(a)}{\sum_a x_i(a)}, \ (i,a) \in S \times A,$$

where $x(\pi)$ is the stationary distribution of the transition matrix $P(\pi)$.
The following result characterizes the set of all average optimal policies.

Theorem 7.12
Let (ϕ, u^*) be an optimal solution of program (7.16). Then, f^∞ is an average optimal policy if and only if $f^\infty \in A^{-1}$, where $A^{-1} = \{f^\infty \mid \phi + \sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\}u_j^* = r_i\big(f(i)\big), \ i \in S\}$.

Proof
If $f^\infty \in A^{-1}$, then $\phi \cdot e + \{I - P(f)\}u^* = r(f)$. Multiplying this equation with the stationary distribution $x^(f)$ gives $\phi = x^T r(f) = \phi(f^\infty)$, i.e. f^∞ is an optimal policy.
Conversely, let f^∞ be an optimal deterministic policy. Then, x^f is an extreme optimal solution of the dual program (7.17) because
$\sum_{(i,a)} r_i(a)x_i^f(a) = \sum_i r_i\big(f(i)\big)x_i^f\big(f(i)\big) = \sum_i r_i\big(f(i)\big)x_i(f) = x(f)^T r(f) = \phi(f^\infty) = \phi,$
the optimum of (7.17). Since $x_i^f\big(f(i)\big) > 0, \ i \in S$, we have by the complementary slackness property of linear programming: $\phi + \sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\}u_j^* = r_i\big(f(i)\big), \ i \in S$, i.e. $f^\infty \in A^{-1}$. □

7.6.2 Bias optimality

We first show that for an average optimal policy f^∞ the second term $u^0(f)$ of the Laurent expansion of $v^\alpha(f^\infty)$ can be obtained from the results of the previous section.

Theorem 7.13
Let f^∞ be an average optimal policy. Then, the bias term $u^0(f) = u^* - P^*(f)u^*$, where u^* is the u-part in an optimal solution of the linear program (7.16).

Proof
Let f^∞ be an average optimal policy. Then, by Theorem 7.12, $\phi \cdot e + \{I - P(f)\}u^* = r(f)$.
Multiplying this equation with $D(f)$ gives
$u^0(f) = D(f)r(f) = D(f)\{\phi \cdot e + \{I - P(f)\}u^*\} = \{I - P^*(f)\}u^* = u^* - P^*(f)u^*.$ □

The policy f^∞ is bias optimal or 0-discount optimal if $u^0(f) = max\{u^0(g) \mid g^\infty \in A^{-1}\}$.
Since $-P^*(g)u^*$ is the average reward of g^∞ with respect to immediate rewards $r_i^{(0)}(a) = -u_i^*$, $(i,a) \in S \times A^{-1}$, the maximization of $u^0(g) = u^* - P^*(g)u^*$ is equivalent to the maximization of the average reward corresponding to immediate rewards $r_i^{(0)}(a) = -u_i^*, \ (i,a) \in S \times A^{-1}$.

Thus, for bias optimality, we can consider a new MDP model with truncated actions sets $A^{-1}(i) = \{a \in A(i) \mid \phi + \sum_j \{\delta_{ij} - p_{ij}(a)\}u_j^* = r_i(a)\},\ i \in S$ and with immediate rewards $r_i^{(0)}(a) = -u_i^*,\ (i,a) \in S \times A^{-1}$.

We now present the primal and dual linear program for this truncated MDP:

$$min\Big\{v^{(0)} \ \Big|\ v^{(0)} + \sum_j \{\delta_{ij} - p_{ij}(a)\}u_j^{(0)} \geq -u_i^*,\ i \in S,\ a \in A^{-1}(i)\Big\} \tag{7.18}$$

and

$$max\left\{-\sum_i u_i^* \sum_{a \in A^{-1}(i)} x_i^{(0)}(a) \ \middle|\ \begin{array}{rcl} \sum_i \sum_{a \in A^{-1}(i)} \{\delta_{ij} - p_{ij}(a)\}x_i^{(0)}(a) & = & 0,\ j \in S \\ \sum_i \sum_{a \in A^{-1}(i)} x_i^{(0)}(a) & = & 1 \\ x_i^{(0)}(a) & \geq & 0,\ i \in S,\ a \in A^{-1}(i) \end{array}\right\} \tag{7.19}$$

respectively. The next theorem is a consequence of above statements.

Theorem 7.14

Let $(v^{(0)}, u^{(0)})$ and $x^{(0)}$ be optimal solutions of the linear programs (7.18) and (7.19) respectively. Furthermore, let $A^0 = \{f^\infty \in A^{-1} \mid v^{(0)} + \sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\}u^{(0)} = -u_i^*,\ i \in S\}$.

Then, f^∞ is a bias optimal policy if and only if $f^\infty \in A^0$.

Example 7.3

$S = \{1,2,3\};\ A(1) = \{1,2\},\ A(2) = \{1\},\ A(3) = \{1\};\ r_1(1) = 3,\ r_1(2) = 4,\ r_2(1) = 2,\ r_3(1) = 1.$

$p_{11}(1) = 0,\ p_{12}(1) = 1,\ p_{13}(1) = 0;\ p_{11}(2) = p_{12}(2) = 0,\ p_{13}(2) = 1;$

$p_{21}(1) = \frac{1}{2},\ p_{22}(1) = 0,\ p_{23}(1) = \frac{1}{2};\ p_{31}(1) = 0,\ p_{32}(1) = 1,\ p_{33}(1) = 0.$

It is easy to verify that this is an irreducible MDP. The primal and dual linear programs for average optimality are:

$$min\Big\{v \ \Big|\ v + u_1 - u_2 \geq 3;\ v + u_1 - u_3 \geq 4;\ v + u_2 - \frac{1}{2}u_1 - \frac{1}{2}u_3 \geq 2;\ v + u_3 - u_2 \geq 1\Big\} \tag{7.20}$$

and

$$max\left\{3x_1(1) + 4x_1(2) + 2x_2(1) + x_3(1) \ \middle|\ \begin{array}{rrrrl} x_1(1) + & x_1(2) - & \frac{1}{2}x_2(1) & & = 0;\ x_1(1) \geq 0 \\ -x_1(1) + & + & x_2(1) & - x_3(1) & = 0;\ x_1(2) \geq 0 \\ - & x_1(2) - & \frac{1}{2}x_2(1) & + x_3(1) & = 0;\ x_2(1) \geq 0 \\ x_1(1) + & x_1(2) + & x_2(1) & + x_3(1) & = 1;\ x_3(1) \geq 0 \end{array}\right\}. \tag{7.21}$$

Optimal solutions of these programs are: for the primal $v^* = \phi = 2,\ u_1^* = 2,\ u_2^* = 1,\ u_3^* = 0$ and for the dual $x_1^*(1) = \frac{1}{4},\ x_1^*(2) = 0,\ x_2^*(1) = \frac{1}{2},\ x_3^*(1) = \frac{1}{4}$. It is simple to verify that $A^{-1}(1) = A(1) = \{1,2\},\ A^{-1}(2) = A(2) = \{1\}$ and $A^{-1}(3) = A(3) = \{1\}$. Hence, the primal and dual programs for bias optimality are:

$$min\left\{v^{(0)} \ \middle|\ \begin{array}{rrrrl} v^{(0)} + u_1^{(0)} & - u_2^{(0)} & & \geq -2; & v^{(0)} - \frac{1}{2}u_1^{(0)} + u_2^{(0)} - \frac{1}{2}u_3^{(0)} \geq -1 \\ v^{(0)} + u_1^{(0)} & & - u_3^{(0)} & \geq -2; & v^{(0)} - u_2^{(0)} + u_3^{(0)} \geq 0 \end{array}\right\} \tag{7.22}$$

and

$$max\left\{ -2x_1^{(0)}(1) - 2x_1^{(0)}(2) - x_2^{(0)}(1) \;\middle|\; \begin{array}{rrrrrl} x_1^{(0)}(1) + & x_1^{(0)}(2) - & \frac{1}{2}x_2^{(0)}(1) & & = 0; & x_1^{(0)}(1) \geq 0 \\ -x_1^{(0)}(1) + & & + \; x_2^{(0)}(1) & - x_3^{(0)}(1) & = 0; & x_1^{(0)}(2) \geq 0 \\ & - \; x_1^{(0)}(2) - & \frac{1}{2}x_2^{(0)}(1) & + x_3^{(0)}(1) & = 0; & x_2^{(0)}(1) \geq 0 \\ x_1^{(0)}(1) + & x_1^{(0)}(2) + & x_2^{(0)}(1) & + x_3^{(0)}(1) & = 1; & x_3^{(0)}(1) \geq 0 \end{array} \right\}. \tag{7.23}$$

Optimal solutions of these programs are: for the primal $v^{(0)} = -\frac{4}{5},\ u_1^{(0)} = -\frac{6}{5},\ u_2^{(0)} = -\frac{4}{5},$ $u_3^{(0)} = 0$ and for the dual $x_1^{(0)}(1) = 0,\ x_1^{(0)}(2) = \frac{1}{5},\ x_2^{(0)}(1) = \frac{2}{5},\ x_3^{(0)}(1) = \frac{2}{5}.$
Hence, $A^0(1) = \{2\},\ A^0(2) = \{1\}$ and $A^0(3) = \{1\}$, i.e. f^∞ with $f(1) = 2,\ f(2) = f(3) = 1$ is the only bias optimal policy.

7.6.3 n-discount optimality

In this subsection we propose an algorithm for determining an n-discount optimal policy based on a system of nested linear programs. This is a generalization of the approach discussed in the preceding subsection. Let us first introduce the pair of dual linear programs for the computation of an n-discount optimal policy. The primal program is

$$min\Big\{ v^{(n)} \;\Big|\; v^{(n)} + \sum_j \{\delta_{ij} - p_{ij}(a)\} u_j^{(n)} \geq -u_i^{(n-1)},\ i \in S,\ a \in A^{n-1}(i) \Big\}, \tag{7.24}$$

where $(v^{(n-1)}, u^{(n-1)})$ is an optimal solution of the primal linear program for an $(n-1)$-discount optimal policy and $A^{n-1}(i) = \{a \in A^{n-2} \mid v^{(n-1)} + \sum_j \{\delta_{ij} - p_{ij}(a)\} u_j^{(n-1)} = -u^{(n-2)}\}$.
The dual linear program of (7.24) is

$$max\left\{ -\sum_i u_i^{(n-1)} \sum_{a \in A^{n-1}(i)} x_i^{(n)}(a) \;\middle|\; \begin{array}{rcl} \sum_i \sum_{a \in A^{n-1}(i)} \{\delta_{ij} - p_{ij}(a)\} x_i^{(n)}(a) & = & 0,\ j \in S \\ \sum_i \sum_{a \in A^{n-1}(i)} x_i^{(n)}(a) & = & 1 \\ x_i^{(n)}(a) & \geq & 0,\ i \in S,\ a \in A^{n-1}(i) \end{array} \right\}. \tag{7.25}$$

By induction on n we will show the following theorem.

Theorem 7.15
Let $\big(v^{(n)}, u^{(n)}\big)$ and $x^{(n)}$ be optimal solutions of the linear programs (7.24) and (7.25) respectively. Furthermore, let $A^{(n)} = \big\{ f^\infty \in A^{(n-1)} \mid v^{(n)} + \sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\} u_j^{(n)} = -u_i^{(n-1)},\ i \in S \big\}$.
(1) Let f^∞ be an $(n-1)$-discount optimal policy. Then, $u^n(f) = u^{(n-1)} - P^*(f)u^{(n-1)}$.
(2) f^∞ is an n-discount optimal policy if and only if $f^\infty \in A^{(n)}$.

Proof
For $n = 0$ we refer to the Theorems 7.13 and 7.14. We proof the induction step for $n \geq 1$.
(1) Let f^∞ be an $(n-1)$-discount optimal policy.

$u^n(f) = -D(f)u^{n-1}(f) = -D(f)\{u^{(n-2)} - P^*(f)u^{(n-2)}\}$, the last equality by induction.

Since $D(f)P^*(f) = 0$, we obtain $u^n(f) = -D(f)u^{(n-2)}$. Because $f^\infty \in A^{(n-1)}$, we can write $u^n(f) = D(f)\{v^{(n-1)} \cdot e + \{I - P(f)\}u^{(n-1)}\}$. Since $D(f)e = 0$ and $D(f)\{I - P(f)\} = I - P^*(f)$, it follows that $u^n(f) = u^{(n-1)} - P^*(f)u^{(n-1)}$.

(2) Let f^∞ be an n-discount optimal policy. Then, according to the induction assumption, f^∞ is average optimal for the MDP model with truncated action sets $A^{(n-1)}(i),\ i \in S$ and rewards $-u_i^{(n-1)}$. We know that there exists a one-to-one correspondence between the deterministic optimal policies and the extreme solutions of the dual program (7.25). Hence, there exists an extreme optimal solution $x^{(n)}$ such that $x_i^{(n)}(f(i)) > 0,\ i \in S$. Then, from the complementary slackness property of linear programming we conclude that

$$v^{(n)} + \textstyle\sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\}u_j^{(n)} = -u_i^{(n-1)},\ i \in S, \text{ i.e. } f^\infty \in A^{(n)}.$$

Conversely, let $f^\infty \in A^{(n)}$. Then, $v^{(n)} \cdot e + \{I - P(f)\}u^{(n)} = -u^{(n-1)}$, implying that $v^{(n)} \cdot e = -P^*(f)u^{(n-1)}$. For any $f^\infty \in A^{(n-1)}$, we derive from the primal program (7.24) that, for every $g^\infty \in A^{(n-1)}$, we have $v^{(n)} \cdot e + \{I - P(g)\}u^{(n)} \geq -u^{(n-1)}$.

Consequently, we have $v^{(n)} \cdot e \geq -P^*(g)u^{(n-1)}$, implying that

$$u^n(f) = u^{(n-1)} - P^*(f)u^{(n-1)} = u^{(n-1)} + v^{(n)} \cdot e \geq u^{(n-1)} - P^*(g)u^{(n-1)} = u^n(g)$$

for every $g^\infty \in A^{(n-1)}$, i.e. f^∞ is an n-discount optimal policy. □

7.7 Blackwell optimality and linear programming

In this section we will show how linear programming in the space of rational functions can be developed to compute optimal policies over the entire range of the discount factor. Furthermore, a procedure is presented for the computation of a Blackwell optimal policy.

Let $\mathbb{R}$ be the ordered field of the real numbers with the usual ordering. By $P(\mathbb{R})$ we denote the set of all polynomials in $x \in \mathbb{R}$ with real coefficients, i.e. the set of elements

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n \text{ where } a_i \in \mathbb{R},\ 1 \leq i \leq n \text{ for some positive integer } n, \tag{7.26}$$

where we assume that $a_n \neq 0$. The field $F(\mathbb{R})$ of rational functions with real coefficients consists of the elements

$$f(x) = \frac{p(x)}{q(x)}, \tag{7.27}$$

where p and q are elements of $P(\mathbb{R})$ having no common linear factors and q is not identically zero. So, each rational function is expressible in the form

$$f(x) = \frac{a_0 + a_1 x + \cdots + a_n x^n}{b_0 + b_1 x + \cdots + b_m x^m}. \tag{7.28}$$

The domain of a rational function consists of all but the finitely many real numbers where the denominator is 0. At these points, the numerator is nonzero, because there are no common linear

factors. So when we compare two rational functions f and g, we can be sure that the common domain $dom\, f \cap dom\, g$ consists of all but finitely many real numbers.

To complete the description of the field, we need specity the addition $(+)$ and multiplication $(\cdot)$ operations on the set $F(\mathbb{R})$, and we need to single out two members 0 and 1 as the 0 and 1 elements. The latter is easy: the elements 0 and 1 are the constant functions having the values 0 and 1, respectively. The operations $+$ and $\cdot$ in $F(\mathbb{R})$ are defined in the usual way:

$$\left(\frac{p}{q}+\frac{r}{s}\right)(x)=\frac{p(x)s(x)+r(x)q(x)}{q(x)s(x)} \text{ and } \left(\frac{p}{q}\cdot\frac{r}{s}\right)(x)=\frac{p(x)r(x)}{q(x)s(x)}, \tag{7.29}$$

with an additional operation to cancelling any common linear factors in the numerator and denominator. As example, let $f(x)=\frac{1}{-1+x^2}$, $g(x)=\frac{x}{-1+x^2}$ and $h(x)=1+x$. Then, $(f+g)(x)=\frac{1}{-1+x^2}+\frac{x}{-1+x^2}=\frac{1+x}{-1+x^2}=\frac{1}{-1+x}$ and $(f\cdot h)(x)=\frac{1}{-1+x^2}\cdot(1+x)=\frac{1+x}{-1+x^2}=\frac{1}{-1+x^2}$.

Next we need to augment this field with an ordering relation making it into an ordered field. We will denote the ordering relation by $>_l$. Since $f>_l g$ if and only if $f-g>_l 0$, it suffices to specify the set of positive rational functions. We have special interest in the value of these functions for x close to 0. Therefore, we define the *dominating coefficient* of a polynomial p given in formula (7.26) as the coefficient $a_k\neq 0$, where k is such that $a_i=0,\ 0\leq i\leq k-1$ (for the function 0, we define the dominating coefficient as 0). In this case we call k the *order* of p: $order(p)=k$. The dominating coefficient of polynomial p is denoted by $d(p)$. Notice that $d(p)=0$ if and only if $p=0$. Let P be the set of positive elements of $F(\mathbb{R})$, defined by

$$f(x)=\frac{p(x)}{q(x)}\in P \text{ if and only if } d(p)d(q)>0. \tag{7.30}$$

We now define $f\geq_l g$ if either $f>_l g$ or $f=g$. With this definition the field $F(\mathbb{R})$ is a total ordered field (the proof is left to the reader as Exercise 7.10). Observe that the function $f(x)=\frac{1}{x}$ is 'infinity large' in the sense that $\frac{1}{x}>_l n$ for any $n\in\mathbb{N}$. Similarly, the reciprocal function $g(x)=x$ is 'infinity small' in the sense that $x<_l \frac{1}{n}$ for any $n\in\mathbb{N}$. Hence, the field is a non-Archimedian ordered field.[1]

The continuity of polynomials implies that the rational function $f=\frac{p}{q}\in P$ if and only if $\frac{p(x)}{q(x)}>0$ for all x sufficiently near to 0. Hence, we obtain the following result.

Lemma 7.9
The rational function $f=\frac{p}{q}\in P$ if and only if there exists a positive real number x_0 such that $f(x)=\frac{p(x)}{q(x)}>0$ for all $x\in(0,x_0)$.

We will apply the above properties on discounted rewards as function of the discount factor α. As before, we will use the parameter $\rho=\frac{1-\alpha}{\alpha}$ instead of α. Note that $\alpha=\frac{1}{1+\rho}$ and that $\alpha\uparrow 1$ is

[1]For more details about ordered fields (Archimedian and non-Archimedian) see: B.L. van der Waerden, *Algebra - Erster Teil*, Springer-Verlag (1966) 235–238.

equivalent to $\rho \downarrow 0$. The total expected discounted reward $v^\rho(f^\infty)$ for a policy $f^\infty \in C(D)$ is the unique solution of the linear system

$$\{(1+\rho)I - P(f)\}x = (1+\rho)r(f). \tag{7.31}$$

Solving (7.31) by Cramer's rule shows that for every $i \in S$, the function $v_i^\rho(f^\infty)$ is an element of $F(\mathbb{R})$, say $v_i^\rho(f^\infty) = \frac{p(\rho)}{q(\rho)}$. The degree of the polynomials p and q is at most N. By Blackwell's theorem, we know that the interval $[0,1)$ of the discount factor α can be broken into a finite number of intervals, say $[0 = \alpha_{s+1}, \alpha_s)$, $[\alpha_s, \alpha_{s-1})$, $\ldots, [\alpha_0, \alpha_{-1} = 1)$, in such a way that there are policies f_k^∞, $k = 0, 1, \ldots, s+1$, where f_k^∞ is α-discounted optimal for all $\alpha \in [\alpha_k, \alpha_{k-1})$. The policy f_0^∞ is a Blackwell optimal policy. Observe that in each interval the components v_i^ρ of the value vector v^ρ are elements of $F(\mathbb{R})$. So, for small ρ corresponding with the interval $[\alpha_0, \alpha_{-1} = 1)$, i.e. $0 < \rho < \frac{1-\alpha_0}{\alpha_0}$, v_i^ρ is an element of $F(\mathbb{R})$.

The optimality equation of discounted rewards implies

$$(1+\rho)v_i^\rho \geq (1+\rho)r_i(a) + \sum_j p_{ij}(a)v_j^\rho, \ (i,a) \in S \times A, \ \rho > 0. \tag{7.32}$$

Since v_i^ρ is an element of $F(\mathbb{R})$ for $\rho \in \left[0, \frac{1-\alpha_0}{\alpha_0}\right)$,we obtain from (7.32) the ordering relations

$$(1+\rho)v_i^\rho \geq_l (1+\rho)r_i(a) + \sum_j p_{ij}(a)v_j^\rho, \ (i,a) \in S \times A. \tag{7.33}$$

An N-vector $w(\rho)$ with elements in $F(\mathbb{R})$ is called *Blackwell-superharmonic* if

$$(1+\rho)w_i(\rho) \geq_l (1+\rho)r_i(a) + \sum_j p_{ij}(a)w_j(\rho), \ (i,a) \in S \times A. \tag{7.34}$$

Theorem 7.16

The discounted value vector v^ρ is the (componentwise) smallest Blackwell-superharmonic vector with components in $F(\mathbb{R})$, i.e. for any Blackwell-superharmonic vector $w(\rho)$, we have $w_i(\rho) \geq_l v_i^\rho$, $i \in S$.

Proof

From (7.33) it follows that the discounted value vector v^ρ is a Blackwell-superharmonic vector. Suppose that $w(\rho)$ is an arbitrary Blackwell-superharmonic vector. Since there are only a finite number of elements in $S \times A$ it follows from Lemma 7.9 that there exists a positive real number ρ_0 such that

$$(1+\rho)w_i(\rho) \geq (1+\rho)r_i(a) + \sum_j p_{ij}(a)w_j(\rho), \ (i,a) \in S \times A, \ \rho \in [0, \rho_0).$$

Hence, for every $\alpha \in \left[\frac{1}{1+\rho_0}\right)$ the vector $w(\rho)$ is α-superharmonic in the sense of (3.30). Therefore, by the results of discounted rewards in Chapter 3, $w_i(\rho) \geq v_i^\rho$, $i \in S$, for all $\rho \in [0, \rho_0)$. Consequently, $w_i(\rho) \geq_l v_i^\rho$, $i \in S$, $\rho \in [0, \rho_0)$. □

Theorem 7.16 implies that the value vector v^ρ for the interval $[0, \rho_0)$ can be found as optimal solution of the following linear program in $F(\mathbb{R})$:

$$min\Big\{ \sum_j w_j(\rho) \;\Big|\; (1+\rho)w_i(\rho) \geq_l (1+\rho)r_i(a) + \sum_j p_{ij}(a)w_j(\rho),\ (i,a) \in S \times A \Big\}. \tag{7.35}$$

Consider also the following linear program in $F(\mathbb{R})$, called the *dual program*:

$$max \left\{ \sum_{(i,a)} (1+\rho)r_i(a) \cdot x_{ia}(\rho) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{(1+\rho)\delta_{ij} - p_{ij}(a)\} \cdot x_{ia}(\rho) & = & 1,\ j \in S \\ x_{ia}(\rho) & \geq_l & 0,\ (i,a) \in S \times A \end{array} \right\}. \tag{7.36}$$

For a fixed positive ρ, the linear programs (7.35) and (7.36) are equivalent to the linear programs of Chapter 3. Therefore, we also have for each fixed ρ a one-to-one correspondence between the basic feasible solutions and the policies of $C(D)$. For the present programs with elements from $F(\mathbb{R})$ we will, as in the simplex method with real numbers, rewrite the equalities

$$\textstyle\sum_{(i,a)} \{(1+\rho)\delta_{ij} - p_{ij}(a)\} \cdot x_{ia}(\rho) = 1,\ j \in S,$$

such that at each iteration there is precisely one positive $x_{ia}(\rho)$ for each state i. Hence, the only difference with the usual simplex method with real numbers is that instead of real numbers the elements in the programs are rational functions.

At any iteration there is an extreme feasible solution $x(\rho)$ of (7.36), correspomnding to a policy f^∞, and a reduced cost vector $w(\rho)$ such that the complementary slackness property is satisfied, i.e.

$$x_{ia}(\rho) \cdot \Big\{ \sum_j \{(1+\rho)\delta_{ij} - p_{ij}(a)\} \cdot w_j(\rho) - (1+\rho)r_i(a) \Big\} = 0,\ (i,a) \in S \times A.$$

Since $x_{jf(j)}(\rho) > 0,\ j \in S$, we have $\sum_j \{(1+\rho)\delta_{ij} - p_{ij}(f)\} \cdot w_j(\rho) = (1+\rho)r_i\big(f(i)\big),\ i \in S$. Hence, $w(\rho) = v^\rho(f^\infty)$. The validation of the above decribed approach and some additional properties follow from the following lemma.

Lemma 7.10

(1) The elements in the simplex tableau can be written as rational functions with the same denominator, say $n(\rho)$, which is the product of the previous pivot elements.

(2) The numerators and common denominator are polynomials with degree at most N; the the numerators of the reduced costs are polynomials with degree at most $N+1$.

(3) The pivot operations in the simplex tableau are as follows:

a. The new common denominator is the numerator of the current pivot element.

b. The new numerator of the pivot element is the current common denominator.

c. The new numerators of the other elements in the pivot row are unchanged.

d. The new numerators of the other elements in the pivot column are the old numerator multiplied by -1.

e. For the other elements, say an element with numerator $p(\rho)$, the new numerator becomes $\frac{p(\rho)t(\rho)-r(\rho)s(\rho)}{n(\rho)}$, which is a polynomial, where $t(\rho)$ is the numerator of the old pivot element, $r(\rho)$ is the numerator of the element in the pivot row and the same column as the element with numerator $p(\rho)$, and $s(\rho)$ is the numerator of the element in the pivot column and the same row as the element with numerator $p(\rho)$.

Proof

(1) We can compute a simplex tableau corresponding to some policy f^∞ as follows. Let $z_j(\rho),\ j \in S$ be artificial variables, i.e. consider the system

$$\sum_{(i,a)} \{(1+\rho)\delta_{ij} - p_{ij}(a)\} \cdot x_{ia}(\rho) + z_j(\rho) = 1,\ j \in S.$$

Then exchange by the usal pivot operations $z_j(\rho)$ with $x_{jf(j)}(\rho)$ for $j = 1, 2, \ldots, N$. The first basis matrix is the identity matrix I corresponding to the artificial variables. Hence, in the first simplex tableau the elements are polynomials (in fact linear functions) in ρ, which may be considered as rational functions with common denominator 1. It is well known from the theory of linear programming (see e.g. [241]) that the elements of a simplex tableau have a common denominator, namely the determinant of the basis matrix which is the product of all previous pivot elements when the first basis matrix is the identity matrix. This result, with a similar proof, is also valid of the elements are rational functions instead of real numbers.

(2) Any basis matrix is of the form $(1+\rho)I - P(f)$. So it has linear functions on the diagonal and constants on the off-diagonal elements. Hence, the determinant of the matrix is a polynomial with degree at most N. By Cramer's rule, the elements of the inverse have numerators which are polynomials with degree at most $N-1$. The elements in a column of the simplex tableau, except the reduced costs, are obtained by multiplication of the inverse of the basis matrix with the right hand side or a nonbasic column. Such columns are constants or linear functions. Hence, the polynomials of the numerators have degree at most N. Since the reduced costs are (rewritten) terms of the objective function $\sum_{(i,a)} (1+\rho) r_i(a) \cdot x_{ia}(\rho)$, they are obtained by multiplying the variables $x_{ia}(\rho)$ with a linear function, so these numerators have degree at most $N+1$.

(3) The transformation rules for the simplex method with rational functions are similar to the transformations rules in case of real numbers. Let $\frac{t(\rho)}{n(\rho)}$ be the pivot element. Then $n(\rho)$ is the product of all previous pivots. The new product of the pivots is $n(\rho) \cdot \frac{t(\rho)}{n(\rho)} = t(\rho)$.
The rules b, c and d are straightforward. Consider an element outsite the pivot row or pivot column, say $\frac{p(\rho)}{n(\rho)}$. Let $\frac{r(\rho)}{n(\rho)}$ be the element in the pivot row and the same column as the element $\frac{p(\rho)}{n(\rho)}$ and let $\frac{s(\rho)}{n(\rho)}$ be the element in the pivot column and the same row as the element $\frac{p(\rho)}{n(\rho)}$. Then the new element becomes: $\frac{p(\rho)}{n(\rho)} - \frac{r(\rho)}{n(\rho)} \cdot \frac{s(\rho)}{n(\rho)} \cdot \frac{n(\rho)}{t(\rho)} = \frac{1}{t(\rho)} \cdot \left\{ \frac{p(\rho) \cdot t(\rho) - r(\rho) \cdot s(\rho)}{n(\rho)} \right\}$.
From the property that $t(\rho)$ is the common denominator of the new tableau it follows that $\frac{p(\rho) \cdot t(\rho) - r(\rho) \cdot s(\rho)}{n(\rho)}$ is a polynomial. □

We shall solve the dual program (7.36) starting with ρ very large, or equivalently α very close to 0. For $\alpha = 0$ the policy f^∞ such that $r_i\big(f(i)\big) = max_a\, r_i(a),\ i \in S$ is an optimal policy. We start with the basic solution corresponding to this policy f^∞. We can compute the first feasible simplex tableau as follows. Let $z_j(\rho),\ j \in S$ be the artificial variales, i.e. we consider the system

$$\sum_{(i,a)} \{(1+\rho)\delta_{ij} - p_{ij}(a)\} \cdot x_{ia}(\rho) + z_j(\rho) = 1,\ j \in S.$$

Then we exchange by the usal pivot operations $z_j(\rho)$ with $x_{jf(j)}(\rho)$ for $j = 1, 2, \ldots, N$. This tableau is optimal for all $\rho \geq \rho_*$, where ρ_* is the smallest ρ for which the reduced costs (the reduced costs are also elements of $F(\mathbb{R})$) are nonnegative. To compute ρ_* we have to compute the zeros of some polynomials. [2] The reduced cost that determines ρ_* determines the next pivot column. After a pivot operation we repeat this approach to obtain a next interval on which the new policy is optimal. In this way we continue until we have an interval that ends with $\rho_* = 0$. That final interval corresponds to a Blackwell optimal policy. This approach determines optimal policies over the entire range $[0, 1)$ of the discount factor α.

Example 7.4

$S = \{1, 2, 3\}:\ A(1) = A(2) = A(3) = \{1, 2\}$.
$r_1(1) = 8,\ r_1(1) = \frac{11}{4};\ r_2(1) = 16,\ r_2(2) = 15;\ r_3(1) = 7,\ r_3(2) = 4..$
$p_{11}(1) = \frac{1}{2},\ p_{12}(1) = \frac{1}{4},\ p_{13}(1) = \frac{1}{4};\ p_{11}(2) = \frac{1}{16},\ p_{12}(2) = \frac{3}{4},\ p_{13}(2) = \frac{3}{16};$
$p_{21}(1) = \frac{1}{2},\ p_{22}(1) = 0,\ p_{23}(1) = \frac{1}{2};\ p_{21}(2) = \frac{1}{16},\ p_{22}(2) = \frac{7}{8},\ p_{23}(2) = \frac{1}{16};$
$p_{31}(1) = \frac{1}{4};\ p_{32}(1) = \frac{1}{4};\ p_{33}(1) = \frac{1}{2};\ p_{31}(2) = \frac{1}{8};\ p_{32}(2) = \frac{3}{4};\ p_{33}(2) = \frac{1}{8}.$
For this example the objective function becomes:
$(1+\rho) \cdot \{8x_{11}(\rho) + \frac{11}{4}x_{12}(\rho) + 16x_{21}(\rho) + 15x_{22}(\rho) + 7x_{31}(\rho) + 4x_{32}(\rho)\}$
The linear constraints are:

$$\begin{array}{ccccccccccccccccc}
(\frac{1}{2}+\rho)x_{11}(\rho) & + & (\frac{15}{16}+\rho)x_{12}(\rho) & - & \frac{1}{2}x_{21}(\rho) & - & \frac{1}{16}x_{22}(\rho) & - & \frac{1}{4}x_{31}(\rho) & - & \frac{1}{8}x_{32}(\rho) = 1 \\
- & \frac{1}{4}x_{11}(\rho) & - & \frac{3}{4}x_{12}(\rho) & + & (1+\rho)x_{21}(\rho) & + & (\frac{1}{8}+\rho)x_{22}(\rho) & - & \frac{1}{4}x_{31}(\rho) & - & \frac{3}{4}x_{32}(\rho) = 1 \\
- & \frac{1}{4}x_{11}(\rho) & - & \frac{3}{16}x_{12}(\rho) & - & \frac{1}{2}x_{21}(\rho) & - & \frac{1}{16}x_{22}(\rho) & + & (\frac{1}{2}+\rho)x_{31}(\rho) & + & (\frac{7}{8}+\rho)x_{32}(\rho) = 1
\end{array}$$

[2] The computation of real zeros of polynomials can be done by Maple. We refer also to the literature on numerical analysis and to [94] in which paper a method based on Sturm's Theorem is discussed.

The first simplex tableau is (the common denominator is the top element in the second column of the simplex tableau; the first common denominator is 1).

	1	$x_{11}(\rho)$	$x_{12}(\rho)$	$x_{21}(\rho)$	$x_{22}(\rho)$	$x_{31}(\rho)$	$x_{32}(\rho)$
$z_1(\rho)$	1	$+\frac{1}{2}+\rho$	$+\frac{15}{16}+\rho$	$-\frac{1}{2}$	$-\frac{1}{16}$	$-\frac{1}{4}$	$-\frac{1}{8}$
$z_2(\rho)$	1	$-\frac{1}{4}$	$-\frac{3}{4}$	$+1+\rho$	$+\frac{1}{8}+\rho$	$-\frac{1}{4}$	$-\frac{3}{4}$
$z_3(\rho)$	1	$-\frac{1}{4}$	$-\frac{3}{16}$	$-\frac{1}{2}$	$+\frac{1}{16}$	$+\frac{1}{2}+\rho$	$+\frac{7}{8}+\rho$
	0	$-8-8\rho$	$-\frac{11}{4}-\frac{11}{4}\rho$	$-16-16\rho$	$-15-15\rho$	$-7-7\rho$	$-4-4\rho$

In the first iteration the pivot column is the column of the variable $x_{11}(\rho)$ and the pivot row is the row of $z_1(\rho)$. The next tableau becomes (with common denominator $\frac{1}{2}+\rho$).

	$\frac{1}{2}+\rho$	$z_1(\rho)$	$x_{12}(\rho)$	$x_{21}(\rho)$	$x_{22}(\rho)$	$x_{31}(\rho)$	$x_{32}(\rho)$
$x_{11}(\rho)$	1	+1	$+\frac{15}{16}+\rho$	$-\frac{1}{2}$	$-\frac{1}{16}$	$-\frac{1}{4}$	$-\frac{1}{8}$
$z_2(\rho)$	$\frac{3}{4}+\rho$	$+\frac{1}{4}$	$-\frac{9}{64}-\frac{1}{2}\rho$	$+\frac{3}{8}+\frac{3}{2}+\rho^2$	$+\frac{3}{16}+\frac{5}{8}\rho+\rho^2$	$-\frac{3}{16}-\frac{1}{4}\rho$	$-\frac{13}{16}-\frac{3}{4}\rho$
$z_3(\rho)$	$\frac{3}{4}+\rho$	$+\frac{1}{4}$	$+\frac{9}{64}+\frac{1}{16}\rho$	$-\frac{3}{8}-\frac{1}{2}\rho$	$-\frac{3}{16}-\frac{1}{16}\rho$	$+\frac{3}{16}+\rho+\rho^2$	$-\frac{3}{16}+\frac{11}{8}\rho+\rho^2$
	$8+8\rho$	$8+8\rho$	$+\frac{49}{8}+\frac{91}{8}\rho+\frac{21}{4}\rho^2$	$-12-28\rho-16\rho^2$	$-8-23\rho-15\rho^2$	$-\frac{11}{2}-\frac{25}{2}\rho-7\rho^2$	$-3-7\rho-4\rho^2$

After inserting $x_{21}(\rho)$ and $x_{31}(\rho)$ into the basis the next tableau is obtained, which is the first feasible tableau in which the common denominator is $\frac{15}{16}\rho+2\rho^2+\rho^3$.

	$\frac{15}{16}\rho+2\rho^2+\rho^3$	$z_1(\rho)$	$x_{12}(\rho)$	$z_2(\rho)$
$x_{11}(\rho)$	$\frac{9}{8}+\frac{9}{4}\rho+\rho^2$	$+\frac{3}{8}+\frac{3}{2}\rho+\rho^2$	$+\frac{87}{67}\rho+\frac{39}{16}\rho^2+\rho^3$	$+\frac{3}{8}+\frac{1}{2}\rho$
$x_{21}(\rho)$	$\frac{9}{16}+\frac{3}{2}\rho+\rho^2$	$+\frac{3}{16}+\frac{1}{4}\rho$	$-\frac{3}{8}\rho-\frac{1}{2}\rho^2$	$+\frac{3}{16}+\rho+\rho^2$
$x_{31}(\rho)$	$\frac{9}{8}+\frac{9}{4}\rho+\rho^2$	$+\frac{3}{8}+\frac{1}{4}\rho$	$-\frac{3}{64}\rho+\frac{1}{16}\rho^2$	$+\frac{3}{8}+\frac{1}{2}\rho$
	$\frac{207}{8}+\frac{669}{8}\rho+\frac{355}{4}\rho^2+31\rho^3$	$+\frac{69}{8}+\frac{221}{8}\rho+\frac{103}{4}\rho^2+8\rho^3$	$+\frac{63}{32}\rho+\frac{269}{32}\rho^2+\frac{187}{16}\rho^3+\frac{21}{4}\rho^4$	$+\frac{69}{8}+\frac{257}{8}\rho+\frac{79}{2}\rho^2+16\rho^3$

$x_{22}(\rho)$	$z_3(\rho)$	$x_{32}(\rho)$
$+\frac{21}{64}\rho+\frac{7}{16}\rho^2$	$+\frac{3}{8}+\frac{1}{4}\rho$	$+\frac{1}{32}\rho+\frac{1}{8}\rho^2$
$+\frac{9}{32}\rho+\frac{9}{8}\rho^2+\rho^3$	$+\frac{3}{16}+\frac{1}{4}\rho$	$-\frac{3}{8}\rho-\frac{1}{2}\rho^2$
$+\frac{21}{64}\rho+\frac{7}{16}\rho^2$	$+\frac{3}{8}+\frac{3}{2}\rho+\rho^2$	$+\frac{41}{32}\rho+\frac{19}{8}\rho^2+\rho^3$
$-\frac{297}{64}\rho-\frac{645}{64}\rho^2-\frac{71}{16}\rho^3+\rho^4$	$+\frac{69}{8}+\frac{201}{8}\rho+\frac{47}{2}\rho^2+7\rho^3$	$-\frac{17}{32}\rho+\frac{35}{32}\rho^2+\frac{37}{8}\rho^3+3\rho^4$

It can be shown that this tableau is optimal for $\rho \in [6.18, \infty)$, or equivalently, $\alpha \in [0, 0.14]$. For $\rho < 6.18$ the column of $x_{22}(\rho)$ becomes the next pivot column. The next tableau, after exchanging $x_{22}(\rho)$ with $x_{21}(\rho)$, is optimal for $\rho \in [0.91, 6.18]$, or equivalently, $\alpha \in [0.14, 0.52]$. For $\rho < 0.91$ the column of $x_{32}(\rho)$ becomes the next pivot column. The next tableau, after exchanging $x_{32}(\rho)$ with $x_{31}(\rho)$, is optimal for $\rho \in [0.27, 0.91]$, or equivalently, $\alpha \in [0.52, 0.79]$. Finally, for $\rho < 0.27$ the column of $x_{12}(\rho)$ becomes the pivot column. The final tableau, after exchanging $x_{12}(\rho)$ with $x_{11}(\rho)$, is optimal for $\rho \in (0, 0.27]$, or equivalently, $\alpha \in [0.79, 1)$. Hence, we have the following results.

$\alpha \in [0, 0.14]$: $f(1) = 1,\ f(2) = 1,\ f(3) = 1$ is the optimal policy.
$\alpha \in [0.14, 0.52]$: $f(1) = 1,\ f(2) = 2,\ f(3) = 1$ is the optimal policy.
$\alpha \in [0.52, 0.79]$: $f(1) = 1,\ f(2) = 2,\ f(3) = 2$ is the optimal policy.
$\alpha \in [0.79, 1)$: $f(1) = 2,\ f(2) = 2,\ f(3) = 2$ is the optimal policy.

The policy f^∞ with $f(1) = f(2) = 2$ is Blackwell optimal.
For the details of this example we refer to [94].

7.8 Bias optimality and linear programming

7.8.1 The general case

In order to compute a bias optimal policy by linear programming, we first solve the linear programs for an average optimal policy. So, we compute optimal solutions $(v^* = \phi, u^*)$ and (x^*, y^*) of the pair of dual linear programs

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{rcll} \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j & \geq & 0, & (i,a) \in S \times A \\ v_i + \sum_j \big(\delta_{ij} - p_{ij}(a)\big) u_j & \geq & r_i(a), & (i,a) \in S \times A \end{array} \right\} \tag{7.37}$$

and

$$max \left\{ \sum_{(i,a)} r_i(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_a x_j(a) \ + \ \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} y_i(a) & = & \beta_j,\ j \in S \\ x_i(a), y_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}, \tag{7.38}$$

where $\beta_j > 0,\ j \in S$, is arbitrarily chosen. Let

$$\begin{aligned} &A_1(i) = \{a \in A(i) \mid \textstyle\sum_j \{\delta_{ij} - p_{ij}(a)\} \phi_j = 0\},\ i \in S; \\ &A_2(i) = \{a \in A_1(i) \mid \phi_i + \textstyle\sum_j \big(\delta_{ij} - p_{ij}(a)\big) u_j^* = r_i(a)\},\ i \in S; \\ &S_1 = \{i \in S \mid A_1(i) \neq \emptyset\};\quad S_2 = \{i \in S \mid A_2(i) \neq \emptyset\}. \end{aligned}$$

Any policy $f^\infty \in C(D)$ induces a Markov chain $P(f)$. Let $R(f)$ and $T(f)$ be the sets of recurrent and transient states, respectively.

Lemma 7.11

Let $f^\infty \in C(D)$ be an average optimal policy. Then,

(1) $f(i) \in A_1(i),\ i \in S$ and $S_1 = S$.

(2) $f(i) \in A_2(i),\ i \in R(f)$.

(3) $u_i^0(f) = u_i^* - \{P^*(f)u^*\}_i,\ i \in R(f)$.

(4) $u_i^0(f) \leq u_i^* - \{P^*(f)u^*\}_i,\ i \in T(f)$.

Proof

(1) $P(f)\phi = P(f)P^*(f)r(f) = P^*(f)r(f) = \phi$. Consequently, $A_1(i) \neq \emptyset,\ i \in S$, i.e. $S_1 = S$.

(2) From Theorem 5.18 it follows that (x^f, y^f), defined by (5.30) and (5.31), is an optimal solution of the dual program (5.24). In the proof of Theorem 5.18 is shown that $R(f) = S_x = \{i \mid x_i^f\big(f(i)\big) > 0\}$ and that $f(i) \in A_2(i),\ i \in S_x$ (see (5.32)).

(3) Since $d_{ij}(f) = 0,\ i \in R(f),\ j \in T(f)$ (see section 5.3 for the structure of the deviation matrix), it follows from part (2) that
$u^0(f)_i = \{D(f)r(f)\}_i = \big\{D(f)\{\phi + \big(I - P(f)\big)u^*\}\big\}_i = \{\big(I - P^*(f)\big)u^*\}_i,\ i \in R(f)$.

(4) Since $d_{ij}(f) \geq 0,\ i,j \in T(f)$ (see section 5.3 for the structure of the deviation matrix), we obtain

$$d_{ij}(f)\{\phi_j + \textstyle\sum_k \big(\delta_{jk} - p_{jk}(f)\big)u_k^*\} \geq d_{ij}(f)r_j(f),\ i,j \in T(f).$$

Part (2) of this lemma implies

$$d_{ij}(f)\{\phi_j + \textstyle\sum_k \big(\delta_{jk} - p_{jk}(f)\big)u_k^*\} = d_{ij}(f)r_j(f),\ i \in T(f),\ j \in R(f).$$

Hence,

$$u_i(f) = \{D(f)r(f)\}_i \leq \big\{D(f)\{\phi + \big(I - P(f)\big)u^*\}\big\}_i = \{\big(I - P^*(f)\big)u^*\}_i,\ i \in T(f). \qquad \square$$

A bias optimal policy is average optimal policy which maximizes $u^0(f)$ over the set of average optimal policies. Lemma 7.11 shows that maximizing $u^0(f)$ over the average optimal policies is, for the states $i \in R(f) \subseteq S_2$, maximizing $-\{P^*(f)u^*\}_i$. Notice that $-P^*(f)u^*$ is the average reward for rewards $\overline{r}_i(a) = -u_i^*$ for all (i,a). Lemma 7.11 also shows that $f(i) \in A_2(i),\ i \in R(f)$ for all average optimal policies. Since the states $S\backslash S_2$ are transient under all average optimal policies, we consider a modified MDP with state space S_2 and action sets $A_2(i),\ i \in S_2$. In order to have a correct MDP we have to remove in the states $i \in S_2$ the actions $a \in A_2(i)$ for which $p_{ij}(a) > 0$ for at least one state $j \in S\backslash S_2$. Since $R(f)$ is a closed set for all average optimal policies f^∞, actions corresponding to average optimal policies are not removed. Furthermore, all policies f^∞ in the modified MDP satisfy $\phi = P(f)\phi$ and $\phi + \{I - P(f)\}u^* = r(f)$. Hence,

$$\phi = P^*(f)r(f) = \phi(f^\infty) \text{ and } u^0(f) = D(f)r(f) = u^* - P^*(f)u^*. \tag{7.39}$$

In the next algorithm the states and actions of the modified MDP are constructed.

Algorithm 7.2 *Construction of the modified MDP*

1. If $p_{ij}(a) = 0$ for all $i \in S_2,\ a \in A_2(i),\ j \in S\backslash S_2$: STOP.

   Otherwise: go to step 2.

2. a. Take some $i \in S_2,\ a \in A_2(i),\ j \in S\backslash S_2$ satisfying $p_{ij}(a) > 0$.

   b. $A_2(i) := A_2(i)\backslash\{a\}$.

   c. If $A_2(i) = \emptyset$, then $S_2 := S_2\backslash\{i\}$.

   d. Return to step 1.

The linear programs for an average optimal policy in the modified MDP are

$$min \left\{ \sum_j \beta_j w_j \;\middle|\; \begin{array}{rcll} \sum_j \{\delta_{ij} - p_{ij}(a)\}w_j & \geq & 0, & (i,a) \in S_2 \times A_2 \\ w_i + \sum_j \big(\delta_{ij} - p_{ij}(a)\big)z_j & \geq & -u_i^*, & (i,a) \in S_2 \times A_2 \end{array} \right\} \tag{7.40}$$

and

$$max \left\{ \sum_{(i,a)} (-u_i^*)t_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}t_i(a) & = & 0,\ j \in S_2 \\ \sum_a t_j(a) + \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}s_i(a) & = & \beta_j,\ j \in S_2 \\ t_i(a), s_i(a) & \geq & 0,\ (i,a) \in S_2 \times A_2 \end{array} \right\}. \tag{7.41}$$

Let (w^*, z^*) be an optimal solution of program (7.40) and let (t^*, s^*) be an extreme optimal solution of the program (7.41). From Chapter 5 we know that $w^* = \phi^*$, the value vector of the modified MDP, and that f_*^∞ satisfying $t_i^*\big(f_*(i)\big) > 0$ if $\sum_a t_i^*(a) > 0$ and $s_i^*\big(f_*(i)\big) > 0$ if $\sum_a t_i^*(a) = 0$ is an average optimal policy for the modified MDP.

In the sequel we will show (see Lemma 7.14) that f_*^∞ is a bias optimal policy for every state i which is recurrent under at least one bias optimal policy. Unfortunatly, this set of states is unknown. Furthermore, we have to determine a policy which is also bias optimal in the other states. Therefore, we use the following observation. Let f^∞ be an average optimal policy and suppose that the value vector of the average rewards is the 0-vector. Then, the α-discounted reward vector $v^\alpha(f^\infty)$ satisfies by the Laurent series expansion $v^\alpha(f^\infty) = u^0(f) + \varepsilon(\alpha)$, where $lim_{\alpha\uparrow 1}\, \varepsilon(\alpha) = 0$. Hence, the bias term $u^0(f)$ may be considered as $lim_{\alpha\uparrow 1}\, v^\alpha(f^\infty)$, the total expected reward. Notice that the value vector of the average reward is the 0-vector if we use $r_i(a) - \phi$ as immediate reward instead of $r_i(a)$ for all (i, a). We will also show (see Lemma 7.15) that for a bias optimal policy f^∞ we have $u_i^0(f) \geq u_i^* + \phi_i^*,\ i \in S_2$. As we have seen in Chapter 4 we can use for total rewards the linear programming formulation for discounted rewards with $\alpha = 1$. We also include in the linear program the inequalities that the total reward is at least $u_i^* + \phi_i^*,\ i \in S_2$. Finally, from Lemma 7.11 part (1) it follows that only actions of $A_1(i),\ i \in S$, may be considered. By these observations the third set of linear programs are

$$min \left\{ \sum_j \beta_j g_j \;\middle|\; \begin{array}{rcll} \sum_j \{\delta_{ij} - p_{ij}(a)\} g_j & \geq & r_i(a) - \phi_i, & (i,a) \in S \times A_1 \\ g_i & \geq & u_i^* + \phi_i^* & i \in S_2 \end{array} \right\} \tag{7.42}$$

and

$$max \left\{ \begin{array}{l} \sum_{(i,a)\in S\times A_1} (r_i(a) - \phi_i) q_i(a) \\ + \sum_{i\in S_2} (u_i^* + \phi_i^*) h_i \end{array} \;\middle|\; \begin{array}{rcll} \sum_{(i,a)\in S\times A_1} \{\delta_{ij} - p_{ij}(a)\} q_i(a) + \sum_{i\in S_2} \delta_{ij} h_i = & & \beta_j, & j \in S \\ q_i(a) \geq & & 0, & (i,a) \in S \times A_1 \\ h_i \geq & & 0, & i \in S_2 \end{array} \right\}. \tag{7.43}$$

Combining the above observation result in the following algorithm.

Algorithm 7.3 *Determination of a bias optimal policy by linear programming*

1. Compute an optimal solution $(v^* = \phi, u^*)$ of linear program (7.37).

2. Determine the following sets:

   $A_1(i) = \{a \in A(i) \mid \sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j = 0\},\ i \in S;$

   $A_2(i) = \{a \in A_1(i) \mid \phi_i + \sum_j \big(\delta_{ij} - p_{ij}(a)\big) u_j^* = r_i(a)\},\ i \in S;$

   $S_2 = \{i \in S \mid A_2(i) \neq \emptyset\}.$

3. Determine the modified MDP with state space S_2 and action sets $A_2(i)$ by Algorithm 7.2.

4. a. Compute an optimal solution $(w^* = \phi^*, z^*)$ of linear program (7.40) and an extreme

optimal solution (t^*, s^*) of (7.41).

b. Take f_*^∞ such that $t_i^*\big(f_*(i)\big) > 0$ if $\sum_a t_i^*(a) > 0$ and $s_i^*\big(f_*(i)\big) > 0$ if $\sum_a t_i^*(a) = 0$.

5. a. Compute an optimal solution g^* of linear program (7.42) and an extreme optimal solution (q^*, h^*) of linear program (7.43).

b. Determine $S_* = \{i \in S_2 \mid g_i^* = u_i^* + \phi_i^*\}$.

c. Take policy f^∞ such that $f(i) = \begin{cases} f_*(i) & i \in S_* \\ q_i^*\big(f(i)\big) > 0 & \in S \backslash S_* \end{cases}$

d. f^∞ is a bias optimal policy (STOP).

Example 7.5

$E = \{1, 2, 3, 4\}$; $A(1) = A(2) = A(3) = A(4) = \{1, 2\}$. Take $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \frac{1}{4}$.

$r_1(1) = 2,\ r_1(2) = 3;\ r_2(1) = 0,\ r_2(2) = 2;\ r_3(1) = 0,\ r_3(2) = -5;\ r_4(1) = 4,\ r_4(2) = 1.$

$p_{12}(1) = p_{13}(2) = p_{21}(1) = p_{23}(2) = p_{32}(1) = p_{34}(2) = p_{41}(1) = p_{43}(1) = 1.$

The linear program (7.37) is:

$min\{\frac{1}{4}v_1 + \frac{1}{4}v_2 + \frac{1}{4}v_3 + \frac{1}{4}v_4\}$

subject to

$$
\begin{array}{rcrcrcrcrcrcrcrcl}
 & v_1 & - & v_2 & & & & & & & & & & & \geq & 0 \\
 & v_1 & & & - & v_3 & & & & & & & & & \geq & 0 \\
- & v_1 & + & v_2 & & & & & & & & & & & \geq & 0 \\
 & & & v_2 & - & v_3 & & & & & & & & & \geq & 0 \\
 & & - & v_2 & + & v_3 & & & & & & & & & \geq & 0 \\
 & & & & & v_3 & - & v_4 & & & & & & & \geq & 0 \\
- & v_1 & & & & & + & v_4 & & & & & & & \geq & 0 \\
 & & & & - & v_3 & + & v_4 & & & & & & & \geq & 0 \\
 & v_1 & & & & & & & + & u_1 & - & u_2 & & & \geq & 2 \\
 & v_1 & & & & & & & + & u_1 & & & - & u_3 & \geq & 3 \\
 & & & v_2 & & & & & - & u_1 & + & u_2 & & & \geq & 0 \\
 & & & v_2 & & & & & & & + & u_2 & - & u_3 & \geq & 2 \\
 & & & & & v_3 & & & & & - & u_2 & + & u_3 & \geq & 0 \\
 & & & & & v_3 & & & & & - & u_2 & + & u_3 & \geq & -5 \\
 & & & & & & & v_4 & - & u_1 & & & & + \ u_4 & \geq & 4 \\
 & & & & & & & v_4 & & & & & - & u_3 + u_4 & \geq & 1
\end{array}
$$

An optimal solution is: $v_1^* = \phi_1 = v_2^* = \phi_2 = v_3^* = \phi_3 = v_4^* = \phi_4 = 1$ (unique) and

$u_1^* = 2,\ u_2^* = 1,\ u_3^* = 0,\ u_4^* = 6$ (not unique).

$A_1(1) = A_1(2) = A_1(3) = A_4(1) = \{1, 2\};\ A_2(1) = A_2(2) = A_2(3) = \{1, 2\},\ A_2(4) = \emptyset.$

$S_1 = \{1, 2, 3, 4\},\ S_2 = \{1, 2, 3\}.$

Modified MDP: $S_2 = \{1,2,3\};\ A_2(1) = A_2(2) = \{1,2\},\ A_2(3) = \{1\}$.

The primal program (7.40) is:

$min\{\frac{1}{4}w_1 + \frac{1}{4}w_2 + \frac{1}{4}w_3\}$

subject to

$$\begin{array}{rcrcrcrcrcrcl}
w_1 & - & w_2 & & & & & & & & \geq & 0 \\
w_1 & & & - & w_3 & & & & & & \geq & 0 \\
- w_1 & + & w_2 & & & & & & & & \geq & 0 \\
& & w_2 & - & w_3 & & & & & & \geq & 0 \\
& - & w_2 & + & w_3 & & & & & & \geq & 0 \\
w_1 & & & & & + & z_1 & - & z_2 & & \geq & -2 \\
w_1 & & & & & + & z_1 & & & - z_3 & \geq & -2 \\
& & w_2 & & & - & z_1 & + & z_2 & & \geq & -1 \\
& & w_2 & & & & & + & z_2 & - z_3 & \geq & -1 \\
& & & & w_3 & & & - & z_2 & + z_3 & \geq & 0
\end{array}$$

An optimal solution is: $w_1^* = \phi_1^* = w_2^* = \phi_2^* = w_3^* = \phi_3^* = -\frac{1}{2}$ (unique) and $z_1^* = \frac{1}{2},\ z_2^* = 0,\ z_3^* = \frac{1}{2}$ (not unique). The dual program (7.41) becomes (without the nonnegativity constraints):

$max\{-2t_1(1) - 2t_1(2) - t_2(1) - t_2(2)\}$

subject to

$$\begin{array}{rcrcrcrcrcrcrcrcrcrcrcl}
t_1(1) & + & t_1(2) & - & t_2(1) & & & & & & & & & & & & & & = 0 \\
- t_1(1) & & & + & t_2(1) & + & t_2(2) & - & t_3(1) & & & & & & & & & & = 0 \\
& - & t_1(2) & & & - & t_2(2) & + & t_3(1) & & & & & & & & & & = 0 \\
t_1(1) & + & t_1(2) & & & & & & & + & s_1(1) & + & s_1(2) & - & s_2(1) & & & & = \frac{1}{4} \\
& & & & t_2(1) & + & t_2(2) & & & - & s_1(1) & & & + & s_2(1) & + & s_2(2) & - & s_3(1) & = \frac{1}{4} \\
& & & & & & & & t_3(1) & & & - & s_1(2) & & & & s_2(2) & + & s_3(1) & = \frac{1}{4}
\end{array}$$

An extreme optimal solution is:

$t_1^*(1) = t_1^*(2) = t_2^*(1) = 0,\ t_2^*(2) = t_3^*(1) = \frac{3}{8};\ s_1^*(1) = \frac{1}{4},\ s_1^*(21) = s_2^*(1) = 0,\ s_2^*(2) = \frac{1}{8},\ s_3^*(1) = 0.$

This gives the policy: $f_*(1) = 1,\ f_*(2) = 2,\ f_*(3) = 1.$

The programs (7.42) and (7.43) are:

$min\{\frac{1}{4}g_1 + \frac{1}{4}g_2 + \frac{1}{4}g_3 + \frac{1}{4}g_4\}$

subject to

$$\begin{array}{rcrcrcrcl}
g_1 & - & g_2 & & & & \geq & 1; \\
g_1 & & & - & g_3 & & \geq & 2; \\
- g_1 & + & g_2 & & & & \geq & -1; \\
& & g_2 & - & g_3 & & \geq & 1; \\
& - & g_2 & + & g_3 & & \geq & -1; \\
& & & & g_3 & - g_4 & \geq & -6; \\
- g_1 & & & & & + g_4 & \geq & 3; \\
& & & - & g_3 & + g_4 & \geq & 0;
\end{array}
\qquad
\begin{array}{rcrcl}
g_1 & & & \geq & \frac{3}{2}; \\
& g_2 & & \geq & \frac{1}{2}; \\
& & g_3 & \geq & -\frac{1}{2};
\end{array}$$

and

$$max\{q_1(1) + 2q_1(2) - q_2(1) + q_2(2) - g_3(1) - 6q_3(2) + 3q_4(1) + \tfrac{3}{2}h_1 + \tfrac{1}{2}h_2 - \tfrac{1}{2}h_3\}$$

subject to

$$\begin{array}{rcrcrcrcrcrcrcrcl}
q_1(1) & + & q_1(2) & - & q_2(1) & & & & & - & q_4(1) & & & + & h_1 & = & \frac{1}{4} \\
-\,q_1(1) & & & + & q_2(1) & + & q_2(2) & - & q_3(1) & & & & & + & h_2 & = & \frac{1}{4} \\
& - & q_1(2) & & & - & q_2(2) & + & q_3(1) & + & q_3(2) & & q_4(2) & + & h_3 & = & \frac{1}{4} \\
& & & & & & & & & - & q_3(2) & + & q_4(1) + q_4(2) & & & = & \frac{1}{4}
\end{array}$$

respectively (program (7.43) without the nonnegativity constraints). The optimal solutions are:

$g_1^* = \frac{3}{2},\ g_2^* = \frac{1}{2},\ g_3^* = -\frac{1}{2},\ g_4^* = \frac{9}{2}$ and $q_1^*(1) = q_1^*(2) = q_2^*(1) = q_2^*(2) = q_3^*(1) = q_3^*(2) = 0,$

$q_4^*(1) = \frac{1}{4},\ q_4^*(2) = 0;\ h_1^* = \frac{1}{2},\ h_2^* = \frac{1}{4},\ h_3^* = \frac{1}{4}.$

$S_* = \{1, 2, 3\};\ f(1) = 1,\ f(2) = 2,\ f(3) = 1,\ f(4) = 1.$

In order to show that Algorithm 7.3 is correct, we need several lemmata.

Lemma 7.12

The policy f^∞, defined in step 5c of the algorithm, is well-defined.

Proof

Take any $j \in S\backslash S_*$, then either $j \in S\backslash S_2$ or $j \in S_2\backslash S_*$. In the last case we have $g_j^* > u_j^* + \phi_j^*$, implying, by the complementary slackness property of linear programming, that $h_j^* = 0$.

So, if $j \in S\backslash S_*$, it follows from the constraints of program (7.43) that

$$\textstyle\sum_a q_j(a) = \beta_j + \sum_{(i,a)\in S\times A_1} p_{ij}(a)\}q_i(a) \geq \beta_j > 0.$$

Hence, policy f^∞ is well-defined. □

Lemma 7.13

Let f^∞ be an average optimal policy and let g be a feasible solution of program (7.43). Then,

$$\begin{aligned}
\big\{\big(I - P(f)\big)g\big\}_i &= r_i(f) - \phi_i \quad , \ i \in R(f); \\
g_i &\geq u_i^0(f) + \{P^*(f)g\}_i \quad , \ i \in T(f).
\end{aligned}$$

Proof

From Lemma 7.11 part (1) it follows that $f(i) \in A_1(i),\ i \in S$. Hence, from the constraints of program (7.42) we obtain $\big(I - P(f)\big)g - r(f) + \phi \geq 0$. Since $p_{ii}^*(f) > 0,\ i \in R(f)$ and $P^*(f)\big\{\big(I - P(f)\big)g - r(f) + \phi\big\} = P^*(f)\{-r(f) + \phi\} = P^*(f)\{-r(f) + \phi(f^\infty)\} = 0$, we have $\big\{\big(I - P(f)\big)g\big\}_i = r_i(f) - \phi_i,\ i \in R(f)$.

Since $d_{ij}(f) \geq 0,\ i, j \in T(f)$ and $\big\{\big(I - P(f)\big)g\big\}_j = r_j(f) - \phi_j,\ j \in R(f)$, we can write

$$\begin{aligned}
0 &\leq \big\{D(f)\{\big(I - P(f)\big)g - r(f) + \phi\}\big\}_i \\
&= \big\{D(f)\{\big(I - P(f)\big)g\}\big\}_i - \{D(f)r(f)\}_i + \{D(f)P^*(f)r(f)\}_i \\
&= \big\{\big(I - P^*(f)\big)g\big\}_i - u_i^0(f),\ i \in T(f).
\end{aligned}$$

□

Let the *bias value vector* u^0 be defined by $u^0 = u^0(f)$ with f^∞ a bias optimal policy.

Lemma 7.14

Let g^∞ be a bias optimal policy and let f_*^∞ be the optimal policy in the modified MDP as defined in step 4b of Algorithm 7.3. Then, $u_i^0 = u_i^0(f_*) = u_i^* - \{P^*(f_*)u^*\}_i = u_i^* + \phi_i^*,\ i \in R(g)$.

Proof

Take any $i \in R(g) \subseteq S_2$. Since g^∞ be a bias optimal policy and f_*^∞ is an optimal policy in the modified MDP, we obtain by Lemma 7.11 part (3) and (4)

$$u_i^0 = u_i^0(g) = u_i^* - \{P(g)u^*\}_i \le u_i^* - \{P(f_*)u^*\}_i \le u_i^0(f_*) \le u_i^0.$$

Hence, $u_i^0 = u_i^0(f_*) = u_i^* - \{P^*(f_*)u^*\}_i = u_i^* + \phi_i^*,\ i \in R(g)$, the last equality because of the optimality of f_*^∞ for the the modified MDP. □

Lemma 7.15

The bias value vector u^0 is the unique optimal solution of program (7.42).

Proof

We first show that u^0 is a feasible solution of program (7.42). Then, by (7.39),

$$u_i^0 \ge u_i^0(f_*) = u_i^* - \{P(f_*)u^*\}_i = u_i^* + \phi_i^*,\ i \in S_2.$$

Let g^∞ be a bias optimal policy and suppose that $\sum_j \{\delta_{ij} - p_{ij}(a)\}u_j^0 < r_i(a) - \phi_i$ for some $(i,a) \in S \times A_1$. Define the policy f^∞ by $f(j) = \begin{cases} g(j) & \text{if } j \ne i; \\ a & \text{if } j = i. \end{cases}$

Since

$$\{I - P(g)\}u^0 = \{I - P(g)\}u^0(g) = \{I - P(g)\}D(g)r(g) = \{I - P^*(g)\}r(g) = r(g) - \phi,$$

we have

$$\{r(f) + P(f)u^0 - u^0 - \phi\}_i > 0 \text{ and } \{r(f) + P(f)u^0 - u^0 - \phi\}_j = 0 \text{ for all } j \ne i. \tag{7.44}$$

Because $f(i) \in A_1(i),\ i \in S$, we have $\phi = P(f)\phi = P^*(f)\phi$. We also have

$$0 \le P^*(f)\{r(f) + P(f)u^0 - u^0 - \phi\} = \phi(f^\infty) - \phi \le 0,$$

implying that $P^*(f)\{r(f) + P(f)u^0 - u^0 - \phi\} = 0$. Hence, $p_{ii}^*(f) = 0$, i.e. $i \in T(f)$.

Since $P(f)$ and $P(g)$ differ only in row i and $i \in T(f)$, we have $R(f) \subseteq R(g)$.

For $j \in R(f)$ and $k \notin R(f)$, we have $d_{jk}(f) = d_{jk}(g) = 0$. Hence,

$$u_j^0(f) = \textstyle\sum_{k \in R(f)} d_{jk}(f) r_k(f) = \sum_{k \in R(g)} d_{jk}(g) r_k(g) = u_j^0(g) = u_j^0,\ j \in R(f).$$

Furthermore,

$$\{P^*(f)u^0\}_i = \sum_{j \in R(f)} p_{ij}^*(f) u_j^0 = \sum_{j \in R(f)} p_{ij}^*(f) u_j^0(f) = \{P^*(f)D(f)r(f)\}_i = 0. \tag{7.45}$$

Since $d_{ii}(f) > 0$ and using (7.44) and (7.45), we can write

$$\begin{aligned} u_i^0(f) &= \{D(f)r(f)\}_i = \textstyle\sum_j d_{ij}(f)r_j(f) \\ &> \textstyle\sum_j d_{ij}(f)\{\big(I - P(f)\big)u^0 + \phi\}_j \;=\; \{D(f)\big(I - P(f)\big)u^0 + D(f)\phi\}_i \\ &= \{\big(I - P^*(f)\big)u^0\}_i + \{D(f)P^*(f)\phi\}_i \;=\; u_i^0, \end{aligned}$$

implying a contradiction. So, we have shown that u^0 is a feasible solution of program (7.42).
Finally we show that u^0 is the, componentwise, smallest solution of (7.42).
Let w be an arbitrary feasible solution of (7.42).

For $j \in R(g)$: Lemma 7.14 implies $u_j^0 = u_j^0(f_*) = u_j^* + \phi_j^* \leq w_j$.
Therefore, $P^*(g)w \geq P^*(g)u^0 = P^*(g)u^0(g) = P^*(g)D^*(g)r(g) = 0$.

For $j \in T(g)$: Lemma 7.13 implies $u_j^0 = u_j^0(g) \leq w_j - \{P^*(g)w\}_j \leq w_j$.

Hence, we have shown that $u_j^0 \leq w_j,\ j \in S$. □

Lemma 7.16
$u_i^0 - \{P(f_*)u^0\}_i = r_i(f_*) - \phi_i,\ i \in S_*$.

Proof
We consider the modified MDP with state space S_2. From the constraints of (7.42) and the optimality of u^0 for (7.42) it follows that

$$u_i^0 - \{P(f_*)u^0\}_i \geq r_i(f_*) - \phi_i,\ i \in S_*.$$

Suppose that $u_i^0 - \{P(f_*)u^0\}_i > r_i(f_*) - \phi_i$ for some $i \in S_*$. Since

$$P^*(f_*)\{u^0 - P(f_*)u^0 - r(f_*) - \phi\} = 0,$$

state $i \notin R(f_*)$, i.e. $i \in T(f_*)$. Because $\{u^0 - P(f_*)u^0 - r(f_*) - \phi\}_j = 0,\ j \in R(f_*)$, and $d_{ij}(f_*) \geq 0,\ j \in T(f_*)$ and $d_{ii}(f_*) > 0$, we can write

$$\begin{aligned} 0 &< \big\{D(f_*)\{u^0 - P(f_*)u^0 - r(f_*) - \phi\}\big\}_i \\ &= \big\{D(f_*)\{u^0 - P(f_*)u^0 - r(f_*) - P^*(f_*)r(f_*)\}\big\}_i \\ &= \{u^0 - P^*(f_*)u^0 - u^0(f_*)\}_i \\ &\leq \{u^0 - P^*(f_*)u^0(f_*) - u^0(f_*)\}_i \\ &= \{u^0 - u^0(f_*)\}_i. \end{aligned}$$

From Lemma 7.15 it follows that $S_* = \{j \in S_2 \mid u_j^0 = u_j^* + \phi_j^*\}$. Relation (7.39) and the optimality of f_*^∞ in the modified MDP imply that $u_j^0(f_*) = u_j^* - \{P^*(f_*)u^*\}_j = u_j^* + \phi_j^*,\ j \in S_2$. Hence,

$$u_j^0(f_*) = u_j^0,\ j \in S_*, \tag{7.46}$$

contradicting the previous statement that $0 < \{u^0 - u^0(f_*)\}_i$. □

Lemma 7.17

S_* is a closed set in the Markov chain $P(f)$, where f^∞ is defined in step 5c of Algorithm 7.3.

Proof

Since $f(i) = f_*(i),\ i \in S_*$ and S_2 is closed in the Markov chain $P(f_*)$, we have to show that S_* is a closed set in the modified MDP. From Lemma 7.16 and relation (7.46) we have for all $i \in S_*$

$$\begin{aligned}
0 &= \{u^0 - P(f_*)u^0 - r(f_*) - P^*(f_*)r(f_*)\}_i \\
0 &= \{u^0(f_*) - P(f_*)u^0 - r(f_*) - P^*(f_*)r(f_*)\}_i \\
0 &= \big\{u^0(f_*) + P(f_*)\big(u^0(f_*) - u^0\big) - P(f_*)u^0(f_*) - r(f_*) - P^*(f_*)r(f_*)\big\}_i \\
0 &= \{P(f_*)\big(u^0(f_*) - u^0\big)\}_i + \{\big(I - P(f_*)\big)D(f_*)r(f_*)\}_i - \{\big(I - P^*(f_*)\big)r(f_*)\}_i \\
0 &= \{P(f_*)\big(u^0(f_*) - u^0\big)\}_i \\
0 &= \textstyle\sum_j p_{ij}(f_*)\{u^0(f_*) - u^0\}_j = \sum_{j \notin S_*} p_{ij}(f_*)\{u^0(f_*) - u^0\}_j.
\end{aligned}$$

Since $u_j^0(f_*) - u_j^0 = u_j^* + \phi_j^* - u_j^0 < 0,\ j \notin S_*$, it follows that $p_{ij}(f_*) = 0,\ i \in S_*,\ j \notin S_*$, i.e. S_* is a closed set in the Markov chain $P(f_*)$. □

Lemma 7.18

The states of $S \backslash S_*$ are transient in the Markov chain $P(f)$, where f^∞ is defined in step 5c of Algorithm 7.3.

Proof

Suppose there is a state $j \in S \backslash S_*$ which is recurrent under $P(f_*)$. Since S_* is closed under $P(f_*)$ there exists a nonempty set $J \subseteq S \backslash S_*$. Let $J = \{j_1, j_2, \ldots, j_m$. The constraints of program (7.43) imply

$$\begin{aligned}
\textstyle\sum_{a \in A_1(j)} q_j^*(a) + h_j^* &= \beta_j + \textstyle\sum_{(i,a) \in S \times A_1} p_{ij}(a)\} q_i^*(a) \geq \beta_j > 0,\ j \in S_2 \\
\textstyle\sum_{a \in A_1(j)} q_j^*(a) &= \beta_j + \textstyle\sum_{(i,a) \in S \times A_1} p_{ij}(a)\} q_i^*(a) \geq \beta_j > 0,\ j \in S \backslash S_2
\end{aligned}$$

Since (q^*, h^*) is an extreme optimal solution and since the linear program has N equality constraints, for each state j either $h_j^* > 0$ (and $q_j^*(a) = 0$ for all $a \in A_1(j)$) or $q_j^*(a) > 0$ for exactly one action, say action a_j (the other variables h_j^* and $q_j^*(a),\ a \neq a_j$, are zero). From the complementary slackness property of linear programming it follows that $h_j^* = 0$ for all $j \in S_2 \backslash S_*$. Hence, in every state j_i of J we have exactly one positive variable, namely $q_{j_i a_{j_i}}^*,\ i = 1, 2, \ldots, m$. The corresponding column vectors of the linear program which have the elements $\delta_{j_i k} - p_{j_i k}(a_{j_i}),\ k = 1, 2, \ldots, N$, are linearly independent. Since J is closed, $\delta_{j_i k} - p_{j_i k}(a_{j_i}) = 0,\ k \notin J$. Therefore, we have

$$\textstyle\sum_{k \in J} \{\delta_{j_i k} - p_{j_i k}(a_{j_i})\} = \sum_{k=1}^N \{\delta_{j_i k} - p_{j_i k}(a_{j_i})\} = 1 - 1 = 0,\ i = 1, 2, \ldots, m,$$

which contradicts the linear independency of these vectors. □

Theorem 7.17

The policy f^∞, defined in step 5c of Algorithm 7.3, is bias optimal.

Proof

Since $q_i^*\big(f(i)\big) > 0,\ i \in S\backslash S_*$, it follows from the complementary slaskness property that

$$u_i^0 - \{P(f)u^0\}_i = r_i(f) - \phi_i,\ i \in S\backslash S_*. \tag{7.47}$$

$P(f)$ and $P(f_*)$ have the same rows on the closed set S_*. Then, by (7.47) and Lemma 7.16

$$u^0 - P(f)u^0 = r(f) - \phi. \tag{7.48}$$

Since $f(i) \in A_1(i),\ i \in S$, we have $\phi = P(f)\phi = P^*(f)\phi$, and consequently, also using (7.48),

$$c0 = P^*(f)\{u^0 - P(f)u^0\} = P^*(f)\{r(f) - \phi\} = \phi(f^\infty) - \phi; \tag{7.49}$$

$$D(f)\phi = D(f)P^*(f)\phi = 0;\ u^0(f) = D(f)r(f) = D(f)\{I - P(f)\}u^0 = u^0 - P^*(f)u^0. \tag{7.50}$$

Because $u^0(f) = u^0 - P^*(f)u^0$ (see (7.50), $R(f) \subseteq S_*$ (see Lemma 7.18), $u_j^0(f_*) = u_j^0,\ j \in S_*$ (see (7.46)) and $D(f_*) = D(f)$ on S_*, we obtain

$$u^0(f) = u^0 - P^*(f)u^0 = u^0 - P^*(f)u^0(f) = u^0, \tag{7.51}$$

implying the bias optimality. □

7.8.2 The unichain case

Linear programming for bias optimality in the irreducible case was discussed in section 7.6.2. This section deals with the unicain case. In the unichain case the value vector ϕ is constant and we consider ϕ as a scalar. Program (7.37) becomes

$$min\ \{v \mid v + \sum_j \big(\delta_{ij} - p_{ij}(a)\big)u_j \geq r_i(a),\ (i,a) \in S \times A\}. \tag{7.52}$$

with as dual program

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & = & 0,\ j \in S \\ \sum_{(i,a)} x_i(a) & = & 1 \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}, \tag{7.53}$$

with optimal solutions $(v^* = \phi, u^*)$ and x^*, respectively.

Let $A_2(i) = \{a \in A(i) \mid \phi + \sum_j \big(\delta_{ij} - p_{ij}(a)\big)u_j^* = r_i(a)\},\ i \in S$ and $S_2 = \{i \in S \mid A_2(i) \neq \emptyset\}$.

Also the programs for the modified MDP simplify and become

$$min\ \{w \mid w + \sum_{j \in S_2} \big(\delta_{ij} - p_{ij}(a)\big)z_j \geq -u_i^*,\ (i,a) \in S_2 \times A_2\} \tag{7.54}$$

and

$$max \left\{ \sum_{(i,a)\in S_2\times A_2} (-u_i^*)t_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)\in S_2\times A_2} \{\delta_{ij} - p_{ij}(a)\}t_i(a) & = & 0,\ j \in S_2 \\ \sum_{(i,a)\in S_2\times A_2} t_i(a) & = & 1 \\ t_i(a) & \geq & 0,\ (i,a) \in S_2 \times A_2 \end{array} \right\}. \tag{7.55}$$

In this unichain case the algorithm becomes as follows (the proof of correctness follows straightforward from the previous section).

Algorithm 7.4 *Determination of a bias optimal policy by linear programming (unichain case)*

1. Compute an optimal solution $(v^* = \phi, u^*)$ of linear program (7.52).
2. Determine the following sets:
   $A_2(i) = \{a \in A(i) \mid \phi + \sum_j \big(\delta_{ij} - p_{ij}(a)\big)u_j^* = r_i(a)\},\ i \in S;$
   $S_2 = \{i \in S \mid A_2(i) \neq \emptyset\}.$
3. Determine the modified MDP with state space S_2 and action sets $A_2(i)$ by Algorithm 7.2.
4. a. Compute an optimal solution $(w^* = \phi^*, z^*)$ of linear program (7.54) and an extreme optimal solution t^* of (7.55).
   b. Take f_*^∞ such that $t_i^*\big(f_*(i)\big) > 0$ if $\sum_a t_i^*(a) > 0$ and arbitrary from $A_2(i)$ if $\sum_a t_i^*(a) = 0$.
5. a. Compute an optimal solution g^* of linear program (7.42) and an extreme optimal solution (q^*, h^*) of linear program (7.43), where $A_1(i) = A(i)$ for all $i \in S$.
   b. Determine $S_* = \{i \in S_2 \mid g_i^* = u_i^* + \phi_i^*\}$.
   c. Take policy f^∞ such that $f(i) = \begin{cases} f_*(i) & i \in S_* \\ q_i^*\big(f(i)\big) > 0 & \in S\backslash S_* \end{cases}$
   d. f^∞ is a bias optimal policy (STOP).

7.9 Overtaking and average overtaking optimality

A policy R_* is *overtaking optimal* if $liminf_{T\to\infty}\{v^T(R_*) - v^T(R)\} \geq 0$ for all policies R.

Example 7.6

$E = \{1,2,3\};\ A(1) = \{1\},\ A(2) = \{1,2\},\ A(3) = \{1\}.$
$r_1(1) = 0,\ r_2(1) = 1;\ r_2(2) = 0,\ r_3(1) = 1.$
$p_{11}(1) = 0,\ p_{12}(1) = 1,\ p_{13}(1) = 0;\ p_{21}(1) = 1,\ p_{22}(1) = 0,\ p_{23}(1) = 0;$
$p_{21}(2) = 0,\ p_{22}(2) = 0,\ p_{23}(2) = 1;\ p_{31}(1) = 0,\ p_{32}(1) = 1,\ p_{33}(1) = 0.$

There are two deterministic stationary policies f_1^∞ with $f_1(2) = 1$ and f_2^∞ with $f_2(2) = 2$.

Observe that $v_2^T(f_1^\infty) = \lceil \frac{T}{2} \rceil$ and $v_2^T(f_2^\infty) = \lfloor \frac{T}{2} \rfloor$. Hence, $v_2^T(f_1^\infty) - v_2^T(f_2^\infty) = \begin{cases} 1 & \text{if } T \text{ is odd;} \\ 0 & \text{if } T \text{ is even.} \end{cases}$

A similar result is obtained when we start in another state. Thus

$liminf_{T\to\infty} \{v_i^T(f_1^\infty) - v_i^T(f_2^\infty)\} = 0,\ i \in S$ and $liminf_{T\to\infty} \{v_i^T(f_2^\infty) - v_i^T(f_1^\infty)\} = -1,\ i \in S$.

So f_1^∞ dominates f_2^∞ in the overtaking optimal sense.

In contrast with other criteria, an overtaking optimal policy doesn't exist in general as the next example shows.

Example 7.7

$E = \{1, 2, 3\};\ A(1) = \{1, 2\},\ A(2) = \{1\},\ A(3) = \{1\}.$

$r_1(1) = 1,\ r_1(2) = 0;\ r_2(1) = 0,\ r_3(1) = 2.$

$p_{11}(1) = 0,\ p_{12}(1) = 1,\ p_{13}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 0,\ p_{13}(2) = 1;$

$p_{21}(1) = 0,\ p_{22}(1) = 0,\ p_{23}(1) = 1;\ p_{31}(1) = 0,\ p_{32}(1) = 1,\ p_{33}(1) = 0.$

There are two deterministic stationary policies f_1^∞ with $f_1(1) = 1$ and f_2^∞ with $f_2(1) = 2$.

Observe that $v_1^T(f_1^\infty) = 2 \cdot \lfloor \frac{T-1}{2} \rfloor + 1$ and $v_1^T(f_2^\infty) = 2 \cdot \lfloor \frac{T}{2} \rfloor$.

Hence, $v_1^T(f_1^\infty) - v_1^T(f_2^\infty) = \begin{cases} 1 & \text{if } T \text{ is odd;} \\ -1 & \text{if } T \text{ is even.} \end{cases}$

Thus $liminf_{T\to\infty} \{v_1^T(f_1^\infty) - v_1^T(f_2^\infty)\} = -1$ and $liminf_{T\to\infty} \{v_1^T(f_2^\infty) - v_1^T(f_1^\infty)\} = -1$.

Hence neither f_1^∞ dominates f_2^∞ nor f_2^∞ dominates f_1^∞ in the overtaking optimal sense.

We next consider a less selective criterion. A policy R_* is calles *average overtaking optimal* if $liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \{v^t(R_*) - v^t(R)\} \geq 0$ for all policies R. Notice that this criterion is the same as 0-average optimality which is also equivalent to bias optimality.

7.10 Bibliographic notes

The material presented in this chapter has its roots in Blackwell's seminal paper [22]. Among other results, Blackwell introduced the criteria we now refer to as 0-discount optimality and Blackwell optimality (he referred to policies which achieve these criteria as *nearly optimal* and *optimal*, respectively). He demonstrated the existence of a Blackwell optimal policy and convergence of the multichain average reward policy iteration method through use of a *partial* Laurent series expansion. That paper raised also the following challenging questions:

a. What is the relationship between 0-discount optimality and Blackwell optimality?

b. When are average optimal policies 0-discount optimal?

c. How does one compute 0-discount optimal and Blackwell optimal policies?

Veinott ([214]), Veinott ([217]) and Miller and Veinott ([138]) addressed these issues. In his 1966 paper, Veinott provided a policy iteration algorithm for finding a 0-discount optimal policy (he referred to such policy as 1-*optimal*). In his comprehensive 1969 paper, Veinott provides the link between 0-discount optimality and Blackwell optimality. Miller and Veinott developed the *complete* Laurent series expansion and related it to the lexicographic ordering and used this ordering to provide a finite policy iteration method for n-discount optimality for any n. They also showed that $N-1$-discount optimality is equivalent to Blackwell optimality. In his 1974 paper ([218]) Veinott provides a more accessible overview of the above work and a simplified presentation of the main results in [217].

Sladky ([186]) generalized overtaking and average overtaking to n-average optimality. He showed the equivalence between n-average optimality and n-discount optimality. An example from Brown ([27]) shows that an overtaking optimal policy need not exists. Denardo and Rothblum ([52]) give an additional assumption under which overtaking policies exist. For the presentation of the lexicographic ordering we refer to Veinott ([217]), Dekker ([41]) and Dekker and Hordijk ([42].

Denardo and Fox ([50]), Denardo [46]) and Kallenberg ([108]) provide linear programming approaches for finding 0-discount optimal policies. The linear programming method for the irreducible case (section 7.6) is due to Avrachenkov and Altman ([4]). The approach for the computation of a Blackwell optimal policy by linear programming, based on asymptotic linear programming, was proposed by Hordijk, Dekker and Kallenberg ([94]). Unfortunately, this method cannot be used for the calculation of n-discount optimal policies that are not Blackwell optimal.

7.11 Exercises

Exercise 7.1

Give a direct proof of Lemma 7.3, i.e. if a policy is n-average optimal, then it is m-average optimal for $m = -1, 0, \ldots, n$.

Exercise 7.2

Sow that for all $n \geq 0,\ T \geq 2$ and $f^{\infty} \in C(D)$:

a. $v^{n,T}(f^{\infty}) = v^{n,T-1}(f^{\infty}) = v^{n-1,T}(f^{\infty})$.

b. $v^{n,T}(f^{\infty}) = \binom{T+n}{n+1} r(f) + P(f) v^{n,T-1}(f^{\infty})$.

Exercise 7.3

Consider the following model:

$S = \{1, 2\};\ A(1) = \{1, 2\},\ A(2) = \{1\};\ r_1(1) = 5,\ r_1(2) = 10,\ r_2(1) = -1.$

$p_{11}(1) = 0.5,\ p_{12}(1) = 0.5;\ p_{11}(2) = 0,\ p_{12}(1) = 1;\ p_{21}(1) = 0,\ p_{22}(1) = 1.$

Determine for both deterministic policies the α-discounted rewards and for which α the policy is optimal.

Exercise 7.4

Consider the following model:

$S = \{1, 2\};\ A(1) = \{1, 2\},\ A(2) = \{1\};\ r_1(1) = 1,\ r_1(2) = 2,\ r_2(1) = 0.$

$p_{11}(1) = 0.5,\ p_{12}(1) = 0.5;\ p_{11}(2) = 0,\ p_{12}(1) = 1;\ p_{21}(1) = 0,\ p_{22}(1) = 1.$

Determine for both deterministic policies the Laurent series expansion in $1 - \alpha$ and derive from this expansion the vectors $u^k(f),\ k = -1, 0, 1, \ldots$.

Exercise 7.5

Consider the following model:

$S = \{1, 2, 3\};\ A(1) = \{1, 2\},\ A(2) = A(3) = \{1\};\ r_1(1) = a,\ r_1(2) = 1,\ r_2(1) = b,\ r_3(1) = 0.$

$p_{11}(1) = 0,\ p_{12}(1) = 1,\ p_{13}(1) = 0;\ p_{11}(2) = 0.5,\ p_{11}(2) = 0.5,\ p_{13}(2) = 0;$

$p_{21}(1) = 0,\ p_{22}(1) = 0,\ p_{23}(1) = 1;\ p_{31}(1) = 0,\ p_{32}(1) = 0,\ p_{33}(1) = 1.$

Determine a and b such that both deterministic policies are (-1)-discount, 0-discount and 1-discount optimal. Which policy is Blackwell optimal?

Exercise 7.6

Let for $f^\infty \in C(D)$ and $\rho > 0$ the *resolvent* $R^\rho(f)$ be defined by $R^\rho(f) = \{\rho I + (I - P(f)\}^{-1}$. Show that

(1) $R^\rho(f) = \alpha\{I - \alpha P(f)\}^{-1}$, where $\alpha = \frac{1}{1+\rho}$.

(2) $lim_{\rho \downarrow 0}\, \rho R^\rho(f) = P^*(f)$.

Exercise 7.7

Determine a 0-optimal policy with Algorithm 7.1 for the following model:

$S = \{1, 2\};\ A(1) = \{1, 2\},\ A(2) = \{1\};\ r_1(1) = 4,\ r_1(2) = 0,\ r_2(1) = 8.$

$p_{11}(1) = 1,\ p_{12}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1;\ p_{21}(1) = 1,\ p_{22}(1) = 0.$

Start with $f(1) = 2,\ f(2) = 1$.

Exercise 7.8

Show that $v^\alpha(R) - v^\alpha(f^\infty) = \sum_{k=-1}^{\infty} \rho^k \psi^k(f, g)$ for the nonstationary policy $R = (g, f, f, f, \ldots)$.

Exercise 7.9

Consider the following model:

$S = \{1, 2, 3, 4\};\ A(i) = \{1, 2\},\ i = 1, 2, 3, 4.$

$p_{11}(1) = 0;\quad p_{12}(1) = 1;\quad p_{13}(1) = 0;\quad p_{14}(1) = 0;\quad r_1(1) = 1$

$p_{11}(2) = 0;\quad p_{12}(2) = \frac{1}{2};\quad p_{13}(2) = \frac{1}{2};\quad p_{14}(2) = 0;\quad r_1(2) = \frac{1}{2}$

$p_{21}(1) = 0;\quad p_{22}(1) = 0;\quad p_{23}(1) = 1;\quad p_{24}(1) = 0;\quad r_2(1) = 0$

$p_{21}(2) = 0;\quad p_{22}(2) = 0;\quad p_{23}(2) = \frac{1}{3};\quad p_{24}(2) = \frac{2}{3};\quad r_2(2) = -\frac{2}{3}$

$p_{31}(1) = 0;\quad p_{32}(1) = 0;\quad p_{33}(1) = 0\quad p_{34}(1) = 1;\quad r_3(1) = 0.$

$p_{31}(2) = \frac{1}{4};\quad p_{32}(2) = 0;\quad p_{33}(2) = 0;\quad p_{34}(2) = \frac{3}{4};\quad r_3(2) = \frac{1}{2}$

$p_{41}(1) = 1;\quad p_{42}(1) = 0;\quad p_{43}(1) = 0;\quad p_{44}(1) = 0;\quad r_4(1) = 3$

$p_{41}(2) = \frac{1}{4};\quad p_{42}(2) = \frac{3}{4};\quad p_{43}(2) = 0;\quad p_{44}(2) = 0;\quad r_4(2) = 3$

With (i,j,k,l), where $i,j,k,l \in \{1,2\}$, we denote the 16 policies, so $(1,2,2,1)$ is the policy that takes action 1 in state 1 and 4, and action 2 in the states 2 and 3.

a. The model is irreducible: show that the policy $(1,2,2,1)$ is irreducible.
b. Formulate the primal and dual linear program for an (-1)-discount optimal policy.
c. Solve the linear programs (use any package that is availabe for you).
d. Determine, using the in c obtained solution, the set of (-1)-discount optimal policies.
e. Formulate the primal and dual linear program for an 0-discount optimal policy.
f. Solve the linear programs for an 0-discount optimal policy.
g. Determine, using the in f obtained solution, the set of 0-discount optimal policies.
h. Formulate the primal and dual linear program for an 1-discounted optimal policy.
i. Solve the linear programs for an 1-discounted optimal policy.
j. Determine, using the in i obtained solution, the set of 1-discounted optimal policies.

Exercise 7.10

Show that the ordering of $F(\mathbb{R})$ given by (7.30) is a correct total ordering.

Exercise 7.11

Consider the following MDP:

$S = \{1,2\};\ A(1) = \{1,2,3\},\ A(2) = \{1\};\ r_1(1) = 1,\ r_1(2) = \frac{3}{4},\ r_1(3) = \frac{1}{2};\ r_2(1) = 0.$

$p_{11}(1) = 0,\ p_{12}(1) = 1;\ p_{11}(2) = \frac{1}{2},\ p_{12}(2) = \frac{1}{2};\ p_{11}(3) = 1,\ p_{12}(3) = 0;,\ p_{21}(1) = 0,\ p_{22}(1) = 1.$

Determine by the linear programming method for rational functions optimal policies for all discount factors $\alpha \in (0.1)$.

Exercise 7.12

Consider the following MDP:

$S = \{1,2,3\};\ A(1) = \{1,2\},\ A(2) = \{1,2\},\ A(3) = \{1\}.$

$r_1(1) = 1,\ r_1(2) = 2;\ r_2(1) = 2,\ r_2(2) = 0;\ r_3(1) = 0.$

$p_{11}(1) = 1,\ p_{12}(1) = 0,\ p_{13}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 0,\ p_{13}(1) = 1;$

$p_{21}(1) = 1,\ p_{22}(1) = 0,\ p_{23}(1) = 0;\ p_{21}(2) = 0,\ p_{22}(2) = 0,\ p_{23}(2) = 1;$

$p_{31}(1) = 1,\ p_{32}(1) = 0,\ p_{33}(1) = 0.$

Determine a bias optimal policy by the linear programming method for unichain MDPs.

Exercise 7.13

Consider the MDP of Example 7.7 with the two deterministic policies f_1^∞ and f_1^∞. Determine $liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \{v_i^t(f_1^\infty) - v_i^t(f_2^\infty)\}$ and $liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \{v_i^t(f_2^\infty) - v_i^t(f_1^\infty)\}$ for each $i \in S$. Is f_1^∞ an average overtaking policy? Is f_2^∞ an average overtaking policy?

Chapter 8

Special models

In this chapter we deal with the models which were introduced in section 1.3. The red-black gambling model was already discussed in section 4.7.1, where we proved the following results:

if $p > \frac{1}{2}$, then timid play, i.e. always bet the amount 1, is optimal.

if $p = \frac{1}{2}$, then any policy is optimal.

if $p < \frac{1}{2}$, then bold play, i.e. betting $min(i, N - i)$ in state i, is optimal.

The model 'How to serve in tennis' is left as Exercise 1.5 to the reader. For the optimal stopping problem we refer to the sections 1.3.3 and 4.7.2. In the next sections we discuss the following models:

1. Replacement problems.
2. Maintenance and repair problems.
3. Production and inventory control.
4. Optimal control of queues.
5. Stochastic scheduling.
6. Multi-armed bandit problems.
7. Separable problems.

8.1 Replacement problems

In this section we discuss four variants of the replacement problem:
- a general replacement model;
- a replacement model with increasing deterioration;
- a skip to the right model with failure costs;
- a separable model.

8.1.1 A general replacement model

In this general replacement model we have state space $S = \{0, 1, \ldots, N\}$, where state 0 corresponds to a new item, and action sets $A(0) = \{1\}$ and $A(i) = \{0, 1\}$, $i \neq 0$, where action 0

means replacing the 'old' item by a new item. We consider in this model costs instead of rewards. Let c be the cost of a new item. Furthermore, assume that an item of state i has trade-in-value s_i and maintenance costs c_i. If in state i action 0 is chosen, then $c_i(0) = c - s_i + c_0$ and $p_{ij}(0) = p_{0j},\ j \in S$; for action 1, we have $c_i(1) = c_i$ and $p_{ij}(1) = p_{ij},\ j \in S$. In contrast with other replacement models, where the state is determined by the age of the item, we allow that the state of the item may change to any other state. In this case the optimal replacement policy is in general no a control-limit rule. As optimality criterion we consider the discounted reward. For this model the primal linear program, which yields the value vector v^α, is:

$$\min\left\{\sum_{j=0}^{N}\beta_j v_j \;\middle|\; \begin{array}{lcll} \sum_{j=0}^{N}(\delta_{ij}-\alpha p_{0j})v_j & \geq & -c+s_i-c_0, & 1\leq i\leq N \\ \sum_{j=0}^{N}(\delta_{ij}-\alpha p_{ij})v_j & \geq & -c_i, & 0\leq i\leq N \end{array}\right\}, \tag{8.1}$$

where $\beta_j > 0,\ j \in S$. Because there is only one action in state 0, namely action 1, we have

$$v_0^\alpha = -c_0 + \alpha\sum_{j=0}^{N} p_{0j}v_j^\alpha \tag{8.2}$$

Hence, instead of $v_i - \alpha\sum_{j=0}^{N} p_{0j}v_j = \sum_{j=0}^{N}(\delta_{ij}-\alpha p_{0j})v_j \geq -c + s_i - c_0$, we can write $v_i - v_0 \geq -c + s_i$, obtaining the equivalent linear program

$$\min\left\{\sum_{j=0}^{N}\beta_j v_j \;\middle|\; \begin{array}{lcll} v_i - v_0 & \geq & r_i, & 1\leq i\leq N \\ \sum_{j=0}^{N}(\delta_{ij}-\alpha p_{ij})v_j & \geq & -c_i, & 0\leq i\leq N \end{array}\right\}, \tag{8.3}$$

where $r_i := -c + s_i,\ i \in S$. The dual linear program of (8.3) is:

$$\max\left\{\sum_{i=1}^{N} r_i x_i - \sum_{i=0}^{N} c_i y_i \;\middle|\; \begin{array}{rcll} -\sum_{i=1}^{N}x_i + \sum_{i=0}^{N}(\delta_{i0}-\alpha p_{i0})y_i & = & \beta_0 & \\ x_j + \sum_{i=0}^{N}(\delta_{ij}-\alpha p_{ij})y_i & = & \beta_j, & 1\leq j\leq N \\ x_i & \geq & 0, & 1\leq i\leq N \\ y_i & \geq & 0, & 0\leq i\leq N \end{array}\right\} \tag{8.4}$$

Theorem 8.1

There is a one-to-one correspondence between the extreme solutions of (8.4) and the set of deterministic policies.

Proof

Let (x, y) be an extreme solution of (8.4). Then, (x, y) has exactly $N + 1$ positive components. Since

$$y_0 = \beta_0 + \sum_{i=1}^{N} x_i + \alpha\sum_{i=0}^{N} p_{i0}y_i \geq \beta_0 > 0$$

and

$$x_j + y_j = \beta_j + \alpha\sum_{i=0}^{N} p_{ij}y_i \geq \beta_j > 0,\ 1 \leq j \leq N,$$

in each state $j,\ 0 \leq j \leq N$, either x_j or y_j is strictly positive.

Hence, (x, y) corresponds to a deterministic policy:

if $x_j > 0$, then action 0 (replacement) is chosen;

if $x_j = 0$, then $y_j > 0$ and action 1 (no replacement) is chosen.

Conversely, let f^∞ be a deterministic policy. Partition the states $\{1, 2, \ldots, N\}$ in $S_0 \cup S_1$, where S_0 and S_1 correspond to the states in which action 0 and action 1, respectively, are chosen. Let $x_j = 0,\ j \in S_1$ and $y_j = 0,\ j \in S_0$. Then, the equations of (8.4) are equivalent to the following system of $N + 1$ equations with $N + 1$ variables:

$$\left\{\begin{array}{rcrcrcl} -\sum_{j\in S_0} x_j & + & y_0 & - & \alpha \sum_{i\in S_1} p_{i0} y_i & = & \beta_0 \\ x_j & & & - & \alpha \sum_{i\in S_1} p_{ij} y_i & = & \beta_j,\ j \in S_0 \\ & & y_j & - & \alpha \sum_{i\in S_1} p_{ij} y_i & = & \beta_j, j \in S_1 \end{array}\right. \tag{8.5}$$

Consider a linear combination of the columns of (8.5) which yields the 0-vector:

$$\left\{\begin{array}{rcrcrcl} -\sum_{j\in S_0} \mu_j & + & \mu_0 & - & \alpha \sum_{i\in S_1} p_{i0} \mu_i & = & 0 \\ \mu_j & & & - & \alpha \sum_{i\in S_1} p_{ij} \mu_i & = & 0,\ j \in S_0 \\ & & \mu_j & - & \alpha \sum_{i\in S_1} p_{ij} \mu_i & = & 0,\ j \in S_1 \end{array}\right. \tag{8.6}$$

i.e. $\mu_0 = \alpha \sum_{i\in S_1} p_{i0}\mu_i + \sum_{j\in S_0} \mu_j = \alpha \sum_{i\in S_1} \{p_{i0} + \sum_{j\in S_0} p_{ij}\}\mu_i$. Hence, we have

$$\mu_j = \alpha \sum_{i\in S_1} q_{ij}\mu_i,\ j \in \{0\} \cup S_1, \tag{8.7}$$

where $q_{ij} = \left\{\begin{array}{ll} p_{i0} + \sum_{j\in S_0} p_{ij} & i \in S_1,\ j = 0, \\ p_{ij} & i \in S_1,\ j \in S_1. \end{array}\right.$

Remark that Q is a transformation matrix on $\{0\} \cup S_1$, because $q_{ij} \geq 0$ for all i and j, and

$$\sum_{j\in\{0\}\cup S_1} q_{ij} = p_{i0} + \sum_{j\in S_0} p_{ij} + \sum_{j\in S_1} p_{ij} = \sum_{j\in S} p_{ij} = 1 \text{ for all } i.$$

Let $\nu_j = \mu_j,\ j \in \{0\} \cup S_1$, then (8.7) implies that $\nu(I - \alpha Q) = 0$. Since $I - \alpha Q$ is nonsingular, we have $\nu = 0$, i.e. $\mu_j = 0,\ j \in \{0\} \cup S_1$. Then, from (8.6) it follows that $\mu_j = 0,\ j \in S_0$, implying that $\mu_j = 0$ for all j. Hence, the columns of (8.5) are linear independent and, consequently, (x, y) is an extreme solution of (8.4) and the correspondence is one-to-one. □

Consider the simplex method to solve (8.4) and start with the basic solution that corresponds to the policy which chooses action 1 (no replacement) in all states. Hence, in the first simplex tableau $y_j,\ 0 \leq j \leq N$, are the basic variables and $x_i,\ 1 \leq i \leq N$, the nonbasic variables. Take the usual version of the simplex method in which the column with the most negative cost is chosen as pivot column. It turns out, as will be shown in Theorem 8.2, that this choice gives the optimal action for that state, i.e. in that state action 0, the replacement action, is optimal. Hence, after interchanging x_i and y_i, the column of y_i can be deleted. Consequently, we obtain the following *greedy simplex algorithm.*

Algorithm 8.1 *The greedy simplex algorithm*

1. Start with the basic solution corresponding to the nonreplacing actions.
2. If the reduced costs are nonnegative: the corresponding policy is optimal (STOP).

   Otherwise:

   (a) Choose the column with the most negative reduced cost as pivot column.

   (b) Execute the usual simplex transformation.

   (c) Delete the pivot column.
3. If all columns are removed: replacement in all states is the optimal policy (STOP).

   Otherwise: return to step 2.

Theorem 8.2 *The greedy simplex algorithm is correct and has complexity* $\mathcal{O}(N^3)$.

Proof

For the correctness of the algorithm it has to be shown that the deletion of the pivot column is allowed. This will be shown by induction on the number of iterations. The first simplex tableau is correct, because no column had been deleted. Suppose that previous iterations were correct and consider the present simplex tableau, corresponding to policy f^∞, with basic variables $y_0, y_i,\ i \in S_1$, and $x_i,\ i \in S_0$. The reduced cost corresponding to state i and action a is in general (see section 3.5):

$$\sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j^\alpha(f^\infty) - r_i(a).$$

For action 0 (replacement) we denote the reduced cost in state i by $w_i(f)$ and this quantity becomes

$$\begin{aligned} w_i(f) &= \sum_j \{\delta_{ij} - \alpha p_{ij}(0)\} v_j^\alpha(f^\infty) + c_i(0) \\ &= v_i^\alpha(f^\infty) - \alpha \sum_j p_{0j} v_j^\alpha(f^\infty) + (c - s_i + c_0) \\ &= v_i^\alpha(f^\infty) - v_0^\alpha(f^\infty) - r_i, \end{aligned}$$

the last equality because $v_0^\alpha(f^\infty) = -c_0 + \alpha \sum_j p_{0j} v_j^\alpha(f^\infty)$.
For action 1 (no replacement) we denote the reduced cost in state i by $z_i(f)$ and we obtain

$$z_i(f) = \sum_j \{\delta_{ij} - \alpha p_{ij}\} v_j^\alpha(f^\infty) + c_i.$$

Since the reduced costs corresponding to basic variables are zero, we have

$$\begin{cases} \sum_j \{\delta_{ij} - \alpha p_{ij}\} v_j^\alpha(f^\infty) = -c_i, & i \in S_1 \cup \{0\} \\ v_i^\alpha(f) - v_0^\alpha(f^\infty) = r_i, & i \in S_0 \end{cases}$$

Let $S_0^* = \{i \mid \text{action 0 is in state } i \text{ optimal}\}$, $S_1^* = S \backslash S_0^*$ and let f_*^∞ be an optimal policy. Then, state 0 is in S_1^*, $S_0 \subseteq S_0^*$ (by the induction hypothesis), and

$$\begin{cases} \sum_j \{\delta_{ij} - \alpha p_{ij}\} v_j^\alpha(f_*^\infty) &= -c_i, \quad i \in S_1^* \\ v_i^\alpha(f_*^\infty) - v_0^\alpha(f_*^\infty) &= r_i, \quad i \in S_0^* \end{cases}$$

Assume that the column of the nonbasic variable x_k is chosen as pivot column. Then, the reduced cost of column k is the most negative reduced cost: $w_k(f) < 0$ and $w_k(f) \leq w_i(f),\ 1 \leq i \leq N$. It is sufficient to show that in state k action 0 is optimal, i.e. $k \in S_0^*$. Let $d(f) = v^\alpha(f) - v^\alpha(f_*)$. Then, because f_*^∞ is optimal, $S_0 \subseteq S_0^*$ and $v_i^\alpha(f_*^\infty) - v_0^\alpha(f_*^\infty) = r_i,\ i \in S_0^*$,

$$\begin{cases} d_i(f) = \alpha \sum_j p_{ij} d_j(f), & i \in S_1^* \\ d_i(f) = d_0(f) + w_i(f) \geq d_0(f) + w_k(f), & i \in S_0^* \end{cases} \tag{8.8}$$

Let $m \in S_1^*$ be such that $d_m(f) = min_{i \in S_1^*}\ d_i(f)$ and suppose that $d_m(f) \leq d_0(f) + w_k(f)$. Then, $d_m(f) \leq d_i(f),\ 0 \leq i \leq N$, and (8.8) implies that $d_m(f) = \alpha \sum_j p_{mj} d_j(f) \geq \alpha d_m(f)$, i.e. $d_m(f) \geq 0$. Since $0 \leq d_m(f) \leq d_i(f),\ 0 \leq i \leq N$, and $d(f) \leq 0$, we have $d(f) = 0$. This means that $v^\alpha(f^\infty) = v^\alpha(f_*^\infty) = v^\alpha$. Hence, the present simplex tableau is optimal which contradicts $w_k(f) < 0$. Therefore, we have shown that

$$d_0(f) + w_k(f) < d_m(f) \leq d_i(f) \text{ for all } i \in S_1^*.$$

Suppose that $k \in S_1^*$. Then,

$$v_k^\alpha(f^\infty) - v_k^\alpha(f_*^\infty) = d_k(f) \geq d_m(f) > d_0(f) + w_k(f) = v_0^\alpha(f^\infty) - v_0^\alpha(f_*^\infty) + w_k(f).$$

Hence,

$$v_k^\alpha = v_k^\alpha(f_*^\infty) < v_k^\alpha(f^\infty) - v_0^\alpha(f^\infty) + v_0^\alpha(f_*^\infty) - v_k^\alpha(f^\infty) + v_0^\alpha(f^\infty) + r_k = v_0^\alpha + r_k,$$

i.e. v^α is infeasible for (8.3): contradiction. This completes the first part of the proof.
In the first step of the algorithm, the simplex tableau for a specific basic solution has to be determined. This is one matrix inversion and has complexity $\mathcal{O}(N^3)$. Since in each iteration one column is removed, there are at most N iterations. In each iteration a simplex tranformation is executed which takes $\mathcal{O}(N^2)$. Hence, the overall complexity of the algorithm is $\mathcal{O}(N^3)$. □

Remark:
An optimal stopping problem may be considered as a special case of a replacement problem with as optimality criterion the total expected reward, i.e. $\alpha = 1$. In an optimal stopping problem there are two actions in each state. The first action is the stopping action and the second action corresponds to continue. If the stopping action is chosen in state i, then a final reward s_i is earned and the process terminates. If the second action is chosen, then a reward r_i is received and the transition probability of being in state j at the next decision time point is $p_{ij},\ j \in S$. This optimal stopping problem can is a special case of the replacement problem with $p_{0j} = 0$ for all $j \in S$, $c_i(0) = -s_i$ and $c_i(1) = -r_i$ for all $i \in S$. Hence, also for the optimal stopping problem, the linear programming approach of this section can be used and the complexity is also $\mathcal{O}(N^3)$.

8.1.2 A replacement model with increasing deterioration

Consider a replacement model with state space $S = \{0, 1, \ldots, N+1\}$. An item is in state 0 if and only if it is new; an item is in state $N+1$ if and only if it is inoperative. In states $1, 2, \ldots, N$ there are two actions: action 0 is to replace the item by a new one and action 1 is not to replace the item. In the states 0 and $N+1$ only one action is possible (not to replace and replace, respectively) and call this action 1. Then, the transition probabilities are:

$$p_{ij}(0) = \begin{cases} 0 & 1 \le i \le N,\ j \neq 0 \\ 1 & 1 \le i \le N,\ j = 0 \end{cases} \text{ and } p_{ij}(1) = p_{ij} \text{ with } p_{i0}(1) = \begin{cases} 0 & i \neq N+1 \\ 1 & i = N+1 \end{cases}$$

We assume two types of cost, the cost $c_0 \ge 0$ to replace an operative item and the cost $c_0 + c_1$, where $c_1 \ge 0$, to replace an inoperative item:

$$r_i(0) = c_0,\ 1 \le i \le N;\ r_i(1) = \begin{cases} 0 & 0 \le i \le N \\ c_0 + c_1 & i = N+1 \end{cases}$$

We state the following equivalent (see Lemma 8.1) assumptions.

Assumption 8.1
The transition probabilities are such that for every nondecreasing function $x_j,\ j \in S$, the function $F(i) = \sum_{j=0}^{N+1} p_{ij} x_j$ is nondecreasing in i.

Assumption 8.2
The transition probabilities are such that for every $k \in S$, the function $G_k(i) = \sum_{j=k}^{N+1} p_{ij}$ is nondecreasing in i.

Lemma 8.1
The Assumptions 8.1 and 8.2 are equivalent.

Proof
Let Assumption 8.1 hold. Take any $k \in S$. Then, for the nondecreasing function $x_j = \begin{cases} 0 & j < k \\ 1 & j \ge k \end{cases}$ the function $F(i) = \sum_{j=0}^{N+1} p_{ij} x_j = \sum_{j=k}^{N+1} p_{ij} = G_k(i)$ is nondecreasing in i.
Conversely, let Assumption 8.2 hold. Take any nondecreasing function $x_j,\ j \in S$. Then, with $c_k = x_k - x_{k-1} \ge 0,\ 1 \le k \le N+1$, we can write $x_j = \sum_{k=1}^{j} c_k + x_0,\ 1 \le j \le N+1$. Therefore, we obtain

$$\begin{aligned} F(i) &= \sum_{j=0}^{N+1} p_{ij} x_j = p_{i0} x_0 + \sum_{j=1}^{N+1} p_{ij} \{\sum_{k=1}^{j} c_k + x_0\} \\ &= x_0 + \sum_{j=1}^{N+1} \sum_{k=1}^{j} c_k p_{ij} = x_0 + \sum_{k=1}^{N+1} c_k \{\sum_{j=k}^{N+1} p_{ij}\}. \end{aligned}$$

Since $\sum_{j=k}^{N+1} p_{ij} = G_k(i)$ is nondecreasing in i and $c_k \ge 0$ for $k = 1, 2, \ldots, N+1$, the function $F(i)$ is also nondecreasing in i. □

We first consider the discounted rewards. The method of value iteration for this model becomes (with $v_i^0 = 0$ for all i) for $n = 0, 1, \ldots$

$$v_i^{n+1} = \begin{cases} \alpha \sum_j p_{0j} v_j^n & , \ i = 0 \\ min\{\alpha \sum_j p_{ij} v_j^n, \ c_0 + \alpha \sum_j p_{0j} v_j^n\} & , \ 1 \le i \le N \\ c_0 + c_1 + \alpha \sum_j p_{0j} v_j^n & , \ i = N+1. \end{cases} \tag{8.9}$$

We assume that Assumption 8.1 (or 8.2) holds. Clearly, v_i^0 is a nondecreasing function in i. Assume v_j^n is nondecreasing in j. Then, it follows from Assumption 8.1 and (8.9) that v_i^{n+1} is also nondecreasing in i. Hence, also $v_i^\alpha = lim_{n\to\infty} v_i^n$ is nondecreasing in i.

Theorem 8.3

Let i_* be such that $i_* = max\{i \mid \alpha \sum_j p_{ij} v_j^\alpha \le c_0 + \alpha \sum_j p_{0j} v_j^\alpha\}$. Then, the control-limit policy f_*^∞ which replaces in the states $i > i_*$ is a discounted optimal policy.

Proof

Since v_i^α is nondecreasing in i and v^α is the unique solution of the optimality equation

$$v_i^\alpha = \begin{cases} \alpha \sum_j p_{0j} v_j^\alpha & , \ i = 0 \\ min\{\alpha \sum_j p_{ij} v_j^\alpha, \ c_0 + \alpha \sum_j p_{0j} v_j^\alpha\} & , \ 1 \le i \le N \\ c_0 + c_1 + \alpha \sum_j p_{0j} v_j^\alpha & , \ i = N+1 \end{cases}$$

by the definition of i_*, we obtain

$$v_i^\alpha = \begin{cases} \alpha \sum_j p_{ij} v_j^\alpha & , \ 0 \le i \le i_* \\ c_0 + \alpha \sum_j p_{0j} v_j^\alpha & , \ i_* < i \le N \\ c_0 + c_1 + \alpha \sum_j p_{0j} v_j^\alpha & , \ i = N+1 \end{cases}$$

implying that the control-limit policy f_*^∞ is optimal. □

Theorem 8.3 implies that the next algorithm computes an optimal control-limit policy for this model. Similar to Algorithm 8.1 it can be shown that the complexity of Algorithm 8.2 is $\mathcal{O}(N^3)$.

Algorithm 8.2 *Computation of an optimal control-limit policy.*

1. (a) Start with the basic solution corresponding to the nonreplacing actions in the states $i = 1, 2, \ldots, N$ and to the only action in the states 0 and $N+1$.
   (b) Let $k = N$ (the number of nonbasic variables corresponding to the replacing actions in the states $i = 1, 2, \ldots, N$).
2. If the reduced costs are nonnegative: the corresponding policy is optimal (STOP).

   Otherwise:
   (a) Choose the column corresponding to state k as pivot column.
   (b) Execute the usual simplex transformation.
   (c) Delete the pivot column.
3. If all columns are removed: replacement in all states is the optimal policy (STOP).

   Otherwise: return to step 2.

Remark

Next, we consider the average reward. By Theorem 8.3 for each $\alpha \in (0,1)$ there exists a control-limit policy f_α^∞ that is α-discounted optimal. Let $\{\alpha_k,\ 1,2,\dots\}$ be any sequence of discount factors such that $lim_{k\to\infty}\alpha_k = 1$. Since there are only a finite number of different control-limit policies, there is a subsequence with one of these policies. Therefore, we may assume that $f_{\alpha_k}^\infty = f_0^\infty$ for all k. Letting $k \to \infty$, we obtain for every $f^\infty \in C(D)$,

$$\phi(f^\infty) = lim_{k\to\infty}(1-\alpha_k)v_k^\alpha(f^\infty) \geq lim_{k\to\infty}(1-\alpha_k)v_k^\alpha(f_0^\infty) = \phi(f_0^\infty).$$

Therefore, also for the average reward criterion there exists a control-limit optimal policy.

8.1.3 Skip to the right model with failure

This model is a slightly different from the previous one. Let the state space $S = \{0,1,\dots,N+1\}$, where state 0 corresponds to a new item and state $N+1$ to failure. The states $i,\ 1 \leq i \leq N$, may be interpreted as the age of the item. The system has in state i a failure probability p_i during the next period. When failure occurs in state i, which is modelled as being transferred to state $N+1$, there is an additional cost f_i. In state $N+1$ the item has to be replaced by a new one. When there is no failure in state i, the next state is state $i+1$: the system skips to the right, i.e. the age of the item increases. The action sets, the cost of a new item, the maintenance costs and the transition probabilities are as follows.

$$S = \{0,1,\dots,N+1\};\ A(0) = \{1\};\ A(0) = \{0,1\},\ 1 \leq i \leq N;\ A(N+1) = \{0\}.$$

$$1 \leq i \leq N+1: \quad p_{ij}(0) = \begin{cases} 1-p_0 & j=1 \\ p_0 & j = N+1 \end{cases}; \quad c_i(0) = c + c_0 + p_0 f_0$$

$$0 \leq i \leq N: \quad p_{ij}(1) = \begin{cases} 1-p_i & j = i+1 \\ p_i & j = N+1 \end{cases}; \quad c_i(1) = c_i + p_i f_i$$

We impose the following assumptions:

(A1) $c \geq 0;\ c_i \geq 0,\ f_i \geq 0,\ 0 \leq i \leq N$.

(A2) $p_0 \leq p_1 \leq \cdots \leq p_N$, i.e. older items have greater failure probability.

(A3) $c_0 + p_0 f_0 \leq c_1 + p_1 f_1 \leq \cdots \leq c_N + p_N f_N$, i.e. the expected maintenance and failure costs grow with the age of the item.

Take any $k \in S$. Since $\sum_{j=k}^{N+1} p_{ij}(1) = \begin{cases} p_i & i \leq k-2 \\ 1 & i \geq k-1 \end{cases}$, this summation is, by assumption A2, nondeceasing in i. Hence, Assumption 8.1, and consequently also Assumption 8.2, is satisfied. This enables us to treat this model in a similar way as the previous one.

The method of value iteration for this model becomes (with $v_i^0 = 0$ for all i) for $n = 0,1,\dots$

$$v_i^{n+1} = \begin{cases} c_0 + p_0 f_0 + \alpha \sum_j p_{0j}(1) v_j^n & ,\ i = 0 \\ min\{c + c_0 + p_0 f_0 + \alpha \sum_j p_{ij}(0) v_j^n,\ c_i + p_i f_i + \alpha \sum_j p_{ij}(1) v_j^n\} & ,\ 1 \leq i \leq N \\ c + c_0 + p_0 f_0 + \alpha \sum_j p_{N+1j}(0) v_j^n & ,\ i = N+1. \end{cases} \tag{8.10}$$

Since $p_{ij}(0) = p_{0j}(1)$ for all i and j, equation (8.10) is equivalent to

$$v_i^{n+1} = \begin{cases} c_0 + p_0 f_0 + \alpha \sum_j p_{0j}(1) v_j^n & , \ i = 0 \\ min\{c + c_0 + p_0 f_0 + \alpha \sum_j p_{0j}(1) v_j^n, \ c_i + p_i f_i + \alpha \sum_j p_{ij}(1) v_j^n\} & , \ 1 \le i \le N \\ c + c_0 + p_0 f_0 + \alpha \sum_j p_{0j}(1) v_j^n & , \ i = N+1. \end{cases} \tag{8.11}$$

Similarly to the analysis in the previous section, we can derive the following results.

Theorem 8.4

(1) v_i^n *is nondecreasing in* i *for every* $n = 0, 1, \ldots$.

(2) *Let* i_* *be such that* $i_* = max\{i \mid c_i + p_i f_i + \alpha \sum_j p_{ij}(1) v_j^\alpha \le c + c_0 + p_0 f_0 + \alpha \sum_j p_{0j}(1) v_j^\alpha\}$. *Then, the control-limit policy* f_*^∞ *which replaces in the states* $i > i_*$ *is an optimal policy.*

Remarks:

1. Algorithm 8.2 is also applicable to this model.
2. Similarly as in the previous section it can be shown that for the average reward criterion there exists also a control-limit optimal policy.

8.1.4 A separable replacement problem

Suppose that the MDP has the following structure:

$S = \{1, 2, \ldots, N\}$; $A(i) = \{1, 2, \ldots, M\}, i \in S$.

$p_{ij}(a) = p_j(a)$, $i, j \in S$, $a \in A(i)$, i.e. the transitions are state independent.

$r_i(a) = s_i + t(a)$, $i \in S$, $a \in A(i)$, i.e. the rewards are separable.

This model is a special case of a separable MDP (see also Section 8.7).

As example, consider the problem of periodically replacing a car. When a car is replaced, it can be replaced not only by a new one, but also by a car in an arbitrary state. Let s_i be the trade-in-value of a car of state i, $t(a)$ the costs of a car of state a. Then, $r_i(a) = s_i - t(a)$ and $p_{ij}(a) = p_j(a)$, where $p_j(a)$ is the probability that a car of state a is in state j at the next decision time point.

We fist consider as optimality criterion the discounted expected rewards. The next theorem shows that a one-step look ahead policy is optimal.

Theorem 8.5

Let a_* *be such that* $-t(a_*) + \alpha \sum_j p_j(a_*) s_j = max_{1 \le a \le M}\{-t(a) + \alpha \sum_j p_j(a) s_j\}$.

Then, the policy f_*^∞, *defined by* $f_*(i) = a_*$ *for every* $i \in S$, *is an* α*-discounted optimal policy.*

Proof

We first show that if an action, say a_1, is optimal in state 1 this action is also optimal in the other states. Let a_1 optimal in state 1, i.e.

$$
\begin{aligned}
r_1(a_1) + \alpha \textstyle\sum_j p_{1j}(a_1) v_j^\alpha &\geq r_1(a) + \alpha \textstyle\sum_j p_{1j}(a) v_j^\alpha,\ a \in A(1) &\Leftrightarrow \\
s_1 - t(a_1) + \alpha \textstyle\sum_j p_j(a_1) v_j^\alpha &\geq s_1 - t(a) + \alpha \textstyle\sum_j p_j(a) v_j^\alpha,\ 1 \leq a \leq M &\Leftrightarrow \\
s_i - t(a_1) + \alpha \textstyle\sum_j p_j(a_1) v_j^\alpha &\geq s_i - t(a) + \alpha \textstyle\sum_j p_j(a) v_j^\alpha,\ i \in S,\ 1 \leq a \leq M &\Leftrightarrow \\
r_i(a_1) + \alpha \textstyle\sum_j p_{i1j}(a_1) v_j^\alpha &\geq r_i(a) + \alpha \textstyle\sum_j p_{ij}(a) v_j^\alpha\ i \in S,\ a \in A(i),
\end{aligned}
$$

i.e. action 1 is optimal in all states. Hence, we may restrict the policies to the set of actions $\{1, 2, \ldots, M\}$ and we have to decide which $a \in \{1, 2, \ldots, M\}$ is the optimal one. For any choice $f^\infty \in C(D)$ with $f(i) = a,\ i \in S$, we have

$$
v_i^\alpha(f^\infty) = s_i - t(a) + \alpha \textstyle\sum_j p_j(a) v_j^\alpha(f^\infty) \ = \ s_i + c(a),\ i \in S,
$$

where $c(a) = -t(a) + \alpha \sum_j p_j(a) v_j^\alpha(f^\infty)$.
A policy $f_*^\infty \in C(D)$ with $f(i) = a_*$ is α-discounted optimal if and only if

$$
\begin{aligned}
s_i - t(a_*) + \alpha \textstyle\sum_j p_j(a_*) v_j^\alpha(f_*^\infty) &\geq s_i - t(a) + \alpha \textstyle\sum_j p_j(a) v_j^\alpha(f_*^\infty)\ \forall (i,a) &\Leftrightarrow \\
s_i - t(a_*) + \alpha \textstyle\sum_j p_j(a_*) \{s_j + c(a_*)\} &\geq s_i - t(a) + \alpha \textstyle\sum_j p_j(a) \{s_j + c(a_*)\}\ \forall (i,a) &\Leftrightarrow \\
-t(a_*) + \alpha \textstyle\sum_j p_j(a_*) s_j &\geq -t(a) + \alpha \textstyle\sum_j p_j(a) s_j,\ 1 \leq a \leq M &\Leftrightarrow \\
-t(a_*) + \alpha \textstyle\sum_j p_j(a_*) s_j &= max_{1 \leq a \leq M} \{-t(a) + \alpha \textstyle\sum_j p_j(a) s_j\}. &\square
\end{aligned}
$$

Corollary 8.1
Let a_0 be such that $t(a_0) + \sum_j p_j(a_0) s_j = max_{1 \leq a \leq M} \{t(a) + \sum_j p_j(a) s_j\}$.
Then, the policy f_0^∞, defined by $f_0(i) = a_0$ for every $i \in S$, is a Blackwell optimal, and therefore also average optimal, policy.

Proof
From Theorem 8.5 it follows that f_0^∞ is an α-discounted optimal policy for all $\alpha \in [\alpha_0, 1)$ for some $\alpha_0 \in [0, 1)$. Therefore, f_0^∞ is a Blackwell optimal policy. $\square$

8.2 Maintenance and repair problems

8.2.1 A surveillance-maintenance-replacement model

Consider a system, in use or in storage, which is deteriorating. Suppose that the deteriorating occurs stochastically and that the conditioning of the system is known only if it is inspected, which is costly.
After inspection the manager of the system has two alternatives: (1) to replace the item by a new item; (2) to keep the item and do some repair on it. Under the second alternative he must decide the extend of repairs to be made and when to make the next inspection. If inspection is put too long the system may fail in the interim, the consequence of which is an incurred cost which is a function of how long the system has been inoperative. Assume that M denotes the upper bound on the number of periods that can elapse without an inspection.

Let us suppose that the uninspected system evolves according to a Markov chain with states $\{0, 1, \ldots, N+1\}$. The state 0 denotes a new system and the state $N+1$ an inoperative system. Let $P = (p_{ij})$ denote the transition matrix of this Markov chain with $p_{iN+1} > 0,\ 0 \leq i \leq N+1$ and $p_{N+1,N+1} = 1$. Assume that when a replacement is made an instantaneous transition to state 0 takes place; when a repair is made an instantaneous transition takes place to one of the states $1, 2, \ldots, N$, depending on the extend of repairs. Replacement or repairs are only made at the time of inspections.

Since there is a failure cost depending how long the system has been inoperative we use additional states $N+1(1), N+1(2), \ldots, N+1(M)$, where $N+1(m)$ denotes the fact that the system is observed to be in state $N+1$ and has been in state $N+1$ for m uninspected periods. Hence, the state space S will consist of the states $0, 1, \ldots, N+1, N+1(1), N+1(2), \ldots, N+1(M)$.

Let c_i denote the cost of inspection when the system is in state i. Let $r_{ij},\ 1 \leq i \leq N+1$, $0 \leq j \leq N$ denote the cost to repair the system from state i to state j. In particular, r_{i0} is the cost to replace the item by a new one. In addition we let $r_{N+1(m)j},\ m = 1, 2, \ldots, M$, denote the cost to place the system in state j from state $N+1$ when prior to discovering the system in state $N+1$, the system has been in state $N+1$ for m uninspected periods. This cost represents, in addition to the repair or replacement costs, the cost associated with undetected failure.

At each state $i = 0, 1, \ldots, N$, an action a_{lm} consists in placing the system in state $l,\ 0 \leq l \leq N$, and deciding to skip m $(0 \leq m \leq M)$ time periods before observing the system again. If the system is observed in one of the states $N+1, N+1(1), \ldots, N+1(M)$, we assume that $a_{l0},\ 0 \leq l \leq N$ are the only possible actions. Hence, when the system is inspected in state i and action a_{jm} is chosen there are inspection and repair costs

$$c_i(a_{jm}) = c_i + r_{ij} \text{ for each } i, j \text{ and } m$$

and transition probabilities

$$p_{ij}(a_{lm}) = p_{lj}^{(m+1)} \text{ for each } i, j, l \text{ and } m.$$

As optimality criterion, we are interested in minimizing the expected average cost per unit time attributed to a surveillance-replacement-maintenance policy. Let $\{X_t, Y_t,\ t = 1, 2 \ldots\}$ be the observed states and actions, respectively, and let the quantities Z_{tijm} and $\overline{Z}_{Tijm}$ be defined by

$$Z_{tijm} = \begin{cases} 1 & \text{if } X_t = i,\ Y_t = a_{jm} \\ 0 & \text{otherwise} \end{cases} \quad t = 1, 2, \ldots;\ \overline{Z}_{Tijm} = \frac{1}{T} \sum_{t=1}^{T} Z_{tijm},\ T = 1, 2, \ldots.$$

If J_T is the average cost up to time T, then for each $i, j \in S$ we have

$$J_T = \frac{\sum_{t=1}^{t}(T) \sum_{i,j,m} (c_i + r_{ij}) Z_{tijm}}{\sum_{t=1}^{t}(T) \sum_{i,j,m} (m+1) Z_{tijm} + \theta},$$

where $t(T)$ is the last inspection time less than or equal to T and $\theta = T - t(T)$. We also have

$$J_T = \frac{\sum_{i,j,m}(c_i + r_{ij})\overline{Z}_{t(T)ijm}}{\sum_{i,j,m}(m+1)\overline{Z}_{t(T)ijm} + \frac{\theta}{T}}.$$

Notice that $t(T) \to \infty$ when $T \to \infty$. It can be shown that

$$lim_{T\to\infty}\,\overline{Z}_{t(T)ijm} = lim_{T\to\infty}\,\overline{Z}_{Tijm} = lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}\{X_t = i,\ Y_t = a_{jm}\}$$

exists for all i, j, m (cf. Chapter 7 in [55] or Section 4.7 in [108]). Denote $lim_{T\to\infty}\,\overline{Z}_{Tijm}$ by $z_{i,j,m}$. Therefore, we obtain for the limiting average cost

$$lim_{T\to\infty}\,J_T = \frac{\sum_{i,j,m}(c_i + r_{ij})z_{ijm}}{\sum_{i,j,m}(m+1)z_{ijm}}.$$

Notice that the underlying Markov chain is a unichain Markov chain with state $N+1$ as the only absorbing state. It also can be shown that in this case (see again Chapter 7 in [55] or Section 4.7 in [108]) that there exists a stationary optimal policy π^∞ which can be derived from the fractional linear program (see also section 6.2 and program (6.3)

$$min\left\{\frac{\sum_{i,j,m}(c_i + r_{ij})x_{ijm}}{\sum_{i,j,m}(m+1)x_{ijm}}\ \middle|\ \begin{array}{lcl}\sum_{i,l,m}\{\delta_{ij} - p_{ij}(a_{lm})\}x_i(a_{lm}) & = & 0 \text{ for all } j\\ \sum_{i,l,m}\ x_i(a_{lm}) & = & 1\\ x_i(a_{jm}) \geq 0 \text{ for all } i,l,m & & \end{array}\right\}. \tag{8.12}$$

Let x be an optimal solution of program (8.12). Then, π^∞ with $\pi_{ia_{jm}} = \frac{x_{ia_{jm}}}{\sum_{j,m}\,x_{ia_{jm}}}$ for all i, j, m is an optimal stationary policy.

The above problem involves minimizing a ratio of linear functions subject to linear constraints where the lower linear form is always positive. Any problem of this form can always be transformed to a linear programming problem. Namely, suppose we have the fractional problem

$$min\left\{\frac{\sum_{i=1}^{n}\,c_i x_i}{\sum_{i=1}^{n}\,d_i x_i}\ \middle|\ \begin{array}{ll}\sum_{i=1}^{n}\,a_{ji}x_i = 0,\ j = 1,2,\ldots,m\\ \sum_{i=1}^{n}\,x_i\ \ = 1\\ x_i \geq 0,\ i = 1,2,\ldots,n\end{array}\right\}, \tag{8.13}$$

where $\sum_{i=1}^{n}\,d_i x_i > 0$ for any feasible x. Set

$$z_i = \frac{x_i}{\sum_{i=1}^{n}\,d_i x_i},\ i = 1,2,\ldots,n;\ z_{n+1} = \frac{1}{\sum_{i=1}^{n}\,d_i x_i}.$$

Then we can write

$$min\left\{\sum_{i=1}^{n}\,c_i x_i\ \middle|\ \begin{array}{ll}\sum_{i=1}^{n}\,a_{ji}z_i\ \ = 0,\ j = 1,2,\ldots,m\\ \sum_{i=1}^{n}\,d_i z_i\ \ = 1\\ \sum_{i=1}^{n}\,z_i - z_{n+1} = 0\\ z_i \geq 0,\ i = 1,2,\ldots,n+1\end{array}\right\}. \tag{8.14}$$

From the one-to-one relation between the fractional and linear program (remark that the reverse mapping is $x_i = \frac{z_i}{z_{n+1}},\ 1 \leq i \leq n$) it follows that if z is an optimal solution of the linear program derived from the fractional program (8.12), then π^∞ with $\pi_{ia_{jm}} = \frac{z_{ia_{jm}}}{\sum_{j,m} z_{ia_{jm}}}$ for all i, j, m is an optimal stationary policy. It can also be shown that, for each state i, $z_{ia_{jm}}$ will be strictly positive for exactly one action a_{jm}. Thus, the optimal policy is stationary and deterministic.

8.2.2 Optimal repair allocation in a series system

Consider the maintenance and repair problem of Section 1.3.5. For this model it can be shown that the optimal policy is irrespective of the repair rates $\mu_i,\ 1 \leq i \leq n$, and is the policy that assigns the repairman to the failed component with the smallest failure rate $\lambda_i,\ 1 \leq i \leq n$, (*SFR policy*), i.e. the longest expected lifetime.

When a policy $f^\infty \in C(D)$ is employed, the time evaluation of the state of the system can be described as a continuous, irreducible Markov chain. Furthermore, the average expected system operation time is equal to the probability of the system being in the functioning state $\underline{1} = (1, 1, \ldots, 1)$ (see [168]). Under a policy $f^\infty \in C(D)$ returns to state $\underline{1}$ generate a renewal process. Employing a renewal argument it can be shown (see [187]) that maximizing the probability of the system being in state $\underline{1}$ is equivalent to minimizing the expected first passage time to state $\underline{1}$ over all initial states.

Let $T_f(x)$ denote the expected first passage time from state x to state $\underline{1}$ when policy f^∞ is employed. The above remarks are formally stated in the following lemma.

Lemma 8.2
A policy $f_*^\infty \in C(D)$ is optimal with respect to the expected average system operation time if and only if $T_{f_*}(x) \leq T_f(x)$ for all $x \neq \underline{1}$ and all policies $f^\infty \in C(D)$.

The action $f(x) = i$ means that the repairman is assigned to component i in state x. We also use the notations:

$$(1_k, x) = (x_1, x_2, \ldots, x_{k-1}, 1, x_{k+1}, \ldots, x_n);\ C_1(x) = \{i \mid x_i = 1\};\ \lambda_1(x) = \textstyle\sum_{i \in C_1(x)}\ \lambda_i;$$
$$(0_k, x) = (x_1, x_2, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n);\ C_0(x) = \{i \mid x_i = 0\}.$$

Given policy f^∞, the Markov chain remains in state x during an exponentially distributed time with rate $\lambda_1(x) + \mu_{f(x)}$. The transition probabilities of the Markov chain satisfy

$$p_{x,(1_{f(x)},x)}(f(x)) = \frac{\mu_{f(x)}}{\lambda_1(x) + \mu_{f(x)}};\ p_{x,(0_k,x)}(f(x)) = \frac{\lambda_k}{\lambda_1(x) + \mu_{f(x)}},\ k \in C_1(x).$$

Notice that by conditioning on the first transition out state x we see that $T_f(x)$ can be obtained as unique solution of the following system of linear equations

$$\begin{cases} T_f(x) &= \frac{1}{\lambda_1(x)+\mu_{f(x)}}\{1 + \mu_{f(x)} T_f(1_{f(x)}, x) + \sum_{i \in C_1(x)}\ \lambda_i T_f(0_i, x)\},\ x \neq \underline{1} \\ T_f(\underline{1}) &= 0. \end{cases} \tag{8.15}$$

A standard result in MDP is that the policy $f_*^\infty \in C(D)$ is optimal if and only if the associated expectd first passage times $T_{f_*}(x)$ satisfy the following functional equation

$$T_{f_*}(x) = min_{j\in C_0(x)}\Big\{\frac{1}{\lambda_1(x)+\mu_j}\{1+\mu_j T_{f_*}(1_j,x)+\sum_{i\in C_1(x)}\lambda_i T_{f_*}(0_i,x)\}\Big\},\ x\in S. \tag{8.16}$$

Since

$$\begin{aligned}\{\lambda_1(x)+\mu_j\}T_{f_*}(x) &\le (=) \quad 1+\mu_j T_{f_*}(1_j,x)+\textstyle\sum_{i\in C_1(x)}\lambda_i T_{f_*}(0_i,x) \ \Leftrightarrow\\ \{\textstyle\sum_{i=1}^n(\lambda_i+\mu_i)\}T_{f_*}(x) &\le (=) \quad 1+\mu_j T_{f_*}(1_j,x)+\textstyle\sum_{k\neq j}\mu_k T_{f_*}(x)+\\ &\qquad \textstyle\sum_{i\in C_1(x)}\lambda_i T_{f_*}(0_i,x)+\sum_{i\notin C_1(x)}\lambda_i T_{f_*}(,x)\end{aligned}$$

the optimality equation (8.16) is equivalent to the optimality equation

$$T_{f_*}(x) = min_{j\in C_0(x)}\Big\{\frac{1}{\sum_{i=1}^n(\lambda_i+\mu_i)}\{1+\mu_j T_{f_*}(1_j,x)+\sum_{k\neq j}\mu_k T_{f_*}(x)+\sum_{i\in C_1(x)}\lambda_i T_{f_*}(0_i,x)+\sum_{i\notin C_1(x)}\lambda_i T_{f_*}(x)\}\Big\}. \tag{8.17}$$

Because

$$\mu_j T_{f_*}(1_j,x)+\textstyle\sum_{k\neq j}\mu_k T_{f_*}(x)+\sum_{i\in C_1(x)}\lambda_i T_{f_*}(0_i,x)+\sum_{i\notin C_1(x)}\lambda_i T_{f_*}(x) =$$
$$\textstyle\sum_{k=1}^n\mu_k T_{f_*}(x)+\mu_j\{T_{f_*}(1_j,x)-T_{f_*}(x)\}+\sum_{i=1}^n\lambda_i T_{f_*}(0_i,x)+\sum_{i\notin C_1(x)}\lambda_i\{T_{f_*}(x)-T_{f_*}(0_i,x)\}$$

and

$$T_{f_*}(x) = T_{f_*}(1_j,x) \text{ for all } j\in C_1(x), \text{ and } T_{f_*}(x) = T_{f_*}(0_i,x) \text{ for all } i\notin C_1(x),$$

equation (8.17) is equivalent to

$$T_{f_*}(x) = \frac{1}{\theta}\Big\{1+\sum_{k=1}^n\mu_k T_{f_*}(x)+\sum_{i=1}^n\lambda_i T_{f_*}(0_i,x)+min_{j\in C_0(x)}\{\mu_j\{T_{f_*}(1_j,x)-T_{f_*}(0_j,x)\}\}\Big\}, \tag{8.18}$$

where $\theta = \sum_{i=1}^n(\lambda_i+\mu_i)$.

Suppose $\lambda_1<\lambda_2<\cdots<\lambda_n$, and let $f_*^\infty\in C(D)$ be the policy which always puts repair on the component of smallest index, i.e. the policy we are trying to prove optimal. From (8.18) it is clear that f_*^∞ is optimal if for all $i<j$ with $i,j\in C_0(x)$ we have

$$\mu_i\{T_{f_*}(1_i,x)-T_{f_*}(0_i,x)\}\le\mu_j\{T_{f_*}(1_j,x)-T_{f_*}(0_j,x)\}. \tag{8.19}$$

Imagine solving the problem using value iteration on the optimality equation written as (8.17), i.e.

$$T^{m+1}(x) = min_{j\in C_0(x)}\Big\{\frac{1}{\theta}\{1+\mu_j T^m(1_j,x)+\sum_{k\neq j}\mu_k T^m(x)+\sum_{i\in C_1(x)}\lambda_i T^m(0_i,x)+\sum_{i\notin C_1(x)}\lambda_i T^m(x)\}\Big\}. \tag{8.20}$$

Consider the following inductive hypothesis:

$$H(m):\ \begin{cases}\mu_i\{T^m(1_i,x)-T^m(0_i,x)\}\le\mu_j\{T^m(1_j,x)-T^m(0_j,x)\}\le 0\\ \text{for all } x \text{ and all } i<j \text{ with } i,j\in C_0(x).\end{cases}$$

We can take $T^0(x) = 0$, so $H(0)$ is true. Assuming $H(m)$ is true for all $m = 0, 1, \ldots$, then the minimizing value is in each iteration $f_*(x) = min_{i \in C_0(x)} \{i\}$. If we succeed in proving the inductive step we have shown the structure of the optimal policy as a SFR-policy.

Assume that $H(m)$ is true. Then, for all states x we have

$$T^{m+1}(x) = \frac{1}{\theta}\big\{1 + \mu_{f_*(x)} T^m(1_{f_*(x)}, x) + \sum_{k \neq f_*(x)} \mu_k T^m(x) + \sum_{k \in C_1(x)} \lambda_k T^m(0_k, x) + \sum_{k \notin C_1(x)} \lambda_k T^m(x)\big\}, \tag{8.21}$$

where $f_*(x) = min_{i \in C_0(x)} \{i\}$. For the inductive step we have to show

$$H(m+1): \begin{cases} \mu_i\{T^{m+1}(1_i, x) - T^{m+1}(0_i, x)\} \leq \mu_j\{T^{m+1}(1_j, x) - T^{m+1}(0_j, x)\} \leq 0 \\ \text{for all } x \text{ and all } i < j \text{ with } i, j \in C_0(x) \end{cases} \tag{8.22}$$

Take any state x and consider first the case in which i and j are the two smallest indices of $C_0(x)$. Notice that $(1_i, x) = (1_i, 0_j, x)$, $(0_i, x) = (0_i, 0_j, x)$, $f_*(1_i, x) = j$ and $f_*(0_i, x) = i$.
Then, we can write

$$\begin{aligned} T^{m+1}(1_i, x) &= \tfrac{1}{\theta}\big\{1 + \mu_j T^m(1_i, 1_j, x) + \mu_i T^m(1_i, 0_j, x) + \textstyle\sum_{k \neq i,j} \mu_k T^m(1_i, 0_j, x) + \\ &\qquad \textstyle\sum_{k \in C_1(1_i, x)} \lambda_k T^m(1_i, 0_k, x) + \sum_{k \in C_0(1_i, x)} \lambda_k T^m(1_i, x)\big\} \end{aligned}$$

Since

$$\textstyle\sum_{k \in C_1(1_i, x)} \lambda_k T^m(1_i, 0_k, x) = \lambda_i T^m(0_i, 0_j, x) + \sum_{k \in C_1(1_i, 0_j, x), k \neq i} \lambda_k T^m(1_i, 0_j, 0_k, x)$$

and

$$\textstyle\sum_{k \in C_0(1_i, x)} \lambda_k T^m(1_i, x) = \lambda_j T^m(1_i, 0_j, x) + \sum_{k \in C_0(1_i, 0_j, x), k \neq j} \lambda_k T^m(1_i, 0_j, 0_k, x),$$

we obtain

$$\begin{aligned} T^{m+1}(1_i, x) &= \tfrac{1}{\theta}\big\{1 + \mu_j T^m(1_i, 1_j, x) + \mu_i T^m(1_i, 0_j, x) + \textstyle\sum_{k \neq i,j} \mu_k T^m(1_i, 0_j, x) + \\ &\qquad \lambda_i T^m(0_i, 0_j, x) + \lambda_j T^m(1_i, 0_j, x) + \textstyle\sum_{k \neq i,j} \lambda_k T^m(1_i, 0_j, 0_k, x)\big\}. \end{aligned}$$

Similarly we obtain

$$\begin{aligned} T^{m+1}(0_i, x) &= \tfrac{1}{\theta}\big\{1 + \mu_i T^m(1_i, 0_j, x) + \mu_j T^m(0_i, 0_j, x) + \textstyle\sum_{k \neq i,j} \mu_k T^m(0_i, 0_j, x) + \\ &\qquad \lambda_i T^m(0_i, 0_j, x) + \lambda_j T^m(0_i, 0_j, x) + \textstyle\sum_{k \neq i,j} \lambda_k T^m(0_i, 0_j, 0_k, x)\big\}. \\ T^{m+1}(1_j, x) &= \tfrac{1}{\theta}\big\{1 + \mu_i T^m(1_i, 1_j, x) + \mu_j T^m(0_i, 1_j, x) + \textstyle\sum_{k \neq i,j} \mu_k T^m(0_i, 1_j, x) + \\ &\qquad \lambda_j T^m(0_i, 0_j, x) + \lambda_i T^m(0_i, 1_j, x) + \textstyle\sum_{k \neq i,j} \lambda_k T^m(0_i, 1_j, 0_k, x)\big\}. \\ T^{m+1}(0_j, x) &= \tfrac{1}{\theta}\big\{1 + \mu_i T^m(1_i, 0_j, x) + \mu_j T^m(0_i, 0_j, x) + \textstyle\sum_{k \neq i,j} \mu_k T^m(0_i, 0_j, x) + \\ &\qquad \lambda_j T^m(0_i, 0_j, x) + \lambda_i T^m(0_i, 0_j, x) + \textstyle\sum_{k \neq i,j} \lambda_k T^m(0_i, 0_j, 0_k, x)\big\}. \end{aligned}$$

Now we have

$$\begin{aligned} &\mu_i\{T^{m+1}(1_i, x) - T^{m+1}(0_i, x)\} - \mu_j\{T^{m+1}(1_j, x) - T^{m+1}(0_j, x)\} = \\ &\mu_i\{\mu_j T^m(1_i, 1_j, x) - \mu_i T^m(1_i, 0_j, x)\} + \qquad (1) \\ &\mu_i\{\mu_i T^m(1_i, 0_j, x) - \mu_j T^m(0_i, 0_j, x)\} + \qquad (2) \end{aligned}$$

$$\mu_i\{\textstyle\sum_{k\neq i,j}\mu_k T^m(1_i,0_j,x)-\sum_{k\neq i,j}\mu_k T^m(0_i,0_j,x)\}+ \tag{3}$$
$$\mu_i\{\lambda_i T^m(0_i,0_j,x)-\lambda_i T^m(0_i,0_j,x)\}+ \tag{4}$$
$$\mu_i\{\lambda_j T^m(1_i,0_j,x)-\lambda_j T^m(0_i,0_j,x)\}+ \tag{5}$$
$$\mu_i\{\textstyle\sum_{k\neq i,j}\lambda_k T^m(1_i,0_j,0_k,x)-\sum_{k\neq i,j}\lambda_k T^m(0_i,0_j,0_k,x)\}+ \tag{6}$$
$$\mu_j\{\mu_i T^m(1_i,0_j,x)-\mu_i T^m(1_i,1_j,x)\}+ \tag{7}$$
$$\mu_j\{\mu_j T^m(0_i,0_j,x)-\mu_j T^m(0_i,1_j,x)\}+ \tag{8}$$
$$\mu_j\{\textstyle\sum_{k\neq i,j}\mu_k T^m(0_i,0_j,x)-\sum_{k\neq i,j}\mu_k T^m(0_i,1_j,x)\}+ \tag{9}$$
$$\mu_j\{\lambda_j T^m(0_i,0_j,x)-\lambda_j T^m(0_i,0_j,x)\}+ \tag{10}$$
$$\mu_j\{\lambda_i T^m(0_i,0_j,x)-\lambda_i T^m(0_i,1_j,x)\}+ \tag{11}$$
$$\mu_j\{\textstyle\sum_{k\neq i,j}\lambda_k T^m(0_i,0_j,0_k,x)-\sum_{k\neq i,j}\lambda_k T^m(0_i,1_j,0_k,x)\}. \tag{12}$$

We show that $(1)+(2)+\cdots+(12)\leq 0$. Therefore, we first remark that $(4)=(10)=0$. Furthermore, we mention

$$\begin{aligned}
(6)+(12) &= \textstyle\sum_{k\neq i,j}\lambda_k\big\{\mu_i\{T^m(1_i,0_j,0_k,x)-T^m(0_i,0_j,0_k,x)\}-\\
&\qquad\qquad \mu_j\{T^m(0_i,1_j,0_k,x)-T^m(0_i,0_j,0_k,x)\}\big\}\\
&\leq 0 \text{ by the inductive hypothesis } H(m).\\
(5)+(10) &= (\lambda_j-\lambda_i)\big\{\mu_i\{T^m(1_i,0_j,x)-T^m(0_i,0_j,x)\}-\mu_j\{T^m(0_i,1_j,x)-T^m(0_i,0_j,x)\}\big\}\\
&\leq 0 \text{ because } \lambda_j>\lambda_i \text{ and by the inductive hypothesis } H(m).\\
(3)+(9) &= \textstyle\sum_{k\neq i,j}\mu_k\big\{\mu_i\{T^m(1_i,0_j,x)-T^m(0_i,0_j,x)\}-\mu_j\{T^m(0_i,1_j,x)-T^m(0_i,0_j,x)\}\big\}\\
&\leq 0 \text{ by the inductive hypothesis } H(m).
\end{aligned}$$

$$\begin{aligned}
(1)+(2)+(7)+(8) &= \mu_i\mu_j T^m(1_i,1_j,x)-\mu_i^2 T^m(1_i,0_j,x)+\mu_i^2 T^m(1_i,0_j,x)-\\
&\quad \mu_i\mu_j T^m(0_i,0_j,x)+\mu_i\mu_j T^m(1_i,0_j,x)-\mu_i\mu_j T^m(1_i,1_j,x)+\\
&\quad \mu_j^2 T^m(0_i,0_j,x)-\mu_j^2 T^m(0_i,1_j,x)\\
&= \mu_j\big\{\mu_j\{T^m(1_i,0_j,x)-T^m(0_i,0_j,x)\}-\mu_j\{T^m(0_i,1_j,x)-T^m(0_i,0_j,x)\}\big\}\\
&\leq 0 \text{ by the inductive hypothesis } H(m).
\end{aligned}$$

This is the hardest case to check. The nonnegativity of $\mu_i\{T^{m+1}(1_i,x)-T^{m+1}(0_i,x)\}$ and the other cases are left to the reader (see also Exercise 8.1).

8.3 Production and inventory control

8.3.1 No backlogging

The basic form of the production control model with no backlogging is as follows. Demand of a single product occurs during each of T consecutive time periods. The demand that occurs during a given period can be satisfied by production during that period or during any earlier period, as inventory is carried forward. This prescribes the case of *no backlogging*. Inventory at epoch 1 is zero, and inventory at the end of period T is also required to be zero. The model includes

production costs and inventory costs. The objective is to schedule the production so as to satisfy demand at minimum total cost.

For the data and variables in the periods $t = 1, 2, \ldots, T$ we use the following notation.

$$\begin{array}{rcl} D_t & = & \text{the demand in period } t \\ c_t(a) & = & \text{the cost of production } a \text{ units in period } t \\ h_t(i) & = & \text{inventory cost when the inventory is } i \text{ at the end of period } t \\ a_t & = & \text{decision variable for the production in period } t \\ I_t & = & \text{the inventory on hand at the beginning of period } t \end{array}$$

When the production and demand occur in integer quantities, the problem of meeting demand at minimal total cost can be formulated as the following integer programming problem.

$$min \left\{ \sum_{t=1}^{T} \{c_t(a_t) + h_t(I_t)\} \;\middle|\; \begin{array}{ll} I_1 = I_{T+1} = 0 & \\ I_t + a_t = D_t + I_{t+1}, & t = 1, 2, \ldots, T \\ a_t \geq 0 \text{ and integer}, & t = 1, 2, \ldots, T \\ I_t \geq 0 \text{ and integer}, & t = 2, 3, \ldots, T \end{array} \right\}. \tag{8.23}$$

To find an optimal production plan by dynamic programming, note that the cost of operating the system during periods t through T depends on the inventory I_n, but not on prior inventories and not on prior production levels. Therefore, we constitute the states as the pairs (i, t), where i denote the inventory and t the period. Let

$$\begin{array}{rcl} f(i,t) & = & \text{the minimum cost of satisfying demand during periods } t \text{ through } T \text{ if the} \\ & & \text{inventory at epoch } t \text{ is } i. \end{array}$$

The cost of an optimal production plan is $f(0, 1)$. The inventory $I_{T+1} = 0$, which constrains inventory and production during period T. We obtain

$$f(i, T) = h_T(i) + c_T(D_T - i),\ i = 0, 1, \ldots, D_T. \tag{8.24}$$

Consider a period $t < T$, and let a_t denote the quantity produced during period t. The production a_t is feasible if $I_t + a_t$ is at least as large as the demand D_t and if $I_t + a_t$ is no larger that the total demand during all remaining periods. Therefore, a_t is a *feasible* production if $a_t \in A(i, t)$, where

$$A(i,t) = \{a \in \mathbb{N}_0 \mid D_t - i \leq a \leq \sum_{s=t}^{T} D_s - i\}. \tag{8.25}$$

The above observations give rise to the functional equation

$$\left\{ \begin{array}{rcl} f(i,T) & = & h_T(i) + c_T(D_T - i),\ i = 0, 1, \ldots, D_T. \\ f(i,t) & = & min_{a \in A(i,t)}\{h_t(i) + c_t(a) + f(i + a - D_t, t + 1)\},\ t = T-1, T-2, \ldots, 1,\ 0 \leq i \leq \sum_{s=t}^{T} D_s \end{array} \right. \tag{8.26}$$

The preceeding functional equation can be solved by backward induction.

In many economic situations the cost functions are concave, reflecting decreasing marginal costs. A function g with domain in $\mathbb{Z}$ is called *concave* if

$$g(n+2) - g(n+1) \leq g(n+1) - g(n) \text{ for all } n. \tag{8.27}$$

Let $\Delta g(n) = g(n+1) - g(n)$ and $\Delta^2 g(n) = \Delta g(n+1) - \Delta g(n) = g(n+2) - 2g(n+1) + g(n),\ n \in \mathbb{Z}$. Then, concavity is equivalent to $\Delta^2 g(n) \leq 0$ for all n. The next theorem shows that for concave production and inventory functions the optimal production plan has a special structure.

Theorem 8.6

If the production and inventory functions, c_t and h_t respectively, are concave for all periods t, then there exists an optimal production plan $(a_1, a_2, \ldots, a_T)$ for which $I_t \cdot a_t = 0$ for $t = 1, 2, \ldots, T$.

Proof

Consider an optimal production plan $(a_1, a_2, \ldots, a_{T+1})$. If there are more optimal plans, take the plan for which $\sum_{t=1}^{T}(a_t + I_t)$ is minimal. Aiming for a contradiction, we assume that $I_t > 0$ and $a_t > 0$ for some $2 \leq t \leq T$. Since, $I_t > 0$ there has been production in at least one of the previous periods and let s the last period prior than t in which $a_s > 0$. Then, $a_{s+1} = a_{s+2} = \cdots = a_{t-1} = 0$. Consider two other production plans, which are - of course - not cheaper than the optimal production plan:

$$a_j^{'} = \begin{cases} a_s + 1 & j = s \\ a_t - 1 & j = t \\ a_j & j \neq s, t \end{cases} \quad \text{and} \quad a_j^{''} = \begin{cases} a_s - 1 & j = s \\ a_t + 1 & j = t \\ a_j & j \neq s, t \end{cases}$$

Note that $I_k^{'} = I_k + 1,\ s < k \leq t$ and $I_k^{''} = I_k - 1,\ s < k \leq t$. Therefore,

$$c_s(a_s) + c_t(a_t) + \textstyle\sum_{k=s+1}^{t} h_k(I_k) \leq c_s(a_s + 1) + c_t(a_t - 1) + \sum_{k=s+1}^{t} h_k(I_k + 1)$$

and

$$c_s(a_s) + c_t(a_t) + \textstyle\sum_{k=s+1}^{t} h_k(I_k) \leq c_s(a_s - 1) + c_t(a_t + 1) + \sum_{k=s+1}^{t} h_k(I_k - 1).$$

Add these two inequalities and rearrange the sum as

$$0 \leq \Delta^2 c_s(a_s - 1) + \Delta^2 c_t(a_t - 1) + \textstyle\sum_{k=s+1}^{t} \Delta^2 h_k(I_k - 1) \leq 0,$$

the last inequality by the concaveness of the production and inventory functions. Hence, the three plans are all optimal. However, $\sum_{t=1}^{T}(a_t^{''} + I_t^{''}) = \sum_{t=1}^{T}(a_t + I_t) - (t - s) < \sum_{t=1}^{T}(a_t + I_t)$. This contradicts the supposed minimality of that sum, which completes the proof. □

Theorem 8.6 demonstrates that production need only occur in period t if the inventory at the start of that period is zero. Consequently the quantity produced during period t must equal the total demand of the periods $t, t+1, \ldots, u-1$ for some $t+1 \leq u \leq T+1$.
Let $d_{tu} = \sum_{k=t}^{u-1} D_k,\ t+1 \leq u \leq T+1$, the demand of the periods $t, t+1, \ldots, u-1$.
This gives rise to a dynamic programming formulation in which state t represents the situation of having no inventory on hand at the start of period t. Let

$f(t)$ = the minimum cost of satisfying demands during periods t through T if the inventory at epoch t is 0.

Let c_{tu} denote the total of costs incurred during periods $t, t+1, \ldots, u-1$ if transition occurs from state t to state u, i.e. the first production after period t is in period u. Inventory at epoch t equals 0, so c_{tu} includes the holding cost $h_t(0)$. Exactly d_{tu} units are produced in period t, so

c_{tu} includes the production cost $c_t(d_{tu})$. Exactly d_{ku} units of inventory remain at any epoch k between $t+1$ and u, so c_{tu} includes the holding cost $h_k(d_{ku})$. This leads to the relation

$$c_{tu} = h_t(0) + c_t(d_{tu}) + \sum_{k=t+1}^{u-1} h_k(d_{ku}). \tag{8.28}$$

and the functional equation

$$\begin{cases} f(T+1) &= 0 \\ f(t) &= min_{\{u \mid t+1 \le u \le T+1\}} \{c_{tu} + f(u)\},\ t = T, T-1, \ldots, 1. \end{cases} \tag{8.29}$$

The tables of $\{d_{tu}\}$ and $\{c_{tu}\}$ can be built with work proportional to T^2. The solution of (8.29) with backward induction is also proportional to T^2. Hence, the overall complexity of this approach is of order $\mathcal{O}(T^2)$.

8.3.2 Backlogging

When backlogging is allowed, demand may accumulate and be satisfied by productions during subsequent periods. The only effect to program (8.23) is to delete the requirement that the variables I_2 through I_T be nonnegative. When I_t is nonnegative, it still represents an inventory of I_t units; when I_t is negative, it now represents a *shortage* of $-I_t$ units of unfilled (backlogged) demand that must be satisfied by production during periods t through T. Hence, the set of states (i,t) and the set $A(i,t)$ of feasible productions are:

$$\{(i,t) \mid 1 \le t \le T;\ -\textstyle\sum_{s=1}^{t-1} D_s \le i \le \sum_{s=t}^{T} D_s\};\ A(i,t) = \{a \in \mathbb{N}_0 \mid 0 \le a \le \textstyle\sum_{s=t}^{T} D_s - i\}.$$

In the case of backlogging h_t is called the *holding/shortage* cost function for period t. When I_t is nonnegative, $h_t(I_t)$ remains equal to the cost of having I_t units of inventory on hand at the start of period t; when I_t is negative, $h_t(I_t)$ becomes the cost of having a shortage of $-I_t$ units of unfilled demand on hand at the start of period t. The functional equation of the production model in which backlogging is allowed becomes:

$$\begin{cases} f(i,T) &= h_T(i) + c_T(D_T - i),\ i = -\sum_{s=1}^{T-1} D_s, \ldots, D_T. \\ f(i,t) &= min_{a \in A(i,t)}\{h_t(i) + c_t(a) + f(i + a - D_t, t+1)\}, \\ & \qquad t = T-1, T-2, \ldots, 1;\ -\textstyle\sum_{s=1}^{t-1} D_s \le i \le \sum_{s=t}^{T} D_s \end{cases} \tag{8.30}$$

Also in this case we consider concave cost functions. The model has *concave shortage cost* if, for any t, the function h_t is convave on $\{0, -1, -2, \ldots\}$, i.e. $h_t(n+2) - h_t(n+1) \le h_t(n+1) - h_t(n)$ for $n = -2, -3, \ldots$. In other words, the model has concave holding and shortage costs if, for each t, the function h_t satisfies

$$h_t(n+2) - h_t(n+1) \le h_t(n+1) - h_t(n),\ n \in \mathbb{Z}\backslash\{-1\}. \tag{8.31}$$

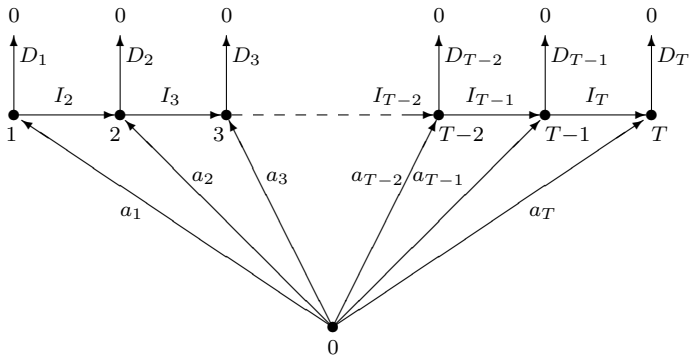
When the production and holding/shortage costs are concave, the solution to the production planning problem has the form given in the following theorem.

Theorem 8.7

If the production and holding/shortage functions, c_t and h_t respectively, are concave for all periods t, then there exists an optimal production plan $(a_1, a_2, \ldots, a_T)$ having this property: if $a_m > 0$ and $a_n > 0$ with $m < n$, then $I_t = 0$ for at least one $t \in \{m+1, m+2, \ldots, n\}$.

Before we prove Theorem 8.7 we model the problem as a network flow problem.

For $t = 1, 2, \ldots, T$ a node arises in the network; furthermore we add a node 0. There are arcs $(t, t+1)$, $t = 1, 2, \ldots, T-1$; $(0, t)$ and $(t, 0)$ for $t = 1, 2, \ldots, T$. From node 0, we send a variable flow a_t into node t, $t = 1, 2, \ldots, T$; from node t we send D_t to node 0 $(t = 1, 2, \ldots, T)$ and I_{t+1} to node $t+1$ $(t = 1, 2, \ldots, T-1)$. Flow into node t equals $I_t + a_t$, where $I_1 = 0$, and flow out of node t equals $D_t + I_{t+1}$, where $I_{T+1} = 0$.



Hence, flow conservation in the nodes t corresponds to the constraints of program (8.23) and flow conservation in node 0 corresponds to a total production of $a_1 + a_2 + \cdots + a_T = D_1 + D_2 + \cdots + D_T$.

We shall call a flow $(a_1, a_2, \ldots, a_T)$ *feasible* if $a_t \geq 0$ and integer for $t = 1, 2, \ldots, T$. Consider a feasible flow. An arc (i, j) is *active* if flow along is not zero. A *loop* is said to exist if one can start at a node and return to it by traversing a sequence of distinct active arcs, not necessarily in the direction of the arcs. Suppose, for instance, that a_2, I_3 and a_3 are all positive. Then, the node sequence $(0, 2, 3, 0)$ prescribes a loop; also the node sequence $(0, 3, 2, 0)$ prescribes a loop.

Proof of Theorem 8.7

Consider an optimal production plan $(a_1, a_2, \ldots, a_{T+1})$. If there are more optimal plans, take the plan for which $\sum_{t=1}^{T}(a_t + I_t)$ is minimal. Aiming for a contradiction, we assume that $a_m > 0$ and $a_n > 0$ with $m < n$, and $I_t \neq 0$ for $t = m+1, m+2, \ldots, n$. Hence, $(0, m, m+1, \ldots, n, 0)$ is a loop. A feasible flow results if we increase a_m by 1, decrease a_n by 1, and consequently, increase I_k by 1 for $m+1 \leq k \leq n$. Similarly, we may decrease a_m by 1, increase a_n by 1, and decrease I_k by 1 for $m+1 \leq k \leq n$. These perturbed flows cannot decrease cost below the costs of the optimal plan. So,

$$c_m(a_m) + c_n(a_n) + \textstyle\sum_{k=m+1}^{n} h_k(I_k) \leq c_m(a_m + 1) + c_n(a_n - 1) + \sum_{k=m+1}^{n} h_k(I_k + 1)$$

and

$$c_m(a_m) + c_n(a_n) + \textstyle\sum_{k=m+1}^{n} h_k(I_k) \leq c_m(a_m - 1) + c_n(a_n + 1) + \sum_{k=m+1}^{m} h_k(I_k - 1).$$

Add these two inequalities and rearrange the sum as

$$0 \leq \Delta^2 c_m(a_m - 1) + \Delta^2 c_n(a_n - 1) + \textstyle\sum_{k=m+1}^{n} \Delta^2 h_k(I_k - 1) \leq 0,$$

the last inequality by the concaveness of the production and holding/shortage functions ($\Delta^2 h_k(I_k - 1)$ is not necessarily nonnegative if $I_k = 0$, but notice that $I_k \neq 0$ for $k = m+1, m+2, \ldots, n$). Hence, the three plans are all optimal. However,

$$\textstyle\sum_{t=1}^{T}(a_t'' + I_t'') = \sum_{t=1}^{T}(a_t + I_t) - (t - s) < \sum_{t=1}^{T}(a_t + I_t).$$

This contradicts the supposed minimality of that sum, which completes the proof. □

Let $a^* = (a_1, a_2, \ldots, a_T)$ be an optimal production plan of the type described in Theorem 8.7. Consider any period $t \leq T$ such that $I_t = 0$ (since $I_1 = 0$ such a t exists). Consider also the lowest-numbered $u > t$ such that $I_u = 0$ (since $I_{T+1} = 0$ such a u exists). Exactly $d_{tu} = \sum_{k=t}^{u-1} D_k$ units must be produced during periods t through $u-1$. We argue by contradiction that production of these d_{tu} units is concentrated in one period k, where $t \leq k \leq u-1$. Suppose not: that is, that a^* splits this production between periods k and $l > k$. Since $a_k > 0$ and $a_l > 0$ with $k < l$, it follows from Theorem 8.7 that $I_p = 0$ for some p with $k+1 \leq p \leq l \leq u-1$. The minimality of u precludes this. Hence, the production of these d_{tu} units is not split.

For $t \leq k \leq u-1$ let $c_{tu}(k)$ denote the total cost incurred during periods t through $u-1$ if the total demand d_{tu} occurring during these periods is satisfied by production in period k, i.e.

$$c_{tu}(k) = h_t(0) + c_k(d_{tu}) + \sum_{s=t+1}^{k} h_s(-d_{ts}) + \sum_{s=k+1}^{u-1} h_s(d_{su}). \tag{8.32}$$

In a dynamic programming formulation state t again denotes the situation of having no inventory on hand at the start of period t. Transition occurs from state t to state u if it is decided to produce d_{tu} units during some intermediate period k. The cheapest transition from state t tot state u costs c_{tu}^*, where

$$c_{tu}^* = min_{\{k \mid t \leq k \leq u-1\}}\, c_{tu}(k). \tag{8.33}$$

As usual, $f(t)$ is defined, for $t = 1, 2, \ldots, T$, by

$f(t)$ = the minimum cost of satisfying demands during periods t through T if $I_t = 0$.

One gets the functional equation

$$\begin{cases} f(T+1) &= 0 \\ f(t) &= min_{\{u \,|\, t+1 \le u \le T+1\}} \{c^*_{tu} + f(u)\}, \ t = T, T-1, \ldots, 1. \end{cases} \tag{8.34}$$

This functonal equation is similar to the functional equation for the concave-cost case without backlogging. The difference is that c^*_{tu} replaces c_{tu}. Once a table is built, (8.34) can be solved with work proportional to T^2, just in the case of backlogging. However, the work needed to build a table of c^*_{tu} from (8.33) is proportional to T^3, not to T^2.

8.3.3 Inventory control and single-critical-number policies

This section concerns an inventory model over a finite horizon of T periods with uncertain demand and without backlogging. The symbols i and a are used consequently throughout this section, where i denotes the inventory on hand at the start of a period, just prior to deciding whether to place an order, and a denotes the inventory on hand at the start of a period, just after deciding whether to place an order (ordering is instantaneous). So $a \ge i$ and the number of units ordered is $a - i$. We assume that stock is indivisible, so a and i are integers.

Let the states be (i, t), depicting the situation of having i units of inventory on hand at the start of period t, just before deciding whether and how much to order. Let $f(i, t)$ be the minimum discounted cost over the remaining periods, given state (i, t). We are interested in $f(0, 1)$ and state $(i, T+1)$ represents the situation with inventory of i units at the end of period T. Hence, $f(i, T+1) = -si, \ i \ge 0$.

The data of the model are as follows:

$$\begin{aligned} D_t &= \text{the (uncertain) demand during period } t. \\ p_t(j) &= \text{the probability that the demand in period } t \text{ is } j, \ j = 0, 1, \ldots. \\ R &= \text{the (retail) unit sales price (independent of the period).} \\ k &= \text{the unit ordering cost (independent of the period).} \\ s &= \text{the unit salvage value at the end of period } T. \\ r &= \text{the interest rate per period.} \\ \alpha &= \text{the discount factor, which equals } \tfrac{1}{1+r}. \end{aligned}$$

The net cost incurred during period t equals:
- ordering cost: $(a - i)k$;
- interest charge on inventory: $(1 - \alpha)ak$;
- expected sales in period t: $\alpha R \, \mathbb{E}\left\{min\{D_t, a\}\right\}$, where $\mathbb{E}\left\{min\{D_t, a\}\right\} = \sum_{j=0}^{a-1} jp_t(j) + a \sum_{j \ge a} p_t(j)$.

Therefore, we obtain the following optimality equation

$$\begin{cases} f(i,T+1) &= -si,\ i \geq 0 \\ f(i,t) &= inf_{\{a\,|\,a\geq i\}}\left\{(a-i)k+(1-\alpha)ak-\alpha R\{\sum_{j=0}^{a-1} jp_t(j)+a\sum_{j\geq a}p_t(j)\}+\right. \\ & \qquad \left.\alpha\,\mathbb{E}\{f((a-D_t)^+,t+1)\}\right\} \\ &= inf_{\{a\,|\,a\geq i\}}\left\{(a-i)k+(1-\alpha)ak-\alpha R\{\sum_{j=0}^{a-1} jp_t(j)+a\sum_{j\geq a}p_t(j)\}+\right. \\ & \qquad \left.\alpha\{\sum_{j=0}^{a-1} p_t(j)f(a-j,t+1)+a\sum_{j\geq a}p_t(j)f(0,t+1)\}\right\},\ i\geq 0,\ t=T,T-1,\ldots,1. \end{cases} \tag{8.35}$$

Define

$$F(i,t)=f(i,t)+ik \text{ for all } i \text{ and } t. \tag{8.36}$$

Then, $f(a-j,t+1)=F(a-j,t+1)-(a-j)k,\ j<a$ and $f(0,t+1)=F(0,t+1)$.
Therefore, the optimality equation (8.35) can be written as

$$\begin{cases} F(i,T+1) &= (k-s)i,\ i\geq 0 \\ F(i,t) &= inf_{\{a\,|\,a\geq i\}}\left\{(2-\alpha)ak-\alpha R\{\sum_{j=0}^{a-1} jp_t(j)+a\sum_{j\geq a}p_t(j)\}+\right. \\ & \qquad \left.\alpha\{\sum_{j=0}^{a-1} p_t(j)F(a-j,t+1)+\sum_{j\geq a}p_t(j)F(0,t+1)\}-\alpha\sum_{j=0}^{a-1} p_t(j)(a-j)k\right\}, \\ & \qquad\qquad i\geq 0,\ t=T,T-1,\ldots,1. \end{cases} \tag{8.37}$$

Since

$$\begin{aligned} \alpha\sum_{j=0}^{a-1} p_t(j)(a-j)k &= \alpha ka\sum_{j=0}^{a-1} p_t(j)-\alpha k\sum_{j=0}^{a-1} jp_t(j) \\ &= \alpha ka\{1-\sum_{j\geq a}p_t(j)\}-\alpha k\sum_{j=0}^{a-1} jp_t(j) \\ &= \alpha ka-\alpha ka\sum_{j\geq a}p_t(j)\}-\alpha k\sum_{j=0}^{a-1} jp_t(j), \end{aligned}$$

we have the optimality equation

$$\begin{cases} F(i,T+1) &= (k-s)i,\ i\geq 0 \\ F(i,t) &= inf_{\{a\,|\,a\geq i\}}\left\{2(1-\alpha)ak-\alpha(R-k)\{\sum_{j=0}^{a-1} jp_t(j)+a\sum_{j\geq a}p_t(j)\}+\right. \\ & \qquad \left.\alpha\{\sum_{j=0}^{a-1} p_t(j)F(a-j,t+1)+\sum_{j\geq a}p_t(j)F(0,t+1)\}\right\},\ i\geq 0,\ t=T,T-1,\ldots,1. \end{cases} \tag{8.38}$$

Notice that the expression in (8.38) for which the infimum is taken over all $a\geq i$ is independent of i. Let

$$\begin{cases} L_t(a) &= 2(1-\alpha)ak-\alpha(R-k)\{\sum_{j=0}^{a-1} jp_t(j)+a\sum_{j\geq a}p_t(j)\}+ \\ & \qquad \alpha\{\sum_{j=0}^{a-1} p_t(j)F(a-j,t+1)+\sum_{j\geq a}p_t(j)F(0,t+1)\}. \end{cases} \tag{8.39}$$

Theorem 8.8

Let $s<k<R,\ k>0$ and $\mathbb{E}\{D_t\}<\infty$ for all t.

(1) The function $L_t(a)$ is convex for all t.

(2) For all t, there exists a nonnegative integer S_t such that

$$L_t(S_t)=min_{\{a\geq 0\}}L_t(a) \text{ and } F(i,t)=\begin{cases} L_t(S_t) & \text{for } i\leq S_t; \\ L_t(i) & \text{for } i>S_t. \end{cases}$$

Proof

Part (1)

We first prove that $2(1-\alpha)ak - \alpha(R-k)\{\sum_{j=0}^{a-1} jp_t(j) + a\sum_{j\geq a} p_t(j)\}$ is convex.

Therefore, it is sufficient to show that $G_t(a) = \sum_{j=0}^{a-1} jp_t(j) + a\sum_{j\geq a} p_t(j)$ is concave.

$$\begin{aligned}\Delta G_t(a) &= \sum_{j=0}^{a} jp_t(j) + (a+1)\sum_{j\geq a+1} p_t(j) - \sum_{j=0}^{a-1} jp_t(j) + a\sum_{j\geq a} p_t(j)\\ &= \sum_{j\geq a+1} p_t(j).\end{aligned}$$

Hence, $\Delta^2 G_t(a) = \Delta G_t(a+1) - \Delta G_t(a) = p_t(a+1) \geq 0$: G_t is concave for all t.

Next, we show that L_T is convex and that $L_T(a) \to \infty$ if $a \to \infty$. Since

$$L_T(a) = G_T(a) + \alpha(k-s)\sum_{j=0}^{a-1}(a-j)p_T(j),$$

it is sufficient to show that $g_T(a) = \sum_{j=0}^{a-1}(a-j)p_T(j)$ is convex.

$$\Delta g_T(a) = \sum_{j=0}^{a}(a+1-j)p_T(j) - \sum_{j=0}^{a-1}(a-j)p_T(j) = \sum_{j=0}^{a} p_T(j),$$

and consequently, $\Delta^2 g_t(a) = \Delta g_t(a+1) - \Delta g_t(a) = p_T(a+1) \geq 0$: g_T is concave.

Furthermore, we have

$$L_T(a) = 2(1-\alpha)ak - \alpha(R-k)\big\{\sum_{j=0}^{a-1} jp_T(j) + a\sum_{j\geq a} p_T(j)\big\} + \alpha(k-s)\sum_{j=0}^{a-1} p_t(j)(a-j).$$

Since

$\sum_{j=0}^{a-1} jp_T(j) + a\sum_{j\geq a} p_T(j) \leq \mathbb{E}\{D_T\}$ and $\sum_{j=0}^{a-1} p_t(j)(a-j) \geq \sum_{j\geq 0} p_t(j)(a-j) = a - \mathbb{E}\{D_T\}$,

we obtain

$L_T(a) \geq a\{2(1-\alpha)k + \alpha(k-s)\} - \alpha(R-s)\,\mathbb{E}\{D_T\}$, implying that $L_T(a) \to \infty$ if $a \to \infty$.

Finally, we show inductively for $t = T, T-1, \ldots, 1$ that L_t is convex and that $L_t(a) \to \infty$ if $a \to \infty$. Assume that L_{t+1} is convex and that $L_{t+1}(a) \to \infty$ if $a \to \infty$.

Then, L_{t+1} attains its minimum at some integer S_{t+1} that satisfies $L_{t+1}(S_{t+1}) = min_{a\geq 0} L_{t+1}(a)$

and $F(i,t+1) = \begin{cases} L_{t+1}(S_{t+1}) & \text{for } i \leq S_{t+1};\\ L_{t+1}(i) & \text{for } i > S_{t+1}.\end{cases}$

For the convexity of $L_t(a)$ it is sufficient to show the convexity of

$$H_t(a) = \sum_{j=0}^{a-1} p_t(j)F(a-j,t+1) + \sum_{j\geq a} p_t(j)F(0,t+1).$$

Therefore, we obtain

$$\begin{aligned}\Delta H_t(a) &= \sum_{j=0}^{a} p_t(j)F(a+1-j,t+1) + \sum_{j\geq a+1} p_t(j)F(0,t+1)\\ &\qquad - \sum_{j=0}^{a-1} p_t(j)F(a-j,t+1) - \sum_{j\geq a} p_t(j)F(0,t+1)\\ &= \sum_{j=0}^{a} p_t(j)\{F(a+1-j,t+1) - F(a-j,t+1)\},\end{aligned}$$

implying

$$\begin{aligned}\Delta^2 H_t(a) &= \sum_{j=0}^{a+1} p_t(j)\{F(a+2-j,t+1) - F(a+1-j,t+1)\} -\\ &\qquad \sum_{j=0}^{a} p_t(j)\{F(a+1-j,t+1) - F(a-j,t+1)\}\\ &= p_t(a+1)\{F(1,t+1) - F(0,t+1)\} + \sum_{j=0}^{a} p_t(j)\big\{\{F(a+2-j,t+1) -\\ &\qquad F(a+1-j,t+1)\} - \{F(a+1-j,t+1) - F(a-j,t+1)\}\big\}\\ &= p_t(a+1)\{F(1,t+1) - F(0,t+1)\} + \sum_{j=0}^{a} p_t(j)\Delta^2 F(a-j,t+1).\end{aligned}$$

Notice that

$$\Delta^2 F(a-j,t+1) = \begin{cases} \Delta^2 L_{t+1}(a-j) \geq 0 & a-j \geq S_{t+1} \\ \Delta L_{t+1}(a+1-j) \geq 0 & a+1-j = S_{t+1} \\ 0 & a+2-j \leq S_{t+1} \end{cases}$$

Since $F(1,t+1) = inf_{\{a\geq 1\}} L_{t+1}(a) \geq inf_{\{a\geq 0\}} L_{t+1}(a) = F(0,t+1)$, we obtain $\Delta^2 H_t(a) \geq 0$ for all $a \geq 0$ and consequently L_t is a convex function.

To complete the proof, we must show that $L_t(a) \to \infty$ as $a \to \infty$.

Pick M large enough that $\mathbb{P}\{D_t \leq M\} \geq 0.5$. Then consider $a \geq M + S_{t+1}$. In the case that $D_t \leq M$, we have $(a-D_t)^+ = a - D_t \geq a - M \geq S_{t+1}$, and the fact that L_{t+1} is nondecreasing for $i \geq S_{t+1}$, assures that $F((a-D_t)^+, t+1) \geq L_{t+1}(a-M)$. In the case $D_t > M$, the fact that $F(i,t+1)$ is nondecreasing assures that $F((a-D_t)^+, t+1) \geq F(0,t+1)$. Then,

$$\mathbb{E}\{F((a-D_t)^+, t+1)\} \geq 0.5 L_{t+1}(a-M) + 0.5F(0,t+1) \text{ for } a \geq M + S_{t+1}.$$

Therefore, we obtain

$$L_t(a) = 2(1-\alpha)ak - \alpha(R-k)\,\mathbb{E}\{D_t\} + 0.5L_{t+1}(a-M) + 0.5F(0,t+1),$$

implying that $L_t(a) \to \infty$ as $a \to \infty$ for all $\alpha \in [0,1]$, even when α equals 0 or 1. □

An optimal policy has the following structure: if the inventory at the start of period t is at least S_t, no order is placed; if the inventory is $i < S_t$ then exactly $S_t - i$ units are ordered. This one-parameter ordering rule is called a *single-critical-number policy.*

8.3.4 Inventory control and (s,S)-policies

This section concerns a model of inventory control with also uncertain demand, but with a fixed charge for placing an order. Ordering is instantaneous and has to be paid at delivery. This model covers both backlogging and no backlogging. The ordering cost includes a cost per unit ordered and a fixed charge of *setup cost* for placing any order. Let $K_t \geq 0$ be the setup cost and h_t the unit ordering cost in period t. The cost-minimizing ordering rule for this type of model is often characterized by two numbers per period, as follows. Each period has an *order-up-to-quantity* S_t and a *reorder point* $s_t \leq S_t$. If the inventory I_t at the start of period t is at least s_t, no order is placed; if $I_t < i_t$ then exactly $S_t - I_t$ units are ordered. This two-parameter ordering rule is called an (s,S)-policy. An (s,S)-policy has the propery that any order in period t must be for more than $S_t - s_t$ units, large enough that the benefit of the added inventory offsets the setup cost. In addition to the notation mentioned previously, we use the following notation:

R_t = the (retail) unit sales price during period t (customers pay at the end of the period in which they place their orders, even if their orders are backlogged).

$e(i^+)$ = the salvage value of having $i^+ = max\{i,0\}$ units of inventory at the end of period T.

$h_t(a)$ = the expectation of the inventory cost during period t, given that the inventory is a at the start of period t, just after deciding whether to place an order.

$I_t(a, D_t)$ = the inventory on hand at the beginning of period $t+1$, given as a function of y and D_t.

To describe the ordering cost we employ the function $H(z) = \begin{cases} 0 & \text{for } z \le 0; \\ 1 & \text{for } z > 0. \end{cases}$

Then, the cost of ordering z units at the start of period t equals $K_t H(z) + z \cdot k_t$. Let $c_t(i, a)$ denote the present value at the start of period t of the expected cost incurred during this period, given in terms of period's before-ordering inventory i and after-ordering inventory a:

$$c_t(i, a) = K_t H(a - i) + (a - i)k_t + h_t(a) - \alpha\big\{a - \mathbb{E}\{I_t(a, D_t)\}\big\}R_t. \tag{8.40}$$

We have $\mathbb{E}\{I_t(a, D_t)\} = \begin{cases} a - \sum_{j=0}^{\infty} p_t(j)j & \text{if backlogging is allowed;} \\ a - \sum_{j=0}^{a} p_t(j)j - a\sum_{j=a+1}^{\infty} p_t(j) & \text{if backlogging is not allowed.} \end{cases}$

The states are (i, t) denoting the situation of having i units of inventory on hand at the start of period t, just before deciding whether and how much to order. Interpret $f(i, t)$ as the present value at the start of period t of the cost incurred from then to the end of the planning horizon if state (i, t) is observed and if an optimal policy is followed. This leads to the following optimality equation

$$\begin{cases} f(i, T+1) &= K_{T+1}H(-i) + k_{T+1}(-i)^+ - e(i^+) \\ f(i, t) &= inf_{\{a \,|\, a \ge i\}} \big\{c_t(i, a) + \alpha\, \mathbb{E}\{f(I_t(a, D_t), t+1)\}\big\},\ t = T, T-1, \ldots, 1. \end{cases} \tag{8.41}$$

where the first equation accounts for the cost of disposing of excess demand $(-i)^+$ by a special end-of-planning-horizon order and for the salvage value $e(i^+)$, converted to a cost. The term $-ik_t$ in (8.40) is independent of the decision a and can be factored out of (8.41). This motivates a change of variables. Let

$$F(i, t) = f(i, t) + ik_t. \tag{8.42}$$

Then, with

$$G_t(a) = h_t(a) + (k_t - \alpha R_t)a + \alpha(R_t - k_{t+1})\, \mathbb{E}\{I_t(a, D_t)\}, \tag{8.43}$$

we obtain (the verification is left to the reader)

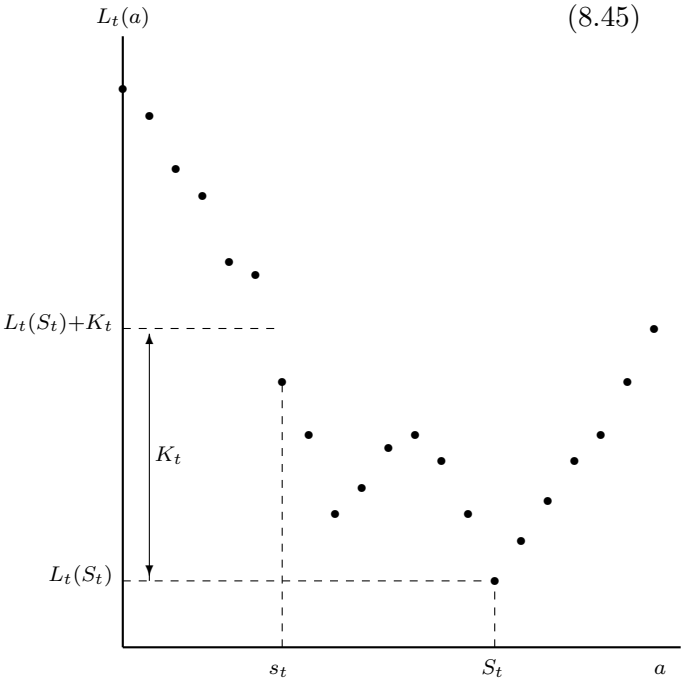
$$\begin{cases} F(i, T+1) &= K_{T+1}H(-i) + k_{T+1}i^+ - e(i^+) \\ F(i, t) &= inf_{\{a \,|\, a \ge i\}} \big\{K_t H(a - i) + G_t(a) + \alpha\, \mathbb{E}\{F(I_t(a, D_t), t+1)\}\big\},\ t = T, T-1, \ldots, 1. \end{cases} \tag{8.44}$$

The quantity $G_t(a)$ is called the *operating cost*; it accounts for all units of cash flow during perod t except the setup cost. The first term in $G_t(a)$ is the inventory carrying cost. To interpret the remaining terms, we imagine that the starting inventory y is purchased from the supplier at unit cost k_t, that ending inventory $I_t(a, D_t)$ is returned to the supplier at unit price k_{t+1}, and that sales of $a - I_t(a, D_t)$ units occur at unit price R_t.

Next, we write the optimality equation as

$$F(i,t) = inf_{\{a \mid a \geq i\}} \{K_t H(a-i) + L_t(a)\}, \text{ where } L_t(a) = G_t(a) + \alpha\, \mathbb{E}\{F(I_t(a, D_t), t+1)\}. \tag{8.45}$$

Up to this point, the development has followed the same pattern as in the previous section, where we showed the convexity of L_t. In the case of fixed setup cost the function L_t is not convec, in general. It may have the structure shown in the picture, which has two local minima. The point identified in the figure as S_t is the point where the function L_t attains its global minimum. The point identified as s_t is the smallest value of a for which $L_t(a) \leq L_t(S_t)+K_t$. Since K_t is nonnegative, one has $s_t \leq S_t$. So S_t and s_t satisfy



$$\begin{aligned} L_t(S_t) &= inf_a \{L_t(a)\} \\ s_t &= inf\{a \mid L_T(a) \leq L_t(S_t) + K_t\} \end{aligned} \tag{8.46}$$

It is now argued that an (s,S)-policy is optimal for the strange shaped function of the figure. Consider state (i,t) with $i \geq s_t$. Note that it is optimal to set $a = i$, because in that case $K_t + L_t(a) \geq L_t(i)$ for all $a \geq i$, i.e. the setup cost cannot be recouped by an $a \geq i$. Now consider a state (i,t) with $i < s_t$. Then, $K_t + L_t(S_t) \leq L_t(i)$ and $K_t + L_t(S_t) \leq K_t + L_t(a)$ for any a. The next theorem gives conditions under which an (s,S)-policy is optimal.

Theorem 8.9

Suppose that the ollowing five conditions hold.

(1) For $t = 1, 2, \ldots, T$, the function $G_t(\cdot)$, defined in (8.43), is convex.

(2) For $t = 1, 2, \ldots, T$, the setup costs K_t satisfies $K_t \geq \alpha K_{t+1}$.

(3) For $t = 1, 2, \ldots, T$ and for each value of D_t, the function $I_t(\cdot, D_t)$ is convex and nondecreasing.

(4) The function $k_{T+1} i^+ - e(i^+)$ is convex and nondecreasing in i.

(5) All expectations are finite and $inf_a\, L_t(a)$ is attained for all $t = 1, 2, \ldots, T$.

Then, for $t = 1, 2, \ldots, T$, there exists S_t and s_t that satisfy $L_t(S_t) = inf_a \{L_t(a)\}$ and $s_t = inf\{a \mid L_t(a) \leq L_t(S_t) + K_t\}$. Moreover, $F(i,t) = \begin{cases} L_t(S_t) + K_t & \text{if } i < s_t; \\ L_t(i) & \text{if } i \geq s_t. \end{cases}$

Before we prove the theorem we make some preparations.
With $K \geq 0$ a function $f : \mathbb{Z} \to \mathbb{R}$ is called *K-convex* if any triple $a < b < c$ satisfies

$$f(c) + K \geq f(b) + (c - b)\Big(\frac{f(b) - f(a)}{b - a}\Big).$$

Hence, the straight line passing through the points $(a, f(a))$ and $(b, f(b))$ has in c a value of at most $f(c) + K$. Since b is between a and c we can write $b = \alpha a + (1 - \alpha)c$ for some $\alpha \in (0, 1)$, and getting the following equivalent form of K-convexity.

$$\begin{aligned} f(c) + K &\geq f\{\alpha a + (1-\alpha)c\} + \alpha(c-a) \cdot \tfrac{f\{\alpha a+(1-\alpha)c\}-f(a)}{(1-\alpha)(c-a)} \;\Leftrightarrow \\ (1-\alpha)\{f(c)+K\} &\geq (1-\alpha)f\{\alpha a + (1-\alpha)c\} + \alpha \cdot \big\{f\{\alpha a + (1-\alpha)c\} - f(a)\big\} \;\Leftrightarrow \\ (1-\alpha)\{f(c)+K\} &\geq f\{\alpha a + (1-\alpha)c\} - \alpha f(a). \end{aligned}$$

Notice that this inequality is also valid if $a = c$ and/or $\alpha \in \{0, 1\}$. Hence, K-convexity is equivalent to $\alpha f(a) + (1 - \alpha)\{f(c) + K\} \geq f\{\alpha a + (1 - \alpha)c\}$ for all $a \leq c$ and all $\alpha \in [0, 1]$.
A function $f : \mathbb{Z} \to \mathbb{R}$ is *K-quasi-convex* if any triple $a < b < c$ with $f(a) < f(b)$ satisfies $f(c) + K \geq f(b)$. Notice that an increase in a K-quasi-convex function cannot be followed by a decrease that exceeds K.

Lemma 8.3
A function $f : \mathbb{Z} \to \mathbb{R}$ that is K-convex is also K-quasi-convex.

Proof
Suppose that $a < b < c$ and $f(a) < f(b)$. Since f is K-convex, we obtain

$$f(c) + K \geq f(b) + (c - b)\Big(\tfrac{f(b)-f(a)}{b-a}\Big) > f(b). \qquad \square$$

Lemma 8.4
Let $f : \mathbb{Z} \to \mathbb{Z}$ be convex and nondecreasing and let $g : \mathbb{Z} \to \mathbb{R}$ be K-convex. Furthermore, we have $g(a) \leq g(c) + K$ for all $a < c$. Then, $g\{f(x)\}$ is K-convex.

Proof
Take any $a < c$ and $\alpha \in (0, 1)$, and let $b = \alpha a + (1 - \alpha)c$. Then, we have to show that

$$\alpha g\{f(a)\} + (1 - \alpha)\big\{g\{f(c)\} + K\big\} \geq g\{f(b)\}.$$

Since f is convex and nondecreasing, we have

$$f(a) \leq f(b) \leq \alpha f(a) + (1 - \alpha)f(c) \leq f(c).$$

Hence, there exists a number $\beta \in [0, 1]$ such that $f(b) = \beta f(a) + (1 - \beta)f(c)$, implying
$\beta f(a) + (1 - \beta)f(c) = f(b) \leq \alpha f(a) + (1 - \alpha)f(c)$, or equivalently, $(\beta - \alpha)\{f(a) - f(b)\} \leq 0$.
Since f is nondecreasing, we obtain $\beta \geq \alpha$. Because g is K-convex on Y, $f(a) \leq f(b) \leq f(c)$ yields $\beta g\{f(a)\} + (1 - \beta)\big\{g\{f(c)\} + K\big\} \geq g\{f(b)\}$.

Therefore,

$$\alpha g\{f(a)\} + (1-\alpha)\big\{g\{f(c)\} + K\big\} - g\{f(b)\} \geq$$
$$\alpha g\{f(a)\} + (1-\alpha)\big\{g\{f(c)\} + K\big\} - \beta g\{f(a)\} - (1-\beta)\big\{g\{f(c)\} + K\big\} =$$
$$(\alpha-\beta)\Big\{g\{f(a)\} - g\{f(c)\} - K\Big\} \geq 0,$$

because $\beta \geq \alpha$ and $g\{f(a)\} \leq g\{f(c)\} + K$, the last inequality since the property of g given in the formulation of the lemma. □

Lemma 8.5

Let the function $L : \mathbb{Z} \to \mathbb{R}$ be K-quasi-convex. Suppose that S such that $L(S) = inf_a\{L(a)\}$ exists. Furthermore, let $F(i) = inf_{\{a \geq i\}}\{KH(a-i) + L(a)\}$.

Then, $F(i) = \begin{cases} L(S)+K & \text{if } i < s \\ L(i) & \text{if } i \geq s \end{cases}$, *where* $s = inf\{a \mid L(a) \leq L(S) + K\}$.

Proof

Since $L(S) \leq L(S) + K$, the definition of s assures that $s \leq S$, possibly $s = -\infty$. We will show the result in 4 cases, depending on the position of i with respect to s and S.

Case 1: $i < s$.

The definition of s assures that $L(i) > L(S) + K$. Since $i < S$, we have $F(i) = L(S) + K$.

Case 2: $i = s > -\infty$.

By the definition of s we have $L(i) \leq L(S) + K$, and consequently, $F(i) = L(i)$ in this case.

Case 3: $s < i \leq S$.

Suppose that $L(i) > L(S) + K$. Then, $i \neq s$. So, $s < i < S$ and $L(s) \leq L(S) + K < L(i)$. Hence, by the K-quasi-convexity of L, $L(S) + K \geq L(i)$, implying a contradiction. Hence, $L(i) \leq L(S) + K$, and we obtain $F(i) = L(i)$ in this case.

Case 4: $i > S$.

Suppose that $F(i) < L(i)$. Then, the definition of F assures that there exists a c that satisfies $c > i$ with $L(c) + K < L(i)$. Then, $S < i < c$ and $L(S) < L(i)$. Hence, by the K-quasi-convexity of L, $L(c) + K \geq L(i)$, implying a contradiction. So, $F(i) = L(i)$ in this case. □

The next lemma provides conditions under which $F : \mathbb{Z} \to \mathbb{R}$ is a K-convex function that satiesfies $F(b) \leq F(c) + K$ if $b < c$.

Lemma 8.6

Let $L : \mathbb{Z} \to \mathbb{R}$ be a K-convex function. Then, F is K-convex, and for all elements $b < c$ we have $F(b) \leq F(c) + K$.

Proof

Take any $b < c$. Since $F(i) = inf_{\{a \mid a \geq i\}}\{KH(a-i) + L(a)\}$, we obtain $F(b) \leq K + L(c)$.

Lemma 8.3 shows that L is K-quasi-convex. Hence, Lemma 8.5 applies, and F satisfies

$$F(i) = \begin{cases} L(S)+K & \text{if } i < s \\ L(i) & \text{if } i \geq s \end{cases}, \text{ where } s = inf\{a \mid L(a) \leq L(S) + K\}.$$

Consider elements $a < b < c$. We have to show $F(b) \le F(c) + K + \frac{c-b}{b-a} \cdot \{F(a) - F(b)\}$.
If $F(a) \ge F(b)$ this follows immediately from $F(b) \le F(c) + K$.
If $a \le s$, Lemma 8.5 assures that $F(a) = L(a)$, $F(b) = L(b)$ and $F(c) = L(c)$, and the result is immediate from the K-convexity of L. In the remaining case, $a < s$ and $F(a) < F(b)$, Lemma 8.5 shows $F(a) = L(S) + K \ge F(s)$. Hence, $F(b) = L(b)$ and $s < b$, implying

$$\frac{c-b}{b-a} \cdot \{F(a) - F(b)\} > \frac{c-b}{b-s} \cdot \{F(a) - F(b)\} \ge \frac{c-b}{b-s} \cdot \{F(s) - F(b)\} \ge F(b) - F(c) - K,$$

the last inequality follows from the case $a = s$. □

The preceding Lemmas are now molded into a proof of Theorem 8.9.

Proof of Theorem 8.9
Condition (4) shows that $F(\cdot, T+1)$ is a K_{T+1}-convex function that satisfies, for all elements $b < c$, $F(b, T+1) \le F(c, T+1) + K_{T+1}$. This initializes the following inductive hypothesis:

> $F(\cdot, t+1)$ is a K_{t+1}-convex function that satisfies $F(b, t+1) \le F(c, t+1) + K_{t+1}$ for all elements $b < c$.

This hypothesis and condition (3) let us supply Lemma 8.4 with $g(\cdot) = F(\cdot, t+1)$ and with $f(\cdot) = I_t(\cdot, D_t)$. Lemma 8.4 shows that $F\{I_t(\cdot, D_t), t+1\}$ is K_{t+1}-convex. Since condition (2) gives $K_t \ge \alpha K_{t+1}$, this shows that $\alpha F\{I_t(\cdot, D_t), t+1\}$ is K_t-convex. Since K-convexity is preserved under convex combinations, (5) suffices for the K_t-convexity of $\alpha\, \mathbb{E}\left\{F\{I_t(\cdot, D_t), t+1\}\right\}$. So, condition (1) shows that $L_t(\cdot)$ is K_t-convex. Lemma 8.3 shows that $L_t(\cdot)$ is K_t-quasi convex, and condition (5) implies that $L_t(S_t) = inf_a \{L_t(a)\}$. Hence, Lemma 8.5 shows that $F(\cdot, t)$ satisfies $F(i,t) = \begin{cases} L_t(S_t) + K_t & \text{if } i < s_t; \\ L_t(i) & \text{if } i \ge s_t. \end{cases}$ Lemma 8.6 shows that $F(\cdot, t)$ is a K_t-convex function that satisfies $F(b,t) \le F(c,t) + K_t$ for all elements $b < c$. This completes the proof. □

8.4 Optimal control of queues

A queueing system includes *servers, customers* and *queues* for the customers awaiting service. The queues are also called *buffers*. We will discuss several types of queueing models.

8.4.1 The single-server queue

Customers enter the queue, wait their turn, are served by the single server, and depart the system. We might place a *controller* at the entrance to the queue to decide which customers to admit to the queues (*admission control*). Or we could impose a control on the server that could adjust the rate at which customers are served (*service rate control*). Both methods of control can be imposed simultaneously.

1. Admission control for batch arrivals

The state of the system is the number of customers in the buffer at the beginning of a time slot, and thus $S = \{0, 1, \dots\}$. At the beginning of each slot a batch of customers arrives and p_j is the probability that j customers arrive ($j = 0, 1, \dots$). In every state there are two actions available: 1 = accept the incoming batch or 0 = reject the incoming batch.

Case a: The action must be chosen before the size of the batch is observed.

There is a nonnegative holding cost $h(i)$ incurred when there are i customers in the buffer (assume $h(0) = 0$). There is a positive rejection cost r incurred whenever a batch is rejected. Hence, the immediate cost is:

$$c(i,a) = \begin{cases} h(i) + r & \text{if } a = 0,\ i \in S; \\ h(i) & \text{if } a = 1,\ i \in S. \end{cases}$$

Service occurs according to a geometric distribution with fixed rate μ, where $0 < \mu < 1$. This means that the probability of a successful service in any slot is μ. If the service is unsuccessful, then another try is made with the same probability of success, and this continues until the customer has been successful served. If a batch arrives to an empty buffer and is accepted, then its customers are available for service at the beginning of the following slot.

Hence, the transition probabilities are:

$$\begin{array}{lllll} i = 0: & p_{00}(0) = 1; & i \geq 1: & p_{i,i-1}(0) = \mu; & p_{i,i-1}(1) = \mu p_0; \\ & p_{0j}(1) = p_j,\ j \geq 0; & & p_{i,i}(0) = 1 - \mu; & p_{i,i+j}(1) = \mu p_{j+1} + (1-\mu)p_j,\ j \geq 0. \end{array}$$

Consider value iteration with $\alpha = 1$, i.e.

$$\begin{aligned} v_0^n &= min\{r + \textstyle\sum_j p_{0j}(0) v_j^{n-1},\ \sum_j p_{0j}(1) v_j^{n-1}\} \\ &= min\{r + v_0^{n-1},\ \textstyle\sum_j p_j v_j^{n-1}\}. \\ v_i^n &= min\{h(i) + r + \textstyle\sum_j p_{ij}(0) v_j^{n-1},\ h(i) + \sum_j p_{ij}(1) v_j^{n-1}\} \\ &= h(i) + min\{r + \mu v_{i-1}^{n-1} + (1-\mu) v_i^{n-1},\ \mu \textstyle\sum_j p_j v_{i-1+j}^{n-1} + (1-\mu) \sum_j p_j v_{i+j}^{n-1}\},\ i \geq 1. \end{aligned}$$

Lemma 8.7

Assume that $h(i)$ is nondecreasing in i and consider value iteration with $v_i^0 = 0,\ i \in S$. Then, v_i^n is nondecreasing in i for all $n \geq 0$.

Proof

The lemma is shown by induction on n. We first show that v_i^n is finite for all n and i by showing that $v_i^n \leq nr + (n+1)h(i),\ i \in S$. For $n = 0$ we have $0 = v_i^0 \leq 0 \cdot r + 1 \cdot (h(i) = h(i),\ i \in S$. Assume that $v_i^n \leq nr + (n+1)h(i),\ i \in S$. Then,

$$\begin{aligned} v_i^{n+1} &\leq h(i) + r + \mu v_{i-1}^n + (1-\mu) v_i^n \\ &\leq h(i) + r + \mu\{nr + (n+1)h(i-1)\} + (1-\mu)\{nr + (n+1)h(i)\} \\ &\leq h(i) + r + \mu\{nr + (n+1)h(i)\} + (1-\mu)\{nr + (n+1)h(i)\} \\ &= h(i) + r + nr + (n+1)h(i) = (n+1)r + (n+2)h(i),\ i \in S. \end{aligned}$$

$v^0 \equiv 0$ is nondecreasing in i. Assume that v_i^n is nondecreasing in i. Since

$$h(1) + r + \mu v_0^n + (1-\mu)v_1^n \geq r + \mu v_0^n + (1-\mu)v_0^n = r + v_0^n$$

and

$$h(1) + \mu \sum_j p_j v_j^n + (1-\mu) \sum_j p_j v_{j+1}^n \geq \mu \sum_j p_j v_j^n + (1-\mu) \sum_j p_j v_j^n = \sum_j p_j v_j^n,$$

each term in the minimum of v_0^{n+1} is bounded above by the corresponding term in v_1^{n+1}, and hence $v_0^{n+1} \leq v_1^{n+1}$. Suppose that the minimum in v_i^{n+1} $(i \geq 1)$ is obtained by the rejection action. By the induction hypothesis v_{i-1}^n and v_i^n are nondecreasing in i and consequently is $h(i) + r + \mu v_{i-1}^n + (1-\mu)v_i^n$ nondecreasing in i. Now suppose that the minimum in v_i^{n+1} $(i \geq 1)$ is obtained by the accepting action. For each fixed j, by the induction hypothesis, v_{i-1+j}^n and v_{i+j}^n are nondecreasing in i. Since $\sum_j p_j v_{i-1+j}^n$ and $\sum_j p_j v_{i+j}^n$ are convex combinations of v_{i-1+j}^n and v_{i+j}^n, $\sum_j p_j v_{i-1+j}^n$ and $\sum_j p_j v_{i+j}^n$ are nondecreasing in i. Hence, $h(i) + \mu \sum_j p_j v_{i-1+j}^n + (1-\mu) \sum_j p_j v_{i+j}^n$ is nondecreasing in i. Thus both terms in the minimum are nondecreasing in i, and consequently v_i^{n+1} is nondecreasing in i. □

Case b: The size of the incoming batch may be observed before the action is chosen.
In this case we take as states the pairs (i,k), where i denotes the number of customers in the buffer and k the size of the incoming batch: $S = \{(i,k) \mid i = 0,1,\dots;\ k = 0,1,\dots\}$. The holding cost is as in Case a and there is a positive rejection cost $r(k)$ incurred whenever a batch of size k is rejected $(r(0) = 0)$. Hence, the cost structure is as follows:

$$c\{(i,k),a) = \begin{cases} h(i) & ,\ i \in S,\ k = 0; \\ h(i) + r & ,\ i \in S,\ k \geq 1,\ a = 0; \\ h(i) & ,\ i \in S,\ k \geq 1,\ a = 1. \end{cases}$$

For the transition probabilities we obtain for all $k \geq 0$ and $j \geq 0$:

$$i = 0: \quad p_{(0,k)(0,j)}(0) = p_j; \quad i \geq 1: \quad p_{(i,k)(i-1,j)}(0) = \mu p_j; \qquad p_{(i,k)(i+k-1,j)}(1) = \mu p_j;$$
$$p_{(0,k)(k,j)}(1) = p_j; \qquad p_{(i,k)(i,j)}(0) = (1-\mu)p_j; \quad p_{(i,k)(i+k,j)}(1) = (1-\mu)p_j.$$

2. Admission control for an $M/M/1$ queue
Assume that customers arrive according to a Poisson process with parameter λ, and assume that the service time is exponentially distributed with parameter μ. We observe the system at each arrival and departure (*semi-Markov model*). As state space we use $S = \{0,1,2,\dots\} \times \{0,1\}$. The system is in state $(i,0)$ if there are i customers in the system and there is a departure; then, the only action $a = 0$ is to continue. The state $(i,1)$ occurs when there are i jobs in the system and a new customer arrives; in state $(i,1)$ the controller may admit $(a = 1)$ or refuse $(a = 0)$ service to the arrival.

In state $(0,0)$ the only action is to continue: with probability 1 the next state is state $(0,1)$ and the time until the next transition is exponentially distributed with rate λ. In state $(0,1)$ there are two actions: if $a = 0$ (refuse) the next state is with probability 1 again state $(0,1)$ and the time until the next transition is exponentially distributed with rate λ; if $a = 1$ (admission) the

time until the next transition is exponentially distributed with rate $\lambda + \mu$, and the next state is with probability $\frac{\lambda}{\lambda+\mu}$ state $(1,1)$ and with probability $\frac{\mu}{\lambda+\mu}$ state $(0,0)$.
In the states $(i,0)$, with $i \geq 1$, the only action $a = 0$ is to continue. Then, the next state is with probability $\frac{\lambda}{\lambda+\mu}$ state $(i,1)$ and with probability $\frac{\mu}{\lambda+\mu}$ state $(i-1,0)$; the time until the next transition is exponentially distributed with rate $\lambda + \mu$.
In the states $(i,1)$, with $i \geq 1$, there are two actions. If $a = 0$ (refuse) the next state is with probability $\frac{\lambda}{\lambda+\mu}$ again state $(i,1)$ and with probability $\frac{\mu}{\lambda+\mu}$ state $(i-1,0)$. If $a = 1$ (admission) the next state is with probability $\frac{\lambda}{\lambda+\mu}$ again state $(i+1,1)$ and with probability $\frac{\mu}{\lambda+\mu}$ state $(i,0)$. The time until the next transition is exponentially distributed with rate $\lambda + \mu$.

Let $\nu_{(i,b)}(a)$ be the parameter of the exponential distribution of the time until the next observation and let $p_{(i,b)(j,c)}(a)$ be the probability that the next state is state (j,c), given the current state (i,b) and the action a. Then, we have

$$\nu_{(i,b)}(a) = \begin{cases} \lambda & \text{if } i = 0,\ b = 0 \text{ or } 1,\ a = 0; \\ \lambda + \mu & \text{if } i = 0,\ b = 1,\ a = 1 \text{ or } i \geq 1. \end{cases}$$

$$p_{(i,b)(j,c)}(a) \begin{cases} 1 & \text{if } (i,b) = (0,0),\ a = 0,\ (j,c) = (0,1) \text{ or } (s,b) = (0,1),\ a = 0,\ (j,c) = (0,1); \\ \frac{\lambda}{\lambda+\mu} & \text{if } i \geq 1,\ b = 1,\ a = 0,\ (j,b) = (i,1) \text{ or } i \geq 0,\ b = 1,\ a = 1,\ (j,b) = (i+1,1); \\ \frac{\mu}{\lambda+\mu} & \text{if } i \geq 0,\ b = 1,\ a = 1,\ (j,b) = (i,0) \text{ or } i \geq 1,\ b = 1,\ a = 0,\ (j,b) = (i-1,0) \\ & \text{or } i \geq 1,\ b = 0,\ a = 0,\ (j,b) = (i-1,0); \\ 0 & \text{otherwise} \end{cases}$$

With the exception of the states $(0,0)$ and $(0,1)$ all transitions occur at rate $\lambda+\mu$. To *uniformize* the system we alter the transition structure in only these states:

$$p'_{(0,0)(0,0)}(0) = \tfrac{\mu}{\lambda+\mu};\ p'_{(0,0)(0,1)}(0) = \tfrac{\lambda}{\lambda+\mu};$$
$$p'_{(0,1)(0,0)}(0) = \tfrac{\mu}{\lambda+\mu};\ p'_{(0,1)(0,1)}(0) = \tfrac{\lambda}{\lambda+\mu};$$
$$p'_{(i,b)(j,c)}(a) = p_{(i,b)(j,c)}(a) \text{ if } i = 0,\ b = 1,\ a = 1 \text{ or } i \geq 1.$$

In the uniformized system, we observe the system more often when it is empty than in the untransformed system, so that this transformation increases the probability that it occupies $(0,0)$ and $(0,1)$ for $a = 0$. We may also interpret this transformation as addding "fictious" service completions at these states.

Furthermore, we assume that each arriving customer contributes r units of revenue and the system incurs a holding cost at rate $h(i)$ per unit time whenever there are i jobs in the systen, where $h(0) = 0$.

As utility function we consider the *discounted model*, in which we assume continuous-time discounting at rate $\alpha > 0$. This means that the present value of one unit received t units in the future equals $e^{-\alpha t}$. For $(i,b) \in S$ and $a = 0$ or 1, let $r'_{(i,b)}(a)$ denote the expected total discounted reward between two decision epochs in the uniformized system, given that the system occupies

state (i,b) and the decision maker chooses action a. The expected discounted holding cost during one epoch, given that the system occupies state (i,b) and the decision maker chooses action a, is per unit:

$$\mathbb{E}^a_{(i,b)}\{\int_0^\tau e^{-\alpha t}dt\} = \tfrac{1}{\alpha}\,\mathbb{E}^a_{(i,b)}\{1-e^{-\alpha\tau}\} = \tfrac{1}{\alpha}\int_0^\infty \{1-e^{-\alpha t}\}\,t\,f(t)\,dt,$$

where $f(t)$ is the density of the exponential distribution with parameter $\lambda+\mu$, i.e. $f(t) = (\lambda+\mu)e^{-(\lambda+\mu)t},\ t \geq 0$. Hence,

$$\begin{aligned}
\mathbb{E}^a_{(i,b)}\{\int_0^\tau e^{-\alpha t}dt\} &= \tfrac{\lambda+\mu}{\alpha}\int_0^\infty \{1-e^{-\alpha t}\}e^{-(\lambda+\mu)t}\,t\,dt \\
&= \tfrac{\lambda+\mu}{\alpha}\Big\{\int_0^\infty e^{-(\lambda+\mu)t}\,t\,dt - \int_0^\infty e^{-(\alpha+\lambda+\mu)t}\,t\,dt\Big\} \\
&= \tfrac{\lambda+\mu}{\alpha}\Big\{\tfrac{1}{\lambda+\mu} - \tfrac{1}{\alpha+\lambda+\mu}\Big\} \\
&= \tfrac{\lambda+\mu}{\alpha}\Big\{\tfrac{\alpha}{(\lambda+\mu)(\alpha+\lambda+\mu)}\Big\} = \tfrac{1}{\alpha+\lambda+\mu}.
\end{aligned}$$

Now it follows that the rewards in the uniformized system satisfy

$$r'_{(0,0)}(0) = 0;\ r'_{(i,0)}(0) = \tfrac{-h(i)}{\alpha+\lambda+\mu},\ i \geq 1;\ r'_{(i,1)}(1) = r + \tfrac{-h(i)}{\alpha+\lambda+\mu},\ i \geq 0.$$

$$r'_{(0,1)}(0) = 0;\ r'_{(i,1)}(0) = \tfrac{-h(i)}{\alpha+\lambda+\mu},\ i \geq 1;$$

The optimaliy equation for this model becomes (cf. Kallenberg [108], Chapter 7):

$$\begin{aligned}
v_{(i,b)} &= max_a\big\{r'_{(i,b)}(a) + \{\int_0^\infty e^{-\alpha t}\,f(t)\,dt\}\textstyle\sum_{(j,c)} p'_{(i,b)(j,c)}(a)v_{(j,c)}\big\} \\
&= max_a\big\{r'_{(i,b)}(a) + \tfrac{\lambda+\mu}{\alpha+\lambda+\mu}\textstyle\sum_{(j,c)} p'_{(i,b)(j,c)}(a)v_{(j,c)}\big\},\ (i,b) \in S.
\end{aligned}$$

Hence, more explicitly,

$$\begin{aligned}
v_{(0,0)} &= \tfrac{\lambda+\mu}{\alpha+\lambda+\mu}\{\tfrac{\mu}{\lambda+\mu}v_{(0,0)} + \tfrac{\lambda}{\lambda+\mu}v_{(0,1)}\}. \\
v_{(0,1)} &= max\Big\{r - \tfrac{h(1)}{\alpha+\lambda+\mu} + \tfrac{\lambda+\mu}{\alpha+\lambda+\mu}\{\tfrac{\mu}{\lambda+\mu}v_{(0,0)} + \tfrac{\lambda}{\lambda+\mu}v_{(1,1)}\},\ \tfrac{\lambda+\mu}{\alpha+\lambda+\mu}\{\tfrac{\mu}{\lambda+\mu}v_{(0,0)} + \tfrac{\lambda}{\lambda+\mu}v_{(0,1)}\}\Big\}. \\
v_{(i,0)} &= -\tfrac{h(1)}{\alpha+\lambda+\mu} + \tfrac{\lambda+\mu}{\alpha+\lambda+\mu}\{\tfrac{\mu}{\lambda+\mu}v_{(i-1,0)} + \tfrac{\lambda}{\lambda+\mu}v_{(i,1)}\},\ i \geq 0. \\
v_{(i,1)} &= max\Big\{r - \tfrac{h(i+1)}{\alpha+\lambda+\mu} + \tfrac{\lambda+\mu}{\alpha+\lambda+\mu}\{\tfrac{\mu}{\lambda+\mu}v_{(i,0)} + \tfrac{\lambda}{\lambda+\mu}v_{(i+1,1)}\}, \\
&\qquad\qquad -\tfrac{h(i)}{\alpha+\lambda+\mu} + \tfrac{\lambda+\mu}{\alpha+\lambda+\mu}\{\tfrac{\mu}{\lambda+\mu}v_{(i-1,0)} + \tfrac{\lambda}{\lambda+\mu}v_{(i,1)}\}\Big\} \\
&= max\{r + v_{(i+1,0)}, v_{(i,1)}\},\ i \geq 1.
\end{aligned}$$

It can be shown that, if $h(i)$ is nondecreasing and convex, there exists an optimal *control limit* policy.

3. Service rate control

As state space we have again $S = \{0,1,\dots\}$. In state 0 there is no control action available since there are no customers to serve. We may think of the action 0 = take no service action. In state $i \geq 1$ actions consist of the allowable service rate $0 < a_1 < a_2 < \cdots < a_m < 1$. This means that the server must serve if the buffer is nonempty ($a_1 > 0$) and that perfect service is unavailable ($a_m < 1$). The holding cost is the same as in arrival control. There is a nonnegative cost $c(k)$

of choosing to serve at rate a_k during a particular slot (the cost in state 0 is 0). Hence, the immediate cost is:

$$c(i,k) = \begin{cases} 0 & \text{if } i = 0; \\ h(i) + c(k) & \text{if } i \geq 1,\ 1 \leq k \leq m. \end{cases}$$

The transition probabilities are:

$$i = 0: \quad p_{0j}(0) = p_j,\ j \geq 0; \quad i \geq 1: \quad p_{i,i-1}(k) = a_k p_0 \qquad 1 \leq k \leq m;$$
$$p_{i,i+j}(k) = a_k p_{j+1} + (1 - a_k) p_j \quad 1 \leq k \leq m,\ j \geq 0.$$

8.4.2 Parallel queues

In parallel queues are a number of K servers with individual queues. Customers arrive at the router and are send to one of these servers. It is assumed that once the routing has taken place, the customer cannot switch from one queue to another. We assume that the service rates of the servers are constant. The control mechanism is involved through the routing decision for an arriving customer. An appropriate state description is the vector $i = (i_1, i_2, \ldots, i_K)$, where i_k is the number of customers in queue k $(k = 1, 2, \ldots, K)$. The cost is then a function of the pair (i, k), where k is the action chosen, i.e. the server to which the customer is routed. This cost consists of a holding cost reflecting the number of customers in each queue and a cost of routing to queue k.

Suppose that the customers that arrived in slot t were routed to queue k but that at the beginning of slot $t+1$ the controller wishes to route the newly arriving customers to queue $l \neq k$. We allow that this switch causes a *switching cost*. To handle this situation, we would enlarge the state description to be (i, k), where the current buffer content vector i is augmented with the previous routing decision. The cost is then a function of the state-action pair $\{(i, k), l\}$.

Let us assume that we have batch arrivals. The problem concerns the routing of an incoming batch to one of the K parallel servers. Each server maintains its own queue, and server k serves its customers at geometric rate μ_k, where $0 < \mu_k < 1,\ k = 1, 2, \ldots, K$. We also assume that the routing decision is made before the size of the incoming batch is observed.

There is a nonnegative holding cost $h_k(i_k)$ associated with the content of queue k. The total holding cost is $h(i) = \sum_{k=1}^{K} h_k(i_k)$. In addition there is a nonnegative cost $c(k, l)$ for changing the routing from server k to server l, where $c(k, k) = 0$ for each k. The cost structure is: $c\{(i, k), l)\} = h(i) + c(k, l)$.

Some thoughtful notation can facilitate the writing of the transition probabilities. Let $j(l)$ be the K-dimensional vector with j in the l-th place and 0 elsewhere. Then,

$$p_{(0,k)(j(l),l)}(l) = p_j,\ 1 \leq k \leq K,\ 1 \leq l \leq K,\ j \geq 0.$$

Now let $i \neq 0$ be a state vector and let $F(i) = \{j \mid i_j > 0\}$. Let $E(i) \subseteq F(i)$ be the subset of $F(i)$ (possibly empty) representing those servers who complete service during the current slot. The probability of this event is

$$\mathbb{P}\{E(i)\} = \prod_{k \in E(i)} \mu_k \prod_{k \in F(i) \setminus E(i)} (1 - \mu_k).$$

Finally let $e(E(i))$ be a vector with 1 in every coordinate of $E(i)$ and 0 elsewhere. If the system is in state (i,k) there is a probability p_j that the next batch contains j customers and there is a probability $\mathbb{P}\{E(i)\}$ that the servers of $E(i)$ complete their services. Hence, we have the following transition probabilities in case the router assigns the next batch to server l:

$$p_{(i,k)(i+j(l)-e(E(i)),l)}(l) = p_j\,\mathbb{P}\{E(i)\},\ i \neq 0,\ E(i) \subseteq F(i),\ 1 \leq k \leq K,\ 1 \leq l \leq K,\ j \geq 0.$$

8.5 Stochastic scheduling

In a scheduling problem, jobs have to be processed on a number of machines. Each machine can only process one job at a time. Each job i has a given processing time T_{ij} on machine j. In stochastic scheduling, these processing times are random variables. At certain time points decisions have to be made, e.g. which job is assigned to which machine. There is a utility function by which different policies can be measured, and we want to find a policy that optimizes the utility function. We will illustrate this in a number of examples.

8.5.1 Maximizing finite-time returns on a single processor

Suppose there are n jobs to be performed sequentially within a fixed time T. The ith job takes an exponentially amount of time with rate μ_i and, if completed within time T, earns the decision maker an amount r_i. At the start and whenever a job is completed the decision maker must decide which of the remaining jobs to process, with his objective being to maximize the total expected earnings.

It follows from the lack-of-memory property of the exponential distribution that, if job i is attempted for a time dt, then it will be completed with probability $\mu_i dt + o(dt)$, thus the expected gain will be $\mu_i r_i dt + o(dt)$. Hence, it seems as if the expected return is the same as if we earned $\mu_i r_i$ per unit time that job i is being performed. To show this formally, suppose that t units of time remain when job i is initiated. If X_i is the time needed to perform this job, then the expected return from job i is

$$\mathbb{E}\{\text{return from job } i\} = r_i \cdot \mathbb{P}\{X_i < t\} = r_i(1 - e^{-\mu_i t}) = \mu_i r_i \cdot \tfrac{1-e^{-\mu_i t}}{\mu_i}.$$

Since for any nonnegative stochastic variable Y with density function $f(y)$ we have

$$\mathbb{E}\{Y\} = \int_0^\infty yf(y)dy = \int_0^\infty \{\int_0^y dx\}f(y)dy = \int_0^\infty \{\int_x^\infty f(y)dy\}dx = \int_0^\infty \mathbb{P}\{Y > x\}dx,$$

and hence,

$$\mathbb{E}\{min(X_i,t)\} = \int_0^\infty \mathbb{P}\{min(X_i,t) > x\}dx = \int_0^t e^{-\mu_i x}dx = \tfrac{1-e^{-\mu_i t}}{\mu_i}.$$

Therefore, we obtain

$$\mathbb{E}\{Y\} = \mu_i r_i\,\mathbb{E}\{min(X_i,t)\} = \mu_i r_i\,\mathbb{E}\{\text{length of time job } i \text{ is worked on}\}.$$

Hence, it follows that, for any policy R,

$$\mathbb{E}_R\{\text{total return}\} = \sum_{i=1}^{n} \mu_i r_i\,\mathbb{E}_R\{\text{length of time job } i \text{ is worked on}\}. \tag{8.47}$$

That is, the total expeced reurn is the same as it would be if we earned money at a rate $\mu_i r_i$ whenever job i is worked on. From this we see that the expected amount earned by time T is maximized by working on jobs in decreasing order of $\mu_i r_i$. So at any decision time point the decision maker chooses job k where

$$\mu_k r_k = max_i\{\mu_i r_i \mid \text{job } i \text{ is not completed}\}.$$

8.5.2 Optimality of the μc-rule

1. One server allocation to parallel queues with preemption
Customers arrive at a system of m parallel queues and one server. The system operates at discrete time points, i.e. arrival times and service times take values in the set $\{1, 2, \dots\}$. Furthermore, the arrival times are arbitrary and the service time T_i, for a customer in queue i, is geometrically distributed with rate μ_i,

$$\mathbb{P}\{T_i = n\} = (1-\mu_i)^{n-1} \cdot \mu_i,\ n \in \mathbb{N},\ \text{with } \mu_i \in (0,1),\ 1 \le i \le m,\ \text{and } \mathbb{E}\{T_i\} = \mu_i^{-1}.$$

At any time point $t = 1, 2, \dots$ the server chooses a customer from one of the queues; this is an example of a server assignment model. Services may be interrupted and resumed later on (*preemption*). For each customer in queue i, a cost c_i is charged per unit of time that this customer is in the system. A policy is a rule to assign each server to one of the queues. Which policy minimizes the total cost in T periods? This model is more interesting than the nonpreemptive model, which is a rather trivial example (cf. Exercise 1.10).

Let $N_i^t(R)$ be the number of customers in period t in queue i, if policy R is used. Then, the performance measure is

$$min_R\ \mathbb{E}\left\{\sum_{t=1}^{T}\sum_{i=1}^{m} c_i \cdot N_i^t(R)\right\}.$$

The next theorem shows that the so-called μc-rule is an optimal policy. This rule assigns the server to queue k, where k is a nonempty queue satisfying

$$\mu_k c_k = max_i\{\mu_i c_i \mid \text{queue } i \text{ is nonempty}\}.$$

Note that $\mu_i c_i$ is the expected cost per unit of service for a customer in queue i, and by using the μc-rule, the largest reduction of the expected cost in the next period is obtained.

Theorem 8.10
The μc-rule is optimal for the preemptive allocation of a single server to parallel queues.

Proof
Assume that the μc-rule is optimal after some time $t \le T$ (any rule is optimal after time T, because we consider a finite horizon of T periods). It will be shown that this rule is also optimal at time t. Then, by backward induction, it is clear that the μc-rule is optimal over the whole horizon.

For any sample path of the states and actions of the stochastic process we make the following observation. Consider a policy that serves a customer in queue j at time t while there is a customer in queue i at time t, where i and j are such that $c_i\mu_i > c_j\mu_j$. Denote by τ the first time after time t that this policy services a customer in queue i (let $\tau = T+1$ if the policy does not serve a customer in queue i during the times $t+1, t+2, \ldots, T$).
Modify the policy by serving a customer in queue i at time t and a customer in queue j at time τ, i.e. interchange the actions at times t and τ. The effect of this modification can be calculated as follows. With probability μ_i the service of the customer in queue i will be completed in epoch t in the modified policy. Thus with probability μ_i the cost of the customer in queue i is reduced by $\sum_{s=t+1}^{\tau} c_i$. Similarly, with probability μ_j the cost of the customer in queue j is increased by $\sum_{s=t+1}^{\tau} c_j$. Thus, the the expected reduction in cost is $(c_i\mu_i - c_j\mu_j)(\tau - t) > 0$. This shows that the μc-rule is an optimal policy. □

2. Serving Poisson arrivals nonpreemptively with a single server
Jobs of different classes arrive as independent Poisson arrivals. The jobs of class i go to queue i, $i = 1, 2, \ldots, m$. A job in queue i has a mean service time equal to $\frac{1}{\mu_i}$ and a waiting cost of c_i per unit of time. All the service times are independent. The problem is to find a nonpreemptive server allocation policy that minimizes the long-term *average waiting cost* per unit of time, i.e.

$$min_R \ \mathbb{E}\left\{\sum_{i=1}^{m} c_i \cdot N_i(R)\right\},$$

where $N_i(R)$ denotes the long-term average number of customers in queue i in the system, given policy R, i.e. $N_i(R) = lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} N_i^t(R)$ with $N_i^t(R)$ be the number of customers in period t in queue i, if policy R is used.

Theorem 8.11
The μc-rule is optimal for serving Poisson arrivals nonpreemptively with a single server.

Proof
The proof is based on a *working-conserving property.*
First one observes that it suffices to consider nonidling policies, i.e. policies under which the server is never idle when there is a customer to serve. Indeed, one can always consider that an idling policy is in fact serving a class $m+1$ of customers with $c_{m+1} = 0$. If the result is true for nonidling policies, the fact that $\mu_{m+1}c_{m+1} = 0$ implies that the class $m+1$ shouls be served last, i.e. that an optimal policy will be nonidling.
Second, consider $\sum_{i=1}^{m} \frac{1}{\mu_i} \cdot N_i(R)$. The term $\frac{1}{\mu_i} \cdot N_i(R)$ is the expected time the server has to work in queue i in the steday state situation, given policy R. So, $\sum_{i=1}^{m} \frac{1}{\mu_i} \cdot N_i(R)$ is the expected service time for the whole system in the steady state situation, i.e. the average workload of the system. Using an argument as in Little's formula, this workload is independent of the policy. So, we write $W = \sum_{i=1}^{m} \frac{1}{\mu_i} \cdot N_i(R)$, from which we obtain $N_1(R) = \mu_1\{W - \sum_{i=2}^{m} \frac{1}{\mu_i} \cdot N_i(R)\}$.

Third, assume that $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_m\mu_m$. Then, we have

$$\begin{aligned} \sum_{i=1}^m c_i \cdot N_i(R) &= c_1 \cdot N_1(R) + \sum_{i=2}^m c_i \cdot N_i(R) \\ &= \mu_1 c_1 W + \sum_{i=2}^m \{c_i - \tfrac{\mu_1 c_1}{\mu_i}\} \cdot N_i(R) \\ &= \mu_1 c_1 W + \sum_{i=2}^m \tfrac{1}{\mu_i}(\mu_i c_i - \mu_1 c_1) \cdot N_i(R). \end{aligned}$$

The coefficients of $N_i(R)$, $i = 2, 3, \ldots, m$ in the above expression are nonnegative. Hence, $\sum_{i=1}^m c_i \cdot N_i(R)$ is minimized by the policy that makes $N_i(R)$ as large as possible ($i = 2, 3, \ldots, m$). Such a policy must necessarily serve a customer of queue 1 whenever it can.

Fourth, consider all the nonidling policies that serve queue 1 whenever it can. The set of times available for those policies to serve the other queues is the same for all these policies. One can check that all these policies have the same value of $\sum_{i=2}^m \frac{1}{\mu_i} \cdot N_i(R)$ (by Little's formula again). Repeating the above argument shows that among these policies, the ones that minimizes $\sum_{i=1}^m c_i \cdot N_i(R)$ must serve queue 2 whenever that can.

Continuing in this way concludes the proof. □

8.5.3 Optimality of threshold policies

Waiting for a fast server or using a slow one

Customers arrive at a service facility that has two servers. The arrival times form a Poisson process with rate λ. The service times are assumed to be exponentially distributed with the respective rates μ_1 (for server 1) and μ_2 (for server 2), where $\mu_1 \geq \mu_2$. Service is nonpreeemptive. When one of the servers becomes available, the decision has to be taken whether or not to send a customer to this server.

This is a customer assignment model. The model is not discrete, but continuous in time. Let $N^t(R)$ be the number of customers in the system at time t. As performance measure the total discounted costs are used, i.e.

$$min_R \ \mathbb{E}\left\{\int_0^\infty e^{-\alpha t} N^t(R) dt\right\},$$

where $\alpha > 0$, which is the continuous analogon of the total discounted costs in the discrete case.

The trade-off is between waiting for the fast server to become available and committing a customer to the slow queue. The next theorem shows that for this model an optimal *threshold policy* exists, namely server 1 will always be used when it becomes available, and the slower server, server 2, is only used when the total number of customers in the queue exceeds some threshold number N.

Theorem 8.12

There is some number N such that the optimal policy is to use the fast server all time and to send a customer to the slow server (when this slow server is available) if and only if the number of customers in the system at that time is at least N.

Proof

We give an ouline of the proof. Decision times are services completion times and arrival times when at least one server is idle. We will rely on the fact that there is a stationary deterministic optimal policy. Consider a stationary deterministic policy. The policy cannot be optimal unless it uses the fast server whenever possible. If the policy uses the fast server whenever possible, then it is specified by a subset A of $\{1, 2, 3, \dots\}$ with the interpretation that a customer is sent to the slower server at decision times when that server is idle and when the queue lenght belongs to A. It then remains to show that the set A must be of the form $A = \{N, N+1, N+2, \dots\}$. This is done by contradiction.

Assume that the set A contains a "gap". That is, assume that $A = \{\dots, M, N, \dots, \dots\}$ with $N \geq M + 2$. Say that at $t = 0$ the fast server is busy, the slow server is idle and there are $M + 1$ customers waiting to be allocated to a server. The policy will then wait until the queue length reaches either M or N before sending a customer to the slow server. It is sufficient to show that the policy can be improved by sending a customer at time $t = 0$ to the slow server.

To see this, denote by σ the service time of that customer sent at time 0 to the slower server. This service time is known at time σ. Pretend that the policy corresponding to A was in fact used, by doing as if the customer had not been sent at time 0 but had been sent only when the queue length hits either M or N, at time τ, say, and by pretending that the slow server is busy during $[\tau, \tau + \sigma]$. This shows that the modified policy behaves as the one corresponding A, except that one customer leaves the queue at time σ instead of time $\tau + \sigma$.

It remains only to show that for all $M \geq 1$ there is some $N > M$ such that $N \in A$. Again, this can be shown by contradiction (the intuition is that if the que length is very large, than it is very likely that a customer at the end of the queue would be served by the slow sever before the fast server could become available). □

8.5.4 Optimality of join-the-shortest-queue policies

Customer allocation to parallel queues

Customers arrive at arbitrary known times at a system consisting of m identical $M/M/1$ queues in parallel. That is, the service times in all queues are independent and exponentially distributed with the same rate μ. The problem is to choose, at each arrival time, which queue the arriving customer should join so as to minimize

$$min_R \ \mathbb{E}\left\{\int_0^\infty e^{-\alpha t} \sum_{i=1}^m N_i^t(R)\, dt\right\}, \tag{8.48}$$

where $\alpha > 0$ is the discount rate and $N_i^t(R)$ is the number of customers at time t in queue i, given policy R. The information available when the decision is made is the evoluation of the vector of queue length up to that time and the set of arrival times. It is assumed that the arrival times are such that $\mathbb{E}\{\int_0^\infty e^{-\alpha t} \sum_{i=1}^m N_i^t(R)\, dt\}$ is finite for at least one policy R.

An *SQP* (*shortest queue policy*) is a policy that sends each arriving customer to the shortest queue. In Theorem 8.13 it will be shown that an optimal *SQP* exists. It shouls be noted that

an SQP is clearly *individually* optimal for each customer for arbitrary decisions of the customers who arrived before him. However, this does not imply that the policy is optimal *socially*, i.e. in the sense of minimizing $\mathbb{E}\left\{\int_0^\infty e^{-\alpha t}\sum_{i=1}^m N_i^t(R)\,dt\right\}$. Indeed, it is often the case that individuals have to accept sacrifices for the benefit of society at large. Mathematically, each customer should take into account not only the personel cost of that customer (here, the discounted waiting time), but also the impact of the decision on the other customers (here, on those who will arrive behind him).

For the proof of this result we use the *forward induction method*. This method can be described as follows. Denote by X_t the state process corresponding to a policy and by Y_t the process corresponding to another policy. Suppose that there exists a partial ordering $\mathbf{B}$ on the set of possible states with the following two properties:

(1) it should be such that it is possible to prove that $X_t\,\mathbf{B}\,Y_t$ implies that $X_s\,\mathbf{B}\,Y_s$ for all $s \geq t$.

(2) the ordering should imply that the cost corresponding to X_t is not larger than the cost corresponding to Y_t.

Then it follows that the policy corresponding to X_t is an optimal policy.

Theorem 8.13

The customer allocation to parallel queues model has an optimal "join-the-shortest-queue" policy.

Proof

For two random variables V and W taking values in $\{0,1,2,\dots\}^m$ we write $V\,\mathbf{B}\,W$ if there exists two random variables V^* and W^* such that:

(a) V^* has the same distribution as V.

(b) W^* has the same distribution as W.

(c) $\mathbb{P}\{S_i(W^*) \geq S_i(V^*),\ 1 \leq i \leq m\} = 1$

where $S_i(V^*)$ denotes the sum of the i largest components of V^* and similarly for $S_i(W^*)$.

Denote by X_t the vector of queue lenghts at time t corresponding to the SQP, and by Y_t the vector of queue lenghts at time t corresponding to an arbitrary policy R. Assume that we have shown $X_t\,\mathbf{B}\,Y_t,\ t \geq 0$. Using the well known and easily verified fact that any $\{0,1,2,\dots\}$-valued random variable X satisfies $\mathbb{E}\{X\} = \sum_{k=0}^\infty \mathbb{P}\{X \geq k\}$, we obtain

$$\begin{aligned}
\mathbb{E}\left\{\sum_{i=1}^m N_i^t(R)\right\} &= \mathbb{E}\left\{\sum_{i=1}^m \{Y_t\}_i\right\} = \sum_{k=0}^\infty \mathbb{P}\{\sum_{i=1}^m \{Y_t\}_i \geq k\} \\
&= \sum_{k=0}^\infty \mathbb{P}\{\sum_{i=1}^m \{Y_t^*\}_i \geq k\} = \sum_{k=0}^\infty \mathbb{P}\{S_m(Y_t^*) \geq k\} \\
&\geq \sum_{k=0}^\infty \mathbb{P}\{S_m(X_t^*) \geq k\} = \sum_{k=0}^\infty \mathbb{P}\{\sum_{i=1}^m \{X_t^*\}_i \geq k\} \\
&= \sum_{k=0}^\infty \mathbb{P}\{\sum_{i=1}^m \{X_t\}_i \geq k\} = \mathbb{E}\left\{\sum_{i=1}^m \{X_t\}_i\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^m N_i^t(SQP)\right\}.
\end{aligned}$$

Hence, the cost (at any time t) in (8.48) corresponding to the SQP is not larger than the cost corresponding to policy R. Thus the partial order $\mathbf{B}$ has the second desired property mentioned in the description of the forward induction method.

To prove that $X_t \mathbf{B} Y_t$ $t \geq 0$, let $0 = t_0 \leq t_1 < t_2 < t_3 < \cdots$ be the values of the arrival and potential service completion times. Assume that $X_t \mathbf{B} Y_t$ for some $t \geq 0$ (for $t = 0$ $X_t \mathbf{B} Y_t$ holds since $X_0 = Y_0$), where $t_{n-1} \leq t < t_n$ for some $n \geq 1$. It then suffices to show that $X_{t_n} \mathbf{B} Y_{t_n}$.
First consider the case when t_n is an arrival time. Let X_t^* and Y_t^* be such that the properties (a), (b) and (c), mentioned in the begin of the proof, hold. Notice that $S_m(X_{t_n}^*) = S_m(X_t) + 1$, while $S_i(Y_{t_n}^*) = S_i(Y_t) + 1$ for all $i \geq k$ if policy R sends the arriving customer to the kth largest queue. This shows $X_{t_n} \mathbf{B} Y_{t_n}$ for this case.
Next consider the case when the event t_n is a potential completion time. Define $X_{t_n}^*$ and $Y_{t_n}^*$ by deciding that if the potential service completion time occurs in the kth longest queue of X_t^*, then the same is true for Y_t^*. This modifies the joint distribution but not the marginals (here one uses the memoryless property of the exponential distribution, implying that the probability that a completion occurs in the kth longest queue is independent of k and the same is true for Y_t^*), so that (a) and (b) will hold for $V^* = X_{t_n}^*$ and $W^* = Y_{t_n}^*$.
To verify (c) one uses the fact that if the potential service completion occurs in the kth longest queue of X_t^*, then $S_i(X_{t_n}^*) = \begin{cases} S_i(X_t^*) & \text{if } i < k \\ S_i(X_t^*) - 1 & \text{if } i \geq k \end{cases}$ and $S_i(Y_{t_n}^*) = \begin{cases} S_i(Y_t^*) & \text{if } i < k \\ S_i(Y_t^*) - 1 & \text{if } i \geq k \end{cases}$

Hence, we conclude that $S_i(Y_{t_n}^*) \geq S_i(X_{t_n}^*)$ for $i = 1, 2, \ldots, m$. □

8.5.5 Optimality of LEPT and SEPT policies

Many results can be shown by the principle of dynamic programming. In this section we present several examples using the optimality equation of dynamic programming.

1. Guessing a diamond
A deck of 52 cards is to be turned over one at a time. Before each card is turned we are given the opportunity to say whether or not it will be a diamond. We are allowed to say that a card is a diamond only one. The objective is to maximize the probability of being correct.

Theorem 8.14
All the decisions rules that select at least one card before all the diamonds are turned over are optimal.

Proof
Denote by $v_n(m)$ the maximum probability when there are n cards left to be turned and when m cards of those n cards are diamonds. Obviously, $v_m(m) = 1,\ 1 \leq m \leq 13$ and $v_n(0) = 0,\ n \geq 1$. The first claim is that

$$v_n(m) = max\Big\{\frac{m}{n}, \frac{m}{n} v_{n-1}(m-1) + \frac{n-m}{n} v_{n-1}(m)\Big\},\ n \geq 2,\ 1 \leq m \leq min\{n, 13\}. \qquad (8.49)$$

To prove this, notice that the first term in the maximization is the probability of being correct if the decision is to declare that the next card is a diamond. We will show that the second term

gives the maximum probability of being correct if the decision is not to declare that the next card is a diamond. Indeed, in the latter case there are two possibilities. With probability $\frac{m}{n}$, the next card is a diamond, so there are $n-1$ cards left with $m-1$ diamonds, with a maximum probability of being correct equal to $v_{n-1}(m-1)$. With probability $\frac{n-m}{n}$, the next card is not a diamond, and there $n-1$ cards left with m diamonds, with a maximum probability of being correct equal to $v_{n-1}(m)$.
The second claim is all the decisions rules that select at least one card before all the diamonds are turned over are optimal. It is sufficient to show that $v_n(m) = \frac{m}{n},\ n \geq 2,\ 1 \leq m \leq min\{n, 13\}$. We apply induction on n (the case $n = 2$ is trivial). Assume that $v_{n-1}(m) = \frac{m}{n-1}$ for all $1 \leq m \leq min\{n-1, 13\}$. Then,

$$\frac{m}{n}v_{n-1}(m-1) + \frac{n-m}{n}v_{n-1}(m) = \frac{m(m-1)}{n(n-1)} + \frac{(n-m)m}{n(n-1)} = \frac{m}{n}. \qquad \square$$

2. Processing a set exponential jobs on parallel machines
A set of n jobs has to be processed, each by one of m identical processors. The jobs have independent and exponentially distributed service times with rates $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$. The n jobs are ready to be processed at tome 0, thus there are no arrivals. Two objectives will be considered: the *expected makespan*

$$MS = \mathbb{E}\left\{max\{T_1, T_2, \ldots, T_n\}\right\}$$

and the *expected flowtime*

$$FT = \mathbb{E}\left\{\sum_{j=1}^{n} T_j\right\}$$

where T_j is the completion time of job $j,\ j = 1, 2, \ldots, n$.

A $LEPT$ policy is a policy that, at time 0 and at each service completion allocates the jobs to available servers in the order $1, 2, \ldots, n$, i.e. largest expected processing times first ($LEPT$). A $SEPT$ policy is a policy that, at time 0 and at each service completion allocates the jobs to available servers in the order $n, n-1, \ldots, 1$, i.e. shortest expected processing times first ($SEPT$). It can we shown that a $LEPT$ policy is optimal for MS, the expected makespan, and that a $SEPT$ policy is optimal for FT, the expected flowtime. We will sketch these results, using the optimality equation of dynamic programming, for the case of two processors ($m = 2$). Furthermore, we present an alternative proof for the optimality of the $LEFT$ policy.

Theorem 8.15
Consider a stochastic scheduling problem in which n jobs with exponential processing times are scheduled on two identical machines. Then, the expected makespan is minimized by the LEPT policy.

Proof (outline)

Assume that the $LEPT$ policy is optimal when there are at most $n-1$ jobs to process (this assumption is verified for $n=2$). It will be shown to be optimal for n jobs. Let $MS(i)$ be the minimum makespan for the jobs $\{1,2,\ldots,n\}\backslash\{i\},\ 1 \leq i \leq n$. By the hypothesis, this makespan $MS(i)$ is achieved by the $LEPT$ policy. We will conditioning on the first of the two jobs initially processed, say the jobs i and j. Notice that the minimum of the exponential distribution for the jobs i and j is also an exponential distribution with parameter $\mu_i+\mu_j$, and that the fractions $\frac{\mu_i}{\mu_i+\mu_j}$ and $\frac{\mu_j}{\mu_i+\mu_j}$ are the probabilities that job i and job j, respectively, is first completed job. At that completion time, the remaining service time for the other job is also an exponential distribution with the same parameter. Hence, we obtain

$$MS = min_{i<j}\left\{\frac{1}{\mu_i+\mu_j} + \frac{\mu_i}{\mu_i+\mu_j}MS(i) + \frac{\mu_j}{\mu_i+\mu_j}MS(j)\right\}, \tag{8.50}$$

or equivalently,

$$0 = min_{i<j}\left\{1+\mu_i\{MS(i)-MS\}+\mu_j\{MS(j)-MS\}\right\}, \tag{8.51}$$

and the minimum in (8.51) is achieved by the same pair (i,j) as in (8.50). To show that $LEPT$ is also optimal when there are n jobs to process, one has to show that the minimum in (8.51) is achieved by $(i,j)=(1,2)$. Let $D_{ij}=\mu_i\{MS(i)-MS\}-\mu_j\{MS(j)-MS\},\ i<j$. Then, it can be shown by induction on n that $D_{ij}\leq 0$ if $i<j$, implying the result. □

Theorem 8.16

Consider a stochastic scheduling problem in which n jobs with exponential processing times are scheduled on two identical machines. Then, the expected flowtime is minimized by the $SEPT$ policy.

Proof (outline)

The proof has the same structure as the proof of Theorem 8.15. Assume that the $SEPT$ policy is optimal when there are at most $n-1$ jobs to process (this assumption is verified for $n=2$). It will be shown to be optimal for n jobs. Let $FT(i)$ be the minimum flowtime for the jobs $\{1,2,\ldots,n\}\backslash\{i\},\ 1 \leq i \leq n$. By the hypothesis, this flowtime $FT(i)$ is achieved by the $SEPT$ policy. We will conditioning on the first of the two jobs initially processed, say the jobs i and j. The completion time of the first completed job has expectation $\frac{1}{\mu_i+\mu_j}$, which will be part of each completion time $T_j,\ j=1,2,\ldots,n$. After that time the remaining $n-1$ jobs have exponential distributions with the original rates. Hence, we obtain

$$FT = min_{i<j}\left\{\frac{n}{\mu_i+\mu_j} + \frac{\mu_i}{\mu_i+\mu_j}FT(i) + \frac{\mu_j}{\mu_i+\mu_j}FT(j)\right\}, \tag{8.52}$$

or equivalently,

$$0 = min_{i<j}\left\{n+\mu_i\{FT(i)-FT\}+\mu_j\{FT(j)-FT\}\right\}, \tag{8.53}$$

and the minimum in (8.53) is achieved by the same pair (i,j) as in (8.52). To show that $SEPT$ is also optimal when there are n jobs to process, one has to show that the minimum in (8.51) is achieved by $(i,j)=(n,n-1)$. This can be done in a similar way as in Theorem 8.15. □

Alternative proof for the optimality of the $LEFT$ policy

It will help our analysis to assume that at time 0 one of the two processors is occupied on a job 0 and will remain accupied for a time X_0, where X_0 is assumed to have an arbitrary distribution and is independent of the other jobs. For any permutation $i_1, i_2, \ldots, i_n$ of $1, 2, \ldots, n$, putting the jobs on the processors in that order defines a schedule. Hence, is policy is a schedule $(0, i_1, i_2, \ldots, i_n)$. Let X_j be the stochastic duration of job j, $j = 0, 1, 2, \ldots, n$ and let D be the amount of time that only one of the processors is busy. That is, at time $MS - D$ one of the processors completes work on a job and finds no other jobs available. As the total amount of work processed is $M + (M - D) = \sum_{j=0}^{n} X_j$, Hence, minimizing the expected difference of the times at which the procesoors become idle also leads to minimizing the expected makespan. The following lemma will be used to show that the $LEPT$ policy is optimal.

Lemma 8.8

Consider the policies $R = (0, 2, 1, 3, 4, \ldots, n)$ and $R_* = (0, 1, 2, \ldots, n)$. Then, $\mathbb{E}_{R_*}\{D\} \leq \mathbb{E}_R\{D\}$.

Proof

Let $p(j)$ and $p_*(j)$, $j = 0, 1, \ldots, n$ be the probabilities that the last job to be completed is job j, under policies R and R_*, respectively. Clearly, $p(0) = p_*(0) = \mathbb{P}\{X_0 > \sum_{j=1}^{n} X_j\}$. We shall prove by induction on n that

$$p_*(1) \leq p(1) \text{ and } p_*(j) \geq p(j), \ j = 2, 3, \ldots, n. \tag{8.54}$$

This is obvious if $n = 1$ ($p_*(1) = p(1) = \mathbb{P}\{X_0 \leq X_1\}$). Assume (8.54) is true whenever there are only $n - 1$ jobs (in addition to job 0) to be scheduled, and let $q_*(j)$ and $q(j)$ be the probabilities that job j is the last of jobs $0, 1, 2, \ldots, n-1$ under policies R^* and R, respectively. Then, by the induction hypothesis

$$q_*(1) \leq q(1) \text{ and } q_*(j) \geq q(j), \ j = 2, 3, \ldots, n-1. \tag{8.55}$$

Now consider the n-job case. However, using the lack of memory of the exponential distribution and the fact that job n is the last to begin processing under both policies, we have

$$p(j) = q(j) \cdot \frac{\mu_n}{\mu_n + \mu_j}, \ p_*(j) = q_*(j) \cdot \frac{\mu_n}{\mu_n + \mu_j}, \ j = 1, 2, \ldots, n-1.$$

Hence, from (8.55), we obtain $p_*(1) \leq p(1)$ and $p_*(j) \geq p(j)$, $j = 2, 3, \ldots, n-1$. Finally, using

$p(n) = 1 - \sum_{j=0}^{n-1} p(j) = 1 - \sum_{j=0}^{n-1} q(j) \cdot \left\{1 - \frac{\mu_j}{\mu_n + \mu_j}\right\} = \sum_{j=0}^{n-1} q(j) \cdot \frac{\mu_j}{\mu_n + \mu_j}$

and similarly $p_*(n) = \sum_{j=0}^{n-1} q_*(j) \cdot \frac{\mu_j}{\mu_n + \mu_j}$, one can write

$$\begin{aligned} p_*(n) - p(n) &= \sum_{j=0}^{n-1} \{p_*(j) - p(j)\} = \sum_{j=0}^{n-1} \{q_*(j) - q(j)\} \cdot \frac{\mu_j}{\mu_n + \mu_j} \\ &= \{q_*(1) - q(1)\} \cdot \frac{\mu_1}{\mu_n + \mu_1} + \sum_{j=2}^{n-1} \{q_*(j) - q(j)\} \cdot \frac{\mu_j}{\mu_n + \mu_j} \\ &\geq \frac{\mu_1}{\mu_n + \mu_1} \sum_{j=1}^{n-1} \{q_*(j) - q(j)\} = 0, \end{aligned}$$

where the inequality follows because $\mu_j \geq \mu_1$ implies that $\frac{\mu_j}{\mu_n + \mu_j} \geq \frac{\mu_1}{\mu_n + \mu_1}$.

Consider any policy $\pi = (0, i_1, i_2, \ldots, i_n)$ and assume that job j, $j \geq 1$, is the last job to be completed. Since at time $M - D$ job j is the last job to be completed, the remaining processing time is exponential distributed with rate μ_j, so we have

$$\mathbb{E}_\pi\{D \mid \text{job } j \text{ is the last job to be completed}\} = \tfrac{1}{\mu_j},\ j = 1, 2, \ldots, n.$$

Furthermore, we have

$$\mathbb{E}_\pi\{D \mid \text{job 0 is the last job to be completed}\} = \mathbb{E}_\pi\{X_0 - \textstyle\sum_{j=1}^n X_j \mid X_0 > \sum_{j=1}^n X_j\}.$$

Therefore, one can write

$$\begin{aligned}
\mathbb{E}_\pi\{D\} &= \textstyle\sum_{j=0}^n p(j) \cdot \mathbb{E}_\pi\{D \mid \text{job } j \text{ is the last job to be completed}\} \\
&= \textstyle\sum_{j=1}^n \frac{p(j)}{\mu_j} + p(0) \cdot \mathbb{E}_\pi\{X_0 - \sum_{j=1}^n X_j \mid X_0 > \sum_{j=1}^n X_j\}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}_{R_*}\{D\} - \mathbb{E}_R\{D\} &= \textstyle\sum_{j=1}^n \frac{1}{\mu_j}\{p_*(j) - p(j)\} = \frac{1}{\mu_1}\{p_*(1) - p(1)\} + \sum_{j=2}^n \frac{1}{\mu_j}\{p_*(j) - p(j)\} \\
&\leq \tfrac{1}{\mu_1}\{p_*(1) - p(1)\} + \textstyle\sum_{j=2}^n \frac{1}{\mu_1}\{p_*(j) - p(j)\} = \frac{1}{\mu_1}\sum_{j=1}^n \{p_*(j) - p(j)\} \\
&= \tfrac{1}{\mu_1}\{\big(1 - p_*(0)\big) - \big(1 - p(0)\big)\} = 0.
\end{aligned}$$

□

Note

From the proof of Lemma 8.8 it follows that the lemma is true for any order of the jobs in which $\mu_1 = min_{1 \leq j \leq n}\, \mu_j$.

Theorem 8.17

The LEFT policy is optimal.

Proof

Consider an arbitrary policy that does not initially process 1, say policy $(0, i_1, i_2, \ldots, i_k, i_{k+1}, 1, \ldots)$. By considering this at the time at which only one of the jobs $0, i_1, i_2, \ldots, i_k$ have not yet finished its processing, we see, using Lemma 8.8, that the schedule $(0, i_1, i_2, \ldots, i_k, 1, i_{k+1}, \ldots)$ has a smaller expected makespan. Continuing in this way we see that $(0, 1, i_1, i_2, \ldots, i_k, i_{k+1}, \ldots)$ is better. If $i_2 \neq 2$ then, repeating this argument, we show that $(0, 1, 2, i_1, i_2, \ldots, i_k, i_{k+1}, \ldots)$ is better. Continuing in this matter shows that the policy $(0, 1, 2, \ldots, n)$ is optimal. Since we may use for job 0 a job with processing time $X_0 = 0$, we have shown that the *LEPT* policy is optimal. □

Remark

Whereas Theorem 8.17 only proved that scheduling tasks in increasing order of their exponential service rates is optimal among the $n!$ policies that determine their ordering in advance, it is also optimal among all policies. That is, it remains optimal even when the choice of tasks to begin processing is allowed to depend on what has occurred up to that time. This is shown by induction as follows. It is immediate when $n = 1$, so assume it to be true whenever there are $n - 1$ tasks to be processed. Now, whichever of the n tasks is initially processed (alongside task 0), at the moment one of the two processors becomes free, it follows by the induction hypothesis that the

remaining tasks should be scheduled in increasing order of their rates. Hence, the only policies we need consider are those n policies for which task i $(i = 1, 2, \ldots, n)$ is scheduled first, and the remaining tasks are scheduled in increasing order of their rates. But Theorem 8.17 shows that the optimal policy of this type is the one that schedules the n tasks in increasing order of their rates. This completes the induction.

Stochastic ordering

We say that the random variable $X \geq_{st} Y$ if $\mathbb{P}\{X > a\} \geq \mathbb{P}\{Y > a\}$ for all a.

Lemma 8.9

If $X \geq_{st} Y$, then $\mathbb{E}\{X\} \geq \mathbb{E}\{Y\}$.

Proof

Assume first that X and Y are nonnegative random variables. Then,

$$\mathbb{E}\{X\} = \int_0^\infty \mathbb{P}\{X > x\}dx \geq \int_0^\infty \mathbb{P}\{Y > x\}dx = \mathbb{E}\{X\}.$$

In general, we can write any random variable Z as the difference of two nonnegative random variables as $Z = Z^+ - Z^-$, where $Z^+ = \begin{cases} Z & \text{if } Z \geq 0 \\ 0 & \text{if } Z < 0 \end{cases}$ and $Z^- = \begin{cases} 0 & \text{if } Z \geq 0; \\ -Z & \text{if } Z < 0. \end{cases}$

We leave it as an exercise (see Exercise 8.6) to show that $X \geq_{st} Y$ implies $X^+ \geq_{st} Y^+$ and $X^- \leq_{st} Y^-$. Hence, $\mathbb{E}\{X\} = \mathbb{E}\{X^+\} - \mathbb{E}\{X^-\} \geq \mathbb{E}\{Y^+\} - \mathbb{E}\{Y^-\} = \mathbb{E}\{Y\}$. □

Lemma 8.10

$X \geq_{st} Y \Leftrightarrow \mathbb{E}\{f(X)\} \geq \mathbb{E}\{f(Y)\}$ for all nondecreasing functions f.

Proof

Suppose first that $X \geq_{st} Y$ and let f be an nondecreasing funcion. Then it is, by Lemma 8.9, sufficient to show that $f(X) \geq_{st} f(Y)$. Letting $f^{-1}(a) = inf\{x \mid f(x) > a\}$, we have

$$\mathbb{P}\{f(X) > a\} = \mathbb{P}\{X > f^{-1}(a)\} \geq \mathbb{P}\{Y > f^{-1}(a)\} = \mathbb{P}\{f(Y) > a\}.$$

Now suppose that $\mathbb{E}\{f(X)\} \geq \mathbb{E}\{f(X)\}$ for all nondecreasing functions f.

For any a, let f_a be the nondecreasing function $f_a(x) = \begin{cases} 1 & \text{if } x > a; \\ 0 & \text{if } x \leq a. \end{cases}$

Then, because $\mathbb{E}\{f_a(X)\} = \mathbb{P}\{X > a\}$ and $\mathbb{E}\{f_a(Y)\} = \mathbb{P}\{Y > a\}$, we see from $\mathbb{E}\{f_a(X)\} \geq \mathbb{E}\{f_a(Y)\}$ that $\mathbb{P}\{X > a\} \geq \mathbb{P}\{Y > a\}$, i.e. $X \geq_{st} Y$. □

Remark

It can be shown that the policy given in Theorem 8.17 has the property that it stochastically minimizes the makespan. That is, that for any a, the probability that the makespan exceeds a is minimized by this policy. This is a stronger result than that in Theorem 8.17, which states only that the policy minimizes the expected makespan. In addition, it can also be shown that the stated policy stochastically maximizes the time until one of the processors becomes idle. That is, in the notation of this section, it maximizes the probabilty that $M - D$ exceeds a for each a.

8.5.6 Maximizing finite-time returns on two processors

Consider the same model as in section 8.5.1, now there are two servers. It follows as in (8.47) that the total expected return under any policy R can be expressed as

$$\mathbb{E}_R\{\text{total return}\} = \sum_{i=1}^{n} \mu_i r_i \, \mathbb{E}_R\{\text{length of time job } i \text{ is worked on}\}. \tag{8.56}$$

Thus, at first glance, it might seem that an optimal policy would be to sequence the tasks in decreasing order of $\mu_i r_i$, as in the case when there is only a single server. To see that this need not be the case, suppose that $\mu_i r_i = 1,\ i = 1, 2, \ldots, n$. Then, the conjecture would imply that all orderings are optimal. Further, assume that $\mu_1 < \mu_2 < \cdots < \mu_n$. The expected return by time T for any policy is equal to the expected total processing time on all tasks by T. Because the *LEPT* policy uniquely stochastically maximizes the time until one of the processors becomes idle (uniquely because the rate are strictly increasing, see the proof of Theorem 8.8), it also uniquely stochastically maximizes the total processing time by T and is thus uniquely optimal under our new objective function. However, this contradicts the conjecture that it is optimal to process tasks in decreasing order of $\mu_i r_i$, for, in the case $\mu_i r_i = 1$ for all i, this conjecture implies that all orderings are optimal. We can, however, prove that the policy that works on the jobs in decreasing order of $\mu_i r_i$ is optimal in a special case. In order to prove this special case we need the following lemma.

Lemma 8.11

Let $T_1, T_2, \ldots, T_n,\ S_1, S_2, \ldots, S_n$ and $c_1, c_2, \ldots, c_n$ be nonnegative numbers such that $\sum_{i=1}^{j} T_i \geq \sum_{i=1}^{j} S_i,\ j = 1, 2, \ldots, n$ and $c_1 \geq c_2 \cdots \geq c_n \geq 0$. Then, $\sum_{i=1}^{n} c_i T_i \geq \sum_{i=1}^{n} c_i S_i$.

Proof

Let $T_{n+1} = 0,\ T = \sum_{i=1}^{n+1} T_i,\ S_{n+1} = T - \sum_{i=1}^{n} S_i \geq 0$. Also, let X and Y be random variables such that $\mathbb{P}\{X = i\} = \frac{T_i}{T},\ \mathbb{P}\{Y = i\} = \frac{S_i}{T},\ i = 1, 2, \ldots, n+1$. Now the hypothesis of the lemma states that

$$\mathbb{P}\{X \leq j\} = \tfrac{1}{T}\textstyle\sum_{i=1}^{j} T_i \geq \tfrac{1}{T}\sum_{i=1}^{j} S_i = \mathbb{P}\{X \leq j\} \text{ for } j = 1, 2, \ldots, n+1, \text{ i.e. } X \leq_{st} Y.$$

Let $c_{n+1} = 0$, then c is a nonincreasing function. Hence, it follows from Lemma 8.10 that $\mathbb{E}\{c_X\} \geq \mathbb{E}\{c_Y\}$, implying $\sum_{i=1}^{n} c_i T_i \geq \sum_{i=1}^{n} c_i S_i$. □

Theorem 8.18

If $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ and $\mu_1 r_1 \geq \mu_2 r_2 \geq \cdots \geq \mu_n r_n \geq 0$, then sequencing the tasks in the order $1, 2, \ldots, n$ maximizes the expected return by T for each $T > 0$.

Proof

Fix T and let T_j denote the expected total processing time of task j by time T. Now, because the policy that sequences according to $1, 2, \ldots, n$ stochastically maximizes the time until one of the processors becomes idle, it follows that it also stochastically maximizes the total processing time

by T. Because this remains true even when the set of tasks is $1, 2, \ldots, j$, it follows that $\sum_{i=1}^{j} T_i$ is, for each j maximized by the plicy under consoderation. The result follows now from Lemma 8.11 with $c_i = \mu_i r_i$ for all i. □

8.5.7 Tandem queues

Each of n jobs needs to be processed on two machines, say A and B. After receiving service on machine A, a job moves to machine B, and upon completion time of service at B it leaves the system. Let A_j and B_j be the service time of job j on machine A and B respectively. The objective is to determine the order in which tpo process jobs at michine A to minimize the expected time until all jobs have been processed on both machines. For the deterministic case, Johnson ([105]) shows that the makespan is minimized if jobs are arranged in the following transitive order on both machines:

$$\text{job } i \text{ precedes job } j \;\Leftrightarrow\; min\{A_i, B_j\} \leq min\{A_j, B_i\}.$$

Here we assume that A_j and B_j are exponentialle distributed with rates λ_j and μ_j respectively. Then,

$$\mathbb{E}\left\{min\{A_i, B_j\}\right\} = \tfrac{1}{\lambda_i+\mu_j} \text{ and } \mathbb{E}\left\{min\{A_j, B_i\}\right\} = \tfrac{1}{\lambda_j+\mu_i}.$$

Taking expectations on both sides of Johnson's rule one obtains the rule

$$\text{job } i \text{ precedes job } j \;\Leftrightarrow\; \lambda_i - \mu_i \geq \lambda_j - \mu_j.$$

We will show that Johnson's rule is also optimal for exponential processing times.
First, to gain some insight, let us consider the case in $n = 2$. If job 1 is processed first on machine A, then the expected completion time, denoted by $\mathbb{E}\{C_{1,2}\}$ is given by

$$\mathbb{E}\{C_{1,2}\} = \frac{1}{\lambda_1} + \frac{1}{\mu_1+\lambda_2} + \frac{\mu_1}{\mu_1+\lambda_2} \cdot \left\{\frac{1}{\lambda_2} + \frac{1}{\mu_1}\right\} + \frac{\lambda_2}{\mu_1+\lambda_2} \cdot \left\{\frac{1}{\mu_1} + \frac{1}{\mu_2}\right\}.$$

This follows because $\frac{1}{\lambda_1}$ is the expected time until job 1 is completed on machine A, at which time job 1 goes to machine B and job 2 goes to A. Then $\frac{1}{\mu_1+\lambda_2}$ is the expected time either job 2 is completed at A (with probability $\frac{\lambda_2}{\mu_1+\lambda_2}$) either job 1 is completed at B (with probability $\frac{\mu_1}{\mu_1+\lambda_2}$). The other two terms are then obtained by conditioning on whichever occuurs first.
Similarly, by reversing the order we have that $\mathbb{E}\{C_{2,1}\}$ is given by

$$\mathbb{E}\{C_{2,1}\} = \frac{1}{\lambda_2} + \frac{1}{\mu_2+\lambda_1} + \frac{\mu_2}{\mu_2+\lambda_1} \cdot \left\{\frac{1}{\lambda_1} + \frac{1}{\mu_2}\right\} + \frac{\lambda_1}{\mu_2+\lambda_1} \cdot \left\{\frac{1}{\mu_2} + \frac{1}{\mu_1}\right\}.$$

With some algebra, left to the reader (see Exercise 8.7) one can show that

$$\mathbb{E}\{C_{1,2}\} \leq \mathbb{E}\{C_{2,1}\} \;\Leftrightarrow\; \lambda_1 - \mu_1 \geq \lambda_2 - \mu_2,$$

i.e. Johnson's rule is true. We now show that this remains true when there are more than two jobs. With A_j the processing time for job j on machine A and C the time until all jobs have been processed on both machines, then $R := C - \sum_{j=1}^{n} A_j$ is the *remainder time*, that is, it represents the amount of work that remains at machine B when machine A has completed its processing.

Hence,

$$\mathbb{E}\{R\} = \mathbb{E}\{C\} - \textstyle\sum_{j=1}^{n} \frac{1}{\lambda_j},$$

so minimizing $\mathbb{E}\{C\}$ is equivalent to minimizing $\mathbb{E}\{R\}$. We shall prove that the policy that schedules jobs at machine A in decreasing order of $\lambda_j - \mu_j$ minimizes $\mathbb{E}\{R\}$. In fact, we shall use an interchange argument to show that this ordering stochastically minimizes R, and thus minimizes $\mathbb{E}\{R\}$.

Consider first the case $n = 2$, and suppose that, initially at time $t = 0$, machine B is occupied with the amount work w. That is, B must spend w units working on prior work before it can start processing either job 1 or job 2. Let $R_{1,2}(w)$ the remainder, i.e. $R_{1,2}(w) = C - A_1 - B_1$, when job 1 is scheduled first, and similarly for $R_{2,1}(w)$. The following lemma shows that the suggested ordering stochastically minimizes $R(w)$ for any w.

Lemma 8.12

If $\lambda_1 - \mu_1 \geq \lambda_2 - \mu_2$, then for any w, $R_{1,2}(w) \leq_{st} R_{2,1}(w)$.

Proof

We have to compare $\mathbb{P}\{R_{1,2}(w) > a\}$ with $\mathbb{P}\{R_{2,1}(w) > a\}$. When $w \geq A_1 + A_2$, then there probabilities are equal, because in either cases $R = w + B_1 + B_2 - A_1 - A_2$, with B_j the processing time for job j on machine B, $j = 1, 2$. Hence, we need only look at $\mathbb{P}\{R_{1,2}(w) > a \mid A_1 + A_2 > w\}$. Now, statingthat $A_1 + A_2 > w$ is equivalent to stating that at some time job 1 will be in machine B and job 2 in machine A. Hence, using the lack of memory of the exponential distribution, and conditioning on which machine finishes first, we see that

$$\begin{aligned}
\mathbb{P}\{R_{1,2}(w) > a \mid A_1 + A_2 > w\} &= \tfrac{\mu_1}{\mu_1+\lambda_2} e^{-\mu_2 a} + \tfrac{\lambda_2}{\mu_1+\lambda_2}\, \mathbb{P}\{e^{\mu_1} + e^{\mu_2} > a\} \\
&= \tfrac{\mu_1}{\mu_1+\lambda_2} e^{-\mu_2 a} + \tfrac{\lambda_2}{\mu_1+\lambda_2} \Big\{ e^{-\mu_1 a} + \textstyle\int_0^a \mu_1 e^{-\mu_1 x} e^{-\mu_2(a-x)}\, dx \Big\} \\
&= \tfrac{\mu_1}{\mu_1+\lambda_2} e^{-\mu_2 a} + \tfrac{\lambda_2}{\mu_1+\lambda_2} \Big\{ e^{-\mu_1 a} + \mu_1 e^{-\mu_2 a} \textstyle\int_0^a e^{-(\mu_1-\mu_2)x}\, dx \Big\} \\
&= \tfrac{\mu_1}{\mu_1+\lambda_2} e^{-\mu_2 a} + \tfrac{\lambda_2}{\mu_1+\lambda_2} \Big\{ e^{-\mu_1 a} + \tfrac{\mu_1}{\mu_1-\mu_2} e^{-\mu_2 a} \cdot \{1 - e^{-(\mu_1-\mu_2)a}\} \Big\} \\
&= \tfrac{\mu_1}{\mu_1+\lambda_2} e^{-\mu_2 a} + \tfrac{\lambda_2}{\mu_1+\lambda_2} \cdot \tfrac{1}{\mu_1-\mu_2} \Big\{ \mu_1 e^{-\mu_2 a} + \mu_2 e^{-\mu_1 a} \Big\} \\
&= \tfrac{\mu_1(\mu_1-\mu_2+\lambda_2) e^{-\mu_2 a} - \mu_2 \lambda_2 e^{-\mu_1 a}}{(\mu_1+\lambda_2)((\mu_1-\mu_2)}.
\end{aligned}$$

Because the expression $\mathbb{P}\{R_{2,1}(w) > a \mid A_1 + A_2 > w\}$ is similar, we have

$$\mathbb{P}\{R_{2,1}(w) > a \mid A_1 + A_2 > w\} = \tfrac{\mu_2(\mu_2-\mu_1+\lambda_1) e^{-\mu_1 a} - \mu_1 \lambda_1 e^{-\mu_2 a}}{(\mu_2+\lambda_1)((\mu_2-\mu_1)}.$$

Hence, we see that

$$\begin{aligned}
&\mathbb{P}\{R_{2,1}(w) > a \mid A_1 + A_2 > w\} - \mathbb{P}\{R_{1,2}(w) > a \mid A_1 + A_2 > w\} \\
&\qquad = \tfrac{\mu_1\mu_2}{(\mu_1+\lambda_2)(\mu_2+\lambda_1)} \cdot \tfrac{e^{-\mu_1 a} - e^{-\mu_2 a}}{\mu_2-\mu_2} \cdot \{(\lambda_1 - \mu_1) - (\lambda_2 - \mu_2)\} \;\geq\; 0,
\end{aligned}$$

which completes the proof of this lemma. □

Theorem 8.19

For any initial workload of machine B, R is stochastically minimized, and thus $\mathbb{E}\{C\}$ is minizized, by scheduling jobs to be processed on A in decreasing order of $\lambda_j - \mu_j$.

Proof

Consider first any of the $n!$ policies in which the ordering is fixed at time 0. Suppose that $\lambda_1 - \mu_1 = max_j \{\lambda_j - \mu_j\}$ and that the ordering calls for job j on A immediately before job 1. Then at the moment at which machine A is to begin on job j, no matter what the remaining work is at machine B at that moment, it follows from Lemma 8.12 that, if we interchange the jobs 1 and j, then the remaining work at machine B when both 1 and j have been processed at A will be stochastically reduced. But it is obvious that, for a given set of jobs to be processed in both machines, the remainder time is a stochastically increasing function of the initial workload of machine B.

Hence, the remainder time is stochastically reduced by the interchange. Repeated use of this interchange argument shows that the suggested policy stochastically minimizes the remainder time among all the $n!$ policies whose ordering is fixed at time 0. Hence, it minimizes the expexted completion time among all such policies.

To show that it is optimal amomg all policies follows by induction (it is immediate for $n = 1$). Assume it whenever there are $n - 1$ jobs to be processed on the two machines no matter the initial workload of machine B. Now no matter which job is initially processed at machine A, at the moment its processing at A is completed, it follows by the induction hypothesis that the remaining jobs are processed in decreasing order of the difference of their rates at machines A and B. Hence, we need only consider fixed-order policies, and thus this policy is optimal. □

8.6 Multi-armed bandit problems

8.6.1 Introduction

The multi-armed bandit problem was introduced in Section 1.3.9. The state space S is the Cartesian product $S = S_1 \times S_2 \times \cdots \times S_n$. Each state $i = (i_1, i_2, \ldots, i_n)$ has the same action set $A = \{1, 2, \ldots, n\}$, where action k means that project k is chosen, $k = 1, 2, \ldots, n$. So, at each stage one can be working on exactly one of the projects. When project k is chosen in state i - the chosen project is called the *active project* - the immediate reward and the transition probabilities only depend on the active project, whereas the states of the remaining projects are frozen. Let r_{i_k} and $p_{i_k j}$, $j \in S_k$ denote these quantities when action k is chosen. As a utility function the total discounted reward is chosen.

Example 8.1

Consider three sequences of nonnegative numbers $\{x_n^1,\ n = 1, 2, 3, \ldots\}$, $\{x_n^2,\ n = 1, 2, 3, \ldots\}$ and $\{x_n^3,\ n = 1, 2, 3, \ldots\}$. At each time one selects one of the sequences and x_n^k is the reward obtained the n-th time that sequence k is chosen. Denote by R_t the reward at time t. The problem is to

find the *optimal order* in which the sequences are chosen so as to maximize $R = \sum_{t=1}^{\infty} \alpha^{t-1} R_t$, where $\alpha \in (0,1)$ is a discount factor such that $\sum_{n=1}^{\infty} \alpha^{t-1} x_n^k < \infty$ for $k = 1,2,3$.
This is a deterministic version of the multi-armed bandit problem: $S = S_1 \times S_2 \times S_3$, where $S_i = \{0,1,2,\dots\}$. The state (i_1,i_2,i_3) means that sequence k was chosen i_k times, $k = 1,2,3$; $r_i(k) = x_{i+1}^k$ and $p_{ij}(k) = 1$ for $j = i+1$ (the other transition probabilities are 0).

Consider sequence k and assume that it has been selected $n_k - 1$ times, so that the next reward from this sequence is $x_{n_k}^k$. Define $G_k(n_k)$ by

$$G_k(n_k) = sup_{\tau \geq n_k} \frac{\sum_{t=n_k}^{\tau} \alpha^{t-1} x_t^k}{\sum_{t=n_k}^{\tau} \alpha^{t-1}}, \ k = 1,2,3. \tag{8.57}$$

The interpretation is that $G_k(n_k)$ is the *maximum discounted reward per unit of discounted time* that can be obtained from the remainder of sequence k. The numbers $G_k(n_k)$ are called the *Gittins indices*. We shall show that the policy that selects in state $(i_1 = n_1 - 1, i_2 = n_2 - 1, i_3 = n_3 - 1)$ the sequence with the largest of the indices $G_1(n_1)$, $G_2(n_2)$, $G_3(n_3)$ is optimal. Such policy is called an *index policy*. Notice that the calculation of the indices is done sequence by sequence. This result is a *decomposition* of the original problem.

8.6.2 A single project with a terminal reward

Consider the one-armed bandit problem with stopping option, i.e. in each state there are two options: action 1 is the stopping option and then one earns a terminal reward M and by action 2 the process continue with in state i an immediate reward r_i and transition probabilities p_{ij}. Let $v^{\alpha}(M)$ be the value vector of this optimal stopping problem. Then, $v^{\alpha}(M)$ is the unique solution of the optimality equation (cf. section 4.7.2)

$$v_i^{\alpha}(M) = max\{M, \ r_i + \alpha \sum_j p_{ij} v_j^{\alpha}(M)\}, \ i \in S. \tag{8.58}$$

and of the linear program

$$min \left\{ \sum_j v_j \ \middle| \ \begin{array}{rcll} \sum_j \{\delta_{ij} - \alpha p_{ij}\} v_j & \geq & r_i, & i \in S \\ v_i & \geq & M, & i \in S \end{array} \right\}. \tag{8.59}$$

Furthermore, we have shown in section 4.7.2 the following result.

Theorem 8.20
Let (x,y) be an extreme optimal solution of the dual program of (8.59), i.e.

$$max \left\{ \sum_j r_i x_i + M \cdot \sum_j y_i \ \middle| \ \begin{array}{rcll} \sum_i \{\delta_{ij} - \alpha p_{ij}\} x_i + y_j & = & 1, & i \in S \\ x_i, y_i & \geq & 0, & i \in S \end{array} \right\}. \tag{8.60}$$

Then, the policy f^{∞} such that $f(i) = \begin{cases} 2 & \text{if } x_i > 0 \\ 1 & \text{if } x_i = 0 \end{cases}$ is an optimal policy.

Lemma 8.13

$v_i^\alpha(M) - M$ is a nonnegative continous nonincreasing function in M, for all $i \in S$.

Proof

The nonnegativity of $v_i^\alpha(M) - M$ is directly from (8.58). By the method of value iteration $v^\alpha(M)$ can be approximated, arbitrary close, by

$$v_i^1(M) = M,\ i \in S;\ v_i^{n+1}(M) = max\{M,\ r_i + \alpha \sum_j p_{ij} v_j^n(M)\},\ i \in S,\ n = 1, 2, \ldots.$$

Hence, $v^\alpha(M) - M \cdot e$ can be approximated, arbitrary close, by

$$w_i^1(M) = 0,\ i \in S;\ w_i^{n+1}(M) = max\{0,\ r_i + \alpha \sum_j p_{ij} w_j^n(M) - (1-\alpha)M\},\ i \in S,\ n = 1, 2, \ldots.$$

By induction on n it is easy to see that $w_i^n(M)$ is continuous and nonincreasing in M. Hence, for all $i \in S$, the limit $v_i^\alpha(M) - M$ is also a continuous and nonincreasing function in M. □

Define the Gittins indices $G_i,\ i \in S$, by

$$G_i = min\{M \mid v_i^\alpha(M) = M\}. \tag{8.61}$$

Hence, $v_i^\alpha(G_i) = G_i$ and, by Lemma 8.13, $v_i^\alpha(M) = M$ for all $M \geq G_i$.

Theorem 8.21

For any M, the policy $f^\infty \in C(D)$ which chooses the stopping action in state i if and only if $M \geq G_i$ is optimal.

Proof

Take any M and let (x, y) be an extreme optimal solution of the dual linear program (8.60). From Theorem 8.20 we see that if $y_i = 0$ (and consequently $x_i > 0$), then the action 'continue' is optimal; when $y_i > 0$ (and consequently $x_i = 0$), then it is optimal to stop in state i.
First, let $M < G_i$, i.e. $v_i^\alpha(M) > M$. Then, by the complementary slackness property of linear programming $y_i = 0$ and it is optimal to continue in state i.
Next, let $M \geq G_i$, implying that $v_i^\alpha(M) = M$. Suppose that the stopping action is not optimal. Then, $v_i^\alpha(M) = r_i + \alpha \sum_j p_{ij} v_j^\alpha(M) > M$, which yields a contradiction. Therefore, it is optimal to stop in state i. □

For $M = G_i$ both actions (stop or continue) are optimal. Hence, an interpretion of the Gittins index G_i is that it is the terminal reward under which in state i both actions are optimal. Therefore, this number is also called the *indifference value*.

8.6.3 Multi-armed bandits

Consider the multi-armed bandit model with an additional option (action 0) in each state. Action 0 is a stopping option and then one earns a terminal reward M. For each state $i = (i_1, i_2, \ldots, i_n)$ we denote i-th component of the value vector by $v_i^\alpha(M)$.

Lemma 8.14

$v_i^\alpha(M)$ is a nondecreasing, convex function in M, for all $i \in S$.

Proof

Choose a fixed state i. It is obvious that $v_i^\alpha(M)$ is a nondecreasing function in M. Consider an arbitrary policy $f^\infty \in C(D)$, and let $\tau(f)$ be the corresponding stopping time, i.e. the stochastic number of steps before stopping. Then, one can write

$$\begin{aligned} v_i^\alpha(f^\infty, M) &= \mathbb{E}_{i,f}\left\{\text{discounted reward until time } \tau(f) + M \cdot \alpha^{\tau(f)}\right\} \\ &= \mathbb{E}_{i,f}\left\{\text{discounted reward until time } \tau(f)\right\} + M \cdot \mathbb{E}_{i,f}\left\{\alpha^{\tau(f)}\right\}. \end{aligned}$$

Hence,

$$v_i^\alpha(M) = max_{f^\infty \in C(D)}\left\{\mathbb{E}_{i,f}\left\{\text{discounted reward until time } \tau(f)\right\} + M \cdot \mathbb{E}_{i,f}\left\{\alpha^{\tau(f)}\right\}\right\}. \quad (8.62)$$

Since $v_i^\alpha(M)$ is the maximum of a finite number of terms, each of which is linear in M, $v_i^\alpha(M)$ is a convex function in M. □

Technical remark:

Since $v_i^\alpha(M)$ is the maximum of a finite number of linear functions, $\frac{\partial}{\partial M} v_i^\alpha(M)$ exists at almost all values of M.

Lemma 8.15

Let i be a fixed initial state and let $\tau(M)$ be the stopping time under the optimal policy $f^\infty(M)$, where M is the terminal reward. Then $\frac{\partial}{\partial M} v_i^\alpha(M) = \mathbb{E}_{i,f(M)}\{alpha^{\tau(M)}\}$.

Proof

Choose any $\varepsilon > 0$. If we employ $f^\infty(M)$ for a problem having terminal reward $M + \varepsilon$, we receive

$$\mathbb{E}_{i,f(M)}\left\{\text{discounted reward until time } \tau(M)\right\} + (M + \varepsilon) \cdot \mathbb{E}_{i,f(M)}\left\{\alpha^{\tau(M)}\right\}.$$

From (8.62) it follows that

$$\begin{aligned} v_i^\alpha(M + \varepsilon) &\geq \mathbb{E}_{i,f(M)}\left\{\text{discounted reward until time } \tau(M)\right\} + (M + \varepsilon) \cdot \mathbb{E}_{i,f(M)}\left\{\alpha^{\tau(M)}\right\} \\ &= v_i^\alpha(M) + \varepsilon \cdot \mathbb{E}_{i,f(M)}\left\{\alpha^{\tau(M)}\right\}. \end{aligned}$$

Similarly, we can derive that $v_i^\alpha(M - \varepsilon) \geq v_i^\alpha(M) - \varepsilon \cdot \mathbb{E}_{i,f(M)}\left\{\alpha^{\tau(M)}\right\}$.

Hence, we obtain $\frac{v_i^\alpha(M+\varepsilon) - v_i^\alpha(M)}{\varepsilon} \geq \mathbb{E}_{i,f(M)}\left\{\alpha^{\tau(M)}\right\} \geq \frac{v_i^\alpha(M) - v_i^\alpha(M-\varepsilon)}{\varepsilon}$, implying

$\frac{\partial}{\partial M} v_i^\alpha(M) = \mathbb{E}_{i,f(M)}\{a^{\tau(M)}\}$. □

Let $v_i^\alpha(M)$ denote the optimal value and let G_i be the indifference value when only a single project is available and its state is i. Now, consider the multiproject case and suppose the state is $i = (i_1, i_2, \ldots, i_j, \ldots, i_n)$ and let us speculate about whether or not we would ever again operate project j.

If $G_{i_j} > M$ then, because it would not be optimal to stop even if project j were the only project available, it is clear that we would never stop before operating project j. On the other hand, what if $G_{i_j} \leq M$? Would we ever want to operate project j under this circumstance? Whereas it is not obvious that we would never operate project j when $G_{i_j} \leq M$, it does seem somewhat intuitive, so let us accept this as a working hypothesis and see where it leads. That is, let us suppose that once a project reaches a state under which it would be optimal to stop if it were the sole project available, then the optimal policy never again operates that project. From this it follows that the optimal policy will stop in state i when $G_{i_j} \leq M$ for all $j = 1, 2, \ldots, n$.
Our speculations lead to the hypothesis that:
(1) project j would never be operated if state i is such that $G_{i_j} \leq M$;
(2) stop should occur if and only if $G_{i_j} \leq M$ for all $j = 1, 2, \ldots, n$.

For a given initial state $i = (i_1, i_2, \ldots, i_j, \ldots, i_n)$, let $\tau_j(M)$ denote the optimal time before we stop when only project j is available, $j = 1, 2, \ldots, n$. That is, $\tau_j(M)$ is the time project j has to be operated upon, when its initial state is i_j, until it reaches a state for which M is at least the indifference value. Also, let $\tau(M)$ denote the optimal stopping time for the multiproject case. Because the changes of state of individual projects are in no way affected by what occurs in other projects, it follows that, under our working hypothesis, $\tau(M) = \sum_{j=1}^{n} \tau_j(M)$. In addition, because $\tau_j(M),\ j = 1, 2, \ldots, n$ are independent random variables, we have

$$\mathbb{E}\{\alpha^{\tau(M)}\} = \mathbb{E}\{\alpha^{\sum_{j=1}^{n} \tau_j(M)}\} = \prod_{j=1}^{n} \mathbb{E}\{\alpha^{\tau_j(M)}\}.$$

Hence, we obtain by Lemma 8.15

$$\frac{\partial}{\partial M} v_i^{\alpha}(M) = \mathbb{E}_{i,f(M)}\{\alpha^{\tau(M)}\} = \prod_{j=1}^{n} \mathbb{E}_{i,f(M)}\{\alpha^{\tau_j(M)}\} = \prod_{j=1}^{n} \frac{\partial}{\partial M} v_{i_j}^{\alpha}(M). \tag{8.63}$$

Let $(1-\alpha)C$ be an upper bound of all one-period rewards. Then, C is an upper bound of the total discounted reward (without the terminal reward). Hence, if $M \geq C$, then the stopping action is optimal in all states, i.e. $v_i^{\alpha}(M) = M$ for $M \geq C$. Integrating (8.63) yields

$$\int_M^C \frac{\partial}{\partial m} v_i^{\alpha}(m) dm = \int_M^C \prod_{j=1}^{n} \frac{\partial}{\partial m} v_{i_j}^{\alpha}(m) dm.$$

or

$$v_i^{\alpha}(M) = C - \int_M^C \prod_{j=1}^{n} \frac{\partial}{\partial m} v_{i_j}^{\alpha}(m) dm. \tag{8.64}$$

We now prove that (8.64) is indeed valid by showing that $C - \int_M^C \prod_{j=1}^{n} \frac{\partial}{\partial m} v_{i_j}^{\alpha}(m) dm$ satisfies the optimality equation. Furthermore, the proof gives also the structure of the optimal policy.

Theorem 8.22

For any state $i = (1_1, i_2, \ldots, i_n)$ and any terminal reward M, we have

(1) $v_i^{\alpha}(M) = C - \int_M^C \prod_{j=1}^{n} \frac{\partial}{\partial m} v_{i_j}^{\alpha}(m) dm,\ M \leq C.$

(2) The optimal policy takes the stopping action if $M \geq M_{i_j}^{\alpha}$ for all $j = 1, 2, \ldots, n$ and continues with project k if $M_{i_k}^{\alpha} = max_j\, M_{i_j}^{\alpha} > M$.

Proof

Let $x_i(M) = C - \int_M^C \prod_{j=1}^n \frac{\partial}{\partial m} v_{ij}^\alpha(m) dm$. We shall show that $x_i(M)$ satisfies the optimality equation. Let $y_i(k, M) = \prod_{j \neq k} \frac{\partial}{\partial M} v_{ij}^\alpha(M)$. Because, from Lemma 8.14, $v_{ij}^\alpha(m)$ is a nondecreasing and convex function of m, it follows that $\frac{\partial}{\partial m} v_{ij}^\alpha(m)$ is a nonnegative (from nondecreasing) and nondecreasing (from convexity) function of m. Hence, $y_i(k, M)$ is also a nonnegative and nondecreasing function of M. Since $x_i(M)$ can be written as $x_i(M) = C - \int_M^C y_i(k, m) \frac{\partial}{\partial m} v_{i_k}^\alpha(m) dm$, we see, by integration by parts, that

$$x_i(M) = C - y_i(k, m) v_{i_k}^\alpha(m) \Big|_{m=M}^{m=C} + \int_M^C v_{i_k}^\alpha(m) dy_i(k, m).$$

Since $v_{ij}^\alpha(M) = M$ for $M \geq C$, we have $\frac{\partial}{\partial M} v_{ij}^\alpha(M) = 1$ for $M \geq C$. Therefore, $y_i(k, M) = 1$ for $M \geq C$. Hence, using that $y_i(k, C) = 1$ and $v_{i_k}^\alpha(C) = C$,

$$x_i(M) = y_i(k, M) v_{i_k}^\alpha(M) + \int_M^C v_{i_k}^\alpha(m) dy_i(k, m). \tag{8.65}$$

Similarly, we have

$$r_{i_k} + \alpha \textstyle\sum_{j \in S_k} p_{i_k j} x_{(i_1, i_2, \ldots, i_{s-1}, j, i_{s+1}, \ldots, i_n)}(M) \;=$$
$$r_{i_k} + \alpha \textstyle\sum_{j \in S_k} p_{i_k j} \big\{ y_i(k, M) v_j^\alpha(M) + \int_M^C v_j^\alpha(m) dy_i(k, m) \big\}.$$

Hence,

$$x_i(M) - \big\{ r_{i_k} + \alpha \textstyle\sum_{j \in S_k} p_{i_k j} x_{(i_1, i_2, \ldots, i_{s-1}, j, i_{s+1}, \ldots, i_n)}(M) \big\} \;=$$
$$y_i(k, M) v_{i_k}^\alpha(M) + \textstyle\int_M^C v_{i_k}^\alpha(m) dy_i(k, m) - \Big\{ r_{i_k} + \alpha \textstyle\sum_{j \in S_k} p_{i_k j} \big\{ y_i(k, M) v_j^\alpha(M) + \int_M^C v_j^\alpha(m) dy_i(k, m) \big\} \Big\}.$$

We can also write

$$r_{i_k} = r_{i_k} \{ y_i(k, M) + y_i(k, C) - y_i(k, M) \} = r_{i_k} y_i(k, M) + r_{i_k} \textstyle\int_M^C dy_i(k, m).$$

Therefore,

$$x_i(M) - \big\{ r_{i_k} + \alpha \textstyle\sum_{j \in S_k} p_{i_k j} x_{(i_1, i_2, \ldots, i_{s-1}, j, i_{s+1}, \ldots, i_n)}(M) \big\} \;=$$
$$y_i(k, M) \big\{ v_{i_k}^\alpha(M) - r_{i_k} - \alpha \sum_{j \in S_k} p_{i_k j} v_j^\alpha(M) \big\} + \int_M^C \big\{ v_{i_k}^\alpha(m) - r_{i_k} - \alpha \sum_{j \in S_k} p_{i_k j} v_j^\alpha(m) \big\} dy_i(k, m). \tag{8.66}$$

Since

$$v_{i_k}^\alpha(M) \geq r_{i_k} + \alpha \sum_{j \in S_k} p_{i_k j} v_j^\alpha(M) \text{ for all actions } k, \tag{8.67}$$

we obtain

$$x_i(M) \geq r_{i_k} + \alpha \sum_{j \in S_k} p_{i_k j} x_{(i_1, i_2, \ldots, i_{s-1}, j, i_{s+1}, \ldots, i_n)}(M) \text{ for all actions } k = 1, 2, \ldots, n. \tag{8.68}$$

Let us see under what conditions equality occurs in (8.68). First, note that (8.67) holds with equality if continuation is optimal when only project k is available, i.e. when $M \leq G_{i_k}$.
In that case we have from (8.66) that

$$x_i(M) - \{r_{i_k} + \alpha \sum_{j \in S_k} p_{i_k j} x_{(i_1,i_2,\ldots,i_{s-1},j,i_{s+1},\ldots,i_n)}(M)\} =$$

$$\int_{G_{i_k}}^{C} \{v_{i_k}^{\alpha}(m) - r_{i_k} - \alpha \sum_{j \in S_k} p_{i_k j} v_j^{\alpha}(m)\} dy_i(k,m) \text{ if } M \le G_{i_k}. \tag{8.69}$$

Since $v_{i_j}^{\alpha}(m) = m$ for $m \ge G_{i_j}$, we have $y_i(k,m) = \prod_{j \ne k} \frac{\partial}{\partial m} v_{i_j}^{\alpha}(m) = 1$ for $m \ge max_{j \ne k} G_{i_j}$.

So, $dy_i(k,m) = 0$ for $m \ge max_{j \ne k} G_{i_j}$.

Hence, using this we see from (8.69) that for $M \le G_{i_k} = max_j G_{i_j}$, we obtain

$$x_i(M) = r_{i_k} + \alpha \sum_{j \in S_k} p_{i_k j} x_{(i_1,i_2,\ldots,i_{s-1},j,i_{s+1},\ldots,i_n)}(M). \tag{8.70}$$

For $M \ge max_j G_{i_j}$, we have $v_{i_j}^{\alpha}(m) = m$ for $j = 1, 2, \ldots, n$, and thus $y_i(k,m) = 1$, implying that $dy_i(k,m) = 0$ for $m \ge M$. Hence, from (8.65) we see that

$$x_i(M) = v_{i_k}^{\alpha}(M) = M \text{ if } M \ge max_j G_{i_j}. \tag{8.71}$$

In addition, also using (8.65) and the monotonicity of $v_k(m)$ in m, we have for all M

$$x_i(M) \ge y_i(k,M) v_{i_k}^{\alpha}(M) + v_{i_k}^{\alpha}(M)\{y_i(k,C) - y_i(k,M)\} = v_{i_k}^{\alpha}(M) \ge M. \tag{8.72}$$

Hence, we have from (8.68) and (8.72)

$$\begin{cases} x_i(M) & \ge & r_{i_k} + \alpha \sum_{j \in S_k} p_{i_k j} x_{(i_1,i_2,\ldots,i_{s-1},j,i_{s+1},\ldots,i_n)}(M), \ i \in S, \ k = 1, 2, \ldots, n \\ x_i(M) & \ge & M, \ i \in S \end{cases}$$

Furthermore, we have from (8.70) and (8.71)

$$\begin{cases} x_i(M) & = & r_{i_k} + \alpha \sum_{j \in S_k} p_{i_k j} x_{(i_1,i_2,\ldots,i_{s-1},j,i_{s+1},\ldots,i_n)}(M), \ i \in S & \text{if } M \le max_j G_{i_j} = G_{i_k} \\ x_i(M) & = & M, \ i \in S & \text{if } M \ge max_j G_{i_j} = G_{i_k} \end{cases}$$

Hence, $x(M)$ satisfies the optimality equation

$$x_i(M) = max\{M, max_{1 \le k \le n}\{r_{i_k} + \alpha \sum_{j \in S_k} p_{i_k j} x_{(i_1,i_2,\ldots,i_{s-1},j,i_{s+1},\ldots,i_n)}(M)\}\}, \ i \in S, \tag{8.73}$$

and the optimal policy is as stated. □

Remark

The preceeding theorem shows that the optimal policy in the multi-project case can be determined by an analysis of the n single-project problems, with the optimal decision in state $i = (i_1, i_2, \ldots, i_n)$ being to operate on that project k having the largest indifference value G_{i_k} if this value is greater than M and to stop otherwise.

Alternative interpretation of the Gittins index

Consider the one-armed bandit problem having initial state i. Because when $M = G_i$ the optimal policy is indifferent between stopping and continuing, so that for any stopping random stopping time τ, $G_i \geq \mathbb{E}\{\text{discounted reward before } \tau\} + G_i \cdot \mathbb{E}\{\alpha^\tau\}$, with equality for the optimal policy. Hence,

$$\begin{aligned}(1-\alpha)G_i &= max_{\tau \geq 1} \frac{\mathbb{E}\{\text{discounted reward before } \tau\}}{\{1-\mathbb{E}\{\alpha^\tau\}\}/(1-\alpha)} \\ &= max_{\tau \geq 1} \frac{\mathbb{E}\{\text{discounted reward before } \tau\}}{\mathbb{E}\{1+\alpha+\cdots+\alpha^{\tau-1}\}} \\ &= max_{\tau \geq 1} \frac{\mathbb{E}\{\text{discounted reward before } \tau\}}{\mathbb{E}\{\text{discounted time before } \tau\}},\end{aligned}$$

where the expextations are conditional on the initial state i. Hence, another way of describing the optimal policy in the multi-armed bandit problem is as follows. For each individual project look for the stopping time τ whose ratio of expected discounted reward and expected discounted time prior to τ is maximal. Then work on the project with the largest ratio. In the case there also is the additional option of stopping, one should stop if all ratios are smaller than $(1-\alpha)M$.

8.6.4 Methods for the computation of the Gittins indices

1. The parametric linear programming method

We have already seen that for a single project with terminal reward M the solution can be obtained from a linear programming problem, namely program (8.60). For M big enough, e.g. for $M \geq C = (1-\alpha) \cdot max_i\, r_i$, we know that $v_i(M) = M$ for all states i. Furthermore, we have seen that the Gittins index $G_i = min\{M|\ v_i(M) = M\}$ (cf. (8.61)).

One can solve program (8.60) as a parametric linear programming problem with parameter M. Starting with $M = C$ one can decrease M and find for each state i the largest M for which it is optimal to keep working on the project, which is in fact $min\{M|\ v_i(M) = M\} = G_i$, in the order of decreasing M-values. One can start with the simplex tableau in which all y-variables are in the basis and in which the x-variables are the nonbasic variables. This tableau is optimal for $M \geq C$. Decrease M until we meet a basis change, say the basic variable y_i will be exchanged with the nonbasic variable x_i. Then, we know the M-value which is equal to G_i. In this way we continue and repeat the procedure N times, where N is the number of states in the current project. The used pivoting row and column do not influence any further pivoting step, so we can delete these row and column from the simplex tableau.

Example 8.2

Consider a project with the following data.

$S = \{1, 2, 3\};\ \alpha = \frac{1}{2};\ r_1 = 8,\ r_2 = 6,\ r_3 = 4.$

$p_{11} = 0,\ p_{12} = 1,\ p_{13} = 0;\ p_{21} = 0,\ p_{22} = 0,\ p_{23} = 1;\ p_{31} = 1,\ p_{32} = 0,\ p_{33} = 0.$

The linear program becomes (the objective function is splitted up into two rows, one for the x-part and one for parametric y-part; the y-variables have to be expressed in the nonbasic x-variables, i.e. we obtain for the last row $y_1 + y_2 + y_3 = 3 - \frac{1}{2}x_1 - \frac{1}{2}x_2 - \frac{1}{2}x_3$).

$$max\{8x_1 + 6x_2 + 4x_3 + My_1 + My_2 + My_3\}$$

subject to

$$\begin{array}{rcrcrcrcrcl}
x_1 & & & - & \frac{1}{2}x_3 & + & y_1 & & & = & 1;\ x_1, y_1 \geq 0; \\
-\frac{1}{2}x_1 & + & x_2 & & & & & + & y_2 & = & 1;\ x_2, y_2 \geq 0; \\
& - & \frac{1}{2}x_2 & + & x_3 & & & & y_3 & = & 1;\ x_3, y_3 \geq 0.
\end{array}$$

The first tableau becomes:

		x_1	x_2	x_3
y_1	1	*1	0	$-\frac{1}{2}$
y_2	1	$-\frac{1}{2}$	1	0
y_3	1	0	$-\frac{1}{2}$	1
x_0	0	-8	-6	-4
M	3	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$

The objective function is $3M+(8-\frac{1}{2}M)x_1+(6-\frac{1}{2}M)x_2+(4-\frac{1}{2}M)x_3$.
Hence, this tableau is optimal for $x_1 = x_2 = x_3 = 0$ if $8 - \frac{1}{2}M \leq 0$, $6 - \frac{1}{2}M \leq 0$ and $4 - \frac{1}{2}M \leq 0$, i.e. if $M \geq 16$.
For $M = 16$ there is indifference in state 1: $G_1 = 16$.
Then, we exchange x_1 and y_1 and obtain a new simplex tableau in which the row of y_1 and the column of x_1 can be deleted.
The second tableau is:

		x_2	x_3
y_2	$\frac{3}{2}$	*1	$-\frac{1}{4}$
y_3	1	$-\frac{1}{2}$	1
x_0	8	-6	-8
M	$\frac{5}{2}$	$\frac{1}{2}$	$\frac{3}{4}$

This tableau is optimal for $x_2 = x_3 = 0$ if $6 - \frac{1}{2}M \leq 0$ and $8 - \frac{3}{4}M \leq 0$, i.e. if $M \geq 12$.
For $M = 12$ there is indifference in state 2: $G_2 = 12$.
Then, we exchange x_2 and y_2 and obtain a new simplex tableau in which the row of y_2 and the column of x_2 can be deleted.
The final tableau is:

		x_3
y_3	$\frac{7}{4}$	$\frac{7}{8}$
x_0	17	$-\frac{19}{2}$
M	$\frac{7}{4}$	$\frac{7}{8}$

This tableau is optimal for $x_3 = 0$ if $\frac{19}{2} - \frac{7}{8}M \leq 0$, i.e. if $M \geq \frac{76}{7}$. Hence, $G_3 = \frac{76}{7}$.

Computational complexity

We can easily determine the computational complexity. Each update of an element in a simplex tableau needs at most two arithmetic operations (multiplication and divisions as well as additions and subtractions): for instance, the value 17 in the last tableau of Example 8.2 is computed by $8 - (-6) \cdot \frac{3}{2} = 17$. Hence, the total number of arithmetic operations in this method is at most $2 \cdot \sum_{k=1}^{N} k^2 = \frac{1}{3}N(N+1)(2N+1) = \frac{2}{3}N^3 + \mathcal{O}(N^2) = \mathcal{O}(N^3)$.

2. The restart-in-k method

We will derive another interpretation for the Gittins index G_k in a fixed state k. The optimality equation for a single project with terminal reward M is, cf. (8.58),

$$v_i^\alpha(M) = max\big\{M,\, r_i + \alpha \sum_j p_{ij} v_j^\alpha(M)\big\},\ i \in S. \tag{8.74}$$

We have seen that G_k is the indifference value, i.e. for $M = G_k$ we have

$$v_k^\alpha(G_k) = G_k = r_k + \alpha \sum_j p_{kj} v_j^\alpha(G_k). \tag{8.75}$$

Using (8.74) and (8.75) yields

$$v_i^\alpha(G_k) = max\big\{r_k + \alpha \sum_j p_{kj} v_j^\alpha(G_k) = G_k,\, r_i + \alpha \sum_j p_{ij} v_j^\alpha(M_k)\big\},\ i \in S. \tag{8.76}$$

With the abbreviation $v_i^k = v_i^\alpha(G_k),\ i \in S$, we get the following expression

$$v_i^k = max\big\{r_i + \alpha \sum_j p_{ij} v_j^k,\, r_k + \alpha \sum_j p_{kj} v_j^k\big\},\ i \in S. \tag{8.77}$$

Hence, G_k is the k-th component of the value vector of the MDP where there are in each state two actions. This problem can be interpreted as follows: in each state there are two options, either to continue working on the project in the given state i, or to restart working in state k, where the total expected discounted reward must be maximized. This gives the problem the name *restart-in-k* problem. By solving this MDP we find $G_k = v_k^k$. Notice that we now have a characterization of the Gittins index without using a terminal reward.

We define C_k for the restart-in-k problem as the set of states i for which it is optimal to continue in that state. If the MDP is solved we find G_k and $v_i^\alpha(G_k),\ i \in S$. The next theorem shows that C_k contains exactly those states j for which $G_j \geq G_k$. When we are in state $(i_1, i_2, \ldots, i_n)$ and decide to work on project k because $G_{i_k} \geq G_{i_l}$ for all $1 \leq l \leq n$, and when we also move in project k from state i_k to a state $j \in C_k$, then we know that the largest Gittins index is still the index of the state j of project k, without knowing the value of G_j. So, the theorem tell us that we only have to calculate a new index when we enter a state which is not in C_k.

Theorem 8.23
Let $C_k = \{j \mid$ for the restart-in-k problem it is in state j is optimal to continue$\}$.
Then, $C_k = \{j \mid G_j \geq G_k\}$.

Proof
$j \notin C_k$ if and only if it is not optimal to continue in state j for the restart-in k problem or, equivalently, for the optimal stopping problem with terminal reward $M = G_k$. Since G_j is the indifference value in state j, it is not optimal to continue in state j if and only if $M > G_j$. Therefore, $G_k > G_j$. So, $j \notin C_k \ \Leftrightarrow\ G_k > G_j$, i.e. $C_k = \{j \mid G_j \geq G_k\}$. □

We can solve the restart-in-k problem by any method for discounted MDPs. If we use the linear programming method the program becomes

$$min \left\{ \sum_j v_j \;\middle|\; \begin{array}{l} \sum_j \{\delta_{ij} - \alpha p_{ij}\} v_j \geq r_i, \;\; i \in S \\ \sum_j \{\delta_{ij} - \alpha p_{kj}\} v_j \geq r_k, \;\; i \in S, \;\; i \neq k \end{array} \right\}. \tag{8.78}$$

Example 8.2 (continued)

The linear program for G_1 is:

$$min \left\{ v_1 + v_2 + v_3 \;\middle|\; \begin{array}{lll} v_1 \geq 8 + \frac{1}{2}v_2; & v_2 \geq 6 + \frac{1}{2}v_3; & v_3 \geq 4 + \frac{1}{2}v_1 \\ v_2 \geq 8 + \frac{1}{2}v_2; & v_3 \geq 8 + \frac{1}{2}v_2 & \end{array} \right\}.$$

The optimal solution is: $v_1 = v_2 = v_3 = 16 \;\rightarrow\; G_1 = v_1 = 16;\; C_1 = \{1\}$.

The linear program for G_2 is:

$$min \left\{ v_1 + v_2 + v_3 \;\middle|\; \begin{array}{lll} v_1 \geq 8 + \frac{1}{2}v_2; & v_2 \geq 6 + \frac{1}{2}v_3; & v_3 \geq 4 + \frac{1}{2}v_1 \\ v_1 \geq 6 + \frac{1}{2}v_3; & v_3 \geq 6 + \frac{1}{2}v_3 & \end{array} \right\}.$$

The optimal solution is: $v_1 = 14;\; v_2 = v_3 = 12 \;\rightarrow\; G_2 = v_2 = 12;\; C_2 = \{1, 2\}$.

The linear program for G_3 is:

$$min \left\{ v_1 + v_2 + v_3 \;\middle|\; \begin{array}{lll} v_1 \geq 8 + \frac{1}{2}v_2; & v_2 \geq 6 + \frac{1}{2}v_3; & v_3 \geq 4 + \frac{1}{2}v_1 \\ v_1 \geq 4 + \frac{1}{2}v_1; & v_2 \geq 4 + \frac{1}{2}v_1 & \end{array} \right\}.$$

The optimal solution is: $v_1 = \frac{96}{7};\; v_2 = \frac{80}{7};\; v_3 = \frac{76}{7};\; \rightarrow\; G_3 = v_3 = \frac{76}{7};\; C_3 = \{1, 2, 3\}$.

Computation on-line

It is interesting to ask what indices must be computed and when this must be done. In the first period, it is necessary to compute the n initial indices, one for each project. Subsequently, it suffices to compute at most *one* index in each period. In particular, one computes the index of a project k in a period only when its state i_k leaves the optimal continuation set C_{i_k}. Thus, by Theorem 8.23, if the indices are computed on-line only as needed, the indices computed for each project will decrease strictly over time.

An alternative linear program

Since the v-variables are unrestricted in sign, one may substitute v_j by $y_j + z$, where z is unrestricted and $y_j \geq 0$ for all j. Then, program (8.23) can be written as

$$min \left\{ \sum_j y_j + N \cdot z \;\middle|\; \begin{array}{l} (1-\alpha)z + \sum_j \{\delta_{ij} - \alpha p_{ij}\} y_j \geq r_i, \;\; i \in S, \;\; i \neq k \\ (1-\alpha)z + \sum_j \{\delta_{ij} - \alpha p_{kj}\} y_j \geq r_k, \;\; i \in S \\ z \text{ unrestricted}, \;\; y_j \geq 0, j \in S \end{array} \right\}. \tag{8.79}$$

Consider the second part of the constraints: $(1-\alpha)z + y_i \geq r_k + \alpha \sum_j p_{kj}\} v_j \geq r_k, \;\; i \in S$, which is equivalent with $(1-\alpha)z + min_i\, y_i \geq r_k + \alpha \sum_j p_{kj}\} v_j \geq r_k$. If the y_j becomes ε smaller for

each j and z becomes ε bigger, then the objective function keeps its value and the constraints remain satisfied. So we can take $min_i\, y_i = 0$ and the linear program becomes

$$min \left\{ \sum_j y_j + N \cdot z \;\middle|\; \begin{array}{lll} (1-\alpha)z + \sum_j \{\delta_{ij} - \alpha p_{ij}\} y_j & \geq & r_i,\ i \in S,\ i \neq k \\ (1-\alpha)z - \alpha \sum_j p_{kj} y_j & \geq & r_k,\ i \in S \\ z \text{ unrestricted},\ y_j \geq 0, j \in S & & \end{array} \right\}. \tag{8.80}$$

For the optimal solution v^* of (8.23) we have $v_i^* \geq r_k + \alpha \sum_j p_{kj} v_j^* = v_k^*,\ i \in S$. Hence, $y_k^* = min_i\, y_i^* = 0$, and consequently, $G_k = v_k^* = min_i\, y_i^* + z^* = z^*$, where (y^*, z^*) is the optimal solution of program (8.79).

Example 8.2 (continued)

The linear program for G_1 is:

$$min\{y_1 + y_2 + y_3 + 3z \mid \tfrac{1}{2}z + y_2 \geq 6 + \tfrac{1}{2}y_3;\ \tfrac{1}{2}z + y_3 \geq 4 + \tfrac{1}{2}y_1;\ \tfrac{1}{2}z \geq 8 + \tfrac{1}{2}y_2;\ y_1, y_2, y_3 \geq 0\}.$$

The optimal solution is: $y_1 = y_2 = y_3 = 0;\ z = 16 \ \rightarrow\ G_1 = z = 16.$

The linear program for G_2 is:

$$min\{y_1 + y_2 + y_3 + 3z \mid \tfrac{1}{2}z + y_1 \geq 8 + \tfrac{1}{2}y_2;\ \tfrac{1}{2}z + y_3 \geq 4 + \tfrac{1}{2}y_1;\ \tfrac{1}{2}z \geq 6 + \tfrac{1}{2}y_3;\ y_1, y_2, y_3 \geq 0\}.$$

The optimal solution is: $y_1 = 2,\ y_2 = y_3 = 0;\ z = 12 \ \rightarrow\ G_2 = z = 12.$

The linear program for G_2 is:

$$min\{y_1 + y_2 + y_3 + 3z \mid \tfrac{1}{2}z + y_1 \geq 8 + \tfrac{1}{2}y_2;\ \tfrac{1}{2}z + y_2 \geq 6 + \tfrac{1}{2}y_3;\ \tfrac{1}{2}z \geq 4 + \tfrac{1}{2}y_1;\ y_1, y_2, y_3 \geq 0\}.$$

The optimal solution is: $y_1 = \frac{20}{7},\ y_2 = \frac{4}{7},\ y_3 = 0;\ z = \frac{76}{7} \ \rightarrow\ G_3 = z = \frac{76}{7}.$

3. The largest-remaining-index method

Theorem 8.24

Suppose that $G_1 \geq G_2 \geq \cdots \geq G_k$ for some k, and $G_k \geq G_i$ for $i = k+1, k+2, \ldots, n$. Let l_k be such that M_{l_k} be such that $G_{l_k} = max_{i>k}\, G_i$. Then, we have

$$(1-\alpha)G_{l_k} = max_{i>k} \frac{\{\big(I - \alpha P(k)\big)^{-1} r\}_i}{\{\big(I - \alpha P(k)\big)^{-1} e\}_i},\ \textit{where}\ \{P(k)\}_{ij} = \begin{cases} p_{ij} & ,\ j \leq k; \\ 0 & ,\ j > k. \end{cases}$$

Proof

Since $v_i^\alpha(G_{l_k}) \geq r_i + \alpha \sum_j p_{ij} v_j^\alpha(G_{l_k})$ and $v_i^\alpha(M) = M$ for $M \geq G_i,\ i \in S$, we can write

$$\begin{aligned} v_i^\alpha(G_{l_k}) &\geq r_i + \alpha \textstyle\sum_{j \leq k} p_{ij} v_j^\alpha(G_{l_k}) + \alpha \sum_{j>k} p_{ij} v_j^\alpha(G_{l_k}) \\ &= r_i + \alpha \textstyle\sum_{j \leq k} p_{ij} v_j^\alpha(G_{l_k}) + \alpha G_{l_k}\{1 - \sum_{j \leq k} p_{ij}\}. \end{aligned}$$

In vector notation, with $v = v^\alpha(G_{l_k})$, this becomes

$$\begin{aligned} v &\geq r + \alpha P(k) v + \alpha G_{l_k} e - \alpha G_{l_k} P(k) e \\ &= r + \alpha P(k) v - (1-\alpha) G_{l_k} e + G_{l_k}\{I - \alpha P(k)\} e. \end{aligned}$$

So,

$$\{I - \alpha P(k)\} v \geq r - (1-\alpha) G_{l_k} e + G_{l_k}\{I - \alpha P(k)\} e.$$

Since $\{I - \alpha P(k)\}$ is nonsingular and $\{I - \alpha P(k)\}^{-1} \geq 0$, we can write

$$v \geq \{I - \alpha P(k)\}^{-1} r - (1-\alpha) G_{l_k} \{I - \alpha P(k)\}^{-1} e + G_{l_k} e.$$

Componentwise, for all $i \geq k$,

$$G_{l_k} = v_i^{\alpha}(G_{l_k}) \geq \{\big(I - \alpha P(k)\big)^{-1} r\}_i - (1-\alpha) G_{l_k} \{\big(I - \alpha P(k)\big)^{-1} e\}_i + G_{l_k},$$

with equality for $i = k$. From this it follows that $(1-\alpha) G_{l_k} \geq \frac{\{\big(I-\alpha P(k)\big)^{-1} r\}_i}{\{\big(I-\alpha P(k)\big)^{-1} e\}_i}$ for all $i \geq k$ with equality for $i = k$. Therefore,

$$(1-\alpha) G_{l_k} = max_{i>k} \frac{\{\big(I-\alpha P(k)\big)^{-1} r\}_i}{\{\big(I-\alpha P(k)\big)^{-1} e\}_i}. \qquad \square$$

To compute G_{l_k}, we have to invert the matrix $\{I - \alpha P(k)\}$, which can be written as $\begin{pmatrix} A_k & 0 \\ B_k & I \end{pmatrix}$. It can now easily be checked that $\{I - \alpha P(k)\}^{-1} = \begin{pmatrix} A_k^{-1} & 0 \\ -B_k A_k^{-1} & I \end{pmatrix}$. The inversion of a matrix of order k can be done in $\mathcal{O}(k^3)$ steps. Hence the computation of the Gittins indices of a project with N states has complexity $\mathcal{O}(N^4)$. Fortunalety, since subsequent matrices $P(k)$ are similar, this can be done efficiently in a recursive way. In this way time can be saved and the computation can be done in $\mathcal{O}(N^3)$ steps, as we will see.

Write $A_{k+1} = \begin{pmatrix} A_k & p \\ q & x \end{pmatrix}$, so $A_{k+1}^{-1} = \begin{pmatrix} N & t \\ s & y \end{pmatrix}$, where $p = (p_{1,k+1}, p_{2,k+1}, \ldots, p_{k,k+1})^T$, $q = (p_{k+1,1}, p_{k+1}, 2, \ldots, p_{k+1,k})^T$ and $x = p_{k+1,k+1}$. Since $A_{k+1} A_{k+1}^{-1} = I$, we get

$$A_k N + ps = I; \tag{8.81}$$

$$qN + xs = 0; \tag{8.82}$$

$$A_k t + py = 0; \tag{8.83}$$

$$qt + xy = 1. \tag{8.84}$$

From (8.83) and (8.84), we obtain

$$t = -y A_k^{-1} p, \; -yqA_k^{-1} p + xy = 1 \;\rightarrow\; y = \frac{1}{x - qA_k^{-1} p}, \tag{8.85}$$

and from (8.81)

$$N = A_k^{-1}(I - ps) = A_k^{-1} - A_k^{-1} ps. \tag{8.86}$$

Insertion into (8.82) gives $0 = qA_k^{-1} - \{qA_k^{-1} p + x\} s = qA_k^{-1} + \frac{1}{y} s \;\rightarrow\; s = -yqA_k^{-1}$, which with (8.85) and (8.86) yields $N = A_k^{-1} + \frac{1}{y} ts$. Therefore, we have shown that

$$A_{k+1}^{-1} = \begin{pmatrix} A_k^{-1} + \frac{1}{y} ts & t \\ s & y \end{pmatrix}, \text{ where } y = \tfrac{1}{x - qA_k^{-1} p}, \; t = -y A_k^{-1} p \text{ and } s = -yqA_k^{-1}.$$

All these calculations can be done in $\mathcal{O}(k^2)$ steps, because at most a vector of k components and a $k \times k$-matrix have to be multiplied. The calculation of the matrix $B_{k+1}A_{k+1}^{-1}$ costs using the standard method $\mathcal{O}(k^3)$ steps, but on this number can also be saved if $B_kA_k^{-1}$ is known.

Write $B_k = \begin{pmatrix} f^T \\ F_k \end{pmatrix}$ and $B_{k+1} = \begin{pmatrix} F_k & g \end{pmatrix}$, where f^T is the top row of B_k. Then, we obtain

$$B_kA_k^{-1} = \begin{pmatrix} f^TA_k^{-1} \\ F_kA_k^{-1} \end{pmatrix} \text{ and } B_{k+1}A_{k+1}^{-1} = \begin{pmatrix} F_k & g \end{pmatrix} \begin{pmatrix} A_k^{-1} + \frac{1}{y}ts & t \\ s & y \end{pmatrix} = \begin{pmatrix} F_kA_k^{-1} + \frac{1}{y}F_kts + gs & F_kt + gy \end{pmatrix}.$$

Because $B_kA_k^{-1}$ is known the matrix $B_{k+1}A_{k+1}^{-1}$ can also be calculated in $\mathcal{O}(k^2)$ steps, so the complexity of this method for the computation of the Gittins indices of one project with N states is $\sum_{k=1}^{N} \mathcal{O}(k^2) = \mathcal{O}(N^3)$.

Example 8.2 (continued)

For the largest Gittins index we have: $(1-\alpha)G_{l_0} = max_i\ r_i = 8$ for $i = 1 \ \rightarrow \ G_{l_0} = G_1 = 16$.

Since the transition matrix $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ and the first index is G_1, we have for $P(1)$ the matrix $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$. Hence, $I - \alpha P(1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}$ and $\{I - \alpha P(1)\}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}$.

Therefore, $\{I - \alpha P(1)\}^{-1}r = (8, 6, 8)$ and $\{I - \alpha P(1)\}^{-1}e = (1, 1, \frac{3}{2})$.

Hence, $(1-\alpha)G_{l_1} = max\left\{\frac{6}{1}, \frac{8}{3/2}\right\} = 6$ for $i = 2 \ \rightarrow \ G_{l_1} = G_2 = 12$.

Since $G_1 \geq G_2$ are the two largest Gittins indices, we have $P(2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$.

Hence, $I - \alpha P(2) = \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}$ and $\{I - \alpha P(2)\}^{-1} = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{pmatrix}$.

Therefore, $\{I - \alpha P(2)\}^{-1}r = (11, 6, \frac{19}{2})$ and $\{I - \alpha P(2)\}^{-1}e = (\frac{3}{2}, 1, \frac{7}{4})$.

Hence, $(1-\alpha)G_{l_2} = \frac{19/2}{7/4} = \frac{38}{7} \ \rightarrow \ G_{l_2} = G_3 = \frac{76}{7}$.

8.7 Separable problems

8.7.1 Introduction

Separable MDPs have the property that for certain pairs $(i, a) \in S \times A$:

(1) the immediate reward is the sum of tow terms, one depends only on the current state and the other depends only on the chosen action: $r_i(a) = s_i + t_a$.

(2) the transition probabilities depend only on the action and not on the state from which the transition occurs: $p_{ij}(a) = p_j(a),\ j \in S$.

Let $S_1 \times A_1$ be the subset of $S \times A$ for which the pairs (i, a) satisfy (1) and (2). We also assume that the action sets of A_1 are *nested*: let $S_1 = \{1, 2, \ldots, m\}$, then $A_1(1) \supseteq A_1(2) \supseteq \cdots \supseteq A_1(m) \neq \emptyset$. Let $S_2 = S \backslash S_1,\ A_2(i) = A(i) \backslash A_1(i),\ 1 \leq i \leq m$ and $A_2(i) = A(i),\ m+1 \leq i \leq N$. We also introduce the notation $B(i) = A_1(i) - A_{i+1}(i),\ 1 \leq i \leq m-1$ and $B(m) = A_1(m)$. Then, $A_1(i) = \bigcup_{j=i}^{m} B(j)$ and the sets $B(j)$ are disjunct. We allow that S_2, A_2 or $B(i)$ is an empty set.

If the system is observed in state $i \in S_1$ and the decision maker will choose an action from $A_1(i)$, then, the decision process can be considered as follows. First, a reward s_i is earned and the system makes a zero-time transition to an additional state $N + i$. In this additional state there are two options: either to take an action $a \in B(i)$ or to take an action $a \in A_1(i) \backslash B(i) = A_1(i+1)$. In the first case the reward t_a is earned and the process moves to state j with probability $p_j(a),\ j \in S$; in the second case we are in the same situation as in state $N + i + 1$, i.e. a zero-time transition is made from state $N + i$ to state $N + i + 1$.

8.7.2 Examples (part 1)

1. Automobile replacement problem

We own a car of a certain age. Our decision problem is to keep it or sell it and, if we sell, what age car to replace it with. Let us agree to review the number of states down, we assume that every car breaks down irreparably as soon as it becomes 10 years old. Number the states from 0 to 40. For $i \leq 39$, state i refers to a car that is $3i$ months old: we say this car is "of age i". State 40 indicates a car that has just become 10 years and, therefore, has just broken down irreparably. At state $i \leq 39$ we can make decision r (for retain) to keep our current car for at least one more time period, or we can trade it in on a car of age k with $0 \leq k \leq 39$. Since there are 41 possible decisions for states 0 through 39, namely $A(i) = \{r, 0, 1, \ldots, 39\}$, and 40 possible decisions for state 40, namely $A(40) = \{0, 1, \ldots, 39\}$, there are nearly 41^{41} different policies.

Consider the following relevant data:

c_j = the cost of buying a car of age $j,\ 0 \leq j \leq 39$;
t_i = the trade-in value of a car of age $i,\ 0 \leq i \leq 40$;
e_i = the expected cost of operating a car of age i for one time period, $0 \leq i \leq 39$;
p_i = the probability that a car of age i will last at least one more time period, $0 \leq i \leq 39$;

To simplify things, assume that if a car fails to survive a time period, it ages at the end of that period to 10 years, causing a transfer to state 40. Trade-ins take place at the beginning of time periods. If the retain decision r is selected in state i, the rewards and transition probabilities are:

$$r_i(r) = -e_i;\ p_{ij}(r) = \begin{cases} p_i & \text{if } j = i+1; \\ 1 - p_i & \text{if } j = 40; \\ 0 & \text{if } j \neq i+1,\ 40. \end{cases} \qquad \text{with } p_{39,40}(r) = 1.$$

Similarly, if a trade-in decision a is selected in state i, the replacement car of age a must be kept for at least one time period and the rewards and transition probabilities are:

$$r_i(a) = t_i - c_a - e_a;\ p_{ij}(a) = \begin{cases} p_a & \text{if } j = a+1; \\ 1-p_a & \text{if } j = 40; \\ 0 & \text{if } j \neq i+1,\ 40. \end{cases} \qquad \text{with } p_{i,40}(39) = 1.$$

For a replacement decision $a \neq r$ in state $0 \leq i \leq 39$, the transition probabilities depend only on a and not on i and the expected reward $r_i(a) = s_i + t_a$ with $s_i = t_i$ and $t_a = -(c_a + e_a)$. Furthermore, we have $A_1(i) = \{0, 1, \ldots, 40\},\ 0 \leq i \leq 40$. Therefore, the problem is separable with $S_1 = S = \{0, 1, \ldots, 40\},\ A_1(i) = \{0, 1, \ldots, 39\},\ 0 \leq i \leq 40,\ A_2(i) = \{r\},\ 0 \leq i \leq 39,\ A_2(40) = \emptyset,$ $B(i) = \emptyset,\ 0, 1, \ldots, 39$ and $B(40) = \{0, 1, \ldots, 39\}$. Ignoring those i's for which $B(i) = \emptyset$, we require a single extra state. Call it state 41. In economic terms, state 41 corresponds to having no car and needing to buy one immediately. Instantaneous transition to state 41 is available from states 0 through 40 by trading in one's current car and this transition is mandatory from state 40.

Hence, we may consider the model as an MDP with state space $S^* = \{0, 1, \ldots, 41\}$, action sets

$$A^*(i) = \begin{cases} \{r, t\} & i = 0, 1, \ldots, 39 \\ \{t\} & i = 40 \\ \{0, 1, \ldots, 39\} & i = 41 \end{cases}$$

with decision t can be interpreted as the trade-in with in state i reward t_i and an immediate transition to the no-car state 41. Purchase decision a from state 41 has expected reward $-(c_a + e_a)$ and transition probability p_a to state $a + 1$ and $1 - p_a$ to state 40. The tranformed problem affords a reduction in the number of policies from 41^{41} to $40 \cdot 2^{40}$ and a reduction in the total number of decisions from 1680 $(40 \cdot 41 + 40)$ to 121 $(40 \cdot 2 + 1 + 40)$, while increasing the number of states from 41 to 42.

The reduction in problem size is so drastic that we can be assured of substantial computational savings in the policy iteration, linear programming and value iteration methods for both discounted and averaging versions of the probem. Note finally that every sequence of two transitions gives rise to at least one change in epoch, a fact that we shall find useful in the analysis.

2. Inventory problem

For a prototype of a variety of inventory models that are separable, consider the model with integer on-hand quantities, instantaneous replenishment, linear ordering cost, a set-up charge if any units are ordered, excess sales lost and a storage capacity of N items. State i denotes i units on hand at the beginning of the period, and a set $A(i)$ of available decisions from state i is given by $A(i) = \{i, i+1, \ldots, N\}$ with $a \in A(i)$ denoting either an order of $a - i$ items if $a > i$ or no order if $a = i$. The ordering cost is 0 if $a = i$ and $K + k(a - i)$ if $a > i$, where K and k are, respectively, the set-up and per item ordering cost. Let p_j be the probability that j sales opportunities appear during the time period, and let h_a be the expected one-period holding and sales cost given a items on hand at the beginning of the period, immediately after delivery of the order.

If no order is placed, i.e. $a = i$, the expexted one-period reward $r_i(i) = -h_i$ and the transition probabilities are given by $p_{ij}(i) = \begin{cases} p_{i-j} & \text{if } 1 \le j \le i; \\ \sum_{k \ge i} p_k & \text{if } j = 0; \\ 0 & \text{if } j > i. \end{cases}$

Similarly, if $a > i$, the expected one-period reward and the transition probabilities are, respectively, given by $r_i(a) = -K - k(a-i) - h_a$ and $p_{ij}(a) = \begin{cases} p_{a-j} & \text{if } 1 \le j \le a; \\ \sum_{k \ge a} p_k & \text{if } j = 0; \\ 0 & \text{if } j > a. \end{cases}$

In the case that $a > i$, the transition probabilities $p_{ij}(a),\ j \in S$ are independent of i and the expected reward $r_i(a) = s_i + t_a$ with $s_i = -K + k \cdot i$ and $t_a = -k \cdot a - h_a$ Then, with $A_1(i) = \{i+1, i+2, \ldots, N\}$ for $0 \le i \le N-1$ and $A_2(i) = \{i\}$ for $0 \le i \le N-1$, the problem satisfies the conditions of separability with $B(i) = \{i+1\}$ for $i = 0, 1, \ldots, N-1$. The reduction in this problem is less dramatic as in the automobile replacement problem: the state space increases from $N+1$ to $2N+1$ and the total number of decisions is reduced from $\sum_{i=0}^{N}(N-i+1) = \frac{1}{2}(N+1)(N+2)$ to $4N+1$ (two options in the states $0, 1, \ldots, N-1$, two options in the N additional states and one option in state N).

8.7.3 Discounted rewards

The description in section 8.7.1 as a problem with zero-time and one-time transitions gives rise to consider the transformed model with $N+m$ states and to the following linear program for the computation of the value vector v^α.

$$min \left\{ \sum_{i=1}^{N} v_i + \sum_{i=1}^{m} y_i \;\middle|\; \begin{array}{lll} v_i \ge r_i(a) + \alpha \sum_{j=1}^{N} p_{ij}(a) v_j & & 1 \le i \le N,\ a \in A_2(i) \\ v_i \ge s_i + y_i & & 1 \le i \le m \\ y_i \ge t_a + \alpha \sum_{j=1}^{N} p_j(a) v_j & & 1 \le i \le m,\ a \in B(i) \\ y_i \ge y_{i+1} & & 1 \le i \le m-1 \end{array} \right\}. \tag{8.87}$$

The first set of inequalities corresponds to the non-separable set $S \times A_2$ with one-time transitions; the second set inequalities to the zero-time transitions from the state i to $N+i,\ 1 \le i \le m$; the third set of inequalities to the set $S_1 \times B$ with one-time transitions and the last set inequalities corresponds to the zero-time transitions from the state $N+i$ to $N+i+1,\ 1 \le i \le m-1$.

The dual of program (8.87), where the dual variables $x_i(a),\ \lambda_i,\ w_i(a),\ \rho_i$ correspond to the four sets of constraints in (8.87), is:

$$max \sum_{i=1}^{N} \sum_{a \in A_2(i)} r_i(a) x_i(a) + \sum_{i=1}^{m} s_i \lambda_i + \sum_{i=1}^{m} \sum_{a \in B(i)} w_i(a) \tag{8.88}$$

subject to the constraints

$$\begin{array}{rcrcrcl} \sum_{i=1}^{N}\sum_{a\in A_2(i)}\{\delta_{ij}-\alpha p_{ij}(a)\}x_i(a) & + & \sum_{i=1}^{m}\delta_{ij}\lambda_i & - & \sum_{i=1}^{m}\sum_{a\in B(i)}p_j(a)w_i(a) & = & 1,\ 1\le j\le N \\ \rho_j-\rho_{j-1} & - & \lambda_j & + & \sum_{a\in B(j)}w_j(a) & = & 1,\ 1\le j\le m-1 \\ -\rho_{m-1} & - & \lambda_m & + & \sum_{a\in B(m)}w_m(a) & = & 1 \end{array}$$

$x_i(a)\ge 0,\ 1\le i\le N,\ a\in A_2(i);\ \lambda_i\ge 0,\ 1\le i\le m;\ w_i(a)\ge 0,\ 1\le i\le m,\ a\in B(i);$
$\rho_i\ge 0,\ 1\le i\le m-1.$

Without using the transformed problem, the linear program to compute the value vector v^α is:

$$min\Big\{\sum_{i=1}^{N}v_i\ \Big|\ v_i\ge r_i(a)+\alpha\sum_{j=1}^{N}p_{ij}(a)v_j,\ 1\le i\le N,\ a\in A(i)\Big\}. \tag{8.89}$$

Lemma 8.16

Let v feasible for (8.89) and define y by $y_i = max_{a\in A_1(i)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j\},\ 1\le i\le m$. Then,

(1) (v,y) is a feasible solution of (8.87).

(2) $\sum_{i=1}^{N}v_i+\sum_{i=1}^{m}y_i\ge\sum_{i=1}^{N}v_i^\alpha+\sum_{i=1}^{m}max_{a\in A_1(i)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j^\alpha\}$.

Proof

First we have to show that (v,y) satisfies the four parts of the constraints of (8.87).
The first and third part are obviously satisfied.
For the second part, notice that for all $1\le i\le m$ and $a\in A_1(i)$ we have

$$v_i\ge s_i+t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j\ \rightarrow\ v_i\ge s_i+max_{a\in A_1(i)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j\}=s_i+y_i.$$

For the fourth part, we write for $i=1,2,\ldots,m-1$

$$y_i-y_{i+1}=max_{a\in A_1(i)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j\}-max_{a\in A_1(i+1)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j\}\ge 0,$$

the last inequality because $A_1(i+1)\subseteq A_1(i)$.
Finally, because v^α is the componentwise smallest solution of (8.89), cf. Theorem 3.16, we have $v_i\ge v_i^\alpha,\ 1\le i\le N$, and consequently,

$$\begin{array}{rcl} \sum_{i=1}^{N}v_i+\sum_{i=1}^{m}y_i & = & \sum_{i=1}^{N}v_i+\sum_{i=1}^{m}max_{a\in A_1(i)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j\} \\ & \ge & \sum_{i=1}^{N}v_i^\alpha+\sum_{i=1}^{m}max_{a\in A_1(i)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j^\alpha\}. \end{array}$$

□

Corollary 8.2

Since v^α is the unique optimal solution of (8.89), we have shown that (v^α,y^α) is the unique optimal solution of (8.87), where $y_i^\alpha=max_{a\in A_1(i)}\{t_a+\alpha\sum_{j=1}^{N}p_j(a)v_j^\alpha\},\ 1\le i\le m$.

Theorem 8.25

Let (x,λ,w,ρ) be an optimal solution of (8.88). Define $S_x=\{j\mid\sum_{a\in A_2(j)}x_j(a)>0\}$ and $k_j=min\{k\ge j\mid\sum_{a\in B(k)}w_k(a)>0\},\ j\in S\backslash S_x$. Take any policy $f^\infty\in C(D)$ such that $x_j\big(f(j)\big)>0$ if $j\in S_x$ and $w_{k_j}\big(f(j)\big)>0$ if $j\in S\backslash S_x$. Then, f^∞ is well-defined and an α-discounted optimal policy.

Proof

From the definition of S_x it follows that $f(j)$ is well-defined if $j \in S_x$. From the first set of the constraints of (8.88) it follows that for $j = 1, 2, \ldots, N$, we have

$\sum_{a\in A_2(j)} x_j(a) + \sum_{i=1}^m \delta_{ij}\lambda_i \geq 1 + \alpha \sum_{i=1}^N \sum_{a\in A_2(i)} p_{ij}(a)x_i(a) + \sum_{i=1}^m \sum_{a\in B(i)} p_j(a)w_i(a) \geq 1.$

Therefore, if $j \notin S_*$, then $1 \leq j \leq m$ and $\lambda_j > 0$. Then, by adding the corresponding last constraints of (8.88), we obtain

$$\begin{aligned} \sum_{k=j}^m \sum_{a\in B(k)} w_k(a) &= \sum_{k=j}^{m-1}\{1 + \lambda_k + (\rho_{k-1} - \rho_k)\} + \{1 + \lambda_m + \rho_{m-1}\} \\ &= \sum_{k=j}^m\{1 + \lambda_k\} + \rho_{j-1} > 0. \end{aligned}$$

Hence, k_j is well-defined, and therefore the policy f^∞ is well-defined.

For the proof of the optimality of f^∞ we first consider a state $i \in S_x = \{j \mid \sum_{a\in A_2(j)} x_j(a) > 0\}$. In such state i, we have $x_i\big(f(i)\big) > 0$ and, by the complementary slackness property of linear programming, $v_i^\alpha = r_i(f) + \alpha \sum_{j=1}^N p_{ij}(f)v_j^\alpha$.

Now, we will show that in a state $i \in S\backslash S_x$ we also have $v_i^\alpha = r_i(f) + \alpha \sum_{j=1}^N p_{ij}(f)v_j^\alpha$.

Consider a state $i \in S\backslash S_x$, i.e. $1 \leq i \leq m$, $\sum_{a\in A_2(j)} x_j(a) = 0$ and $\lambda_i > 0$.

Let $w_k\big(f(i)\big) > 0$, i.e. $\sum_{a\in B(j)} w_j(a) = 0$ for $j = i, i+1, \ldots, k-1$ and $\sum_{a\in B(k} w_k(a) > 0$.

Then, by the complementary slackness property of linear programming, we have

$$\begin{aligned} &\lambda_i > 0 && \rightarrow \quad v_i^\alpha = s_i + y_i^\alpha \\ &\textstyle\sum_{a\in B(j)} w_j(a) = 0,\ j = i, i+1, \ldots, k-1 && \rightarrow \quad y_i^\alpha = y_{i+1}^\alpha = \cdots = y_k^\alpha \\ &w_k\big(f(i)\big) > 0 && \rightarrow \quad y_k^\alpha = t_{f(i)} + \alpha \textstyle\sum_{j=1}^N p_j(f)v_j^\alpha \end{aligned}$$

Hence,

$$\begin{aligned} v_i^\alpha &= s_i + y_i^\alpha = v_i^\alpha = s_i + y_k^\alpha = s_i + t_{f(i)} + \alpha \textstyle\sum_{j=1}^N p_j(f)v_j^\alpha \\ &= r_i\big(f(i)\big) + \alpha \textstyle\sum_{j=1}^N p_{ij}(f)v_j^\alpha. \end{aligned}$$

Therefore, we have shown that $v_i^\alpha = r_i(f) + \alpha \sum_{j=1}^N p_{ij}(f)v_j^\alpha,\ i \in S$, in vector notation

$v^\alpha = r(f) + \alpha P(f)v^\alpha \ \rightarrow\ \{I - P(f)\}v^\alpha = r(f) \ \rightarrow\ v^\alpha = \{I - P(f)\}^{-1}r(f) = v^\alpha(f^\infty),$

i.e. f^∞ is an α-discounted optimal policy. □

8.7.4 Average rewards - unichain case

Consider the problem again in the transformed model with $N + m$ states and with zero-time and one-time transitions. This interpretation gives rise to the following linear program for the computation of the value vector ϕ.

$$min \left\{ x \;\middle|\; \begin{array}{lllll} x + y_i \geq r_i(a) + \sum_{j=1}^N p_{ij}(a)y_j & & & & 1 \leq i \leq N,\ a \in A_2(i) \\ y_i \geq s_i + z_i & & & & 1 \leq i \leq m \\ x + z_i \geq t_a + \sum_{j=1}^N p_j(a)y_j & & & & 1 \leq i \leq m,\ a \in B(i) \\ z_i \geq z_{i+1} & & & & 1 \leq i \leq m-1 \end{array} \right\}. \tag{8.90}$$

The dual of program (8.90), where the dual variables $x_i(a),\ \lambda_i,\ w_i(a),\ \rho_i$ correspond to the four sets of constraints in (8.90), is:

$$max \sum_{i=1}^{N} \sum_{a \in A_2(i)} r_i(a) x_i(a) + \sum_{i=1}^{m} s_i \lambda_i + \sum_{i=1}^{m} \sum_{a \in B(i)} w_i(a) \tag{8.91}$$

subject to the constraints

$$\begin{array}{rcrcrcl}
\sum_{i=1}^{N} \sum_{a \in A_2(i)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & + & \sum_{i=1}^{m} \delta_{ij} \lambda_i & - & \sum_{i=1}^{m} \sum_{a \in B(i)} p_j(a) w_i(a) & = & 0,\ 1 \le j \le N \\
\rho_j - \rho_{j-1} & - & \lambda_j & + & \sum_{a \in B(j)} w_j(a) & = & 0,\ 1 \le j \le m-1 \\
-\rho_{m-1} & - & \lambda_m & + & \sum_{a \in B(m)} w_m(a) & = & 0 \\
\sum_{i=1}^{N} \sum_{a \in A_2(i)} x_i(a) & & & + & \sum_{i=1}^{m} \sum_{a \in B(i)} w_i(a) & = & 1
\end{array}$$

$x_i(a) \ge 0,\ 1 \le i \le N,\ a \in A_2(i);\ \lambda_i \ge 0,\ 1 \le i \le m;\ w_i(a) \ge 0,\ 1 \le i \le m,\ a \in B(i);$

$\rho_0 = 0;\ \rho_i \ge 0,\ 1 \le i \le m-1.$

Without using the transformed problem, the linear program to compute the value ϕ is:

$$min \Big\{ x \ \Big|\ x + y_i \ge r_i(a) + \sum_{j=1}^{N} p_{ij}(a) y_j,\ 1 \le i \le N,\ a \in A(i) \Big\}. \tag{8.92}$$

Lemma 8.17

Let (x, y) feasible for (8.92) and define z by $z_i = max_{a \in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a) y_j\} - x,\ 1 \le i \le m$. Then, (x, y, z) is a feasible solution of (8.90) and $x \ge \phi$.

Proof

First we have to show that (x, y, z) satisfies the four parts of the constraints of (8.90).

The first and third part are obviously satisfied.

For the second part, notice that for all $i = 1, 2, \ldots, m$ and $a \in A_1(i)$ we have

$$x + y_i \ge s_i + t_a + \sum_{j=1}^{N} p_j(a) y_j \ \rightarrow\ y_i \ge s_i + max_{a \in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a) y_j\} - x = s_i + z_i.$$

For the fourth part, we write for $i = 1, 2, \ldots, m-1$

$$z_i - z_{i+1} = max_{a \in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a) y_j\} - max_{a \in A_1(i+1)} \{t_a + \sum_{j=1}^{N} p_j(a) y_j\} \ge 0,$$

the last inequality because $A_1(i+1) \subseteq A_1(i)$.

Finally, because ϕ is the optimal solution of (8.92), we have $x \ge \phi$. □

Corollary 8.3

Since any optimal solution (x^*, y^*) of (8.92) satisfies $x^* = \phi$, the optimum value of (8.90) is also ϕ. Furthermore, $(x^* = \phi, y^*, z^*)$ is an optimal solution of program (8.90), where $z_i^* = max_{a \in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a) y_j^*\} - \phi,\ 1 \le i \le m.$

Theorem 8.26

Let (x, λ, w, ρ) be an optimal solution of (8.91). Define $S_x = \{j \mid \sum_{a\in A_2(j)} x_j(a) > 0\}$ and $k_j = min\{k \geq j \mid \sum_{a\in B(k)} w_k(a) > 0\}$, $j \in S_w$, where $S_w = \{j \in S\backslash S_x \mid \sum_{a\in A_1(j)} w_j(a) > 0\}$. Take any policy $f^\infty \in C(D)$ such that $x_j\big(f(j)\big) > 0$ if $j \in S_x$, $w_{k_j}\big(f(j)\big) > 0$ if $j \in S_w$ and $f(j)$ arbitrarily chosen if $j \notin S_x \cup S_w$. Then, f^∞ is an average optimal policy.

Proof

Let (ϕ, y, z) be an optimal solution of (8.91). Then, by the complementary slackness property of linear programming, we have

$$x_i(a) \cdot \{\phi + y_i - r_i(a) - \sum_{j=1}^{N} p_{ij}(a)y_j\} = 0,\ 1 \leq i \leq N;\ a \in A_2(i) \tag{8.93}$$

$$\lambda_i \cdot \{y_i - s_i - z_i\} = 0,\ 1 \leq i \leq m \tag{8.94}$$

$$w_i(a) \cdot \{\phi + z_i - t_a - \sum_{j=1}^{N} p_j(a)y_j\} = 0,\ 1 \leq i \leq m;\ a \in B(i) \tag{8.95}$$

$$\rho_i \cdot \{z_i - z_{i+1}\} = 0,\ 1 \leq i \leq m-1 \tag{8.96}$$

Let $S_+ = \{j \in S \mid \sum_{a\in A_2(j)} x_j(a) + \lambda_j > 0\}$ and take any $i \in S_+$.

If $\sum_{a\in A_2(i)} x_i(a) > 0$, then from equation (8.93), we obtain

$$\phi = r_i(f) - y_i + \sum_{j=1}^{N} p_{ij}(f)y_j,\ i \in S_x. \tag{8.97}$$

If $\sum_{a\in A_2(i)} x_i(a) = 0$, then $1 \leq i \leq m$, $\lambda_i > 0$, and equation (8.94) gives $y_i = s_i + z_i$.

Furthermore, we have $\sum_{a\in A_1(i)} w_i(a) > 0$, namely: adding the corresponding constraints of (8.91) yields $\sum_{a\in A_1(i)} w_i(a) = \sum_{j=i}^{m} \sum_{a\in B(j)} w_j(a) = \sum_{j=i}^{m} \lambda_j + \rho_{i-1} \geq \lambda_i > 0$.

The definition of k_i implies, denoting k_i by k, that $\sum_{a\in B(j)} w_j(a) = 0$ for $j = i, i+1, \ldots, k-1$.

Hence, by the constraints of program (8.91), we obtain $\rho_j = \lambda_j + \rho_{j-1}$ for $j = i, i+1, \ldots, k-1$, implying $\rho_i = \lambda_i + \rho_{i-1} \geq \lambda_i > 0$, $\rho_{i+1} = \lambda_{i+1} + \rho_i \geq \rho_i > 0$, $\ldots$, $\rho_{k-1} = \lambda_{k-1} + \rho_{k-2} \geq \rho_{k-2} > 0$.

Then, it follows from (8.96) that $z_i = z_{i+1} = \cdots = z_k = 0$.

Since $w_k\big(f(i)\big) > 0$, by (8.95), we can write

$$\begin{aligned} \phi &= t_{f(i)} + \textstyle\sum_{j=1}^{N} p_j(f(i))y_j - z_k = t_{f(i)} + \sum_{j=1}^{N} p_j(f(i))y_j - z_i \\ &= s_i + t_{f(i)} + \textstyle\sum_{j=1}^{N} p_j(f(i))y_j - y_i, \end{aligned}$$

implying

$$\phi = r_i(f) - y_i + \sum_{j=1}^{N} p_{ij}(f)y_j,\ i \in S_+\backslash S_x. \tag{8.98}$$

Combining (8.97) and (8.98) yields

$$\phi = r_i(f) - y_i + \sum_{j=1}^{N} p_{ij}(f)y_j,\ i \in S_+. \tag{8.99}$$

Next we show that S_+ is closed under $P(f)$, i.e. $p_{ij}(f) = 0,\ i \in S_+,\ j \notin S_+$. Suppose that $p_{ij}(f) > 0$ for some $i \in S_+$ and $j \notin S_+$. Since $j \notin S_+$, $\sum_{a \in A_2(j)} x_j(a) + \sum_{i=1}^m \delta_{ij}\lambda_i = 0$.
If $i \in S_x$, then, from the constraints of program (8.91) it follows that,

$$\begin{aligned} 0 &= \textstyle\sum_{a \in A_2(j)} x_j(a) + \sum_{i=1}^m \delta_{ij}\lambda_i \\ &= \textstyle\sum_{i=1}^N \sum_{a \in A_2(i)} p_{ij}(a)\}x_i(a) + \sum_{i=1}^m \sum_{a \in B(i)} p_j(a)w_i(a) \geq p_{ij}(f)x_i(f(i)) > 0, \end{aligned}$$

implying a contradiction.
If $i \in S_+ \backslash S_x$, then, from the constraints of program (8.91) it follows that,

$$\begin{aligned} 0 &= \textstyle\sum_{a \in A_2(j)} x_j(a) + \sum_{i=1}^m \delta_{ij}\lambda_i \\ &= \textstyle\sum_{i=1}^N \sum_{a \in A_2(i)} p_{ij}(a)\}x_i(a) + \sum_{i=1}^m \sum_{a \in B(i)} p_j(a)w_i(a) \geq p_j(f(i))w_{k_i}(f(i)) > 0, \end{aligned}$$

which also implies a contradiction.
Since $P(f)$ is a unichain Markov chain and S_+ is closed, the states of $S \backslash S_+$ are transient.
Let $\pi(f)$ be the stationary distribution of the unichain Markov chain $P(f)$, then it follows from (8.99) that $\phi \cdot e = \phi \cdot \pi(f) = \pi(f)\{r(f) - \pi(f)y + \pi(f)P(f)y = \phi(f^\infty)$, i.e. f^∞ is an average optimal policy. □

8.7.5 Average rewards - general case

Again, the interpretation of the transformed model gives rise to consider the following linear program in order to compute the value vector ϕ.

$$min \left\{ \sum_{j=1}^N x_j + \sum_{j=1}^m w_j \;\middle|\; \begin{array}{rcl} x_i & \geq & \sum_{j=1}^N p_{ij}(a)x_j \quad 1 \leq i \leq N,\ a \in A_2(i) \\ x_i & \geq & w_i \quad 1 \leq i \leq m \\ w_i & \geq & \sum_{j=1}^N p_j(a)x_j \quad 1 \leq i \leq m,\ a \in B(i) \\ w_i & \geq & w_{i+1} \quad 1 \leq i \leq m-1 \\ x_i + y_i & \geq & r_i(a) + \sum_{j=1}^N p_{ij}(a)y_j \quad 1 \leq i \leq N,\ a \in A_2(i) \\ y_i & \geq & s_i + z_i \quad 1 \leq i \leq m \\ w_i + z_i & \geq & t_a + \sum_{j=1}^N p_j(a)y_j \quad 1 \leq i \leq m,\ a \in B(i) \\ z_i & \geq & z_{i+1} \quad 1 \leq i \leq m-1 \end{array} \right\}. \tag{8.100}$$

The dual of program (8.100), where the dual variables $y_i(a),\ \mu_i,\ z_i(a),\ \sigma_i,\ x_i(a),\ \lambda_i,\ w_i(a),\ \rho_i$ correspond to the eight sets of constraints in (8.100), is:

$$max \sum_{i=1}^N \sum_{a \in A_2(i)} r_i(a)x_i(a) + \sum_{i=1}^m s_i\lambda_i + \sum_{i=1}^m \sum_{a \in B(i)} t_a w_i(a) \tag{8.101}$$

subject to the constraints

$$\begin{array}{rcrcrcrcl} \sum_{i=1}^N \sum_{a \in A_2(i)} \{\delta_{ij} - p_{ij}(a)\}y_i(a) & + & \sum_{i=1}^m \delta_{ij}\mu_i & - & \sum_{i=1}^m \sum_{a \in B(i)} p_j(a)z_i(a) & + & \sum_{a \in A_2(i)} x_j(a) & = & 1,\ 1 \leq j \leq N \\ \sigma_j - \sigma_{j-1} & - & \mu_j & + & \sum_{a \in B(j)} w_j(a) & + & \sum_{a \in B(j)} z_j(a) & = & 1,\ 1 \leq j \leq m \\ \sum_{i=1}^N \sum_{a \in A_2(i)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & + & \sum_{i=1}^m \delta_{ij}\lambda_i & - & \sum_{i=1}^m \sum_{a \in B(i)} p_j(a)w_i(a) & & & = & 0,\ 1 \leq j \leq N \\ \rho_j - \rho_{j-1} & - & \lambda_j & + & \sum_{a \in B(j)} w_j(a) & & & = & 0,\ 1 \leq j \leq m \end{array}$$

$\rho_0 = \rho_m = \sigma_0 = \sigma_m = 0;\ x_i(a),\ y_i(a),\ z_i(a),\ w_i(a),\ \lambda_i,\ \mu_i,\ \rho_i,\ \sigma_i \geq 0$ for all i and a.

Without using the transformed problem, the linear program to compute the value ϕ is:

$$min \left\{ \sum_{j=1}^{N} x_j \;\middle|\; \begin{array}{rcl} \sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}x_j & \geq & 0 \quad 1 \leq i \leq N,\ a \in A(i) \\ x_i + \sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}u_j & \geq & r_i(a) \quad 1 \leq i \leq N,\ a \in A(i) \end{array} \right\}. \tag{8.102}$$

Lemma 8.18

Let (x,u) be a feasible solution of (8.102) and define w,y,z by $y_i = x_i + u_i,\ 1 \leq i \leq N$, $w_i = max_{a\in A_1(i)} \sum_{j=1}^{N} p_j(a)x_j,\ 1 \leq i \leq m$ and $z_i = max_{a\in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a)u_j\},\ 1 \leq i \leq m$. Then,

(1) (x,w,y,z) is a feasible solution of (8.100).

(2) $\sum_{j=1}^{N} x_j + \sum_{j=1}^{m} w_j \geq \sum_{j=1}^{N} \phi_j + \sum_{j=1}^{m} max_{a\in A_1(j)} \sum_{k=1}^{N} p_k(a)\phi_k$.

Proof

The proof of part (1) consists of the verification of the eight sets of the constraints of (8.100).

a. $x_i \geq \sum_{j=1}^{N} p_{ij}(a)x_j,\ 1 \leq i \leq N,\ a \in A_2(i)$.

b. $$\begin{array}{rcl} x_i - w_i & = & x_i - max_{a\in A_1(i)} \sum_{j=1}^{N} p_j(a)x_j \\ & \geq & max_{a\in A(i)} \sum_{j=1}^{N} p_j(a)\, x_j - max_{a\in A_1(i)} \sum_{j=1}^{N} p_j(a)\, x_j \geq 0,\ 1 \leq i \leq m. \end{array}$$

c. $w_i - \sum_{j=1}^{N} p_j(a)x_j = max_{a\in A_1(i)} \sum_{j=1}^{N} p_j(a)x_j - \sum_{j=1}^{N} p_j(a)x_j \geq 0,\ 1 \leq i \leq m,\ a \in B(i)$

d. $w_i - w_{i+1} = max_{a\in A_1(i)} \sum_{j=1}^{N} p_j(a)x_j - max_{a\in A_1(i+1)} \sum_{j=1}^{N} p_j(a)x_j \geq 0,\ 1 \leq i \leq m-1$ because $A_1(i+1) \subseteq A_1(i)$.

e. $$\begin{array}{rcl} x_i + \sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}y_j & = & x_i + \sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}(x_i + u_i) \\ & \geq & x_i + \sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}u_i \geq r_i(a),\ 1 \leq i \leq N,\ a \in A_2(i). \end{array}$$

f. $$\begin{array}{rcl} y_i - z_i & = & x_i + u_i - max_{a\in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a)u_j\} \\ & = & min_{a\in A_1(i)} \{x_i + u_i - t_a - \sum_{j=1}^{N} p_j(a)u_j\} \\ & \geq & min_{a\in A_1(i)}(r_i(a) - t_a) = s_i,\ 1 \leq i \leq m. \end{array}$$

g. $$\begin{array}{rcl} w_i + z_i - \sum_{j=1}^{N} p_j(a)y_j & = & max_{a\in A_1(i)} \sum_{j=1}^{N} p_j(a)x_j + \\ & & max_{a\in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a)u_j\} - \sum_{j=1}^{N} p_j(a)y_j \\ & \geq & \sum_{j=1}^{N} p_j(a)x_j + t_a + \sum_{j=1}^{N} p_j(a)(u_j - u_j) = t_a,\ 1 \leq i \leq m,\ a \in B(i) \end{array}$$

h. $$\begin{array}{rcl} z_i - z_{i+1} & = & max_{a\in A_1(i)} \{t_a + \sum_{j=1}^{N} p_j(a)u_j\} - max_{a\in A_1(i+1)} \{t_a + \sum_{j=1}^{N} p_j(a)u_j\} \\ & \geq & 0, \text{ because } A_1(i+1) \subseteq A_1(i). \end{array}$$

For the proof of part (2), we can use $x \geq \phi$ and write

$$\begin{aligned} \sum_{j=1}^{N} x_j + \sum_{j=1}^{m} w_j &= \sum_{j=1}^{N} x_j + \sum_{j=1}^{m} max_{a \in A_1(j)} \sum_{k=1}^{N} p_k(a) x_k \\ &\geq \sum_{j=1}^{N} \phi_j + \sum_{j=1}^{m} max_{a \in A_1(j)} \sum_{k=1}^{N} p_k(a) \phi_k. \end{aligned}$$

□

Lemma 8.19

Let (x, w, y, z) is a feasible solution of (8.100). Then,

(1) $w_i \geq \sum_{j=1}^{N} p_j(a) x_j$ for all $1 \leq i \leq m,\ a \in A_1(i)$.

(2) (x, y) is a feasible solution of (8.102) and $x \geq \phi$.

Proof

For the proof of part (1) take any $1 \leq i \leq m$ and $a \in A_1(i)$, say $a \in B(k)$ for some $i \leq k \leq m$. Then, we have $w_i \geq w_k \geq \sum_{j=1}^{N} p_j(a) x_j$.

For the proof of part (2) we have to verify the constraints of (8.102). It is obvious that

$$\sum_{j=1}^{N} \{\delta_{ij} - p_{ij}(a)\} x_j \geq 0,\ 1 \leq i \leq N,\ a \in A_2(i).$$

and

$$x_i + \sum_{j=1}^{N} \{\delta_{ij} - p_{ij}(a)\} y_j \geq r_i(a),\ 1 \leq i \leq N,\ a \in A_2(i).$$

Take any $1 \leq i \leq m$ and $a \in A_1(i)$, say $a \in B(k)$ for some $i \leq k \leq m$. Then,

$$\sum_{j=1}^{N} \{\delta_{ij} - p_{ij}(a)\} x_j = x_i - \sum_{j=1}^{N} p_j(a) x_j \geq w_i - w_k \geq 0$$

and

$$\begin{aligned} x_i + \sum_{j=1}^{N} \{\delta_{ij} - p_{ij}(a)\} y_j &= x_i + y_i - \sum_{j=1}^{N} p_j(a) y_j \\ &\geq w_i + s_i + z_i - \sum_{j=1}^{N} p_j(a) y_j \\ &\geq s_i + t_a + \sum_{j=1}^{N} p_j(a) y_j - \sum_{j=1}^{N} p_j(a) y_j = r_i(a). \end{aligned}$$

Hence, (x, y) is a feasible solution of (8.102). Furthermore, $x \geq \phi$, because ϕ is the componentwise smallest superharmonic vector.

□

Theorem 8.27

(1) The linear programs (8.100) and (8.101) have finite optimal solutions.

(2) If (x, w, y, z) is an optimal solution of (8.100), then $x = \phi$ and $w_j = max_{a \in A_1(j)} \sum_{k=1}^{N} p_k(a) \phi_k$.

Proof

We know that (8.102) has a finite optimal solution (x^*, u^*) with $x^* = \phi$. Hence, program (8.100) is feasible and bounded, implying that (8.100) and its dual (8.101) have finite optimal solutions. Consider an optimal solution $(x = \phi, u)$ of (8.102). Then, it follows from Lemma 8.18 part (2) that the corresponding solution $(x = \phi, w, y, z)$ is an optimal solution of (8.100), because

$$\sum_{j=1}^{N} x_j + \sum_{j=1}^{m} w_j = \sum_{j=1}^{N} \phi_j + \sum_{j=1}^{m} max_{a \in A_1(j)} \sum_{k=1}^{N} p_k(a) \phi_k.$$

Let (x, w, y, z) is an optimal solution of (8.100). Then,

$$\sum_{j=1}^{N} x_j + \sum_{j=1}^{m} w_j = \sum_{j=1}^{N} \phi_j + \sum_{j=1}^{m} max_{a \in A_1(j)} \sum_{k=1}^{N} p_k(a) \phi_k.$$

By Lemma 8.19, $x \geq \phi$ and $w_j \geq max_{a \in A_1(j)} \sum_{k=1}^{N} p_k(a) \phi_k,\ 1 \leq j \leq m$.

Hence, $x = \phi$ and $w_j = max_{a \in A_1(j)} \sum_{k=1}^{N} p_k(a) \phi_k$.

□

Lemma 8.20

For any pair of feasible solutions (x, w, y, z) *and* $(y, \mu, z, \sigma, x, \lambda, w, \rho)$ *of (8.100) and (8.101), respectively, the following equalities hold:*

$$\Big\{\sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}x_j\Big\} \cdot x_i(a) = 0,\ 1 \le i \le N,\ a \in A_2(i) \tag{8.103}$$

$$\{x_i - w_i\} \cdot \lambda_i = 0,\ 1 \le i \le m \tag{8.104}$$

$$\Big\{w_i - \sum_{j=1}^{N} p_j(a)x_j\Big\} \cdot w_i(a) = 0,\ 1 \le i \le m,\ a \in B(i) \tag{8.105}$$

$$\{w_i - w_{i+1}\} \cdot \rho_i = 0,\ 1 \le i \le m-1 \tag{8.106}$$

Proof

From the constraints of (8.100) and the nonnegativity of the variables of (8.101) it follows that the right hand sides of (8.103), (8.104), (8.105) and (8.106) are at least 0. If we add all left hand sides, we obtain

$$\begin{aligned}
&\textstyle\sum_{i=1}^{N}\sum_{a\in A_2(i)}\sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}x_j \cdot x_i(a) + \sum_{i=1}^{m}\{x_i - w_i\}\cdot\lambda_i + \\
&\qquad \textstyle\sum_{i=1}^{m}\sum_{a\in B(i)}\{w_i - \sum_{j=1}^{N} p_j(a)x_j \cdot w_i(a) + \sum_{i=1}^{m-1}\{w_i - w_{i+1}\}\cdot\rho_i = \\
&\textstyle\sum_{j=1}^{N} x_j \cdot \Big\{\sum_{i=1}^{N}\sum_{a\in A_2(i)}\{\delta_{ij} - p_{ij}(a)\}x_i(a) + \sum_{i=1}^{m}\delta_{ij}\lambda_i - \sum_{i=1}^{m}\sum_{a\in B(i)} p_j(a)w_i(a)\Big\} + \\
&\qquad \textstyle\sum_{i=1}^{m} w_i\Big\{-\lambda_i + \sum_{a\in B(i)} w_i(a) + \rho_i - \rho_{i-1}\Big\} = 0,
\end{aligned}$$

since the terms between brackets are zero by the constraints of (8.101). Hence, the relations (8.103), (8.104), (8.105) and (8.106) are proven. □

From the complementary slackness property of linear programming is follows that optimal solutions (x, w, y, z) and $(y, \mu, z, \sigma, x, \lambda, w, \rho)$ of (8.100) and (8.101), respectively, satisfy

$$\Big\{\sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}x_j\Big\} \cdot y_i(a) = 0,\ 1 \le i \le N,\ a \in A_2(i) \tag{8.107}$$

$$\{x_i - w_i\} \cdot \mu_i = 0,\ 1 \le i \le m \tag{8.108}$$

$$\Big\{w_i - \sum_{j=1}^{N} p_j(a)x_j\Big\} \cdot z_i(a) = 0,\ 1 \le i \le m,\ a \in B(i) \tag{8.109}$$

$$\{w_i - w_{i+1}\} \cdot \sigma_i = 0,\ 1 \le i \le m-1 \tag{8.110}$$

$$\Big\{x_i + \sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(a)\}y_j - r_i(a)\Big\} \cdot x_i(a) = 0,\ 1 \le i \le N,\ a \in A_2(i) \tag{8.111}$$

$$\{y_i - z_i - s_i\} \cdot \lambda_i = 0,\ 1 \le i \le m \tag{8.112}$$

$$\Big\{w_i + z_i - \sum_{j=1}^{N} p_j(a)y_j - t_a\Big\} \cdot w_i(a) = 0,\ 1 \le i \le m,\ a \in B(i) \tag{8.113}$$

$$\{z_i - z_{i+1}\} \cdot \rho_i = 0,\ 1 \le i \le m-1 \tag{8.114}$$

Lemma 8.21

Let (x, w, y, z) *and* $(y, \mu, z, \sigma, x, \lambda, w, \rho)$ *be optimal solutions of (8.100) and (8.101), respectively. Let* $m_i = min\{j \geq i \mid \sum_{a\in B(j)} w_j(a) > 0\}$ *and* $n_i = min\{j \geq i \mid \sum_{a\in B(j)} \{w_j(a) + z_j(a)\} > 0\}$. *Define the policy* $f^\infty \in C(D)$ *by*

$$x_i\big(f(i)\big) > 0 \quad if \quad \sum_{a\in A_2(i)} x_i(a) > 0 \tag{8.115}$$

$$w_{m_i}\big(f(i)\big) > 0 \quad if \quad \sum_{a\in A_2(i)} x_i(a) = 0 \ and \ \lambda_i > 0 \tag{8.116}$$

$$y_i\big(f(i)\big) > 0 \quad if \quad \sum_{a\in A_2(i)} x_i(a) = \lambda_i = 0 \ and \ y_i\big(f(i)\big) > 0 \tag{8.117}$$

$$w_{n_i}\big(f(i)\big) > 0 \quad if \quad \sum_{a\in A_2(i)} x_i(a) = \lambda_i = \sum_{a\in A_2(i)} y_i(a) = 0 \ and \ \sum_{a\in A_1(i)} w_{n_i}(a) > 0 \tag{8.118}$$

$$z_{n_i}\big(f(i)\big) > 0 \quad if \quad \sum_{a\in A_2(i)} x_i(a) = \lambda_i = \sum_{a\in A_2(i)} y_i(a) = \sum_{a\in A_1(i)} w_{n_i}(a) = 0 \tag{8.119}$$

Then,

(1) f^∞ *is well-defined.*

(2) $x_i + \sum_{j=1}^N \{\delta_{ij} - p_{ij}(f)\}y_j = r_i(f),\ i \in S_+ = \{j \in S \mid \sum_{a\in A_2(i)} x_j(a) + \lambda_j > 0\}$.

(3) $\sum_{j=1}^N \{\delta_{ij} - p_{ij}(f)\}x_j = 0,\ i \in S$.

Proof

(1) Suppose that $\sum_{a\in A_2(i)} x_i(a) = 0,\ \lambda_i > 0$ and $\sum_{a\in B(j)} w_j(a) = 0$ for all $j \geq i$. Then, by the constraints of (8.101), we obtain $0 = \sum_{j=i}^m \{\rho_j - \rho_{j-1} - \lambda_j\} = -\rho_{i-1} - \sum_{j=i}^m \lambda_j \leq \lambda_i < 0$, implying a contradiction. Hence, f^∞ is well-defined if $\sum_{a\in A_2(i)} x_i(a) = 0$ and $\lambda_i > 0$.

Because $\sum_{a\in B(m)} \{w_m(a) + z_m(a)\} = 1 + \mu_m + \sigma_{m-1} > 0$, n_i is well-defined for all i.

Therefore, the policy f^∞ is well-defined.

(2) Take any $i \in S_+$.

If $\sum_{a\in A_2(i)} x_i(a) > 0$, then by (8.111), $x_i + \sum_{j=1}^N \{\delta_{ij} - p_{ij}(f)\}y_j = r_i(f)$.

If $\sum_{a\in A_2(i)} x_i(a) = 0$, then $\lambda_i > 0$, and by (8.104) and (8.112), $x_i = w_i$ and $y_i = s_i + z_i$.

The definition of m_i implies that $\sum_{a\in B(j)} w_j(a) = 0,\ j = i, i+1, \ldots, m_i - 1$ and $w_{m_i}\big(f(i)\big) > 0$.

Hence, by the constraints of (8.101), we obtain $\rho_j = \rho_{j-1} + \lambda_j,\ j = i, i+1, \ldots, m_i - 1$, i.e. $\rho_j = \rho_{i-1} + \sum_{k=i}^j \geq \lambda_i > 0,\ j = i, i+1, \ldots, m_i - 1$. Then, by (8.106) and (8.114), $w_j = w_{j+1},\ z_j = z_{j+1}$ for $j = i, i+1, \ldots, m_i - 1$, implying $w_i = w_{m_i},\ z_i = z_{m_i}$.

Since $w_{m_i}\big(f(i)\big) > 0$, by (8.113), we have $w_{m_i} + z_{m_i} - \sum_{j=1}^N p_j(f)y_j - t_{f(i)} = 0$.

Hence,

$$\begin{aligned} r_i(f) &= s_i + t_{f(i)} = y_i - z_i + w_{m_i} + z_{m_i} - \textstyle\sum_{j=1}^N p_j(f)y_j \\ &= y_i - z_i + w_i + z_i - \textstyle\sum_{j=1}^N p_j(f)y_j = w_i + \sum_{j=1}^N \{\delta_{ij} - p_{ij}(f)\}y_j \\ &= x_i + \textstyle\sum_{j=1}^N \{\delta_{ij} - p_{ij}(f)\}y_j. \end{aligned}$$

(3) If $f(i)$ is determined by (8.115) or (8.117), the result follows from (8.103) and (8.107), respectively. Suppose that $f(i)$ is determined by (8.116), then $\lambda_i > 0$ and $w_{m_i}\big(f(i)\big) > 0$. By, (8.105), $w_{m_i} = \sum_{j=1}^{N} p_j(f)x_j$. In the proof of part (2) is shown that $\lambda_i > 0$ implies $w_{m_i} = w_i = x_i$. Therefore, $\sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(f)\}x_j = w_{m_i} - \sum_{j=1}^{N} p_j(f)x_j = 0$.
Finally, suppose that $f(i)$ is determined by (8.118) and (8.119). Then, by (8.105) or (8.109), $w_{n_i} = \sum_{j=1}^{N} p_j(f)x_j$. Since $\sum_{a\in A_2(i)} x_i(a) = \sum_{a\in A_2(i)} y_i(a) = 0$, it follows from the constraints of (8.101) that $\mu_i > 0$, implying by (8.108), that $w_i = x_i$. Furthermore, from the definition of n_i, it follows that $\sum_{a\in B(j)}\{w_j(a) + z_j(a)\} = 0,\ \ j = i, i+1, \ldots, n_i - 1$.
Then, by the constraints of (8.101) it follows that $\sigma_j > 0,\ \ j = i, i+1, \ldots, n_i - 1$, and consequently, by (8.110), $w_j = w_{j+1},\ \ j = i, i+1, \ldots, n_i - 1$. Combining these results yields $\sum_{j=1}^{N}\{\delta_{ij} - p_{ij}(f)\}x_j = w_i - \sum_{j=1}^{N} p_j(f)x_j = w_{n_i} - \sum_{j=1}^{N} p_j(f)x_j = 0$. □

Lemma 8.22

Let $(y, \mu, z, \sigma, x, \lambda, w, \rho)$ be an optimal solution of (8.101), and let S_+ and policy f^∞ be defined as in Lemma 8.21. Then, S_+ is closed under $P(f)$, i.e. $p_{ij}(f) = 0$ for every $i \in S_+$ and $j \notin S_+$.

Proof

Suppose that $p_{ij}(f) > 0$ for some $i \in S_+$ and $j \notin S_+$. Since $i \in S_+$, the action $f(i)$ is defined either by (8.115) or by (8.116). Furthermore, since $j \notin S_+$, $\sum_{a\in A_2(j)} x_j(a) + \sum_{i=1}^{m} \delta_{ij}\lambda_i = 0$.
If $f(i)$ is defined by (8.115), then by the constraints of (8.101), we can write

$$\begin{aligned} 0 &= \textstyle\sum_{a\in A_2(j)} x_j(a) + \sum_{i=1}^{m} \delta_{ij}\lambda_i = \sum_{i=1}^{N}\sum_{a\in A_2(i)} p_{ij}(a)x_i(a) + \sum_{i=1}^{m}\sum_{a\in B(i)} p_j(a)w_i(a) \\ &\geq p_{ij}(f)x_i\big(f(i)\big) > 0, \end{aligned}$$

implying a contradiction.

If $f(i)$ is defined by (8.116), then we obtain

$$\begin{aligned} 0 &= \textstyle\sum_{a\in A_2(j)} x_j(a) + \sum_{i=1}^{m} \delta_{ij}\lambda_i = \sum_{i=1}^{N}\sum_{a\in A_2(i)} p_{ij}(a)x_i(a) + \sum_{i=1}^{m}\sum_{a\in B(i)} p_j(a)w_i(a) \\ &\geq p_j(f)w_{m_i}\big(f(i)\big) > 0, \end{aligned}$$

which also gives a contradiction. □

Lemma 8.23

Let $(y, \mu, z, \sigma, x, \lambda, w, \rho)$ be an extreme optimal solution of (8.101), and let S_+ and policy f^∞ be defined as in Lemma 8.21. Then, the states of $S\backslash S_+$ are transient in the Markov chain with transition matrix $P(f)$.

Proof

Suppose that there exists a state $j \in S\backslash S_+$, which is recurrent under $P(f)$. Since S_+ is closed, there is an ergodic class $J \subseteq S\backslash S_+$. Let $J = J_1 \cup J_2 \cup J_3$, where J_1, J_2 and J_3 are the states of J

in which $f(i)$ is determined by (8.117), (8.118) and (8.119), respectively.

We first show that $J_2 = \emptyset$. From the constraints of (8.101) it follows that for any $j \in S \backslash S_+$, we have $\sum_{i=1}^{N} \sum_{a \in A_2(i)} p_{ij}(a) x_i(a) + \sum_{i=1}^{m} \sum_{a \in B(i)} p_j(a) w_i(a) = 0$, i.e. every term in this equation is 0. Suppose that $i \in J_2$. Then, $w_{n_i}\big(f(i)\big) > 0$, and consequently $p_{ij}(f) = 0$ for every $j \in S \backslash S_+$. This contradicts that $S \backslash S_+$ contains an ergodic class J. For $i \in J_1$, we have $y_i\big(f(i)\big) > 0$. For $i \in J_3$, we have $\sum_{a \in A_2(i)} \{x_i(a) + y_i(a)\} = 0$, $\sum_{a \in B(j)} \{w_j(a) + z_j(a)\} = 0$, $i \le j \le n_i - 1$, and $z_{n_i}\big(f(i)\big) > 0$. From the constraints of (8.101) it follows that $\mu_i > 0$ and $\sigma_j > 0$, $i \le j \le n_i - 1$. Next, consider the columns, denoted by $a^{(i)}, b^{(i)}, c^{(i)}$ and $d^{(j)}$, of the matrix of the constraints of (8.101) corresponding to the positive variables: $a^{(i)}$ corresponds to $y_i\big(f(i)\big)$, $i \in J_1$, $b^{(i)}$ to $z_{n_i}\big(f(i)\big)$, $i \in J_3$, $c^{(i)}$ to μ_i, $i \in J_3$ and $d^{(j)}$ to σ_j, $j = i, i+1, \ldots, n_i - 1$ for $i \in J_3$. These columns have the following $2N + 2m$ components:

$$a_k^{(i)} = \begin{cases} \delta_{ik} - p_{ik}(f) & k = 1, 2, \ldots, N \\ 0 & k = N+1, N+2, \ldots, 2N+2m \end{cases} \tag{8.120}$$

$$b_k^{(i)} = \begin{cases} -p_k(f) & k = 1, 2, \ldots, N \\ \delta_{n_i, k-N} & k = N+1, N+2, \ldots, N+m \\ 0 & k = N+m+1, N+m+2, \ldots, 2N+2m \end{cases} \tag{8.121}$$

$$c_k^{(i)} = \begin{cases} \delta_{ik} & k = 1, 2, \ldots, N \\ -\delta_{i,k-N} & k = N+1, N+2, \ldots, N+m \\ 0 & k = N+m+1, N+m+2, \ldots, 2N+2m \end{cases} \tag{8.122}$$

$$d_k^{(j)} = \begin{cases} 0 & k = 1, 2, \ldots, N \\ \delta_{j,k-N} - \delta_{j,k-N-1} & k = N+1, N+2, \ldots, N+m \\ 0 & k = N+m+1, N+m+2, \ldots, 2N+2m \end{cases} \qquad j = i, i+1, \ldots, n_i - 1 \tag{8.123}$$

Since $(y, \mu, z, \sigma, x, \lambda, w, \rho)$ is an extreme optimal solution of (8.101) and the above columns correspond to strictly positive variables, these columns are linearly independent.

Let p be the number of elements in $\cup_{i \in J_3} \{i, i+1, \ldots, n_i - 1\}$. Then, there are $q = |J_1| + 2 \cdot |J_3| + p$ independent vectors. Notice that all columns have zeros in the last $N + m$ components. Since an ergodic class is closed, the components $k \notin J$, $1 \le k \le N$ of the vectors are zero, because for $i \in J$, $\delta_{ik} = p_{ik}(f) = 0$. Furthermore, we observe that the components $N + k$, $1 \le k \le m$ are zero, except $\{n_i \mid i \in J_3\}$ (in $b^{(i)}$), $\{i \mid i \in J_3\}$ (in $c^{(i)}$) and $\cup_{i \in J_3} \{i, i+1, \ldots, n_i - 1\}$ (in $d^{(j)}$). Hence, there are at most $|J| + |J_3| + p = |J_1| + 2 \cdot |J_3| + p = q$ components (of the $2N = 2m$ components) which can be positive.

Consider the contracted vectors, obtained from $a^{(i)}, b^{(i)}, c^{(i)}$ and $d^{(j)}$, by deleting the components that are 0 in all vectors. Then, the q contracted vectors have at most q components and are still independent, i.e. they have exactly q components and the corresponding matrix is nonsingular. On the other hand, the components of each vector add up to 0, which contradicts the nonsingularity. This completes the proof of this lemma. □

Theorem 8.28

The policy f^∞, defined in Lemma 8.21, is an average optimal policy.

Proof

Let (x, w, y, z) be optimal solution of (8.100). Then, by Theorem 8.27, $x = \phi$, and, by Lemma 8.21 part (3), $\phi = P(f)\phi$. Consequenty, $\phi = P^*(f)\phi$, where $P^*(f)$ is the stationary matrix of $P(f)$. Since the states of $S \backslash S_+$ are transient in the Markov chain with transition matrix $P(f)$, see Lemma 8.23, we have $p^*_{ik}(f) = 0$ for every $i \in S,\ k \notin S_+$. Hence, we can write by Lemma 8.21 part (2),

$$\begin{aligned}
\phi_i(f^\infty) &= \{P^*(f)r(f)\}_i = \textstyle\sum_{k\in S} p^*_{ik}(f)r_k(f) = \sum_{k\in S_+} p^*_{ik}(f)r_k(f) \\
&= \textstyle\sum_{k\in S_+} p^*_{ik}(f)\{\phi_k + \sum_{j=1}^N \{\delta_{kj} - p_{kj}(f)\}y_j\} \\
&= \{P^*(f)\phi\}_i + \big\{P^*(f)\{I - P(f)\}y\big\}_i = \{P^*(f)\phi\}_i = \phi_i,\ i \in S,
\end{aligned}$$

i.e. f^∞ is an average optimal policy. □

8.7.6 Examples (part 2)

1. Replacement problem

Consider the following replacement problem:
state space $S = \{0, 1, \ldots, N\}$; action sets $A(i) = \{r, 1, \ldots, N\}$, where r is the action 'retain' (keep the item for at least one more time period) and action $1 \le a \le N$ means that we replace the item for another item of state a. The automobile replacement problem of Section 8.7.2 is an example of a replacement problem.

Consider the following relevant data:

s_a = the cost of buying an item of state $a,\ 1 \le a \le N$;
u_i = the trade-in value of an item of state $i,\ 1 \le i \le N$;
t_a = the expected cost of operating an item of state a for one time period, $1 \le a \le N$;
p_{ij} = the probability that an item of state i is tranfered to state j in one time period, $1 \le i, j \le N$.

The standard MDP for this model has the rewards and transition probabilities:

$$r_i(a) = \begin{cases} -t_i & i = 1, 2, \ldots, N;\ a = r \\ u_i - (s_a + t_a) & i = 1, 2, \ldots, N;\ a = 1, 2, \ldots, N \end{cases}$$

$$p_{ij}(a) = \begin{cases} p_{ij} & i, j = 1, 2, \ldots, N;\ a = r \\ p_{aj} & i, j = 1, 2, \ldots, N;\ a = 1, 2, \ldots, N \end{cases}$$

The standard linear progam (8.102) has $2 \cdot \sum_{i \in S} \sum_{a \in A(i)} = 2N(N+1)$ constraints and $2N$ variables. For the reduced formulation as separable problem, we obtain:

$$S_1 = S;\ S_2 = \emptyset;\ A_1(i) = \{1, 2, \ldots, N\},\ 1 \leq i \leq N;\ A_2(i) = \{r\},\ 1 \leq i \leq N.$$

Notice that $m = N,\ B(i) = \emptyset,\ 1 \leq i \leq N-1$ and $B(N) = \{1, 2, \ldots, N\}$. From the constraints of (8.101) it follows that

$$\rho_j = \sum_{k=1}^{j} \lambda_k \text{ and } \sigma_j = \sum_{k=1}^{j} \mu_k + j \text{ for } j = 1, 2, \ldots, N-1.$$

Hence, the variables ρ_j and σ_j can be deleted from (8.101) for all j. Then, program (8.101) can be formulated as

$$max \sum_{i=1}^{N} -t_i x_i + \sum_{i=1}^{N} u_i \lambda_i - \sum_{a=1}^{N} (s_a - t_a) w_a \tag{8.124}$$

subject to the constraints

$$\begin{array}{rcrcrcrcl}
\sum_{i=1}^{N}\{\delta_{ij} - p_{ij}\} y_i & + & \mu_j & - & \sum_{a=1}^{N} p_j(a) z_a & + & x_j & = & 1,\ 1 \leq j \leq N \\
& - & \sum_{j=1}^{N} \mu_j & + & \sum_{a=1}^{N} w_a & + & \sum_{a=1}^{N} z_a & = & N \\
\sum_{i=1}^{N}\{\delta_{ij} - p_{ij}\} x_i & + & \lambda_j & - & \sum_{a=1}^{N} p_j(a) w_a & & & = & 0,\ 1 \leq j \leq N \\
& - & \sum_{j=1}^{N} \lambda_j & + & \sum_{a=1}^{N} w_a & & & = & 0
\end{array}$$

$x_i,\ y_i,\ z_a,\ w_a,\ \lambda_i,\ \mu_i \geq 0$ for all i and a.

The last relation $-\sum_{j=1}^{N} \lambda_j + \sum_{a=1}^{N} w_a = 0$ can be deleted, because is implied by the previous set of equalities, namely: $\sum_{j=1}^{N} \{ \sum_{i=1}^{N}\{\delta_{ij} - p_{ij}\} x_i + \lambda_j - \sum_{a=1}^{N} p_j(a) w_a \} = \sum_{j=1}^{N} \lambda_j - \sum_{a=1}^{N} w_a$. Hence, the linear program becomes

$$max \sum_{i=1}^{N} -t_i x_i + \sum_{i=1}^{N} u_i \lambda_i - \sum_{a=1}^{N} (s_a - t_a) w_a \tag{8.125}$$

subject to the constraints

$$\begin{array}{rcrcrcrcl}
\sum_{i=1}^{N}\{\delta_{ij} - p_{ij}\} y_i & + & \mu_j & - & \sum_{a=1}^{N} p_j(a) z_a & + & x_j & = & 1,\ 1 \leq j \leq N \\
& - & \sum_{j=1}^{N} \mu_j & + & \sum_{a=1}^{N} w_a & + & \sum_{a=1}^{N} z_a & = & N \\
\sum_{i=1}^{N}\{\delta_{ij} - p_{ij}\} x_i & + & \lambda_j & - & \sum_{a=1}^{N} p_j(a) w_a & & & = & 0,\ 1 \leq j \leq N
\end{array}$$

$x_i,\ y_i,\ z_a,\ w_a,\ \lambda_i,\ \mu_i \geq 0$ for all i and a.

This linear program has $6N$ variables and $2N+1$ constraints. Let $(y, \mu, z, x, \lambda, w)$ be an extreme optimal solution of program 8.125. An optimal action in state $i \in S$, as defined in Lemma 8.21, becomes for this replacement problem:

If $x_i > 0$ or $x_i = \lambda_i = 0$ and $y_i > 0$: take the action r (retain).

If $x_i = 0$ and $\lambda_i > 0$ or $x_i = \lambda_i = y_i = 0$ and $\sum_{a=1}^{N} w_a > 0$: take an action a with $w_a > 0$.

If $x_i = \lambda_i = y_i = \sum_{a=1}^{N} w_a = 0$: take an action a with $z_a > 0$.

2. Inventory problem

Consider the inventory problem of Section 8.7.2. We have seen that there are $N+1$ states and that the total number of decisions is equal to $\frac{1}{2}(N+1)(N+2)$. Therefore, the standard LP formulation (8.102) has $(N+1)(N+2)$ constraints and $2(N+1)$ variables. In the reduced formulation of this separable problem, we have

$$S_1 = \{0, 1, \ldots, N-1\};\ S_2 = \{N\}.$$

$$A_1(i) = \{i+1, i+2, \ldots, N\} \ \rightarrow\ B(i) = \{i+1\},\ 0 \leq i \leq N-1;\ A_2(i) = \{i\},\ 0 \leq i \leq N.$$

The dual linear program (8.101) for this inventory problem becomes

$$max \sum_{i=0}^{N} -h_i x_i + \sum_{i=0}^{N-1} (-K + ci)\lambda_i + \sum_{i=0}^{N-1} \{-c(i+1) - h_{i+1}\} w_{i+1} \tag{8.126}$$

subject to the constraints

$$\begin{array}{rcrcrcrcl}
y_0 - \sum_{i=0}^{N} \{\sum_{k \geq i} p_k\} y_i & + & \mu_0 & - & \sum_{i=0}^{N-1} \{\sum_{k \geq i+1} p_k\} z_{i+1} & + & x_0 & = & 1 \\
y_j - \sum_{i=j}^{N} p_{i-j} y_i & + & \mu_j & - & \sum_{i=j}^{N} p_{i-j} z_i & + & x_j & = & 1,\ 1 \leq j \leq N-1 \\
(1-p_0) y_N & & & - & p_0 z_N & + & x_N & = & 1 \\
\sigma_j - \sigma_{j-1} & - & \mu_j & + & w_{j+1} & + & z_{j+1} & = & 1,\ 0 \leq j \leq N-1 \\
x_0 - \sum_{i=0}^{N} \{\sum_{k \geq i} p_k\} x_i & + & \lambda_0 & - & \sum_{i=0}^{N-1} \{\sum_{k \geq i+1} p_k\} w_{i+1} & & & = & 0 \\
x_j - \sum_{i=j}^{N} p_{i-j} x_i & + & \lambda_j & - & \sum_{i=j}^{N} p_{i-j} w_i & & & = & 0,\ 1 \leq j \leq N-1 \\
(1-p_0) x_N & & & - & p_0 w_N & & & = & 0 \\
\rho_j - \rho_{j-1} & - & \lambda_j & + & w_{j+1} & & & = & 0,\ 0 \leq j \leq N-1
\end{array}$$

$\rho_{-1} = \rho_{N-1} = \sigma_{-1} = \sigma_{N-1} = 0;\ x_i,\ y_i,\ z_i,\ w_i,\ \lambda_i,\ \mu_i,\ \rho_i,\ \sigma_i \geq 0$ for all i.

This linear program has $8N$ variables and $2(2N+1)$ constraints. Let $(y, \mu, z, \sigma, x, \lambda, w, \rho)$ be an extreme optimal solution of program (8.126). An optimal action in state $i \in S$, as defined in Lemma 8.21, where $m_i = min\{j \geq i+1 \mid w_j > 0\}$ and $n_i = min\{j \geq i+1 \mid w_j + z_j > 0\}$, becomes for this inventory problem:

If $x_i > 0$: no order.

If $x_i = 0$ and $\lambda_i > 0$: order $m_i - 1$ items.

If $x_i = \lambda_i = 0$ and $y_i > 0$: no order.

If $x_i = \lambda_i = y_i = 0$: order $n_i - 1$ items.

Remark

In the case that the optimal policy is an (s, S)-policy, the underlying Markov chain is unichained. Then, a linear program with $4N$ variables and $2N+2$ constraints suffices (see Exercise 8.11).

8.8 Exercises

Exercise 8.1

a. Show for the case in which i and j are the two smallest indices of $C_0(x)$ the nonpositivity of the inductive hypothesis $H(m+1)$ (see (8.22)), i.e. assuming $H(m)$ show that $\mu_j\{T^{m+1}(1_j, x) - T^{m+1}(0_j, x)\} \leq 0$.

b. Show for the case $f_*(x) < i < j$ of the inductive hypothesis $H(m+1)$.

Exercise 8.2

Consider the following production and inventory control model without backlogging:

$T = 5;\ D_1 = 1,\ D_2 = 4,\ D_3 = 5,\ D_4 = 3,\ D_5 = 1.$

$$c_t(a) = \begin{cases} 0 & \text{if } a = 0 \\ 7 & \text{if } a \geq 1 \end{cases},\ 1 \leq t \leq T;\ h_t(i) = i,\ i \geq 0,\ 1 \leq t \leq T.$$

Compute an optimal production plan.

Exercise 8.3

Consider the production and inventory control model of Exercise 8.2 with backlogging.

Let the shortage cost functions be $h_t(i) = -2i,\ i \leq 0,\ 1 \leq t \leq T$.

Compute an optimal production plan.

Exercise 8.4

Consider the following inventory control model with a single-critical-number optimal policy.

Let $s = 2;\ k = 3;\ R = 5;\ \alpha = 0.9;\ T = 4$.

The demand is as follows:

t	$p_t(0)$	$p_t(1)$	$p_t(2)$	$p_t(3)$	$p_t(4)$	$p_t(5)$
1	0.2	0.3	0.3	0.1	0.1	0.0
2	0.0	0.1	0.2	0.3	0.3	0.1
3	0.0	0.1	0.2	0.3	0.3	0.1
4	0.2	0.3	0.3	0.1	0.1	0.0

Compute an optimal single-critical-number policy.

Exercise 8.5

Consider the following inventory control model with fixed ordering cost, in which backlogging is not allowed.

$T = 4;\ \alpha = 0.9;\ k_t = 3,\ 1 \leq t \leq 4;\ K_t = 1,\ 1 \leq t \leq 5;\ R_t = 5,\ 1 \leq t \leq 4;\ e(i) = 2i,\ i \geq 0.$

$h_t(a) = (1-\alpha)k_t a,\ 1 \leq t \leq 4.$

The demand is as follows:

t	$p_t(0)$	$p_t(1)$	$p_t(2)$	$p_t(3)$	$p_t(4)$	$p_t(5)$
1	0.2	0.3	0.3	0.1	0.1	0.0
2	0.0	0.1	0.2	0.3	0.3	0.1
3	0.0	0.1	0.2	0.3	0.3	0.1
4	0.2	0.3	0.3	0.1	0.1	0.0

Compute an optimal policy.

Exercise 8.6

Show that $X \geq_{st} Y$ implies $X^+ \geq_{st} Y^+$ and $X^- \leq_{st} Y^-$.

Exercise 8.7

Show that $\mathbb{E}\{C_{1,2}\} \leq \mathbb{E}\{C_{2,1}\} \Leftrightarrow \lambda_1 - \mu_1 \geq \lambda_2 - \mu_2$, where $\mathbb{E}\{C_{1,2}\}$ is defined in section 8.5.7.

Exercise 8.8

Consider the model of Example 8.1 with N sequences of nonnegative numbers. Let for any k the sequence $\{x_n^k \mid n = 1, 2, \dots\}$ be nonincreasing in n. Show that the policy that chooses the sequence with the largest next reward (such policy is called a *myopic policy*) is optimal.

Exercise 8.9

Consider the model of Example 8.1 with three sequences: $x^1 = \{3, 2, 4, 0, 0, \dots\}$, $x^2 = \{2, 3, 2, 0, 0, \dots\}$ and $x^3 = \{2, 1, 4, 0, 0, \dots\}$. Let $\alpha = 0.5$. What is the optimal order of the sequences to maximize $\sum_{t=1}^{\infty} \alpha^{t-1} R_t$.

Exercise 8.10

Consider a multi-armed bandit problem with three projects and with discount factor $\alpha = \frac{1}{2}$.
The data of the projects are:

Project 1: $S_1 = \{1, 2, 3, 4\}$; $r_1^1 = 4$, $r_2^1 = 2$, $r_3^1 = 4$, $r_4^1 = 0$.
$p_{12}^1 = p_{23}^1 = p_{34}^1 = p_{44}^1 = 1$ (the other transition probabilities are 0).

Project 2: $S_2 = \{1, 2, 3, 4\}$; $r_1^2 = 2$, $r_2^2 = 6$, $r_3^2 = 2$, $r_4^2 = 0$.
$p_{12}^2 = p_{23}^2 = p_{34}^2 = p_{44}^2 = 1$ (the other transition probabilities are 0).

Project 3: $S_3 = \{1, 2, 3, 4\}$; $r_1^3 = 3$, $r_2^3 = 3$, $r_3^3 = 4$, $r_4^4 = 0$.
$p_{12}^3 = p_{23}^3 = p_{34}^3 = p_{44}^3 = 1$ (the other transition probabilities are 0).

a. Determine the 12 Gittins indices by the interpretation with stopping times.

b. Determine the 12 Gittins indices by the parametric linear programming method.

c. Determine the 12 Gittins indices by the restart-in-k method.

d. Determine the 12 Gittins indices by the largest-remaining-index method.

e. If the starting state is $(1, 1, 1)$, i.e. in each project we start in state 1, what will be the sequence of the projects in an optimal policy?

Exercise 8.11
Consider the inventory model as described in Section 8.7.6. Show that in the unichain case a linear program with $4N$ variables and $2N + 2$ constraints suffices.

Exercise 8.12
Consider the *totally separable problem*, i.e. an MDP with $S = \{1, 2, \ldots, N\}$; $A(i) = \{1, 2, \ldots, M\}$, $i \in S$; $r_i(a) = s_i + t_a$, $(i, a) \in S \times A$ and $p_{ij}(a) = p_j(a)$, $(i, a) \in S \times A$, $j \in S$.
Let the action a_* be defined by $t_{a_*} + \sum_{j=1}^{N} p_{a_* j} s_j = max_{1 \le a \le M} \{t_a + \sum_{j=1}^{N} p_{aj} s_j\}$.
Show that the policy f_*^∞ with $f(i) = a_*$, $i \in S$, is an average optimal policy for this totally separable problem.

8.9 Bibliographic notes

The general replacement model of Section 8.1.1 is strongly related to a paper by Gal ([77]), in which paper the method of policy iteration was considered. With the same approach the average reward case for an irreducible MDP can be treated. The replacement model with increasing deterioration, that has a control-limit optimal policy, appears in Derman ([54]). The separable replacement model was discussed in Sobel ([188]). It may also be viewed as a special case of the SER-SIT game (see ([147]).

The surveillance-maintenance-replacentent problem is taken Section 9.2 from Derman ([55]). The problem of optimal repair allocation in a series system appears in [115]. We follow a proof contributed by Weber (personal communication).

The section production and inventory control is taken from Denardo ([49]): Chapter 5 (for our sections 8.3.1 and 8.3.2), Chapter 6 (for our section 8.3.3) and Chapter 7 (for our section 8.3.4). The production control problem has received considerable attention in the literature. Dynamic programming formulations for the concave-cost case were due initially to Wagner and Whithin ([223]), and, independently, to Manne ([133]). Extensions to backlogging are due to Zangwill ([239], [240]) and Manne and Veinott ([135]). Work on single-critical-number policies include Bellman, Glicksberg and Gross ([12]), Karlin ([114]) and Veinott ([213], [215]. The notion that the ordering cost can be absorbed into the holding cost may be implicit in Beckmann ([10]). Scarf ([172]) analyzed an inventory model with set-up costs, backlogging and convex operating costs. He introduced K-convexity and used it to show that an (s, S) policy is optimal. Veinott ([215]) analysed this model with quasi-convex costs. Porteus [153]) introduced K-quasi-convexity.

The queueing control models are taken from Sennott ([179]) with the exception of the admission control of an $M/M/1$ queue model which can be found in Puterman ([157]. Lippman ([129]) applies uniformization to characterize optimal policies in several exponential queueing control

systems. Serfozo ([181]) formalizes this approach in the context of countable-state continuous-time models. Bertsekas ([15]) and Walrand ([224]) contain many interesting applications of the use of uniformization in queueing control models.

The material of section 8.5 is taken from Ross [170] (sections 8.5.1, 8.5.5, 8.5.6 and 8.5.7) and Walrand [224] (sections 8.5.2, 8.5.3, 8.5.4 and 8.5.5). The work of section 8.5.1 appeared in [58]. The classical μc-rule is given by Cox and Smith ([37]). The proof of Theorem 8.10 for the optimality of the μc-rule is due to Buyukkoc, Varaiya and Walrand ([29]). The optimality of the threshold policy in section 8.5.3 was shown by Lin and Kumar ([128]). The first to prove the optimality of the SQP was Winston ([238]), in 1977. The proof of Theorem 8.13 is a variant of a proof given by Ephremides, Varaiya and Walrand ([64]). Theorem 8.14 is from Ross ([170]). Theorem 8.15 is due to Bruno, Downey and Frederickson [28]) and Theorem 8.16 to Glazebrook ([81]). The alternative proof of the optimality of the $LEFT$ policy is from Pinedo and Weiss ([149]). For the work on stochastic minimizing the makespan or the time until one of the processors is idle we refer to Weber ([225]). The tandem queue model can be found in Weiss [227]), which presents a nice survey of multiserver scheduling models. Another survey of such models is given by Pinedo and Schrage ([148]). For dynamic programming and stochastic scheduling we refer also to Koole ([122]).

The fundamental contribution on the multi-armed bandit problem, the optimality of the index policy, is due to Gittins ([79] and [80]). The presenttion of this result as formulated in the proofs of Lemma 8.15 and Theorem 8.22 is taken from Ross ([170]). Other proofs of this theorem are given by Whittle ([236] and [237]), who introduced the term Gittins index in honour to Gittins, Katehakis and Veinott ([117]), Weber ([226]), Tsitsiklis ([197] and [198]) and Weiss ([228]), who in fact established an index theorem for the more general branching bandits model. Bertsimas and Nino-Mora ([16]) provided a proof for many other classes of multi-armed bandit problems.
The parametric linear programming method with complexity $\mathcal{O}(N^3)$ was proposed by Kallenberg ([109]). He improved an order $\mathcal{O}(N^4)$ method of Chen and Katehakis ([31]), who have introduced the linear program (8.80). In [142] Nino-Mora presents a fast-pivoting algorithm that computes the N Gittins indices in the discounted and undiscounted case by performing $\frac{2}{3}N^3 + \mathcal{O}(N^2)$ arithmetic operations.
The interpretation as restart-in-k problem is given by Katehakis and Veinott ([117]) and the method of the largest-remaining-index rule is due to Varaiya, Walrand and Buyukkoc ([212]).
Other contributions on this subject are e.g. from Ben-Israel and Fläm ([13]), who use a bisection method to solve the equation (8.61), Katehakis and Rothblum ([116], who considered the problem under alternative optimality criteria, namely sensitive discount optimality, average reward optimality and average overtaking optimality, and Glazebrook and Owen ([82]).

De Ghellinck and Eppen ([40]) examined separable MDPs with the discounted rewards as optimality criterion. They streamlined the linear program of D'Epenoux ([53]). Denardo introduced in [45] the notion of zero-time transitions. Discounted and averaging versions (for the unichain

case) are then shown to yield policy iteration and linear programming formulations. In the discounted case, the linear program is identical to that of De Ghellinck and Eppen. Kallenberg ([110]) has shown that for the average reward criterion also in the multichain case a simpler linear program can be used to solve the original problem. The automobile replacement problem was first considered by Howard ([101]). The totally separable problem of Exercise 8.12 is a special case of a stochastic game studied in [147] and [188].

Chapter 9

Other topics

9.1 Additional constraints

9.1.1 Introduction

Formulating MDPs only in terms of the standard utility functions can be quite insufficient. Instead of introducing a single utility that has to be maximized (minimized) we often consider a situation where one type of profit (costs) has to be maximized (minimized) while keeping other types of rewards (costs) above (below) some given bounds. The first reference in this area is a paper of Derman and Klein ([56]). They consider an inventory problem for which the total costs are minimized under the constraint that the shortage is bounded by a given number. We will first present two examples of constrained problems from telecommunication.

Telecommunication networks are designed to enable simultaneous transmission of heterogeneous types of information. At the access to the network, or at the nodes within the network itself, the different types of traffic typically compete for a shared resource. Typical performance measures are the transmission delay, the throughputs, probabilities of losses of packets, etc.. Then, several constrained MDP problems can be considered, e.g.

(1) The maximization of the throughput, subject to constraints on its delay.
A tradeoff exists between achieving high throughput, on the one hand, and low expected delays on the other.

(2) Dynamic control of access of different traffic types.
In this model the problem is considered where several different traffic types compete for some resource; some weighted sum of average delays of some traffic is to be minimized, whereas for some other traffic types, a weighted sum of average delays should be bounded by some given limit.

For constrained Markov decision problems, for short CMDP, the nice property for the standard utility functions that there exists a deterministic optimal policy doesn't hold, in general. Even optimality simultaneously for all starting states is no longer valid. Therefore, we will optimize

with respect to a given initial distribution β, i.e. β_j is the probability that state j is the starting state, $j \in S$. A special case is $\beta_j = \delta_{ij}$, i.e. that state i is the (fixed) starting state.

In many cases the reward or costs functions are specified in terms of expectations of some functions of the *state-action probabilities* $x_{ia}^R(t)$, defined for any policy R by

$$x_{ia}^R(t) = \sum_{j \in S} \beta_j \cdot \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\},\ t = 1, 2, \ldots. \tag{9.1}$$

9.1.2 Infinite horizon and discounted rewards

For the additional constraints we assume that, besides the immediate costs $r_i(a)$, there are for $k = 1, 2, \ldots, m$ also certain immediate costs $c_i^k(a),\ (i, a) \in S \times A$. A policy R is called a *feasible* policy for a CMDP if the total expected discounted costs over the infinite horizon, denoted for the k-th cost function as $c_k^\alpha(R)$ and defined by

$$c_k^\alpha(R) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{i,a} x_{ia}^R(t) c_i^k(a),$$

is at most $b_k,\ k = 1, 2, \ldots, m$. An *optimal policy* R^* is a feasible policy that maximizes $v^\alpha(R)$, defined by

$$v^\alpha(R) = \sum_j \beta_j \cdot v_j^\alpha(R) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{i,a} x_{ia}^R(t) r_i(a),$$

over all feasible policies R, i.e.

$$v^\alpha(R^*) = sup_R \{v^\alpha(R) \mid c_k^\alpha(R) \le b_k,\ k = 1, 2, \ldots, m\}. \tag{9.2}$$

Define $x_{ia}^\alpha(R) = \sum_{t=1}^{\infty} \alpha^{t-1} x_{ia}^R(t),\ (i, a) \in S \times A$, as the *total discounted state-action frequenties*. Then, $v^\alpha(R) = \sum_{i,a} x_{ia}^\alpha(R) r_i(a)$ and $c_k^\alpha(R) = \sum_{i,a} x_{ia}^\alpha(R) c_i^k(a),\ k = 1, 2, \ldots, m$.

Define the vector sets $K,\ K(M),\ K(S),\ K(D)$ and P, with components $(i, a) \in S \times A$ by

$$\begin{aligned}
K &= \{x^\alpha(R) \mid R \text{ is an arbitrary policy}\};\\
K(M) &= \{x^\alpha(R) \mid R \text{ is a Markov policy}\};\\
K(S) &= \{x^\alpha(R) \mid R \text{ is a stationary policy}\};\\
K(D) &= \{x^\alpha(R) \mid R \text{ is a deterministic policy}\};\\
P &= \left\{ x \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_{ia} &=& \beta_j,\ j \in S \\ x_{ia} &\ge& 0,\ (i, a) \in S \times A \end{array} \right\}.
\end{aligned}$$

For any $|S \times A|$-vector $x \in P$, we define an $|S|$-vector, also denoted by x, by $x_i = \sum_a x_{ia},\ i \in S$. From the context it will be clear whether an $|S \times A|$-vector x or an $|S|$-vector x is meant.

Theorem 9.1
$K = K(M) = K(S) = \overline{K(D)} = P$, where $\overline{K(D)}$ is the closed convex hull of the finite set of vectors $K(D)$.

Proof

The equality $K = K(M)$ follows directly from Theorem 1.1. Furthermore, it is obvious that $K(D) \subseteq K(S) \subseteq K(M) \subseteq K$. We first show that $K \subseteq \overline{K(D)}$, then $K = K(M) = \overline{K(S)} = \overline{K(D)}$, where $\overline{K(S)}$ is the closed convex hull of the infinite set of vectors $K(S)$, and finally we show that $K(S) = P$, which implies - because P is a closed convex set - that $\overline{K(S)} = K(S)$.

For the proof of $K \subseteq \overline{K(D)}$, suppose the contrary. Then, there exists a policy R such that $x^\alpha(R) \in K$ and $x^\alpha(R) \notin \overline{K(D)}$. Since $\overline{K(D)}$ is a closed convex set, it follows from the Separating Hyperplane Theorem (see e.g. [113] pp.397–398) that there are coefficients $r_i(a),\ (i,a) \in S \times A$, such that

$$\sum_{i,a} x^\alpha_{ia}(R) r_i(a) > \sum_{i,a} x_{ia} r_i(a) \text{ for all } x \in \overline{K(D)}. \tag{9.3}$$

Consider the discounted MDP with immediate rewards $r_i(a),\ (i,a) \in S \times A$. We have seen in Chapter 3 that there exists an optimal policy $f^\infty \in C(D)$. Because $x^\alpha(R) \in K$, we can write

$$\textstyle\sum_{i,a} x^\alpha_{ia}(R) r_i(a) = v^\alpha(R) \le v^\alpha(f^\infty) = \sum_{i,a} x^\alpha_{ia}(f^\infty) r_i(a),$$

which contradicts (9.3). Hence, we have shown that $K \subseteq \overline{K(D)}$, and consequently

$$K(D) \subseteq K(S) \subseteq K(M) = K \subseteq \overline{K(D)} \text{ and } \overline{K(S)} = \overline{K(D)}.$$

Next, we will show that $\overline{K(D)} \subseteq K(M)$, implying $K = K(M) = \overline{K(S)} = \overline{K(D)}$. Take any $x \in \overline{K(D)}$. Let $C(D) = \{f^\infty_1, f^\infty_2, \ldots, f^\infty_n\}$. Then, $x_{ia} = \sum_{k=1}^n p_k\, x^\alpha_{ia}(f^\infty_k),\ (i,a) \in S \times A$ for certain $p_k \ge 0$ with $\sum_{k=1}^n p_k = 1$. By Theorem 1.1, there exists a policy $R \in C(M)$ satisfying

$$\textstyle\sum_{j \in S} \beta_j \cdot \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} = \sum_{j \in S} \beta_j \cdot \sum_{k=1}^n p_k\, \mathbb{P}_{f^\infty_k}\{X_t = i,\ Y_t = a \mid X_1 = j\},$$

for all $(i,a) \in S \times A$ and $t = 1, 2, \ldots$. Hence,

$$\begin{aligned} x_{ia} &= \textstyle\sum_{k=1}^n p_k\, x^\alpha_{ia}(f^\infty_k) = \sum_{k=1}^n p_k \sum_{t=1}^\infty \alpha^{t-1} x^{f^\infty_k}_{ia}(t) = \sum_{t=1}^\infty \alpha^{t-1} \sum_{k=1}^n p_k\, x^{f^\infty_k}_{ia}(t) \\ &= \textstyle\sum_{t=1}^\infty \alpha^{t-1} \sum_{k=1}^n p_k \sum_{j \in S} \beta_j \cdot \mathbb{P}_{f^\infty_k}\{X_t = i,\ Y_t = a \mid X_1 = j\} \\ &= \textstyle\sum_{t=1}^\infty \alpha^{t-1} \sum_{j \in S} \beta_j \cdot \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} = x_{ia}(R),\ (i,a) \in S \times A. \end{aligned}$$

Therefore, $x = x(R) \in K(M)$.

Finally, we we show that $K(S) = P$. For each $x \in P$, let $\pi^\infty \in C(S)$ be defined by

$$\pi_{ia} = \frac{x_{ia}}{x_i} \text{ if } x_i = \textstyle\sum_a x_{ia} > 0 \text{ and arbitrary if } x_i = 0. \tag{9.4}$$

Then, $\pi_{ia} x_i = x_{ia}$ for all $(i,a) \in S \times A$. Since $x \in P$, we can write

$$\begin{aligned} \beta_j &= \textstyle\sum_a x_{ja} - \alpha \sum_{(i,a)} p_{ij}(a) x_{ia} = x_j - \alpha \sum_{(i,a)} p_{ij}(a) \pi_{ia} x_i \\ &= \textstyle x_j - \alpha \sum_i p_{ij}(\pi) x_i,\ j \in S, \end{aligned}$$

or, in vector notation, $\beta^T = x^T\{I - \alpha P(\pi)\}$, implying $x^T = \beta^T\{I - \alpha P(\pi)\}^{-1}$, i.e.

$$x_i = \textstyle\sum_{t=1}^\infty \alpha^{t-1} \sum_{j \in S} \beta_j \cdot \mathbb{P}_{\pi^\infty}\{X_t = i \mid X_1 = j\},\ i \in S.$$

Hence,

$$\begin{aligned} x_{ia} &= x_i\pi_{ia} = \sum_{t=1}^{\infty}\alpha^{t-1}\sum_{j\in S}\beta_j\cdot\mathbb{P}_{\pi^\infty}\{X_t=i,\ Y_t=a\mid X_1=j\} \\ &= x_{ia}^{\alpha}(\pi^\infty),\ (i,a)\in S\times A. \end{aligned}$$

showing $P\subseteq K(S)$. Conversely, take any $x^\alpha(\pi^\infty)\in K(S)$. Then,

$$x_{ia}^{\alpha}(\pi^\infty)=\sum_{t=1}^{\infty}\alpha^{t-1}\sum_{j\in S}\beta_j\cdot\{P(\pi)^{t-1}\}_i\cdot\pi_{ia}=\left\{\beta^T\cdot\{\sum_{t=1}^{\infty}\{\alpha P(\pi)\}^{t-1}\right\}_i\cdot\pi_{ia}\geq 0,\ (i,a)\in S\times A,$$

which can be written as

$$x_{ia}^{\alpha}(\pi^\infty)=\left\{\beta^T\cdot\{I-\alpha P(\pi)\}^{-1}\right\}_i\cdot\pi_{ia}\ \text{ or }\ x_{ia}^{\alpha}(\pi^\infty)=x_i^{\alpha}(\pi^\infty)\cdot\pi_{ia},$$

where $x_i^\alpha(\pi^\infty)=\left\{\beta^T\cdot\{I-\alpha P(\pi)\}^{-1}\right\}_i$. From this expression it follows that

$$\begin{aligned} \sum_{(i,a)}\{\delta_{ij}-\alpha p_{ij}(a)\}x_{ia}^{\alpha}(\pi^\infty) &= x_j^{\alpha}(\pi^\infty)-\alpha\sum_i\{\sum_a p_{ij}(a)\pi_{ia}\}x_i^{\alpha}(\pi^\infty) \\ &= x_j^{\alpha}(\pi^\infty)-\alpha\sum_i p_{ij}(\pi)x_i^{\alpha}(\pi^\infty) = \left\{\left(x^\alpha(\pi)\right)^T\{I-\alpha P(\pi)\}\right\}_j \\ &= \left\{\beta^T\{I-\alpha P(\pi)\}^{-1}\{I-\alpha P(\pi)\}\right\}_j = \beta_j,\ j\in S. \end{aligned}$$

Hence, $x^\alpha(\pi^\infty)\in P$, completing the proof that $P=K(S)$. □

In order to solve the CMDP (9.2), we consider the following linear program

$$max\left\{\sum_{(i,a)}r_i(a)x_i(a)\ \middle|\ \begin{array}{rcl}\sum_{(i,a)}\{\delta_{ij}-\alpha p_{ij}(a)\}x_i(a) &=& \beta_j,\ j\in S\\ \sum_{(i,a)}c_i^k(a)x_i(a) &\leq& b_k,\ k=1,2,\ldots,m\\ x_i(a) &\geq& 0,\ (i,a)\in S\times A\end{array}\right\}. \tag{9.5}$$

Theorem 9.2

(1) The linear program (9.5) is infeasible if and only if the CMDP (9.2) is infeasible.

(2) If x is an optimal solution of program (9.5), then π^∞, defined by (9.4), is a stationary optimal policy for the CMDP (9.2).

Proof

(1) Assume that the linear program (9.5) is infeasible and that the CMDP (9.2) is feasible, i.e. there exists a policy R satisfying $c_k^\alpha(R)=\sum_{i,a}x_{ia}^\alpha(R)c_i^k(a)\leq b_k,\ k=1,2,\ldots,m$. Since $K=P$, there exists an $x\in P$ with $x=x(R)$. Hence x is a feasible solution of the linear program (9.5), which yields a contradiction. The reverse statement can be shown in a similar way.

(2) Let x be an optimal solution of program (9.5) and let π^∞ be defined by (9.4). Then, π^∞ is a feasible solution of the CMDP (9.2) with $v^\alpha(\pi^\infty)=\sum_{i,a}x_{ia}^\alpha(\pi^\infty)r_i(a)=\sum_{i,a}x_{ia}r_i(a)$ as value of the objective function. Let R_* be an arbitrary feasible sulution of (9.2). Then, $x^\alpha(R_*)$ is a feasible solution of (9.5) with $v^\alpha(R_*)=\sum_{i,a}x_{ia}^\alpha(R_*)r_i(a)\leq\sum_{i,a}x_{ia}r_i(a)=v^\alpha(\pi^\infty)$, i.e. π^∞ is an optimal policy of the CMDP (9.2). □

Remarks

1. If $\beta_j > 0,\ j \in S$, then it is shown in Theorem 3.18 that the mapping $\pi_{ia}(x) = \frac{x_{ia}}{\sum_a x_{ia}}$, $(i,a) \in S \times A$ is a bijection between P and $K(S)$ with as inverse mapping $x(\pi)$, defined by $x_{ia}(\pi) = \{\beta^T \cdot \{I - \alpha P(\pi)\}^{-1}\}_i \cdot \pi_{ia},\ (i,a) \in S \times A$. Furthermore, the extreme points of P correspond to the deterministic policies of $C(D)$.
2. If the linear program (9.5) is feasible, an extreme optimal solution has at most $N + m$ (the number of the constraints in (9.5)) strictly positive variables. Hence, there exists an optimal stationary policy with in at most m states a randimization.
3. Similar results can be derived for total expected rewards. For details see ([98]).

9.1.3 Infinite horizon and average rewards

Consider a similar problem as in the discounted case, but with average rewards, with respect to immediate rewards $r_i(a)$, and costs, with respect to immidiate costs $c_i^k(a),\ (i,a) \in S \times A$, for $k = 1, 2, \ldots, m$. Let β be an arbitrary initial distribution. For any policy R, let the average reward and the average k-th cost function with respect to the initial distribution β be defined by

$$\phi(R) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j\in S} \beta_j \cdot \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i(a)$$

and

$$c^k(R) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j\in S} \beta_j \cdot \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot c_i^k(a),$$

respectively. A policy R is a feasible policy for a CMDP with average rewards and costs if $c^k(R) \le b_k,\ k = 1, 2, \ldots, m$. An *optimal policy* R^* for this criterion is a feasible policy that maximizes $\phi(R)$, i.e.

$$\phi(R^*) = sup_R \{\phi(R) \mid c^k(R) \le b_k,\ k = 1, 2, \ldots, m\}. \tag{9.6}$$

For any policy R and any $T \in \mathbb{N}$, we denote the *expected state-action frequenties* in the first T periods by

$$x_{ia}^T(R) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j\in S} \beta_j \cdot \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\},\ (i,a) \in S \times A. \tag{9.7}$$

By $X(R)$ we denote the set of all limit points of the vectors $\{x^T(R),\ T = 1, 2, \ldots\}$. These limit points are limit points in the $S \times A$-dimensional vector space of vectors $x^T(R)$ with components $x_{ia}^T(R),\ (i,a) \in S \times A$. Any $x^T(R)$ satisfies $\sum_{(i,a)} x_{ia}^T(R) = 1$ and therefore also $\sum_{(i,a)} x_{ia}(R) = 1$ for all $x(R) \in X(R)$. For $\pi^\infty \in C(S)$ we have $\mathbb{P}_{\pi^\infty}\{X_t = i,\ Y_t = a \mid X_1 = j\} = \{P^{t-1}(\pi)\}_{ji} \cdot \pi_{ia}$ for all (i,a) and consequently, $lim_{T\to\infty} x_{ia}^T(\pi^\infty) = \sum_{j\in S} \beta_j \{P^*(\pi)\}_{ji} \cdot \pi_{ia}$, i.e. $X(\pi^\infty)$ consists of one element, namely $x(\pi)$, where $x_{ia}(\pi) = \{\beta^T P^*(\pi)\}_i \cdot \pi_{ia},\ (i,a) \in S \times A$. Let the policy set C_1 be the set of *convergent policies*, defined by $C_1 = \{R \mid X(R)$ consists of one element$\}$.

Hence, $C(S) \subseteq C_1$. Furthermore, define the vector sets L, $L(M)$, $L(C)$, $L(S)$ and $L(D)$ by

$$\begin{aligned} L &= \{x(R) \in X(R) \mid R \text{ is an arbitrary policy}\}; \\ L(M) &= \{x(R) \in X(R) \mid R \text{ is a Markov policy}\}; \\ L(C) &= \{x(R) \in X(R) \mid R \text{ is a convergent policy}\}; \\ L(S) &= \{x(R) \in X(R) \mid R \text{ is a stationary policy}\}; \\ L(D) &= \{x(R) \in X(R) \mid R \text{ is a deterministic policy}\}. \end{aligned}$$

General case

In the general case the Markov chain $P(f)$ for any $f^\infty \in C(D)$ may be irreducible, unichain or multichain. We will show that $L = L(M) = L(C) = \overline{L(S)} = \overline{L(D)}$. Therefore, we need that there exists a deterministic optimal policy with respect to the average rewards $\overline{\phi}(R)$, defined by

$$\overline{\phi}_j(R) = limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i(a),\ j \in S. \tag{9.8}$$

Lemma 9.1
Let $f^\infty \in C(D)$ be an optimal policy with respect to the average rewards $\phi(R)$. Then, f^∞ is also an optimal policy with respect to the average rewards $\overline{\phi}(R)$.

Proof
From Theorem 1.1 it follows that it is sufficient to prove that $\overline{\phi}(f^\infty) \geq \overline{\phi}(R)$ for all Markov policies R. Let $R = (\pi^1, \pi^2, \dots)$ be an arbitrary Markov policy. Since the value vector ϕ is superharmonic (cf. Theorem 5.15), there exists a vector $u \in \mathbb{R}^N$ such that $\phi_i \geq \sum_j p_{ij}(a)\phi_j$ and $\phi_i + u_i \geq r_i(a) + \sum_j p_{ij}(a)u_j$ for all $(i,a) \in S \times A$. Hence, $\phi \geq P(\pi^t)\phi$ and $\phi + u - P(\pi^t)\phi \geq r(\pi^t)$ for $t = 1, 2, \dots$. Consequently,

$$\begin{aligned} \textstyle\sum_{t=1}^{T} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t) &\leq \textstyle\sum_{t=1}^{T} P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1}) \cdot \{\phi + u - P(\pi^t)u\} \\ &\leq T \cdot \phi + u - P(\pi^1)P(\pi^2)\cdots P(\pi^T)u,\ T \in \mathbb{N}. \end{aligned}$$

Since $\frac{1}{T}\{u - P(\pi^1)P(\pi^2)\cdots P(\pi^T)u\} \to 0$ for $T \to \infty$, we can write

$$\begin{aligned} \overline{\phi}_j(R) &= limsup_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \{P(\pi^1)P(\pi^2)\cdots P(\pi^{t-1})r(\pi^t)\}_j \\ &\leq limsup_{T\to\infty} \tfrac{1}{T}\{T \cdot \phi + u - P(\pi^1)P(\pi^2)\cdots P(\pi^T)u\}_j = \phi_j = \phi_j(f^\infty),\ j \in S. \end{aligned}$$

□

Theorem 9.3
$L = L(M) = L(C) = \overline{LS)} = \overline{L(D)}$.

Proof

The proof has the same structure as the proof as Theorem 9.1. The equality $L = L(M)$ follows directly from Theorem 1.1. Furthermore, it is obvious that $L(D) \subseteq L(S) \subseteq L(C) \subseteq L$. We first show that $L \subseteq \overline{L(D)}$. Suppose the contrary. Then, there exists a policy R such that $x(R) \in L$ and $x(R) \notin \overline{L(D)}$. Since $\overline{L(D)}$ is a closed convex set, it follows from the Separating Hyperplane Theorem, that there are coefficients $r_i(a),\ (i,a) \in S \times A$, such that

$$\sum_{i,a} x_{ia}(R) r_i(a) > \sum_{i,a} x_{ia} r_i(a) \text{ for all } x \in \overline{L(D)}. \tag{9.9}$$

Consider the MDP with immediate rewards $r_i(a),\ (i,a) \in S \times A$. We have seen in Chapter 5 that there exists an average optimal policy $f^\infty \in C(D)$ with respect to $\phi(R)$. By Lemma 9.1, f^∞ is also average optimal with respect to $\overline{\phi}(R)$. Because $x(R) \in L$, there exists a sequence $\{T_k,\ k = 1, 2, \dots\}$ such that $x_{ia}(R) = lim_{k\to\infty}\, x_{ia}^{T_k}(R),\ (i,a) \in S \times A$. Hence,

$$\begin{aligned}
\textstyle\sum_{(i,a)} r_i(a) x_{ia}(R) &= \textstyle\sum_{(i,a)} r_i(a) \cdot lim_{k\to\infty}\, x_{ia}^{T_k}(R) \\
&= lim_{k\to\infty}\, \tfrac{1}{T_k} \textstyle\sum_{t=1}^{T_k} \sum_{j\in S} \beta_j \cdot \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i(a) \\
&\le \textstyle\sum_{j\in S} \beta_j \cdot limsup_{k\to\infty}\, \tfrac{1}{T_k} \sum_{t=1}^{T_k} \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i(a) \\
&= \textstyle\sum_{j\in S} \beta_j \cdot \overline{\phi}(R) \le \sum_{j\in S} \beta_j \cdot \overline{\phi}(f^\infty) = \sum_{(i,a)} r_i(a) x_{ia}(f^\infty),
\end{aligned}$$

which contradicts (9.9), completing the proof that $L \subseteq \overline{L(D)}$. Since $L(D) \subseteq L(S) \subseteq L(C) \subseteq L$, we obtain $\overline{L(S)} = \overline{L(D)}$. From $L(C) \subseteq L = L(M) \subseteq \overline{L(S)} = \overline{L(D)}$, it remains to show that $\overline{L(D)} \subseteq L(M) \cap L(C)$. Take any $x \in \overline{L(D)}$. Let $C(D) = \{f_1^\infty, f_2^\infty, \dots, f_n^\infty\}$. Then, $x_{ia} = \sum_{k=1}^n p_k\, x_{ia}(f_k^\infty),\ (i,a) \in S \times A$ for certain $p_k \ge 0$ with $\sum_{k=1}^n p_k = 1$.

By Theorem 1.1, there exists a policy $R \in C(M)$ satisfying

$$\textstyle\sum_{j\in S} \beta_j \cdot \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} = \sum_{j\in S} \beta_j \cdot \sum_{k=1}^n p_k\, \mathbb{P}_{f_k^\infty}\{X_t = i,\ Y_t = a \mid X_1 = j\},$$

for all $(i,a) \in S \times A$ and $t = 1, 2, \dots$. Hence,

$$\begin{aligned}
x_{ia} &= \textstyle\sum_{k=1}^n p_k\, x_{ia}(f_k^\infty) \\
&= \textstyle\sum_{k=1}^n p_k\, lim_{T\to\infty}\, \tfrac{1}{T} \sum_{t=1}^{T} \sum_{j\in S} \beta_j \cdot \mathbb{P}_{f_k^\infty}\{X_t = i,\ Y_t = a \mid X_1 = j\} \\
&= lim_{T\to\infty}\, \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \sum_{j\in S} \beta_j \cdot \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \\
&= x_{ia}(R),\ (i,a) \in S \times A.
\end{aligned}$$

Therefore, $x = x(R) \in L(M)$, and $x = lim_{T\to\infty}\, x^T(R) \in L(C)$, which completes the proof of the theorem. □

Analogously to the discounted case we introduce a polyhedron, namely

$$Q = \left\{ x \;\middle|\; \begin{array}{rcl} \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\} x_{ia} & = & 0,\ j \in S \\ \sum_a x_{ja} \;+\; \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\} y_{ia} & = & \beta_j,\ j \in S \\ x_{ia},\ y_{ia} & \ge & 0,\ (i,a) \in S \times A \end{array} \right\}.$$

Hence, Q is the projection (on the x-space) of the feasible solutions (x, y) of the dual linear program (5.24) for the computation of an average optimal policy.

Theorem 9.4
$L = Q$.

Proof
Theorem 9.3 implies that it is sufficient to show that $\overline{L(D)} = Q$. For $\pi^\infty \in C(S)$, we have $x_{ia}(\pi) = \left\{\beta^T P^*(\pi)\right\}_i \cdot \pi_{ia},\ (i,a) \in S \times A$. Then, with $y(\pi)$ defined by (5.31), we have shown in Theorem 5.17 that $\big(x(\pi), y(\pi)\big)$ is a feasible solution of dual linear program (5.24) (it can easily be checked that the proof of Theorem 5.17 is also valid when $\beta_j = 0$ for some $j \in S$).
Hence, $L(D) \subseteq L(S) \subseteq Q$. Since Q is the projection of a polyhedron, Q is also a polyhedron and consequently $\overline{L(D)} \subseteq Q$. If $x \in Q$, then it follows from the definition of Q that $x_{ia} \geq 0$ for all $(i,a) \in S \times A$ and $\sum_{(j,a)} x_{ja} = \sum_j \beta_j = 1$. Therefore, Q is a polytope, i.e. a bounded polyhedron. Hence, Q is the closed convex hull of a finite number of extreme points, and consequently it is sufficient to show that any extreme point of Q belongs to $L(D)$.
Let x^* be an arbitrary extreme point of Q and let Q^* be the closed convex hull of the extreme points of Q that are different from x^*. Then, $x^* \notin Q^*$ and, by the Separating Hyperplane Theorem, there are coefficients $r_i(a),\ S \times A$, such that

$$\sum_{i,a} r_i(a) x^*_{ia} > \sum_{i,a} r_i(a) x_{ia} \text{ for all } x \in Q^*. \tag{9.10}$$

From (9.10) it follows that any optimal solution $(\overline{x}, \overline{y})$ of

$$max \left\{ \sum_{(i,a)} r_i(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_a x_j(a) + \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} y_i(a) & = & \beta_j,\ j \in S \\ x_i(a), y_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\} \tag{9.11}$$

satisfies $\overline{x} = x^*$. Let $f^\infty_* \in C(D)$ be an average optimal policy for the MDP with immediate rewards $r_i(a),\ S \times A$. Then, by Theorem 5.18, $\big(x(f), y(f)\big)$ - defined by (5.30) and (5.31) - is an optimal solution of (9.11). Hence, $x^* = x(f) \in C(D)$, which completes the proof. □

Example 9.1
From the Theorems 9.3 and 9.4 it follows that any extreme piint of Q is an element of $L(D)$. This example will show that the converse statement is not true, in general. Furthermore, this example shows that $L(S) \neq Q$ is possible and that Q can be a real subset of

$$Q_0 = \left\{ x \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_{ia} & = & 0,\ j \in S \\ \sum_{(i,a)} x_{ia} & = & 1 \\ x_{ia} \geq 0,\ (i,a) \in S \times A & & \end{array} \right\}.$$

Consider the following MDP: $S = \{1,2,3\};\ A(1) = \{1,2\},\ A(2) = \{1,2\},\ A(3) = \{1\};$

$p_{11}(1)=0,\ p_{12}(1)=1,\ p_{13}(1)=0;\ p_{11}(2)=0,\ p_{12}(2)=0,\ p_{13}(2)=1;$
$p_{21}(1)=0,\ p_{22}(1)=1;\ p_{23}(1)=0;\ p_{21}(2)=1,\ p_{22}(2)=0;\ p_{23}(2)=0;$
$p_{31}(1)=0,\ p_{32}(1)=0,\ p_{33}(1)=1$. Take $\beta_1=\beta_2=\beta_3=\frac{1}{3}$.
Any stationary policy π^∞ induces a Markov chain with transition matrix

$$P(\pi)=\begin{pmatrix} 0 & \pi_1 & 1-\pi_1 \\ \pi_2 & 1-\pi_2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For the computation of $P^*(\pi)$ and $x(\pi)$ we distinguish between the following three cases.

Case 1: $\pi_1=1$:

$$P^*(\pi)=\begin{pmatrix} \frac{\pi_2}{1+\pi_2} & \frac{1}{1+\pi_2} & 0 \\ \frac{\pi_2}{1+\pi_2} & \frac{1}{1+\pi_2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \begin{array}{l} x_{11}(\pi)=\frac{2}{3}\cdot\frac{\pi_2}{1+\pi_2};\ x_{12}(\pi)=0 \\ x_{21}(\pi)=\frac{2}{3}\cdot\frac{1-\pi_2}{1+\pi_2};\ x_{22}(\pi)=\frac{2}{3}\cdot\frac{\pi_2}{1+\pi_2} \\ x_{31}(\pi)=\frac{1}{3} \end{array}$$

Case 2: $\pi_1\neq 1$ and $\pi_2=0$:

$$P^*(\pi)=\begin{pmatrix} 0 & \pi_1 & 1-\pi_1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \begin{array}{l} x_{11}(\pi)=0;\ x_{12}(\pi)=0 \\ x_{21}(\pi)=\frac{1}{3}\cdot(1+\pi_1);\ x_{22}(\pi)=0 \\ x_{31}(\pi)=\frac{1}{3}\cdot(2-\pi_1) \end{array}$$

Case 3: $\pi_1\neq 1$ and $\pi_2\neq 0$:

$$P^*(\pi)=\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \begin{array}{l} x_{11}(\pi)=0;\ x_{12}(\pi)=0 \\ x_{21}(\pi)=0;\ x_{22}(\pi)=0 \\ x_{31}(\pi)=1 \end{array}$$

Since in each case $x_{11}(\pi)=x_{22}(\pi),\ x_{12}(\pi)=0,\ x_{11}(\pi)+x_{12}(\pi)+x_{21}(\pi)+x_{22}(\pi)+x_{31}(\pi)=1$, we can describe $L(S)$ in the space with the nonnegative variables $x_{11}(\pi),\ x_{21}(\pi)$ and $x_{31}(\pi)=0$:
$L(S)=\{2x_{11}+x_{31}=\frac{2}{3}; x_{31}=\frac{1}{3}\}\cup\{x_{11}=0; x_{21}+x_{31}=1, \frac{1}{3}\le x_{31}\le\frac{2}{3}\}\cup\{x_{11}=x_{21}=0; x_{31}=1\}$.
The four deterministic policies correspond to $\pi_1=1,\ \pi_2=1$, $\pi_1=1,\ \pi_2=0$, $\pi_1=0,\ \pi_2=1$ and $\pi_1=0,\ \pi_2=0$: $x_{11}(f_1)=\frac{1}{3}$, $x_{21}(f_1)=0$, $x_{31}(f_1)=\frac{1}{3}$, $x_{11}(f_2)=0$, $x_{21}(f_2)=\frac{2}{3}$, $x_{31}(f_2)=\frac{1}{3}$, $x_{11}(f_3)=0$, $x_{21}(f_3)=0$, $x_{31}(f_3)=1$ and $x_{11}(f_4)=0$, $x_{21}(f_4)=\frac{1}{3}$, $x_{31}(f_4)=\frac{2}{3}$. Q is the closed convex hull of $x(f_1),\ x(f_2),\ x(f_3)$ and $x(f_4)$. Hence $x=\frac{1}{4}\{x(f_1)+x(f_2)+x(f_3)+x(f_1)\}\in Q$ and $x_{11}=\frac{1}{12},\ x_{21}=\frac{1}{4},\ x_{31}=\frac{7}{12}$ and it can easily verified that x is not an element of $L(S)$.
Since Q is the closed convex hull of $x(f_1),\ x(f_2),\ x(f_3)$ and $x(f_4)$, we have

$$Q=\left\{\begin{array}{l} x_{11}+x_{12}+x_{21}+x_{22}+x_{31}\ =\ 1 \\ x_{12}=0;\ x_{11}=x_{22};\ x_{31}\ \ge\ \frac{1}{3} \\ x_{11},x_{12},x_{21},x_{22},x_{31}\ \ge\ 0 \end{array}\right\} \text{ and } Q_0=\left\{\begin{array}{rcrcrcl} & x_{11} & + & x_{12} & - & x_{22} & = 0 \\ - & x_{11} & & & + & x_{22} & = 0 \\ & & - & x_{12} & & & = 0 \\ \multicolumn{7}{c}{x_{11}+x_{12}+x_{21}+x_{22}+x_{31}\ =\ 1} \end{array}\right\}.$$

Since $x_{11}=\frac{1}{2},\ x_{12}=0,\ x_{21}=0,\ x_{22}=\frac{1}{2},\ x_{31}=0$ belongs to Q_0 and not to Q, i.e. Q is a real subset of Q_0.

In order to solve the CMDP (9.6) we consider the linear program

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{lcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & = & 0,\ j \in S \\ \sum_a x_j(a) \ + \ \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}y_i(a) & = & \beta_j,\ j \in S \\ \sum_{(i,a)} c_i^k(a)x_i(a) & \leq & b_k,\ k = 1,2,\ldots,m \\ x_i(a), y_i(a) \geq 0,\ (i,a) \in S \times A & & \end{array} \right\}. \tag{9.12}$$

Theorem 9.5

(1) Problem (9.6) is feasible if and only if problem (9.12) is feasible.

(2) The optima of (9.6) and (9.12) are equal.

(3) If R is optimal for problem (9.6), then $x(R)$ is optimal for (9.12).

(4) Let (x,y) be an optimal solution of problem (9.12) and let $x = \sum_{k=1}^n p_k x(f_k)$, where $p_k \geq 0$ and $\sum_{k=1}^n p_k = 1$ and $f_1^\infty, f_2^\infty, \ldots, f_n^\infty$ are the stationary policies. Let $R \in C(M)$ be the policy of Theorem 1.1 that satisfies $\sum_j \beta_j \cdot \mathbb{P}_R\{X_t = i, Y_t = a \mid X_1\} = \sum_j \beta_j \cdot \sum_k p_k \cdot \mathbb{P}_{f_k^\infty}\{X_t = i, Y_t = a \mid X_1\} = \beta_j\}$ for all $(i,a) \in S \times A$ and all $t \in \mathbb{N}$. Then, R is an optimal solution of problem (9.6).

Proof

The theorems 9.3 and 9.4 imply that $Q = L(C)$. Moreover, $\phi(R) = \sum_{i,a} x_{ia}(R)r_i(a)$ for any $R \in C_1$. By these observations, the parts (1), (2) and (3) are straightforward. For the proof of part (4) we can similarly as in the proof of Theorem 9.3 show that $x = x(R)$ and $R \in C_1$. Therefore, $\phi(R) = \sum_{i,a} x_{ia}(R)r_i(a) = \sum_{i,a} x_{ia}r_i(a) =$ optimum of problem (9.12). Hence, R is an optimal policy for problem (9.6). □

To compute an optimal policy from an optimal solution (x,y) of the linear program (9.12), we first have to express x as $x = \sum_{k=1}^n p_k x(f_k)$, where $p_k \geq 0$ and $\sum_{k=1}^n p_k = 1$. Next, we have to determine the policy $R = (\pi_1, \pi^2, \ldots) \in C(M)$ such that R satisfies $\sum_j \beta_j \cdot \mathbb{P}_R\{X_t = i, Y_t = a \mid X_1\} = \sum_j \beta_j \cdot \sum_k p_k \cdot \mathbb{P}_{f_k^\infty}\{X_t = i, Y_t = a \mid X_1\} = \beta_j\}$ for all $(i,a) \in S \times A$ and all $t \in \mathbb{N}$. The decision rules $\pi^t,\ t \in \mathbb{N}$, can be determined by formula (1.7) in Theorem 1.1.

Algorithm 9.1 *Construction of an optimal policy $R \in L(M) \cap L(C)$ for CMDP problem (9.6)*

1. Determine an optimal solution (x,y) of linear program (9.12); if (9.12) is infeasible, then problem (9.6) is also infeasible.

2. (a) Let $C(D) = \{f_1^\infty, f_2^\infty, \cdots, f_n^\infty\}$ and compute $P^*(f_k)$ for $k = 1,2,\ldots,n$.

   (b) Take $x_{ia}^k = \begin{cases} \sum_j \beta_j \{P^*(f_k)\}_{ji} & a = f_k(i) \\ 0 & a \neq f_k(i) \end{cases}, \quad i \in S,\ k = 1,2,\ldots,n.$

3. Determine $p_k,\ k = 1, 2, \ldots, n$ as feasible solution of the following linear system

   (this computation can be performed by the Phase I technique of the simplex method)

$$\begin{cases} \sum_{k=1}^{n} p_k x_{ia}^k &= x_{ia} \quad a \in A(i),\ i \in S \\ \sum_{k=1}^{n} p_k &= 1 \\ p_k &\geq 0 \quad k = 1, 2, \ldots, n \end{cases}$$

4. $R = (\pi_1, \pi^2, \ldots)$, defined by

$$\pi_{ia}^t = \begin{cases} \frac{\sum_j \beta_j \sum_k p_k \{P^{t-1}(f_k)\}_{ji} \cdot \delta_{a f_k(i)}}{\sum_j \beta_j \sum_k p_k \{P^{t-1}(f_k)\}_{ji}} & \text{if } \sum_j \beta_j \sum_k p_k \{P^{t-1}(f_k)\}_{ji} \neq 0; \\ \text{arbitrary} & \text{if } \sum_j \beta_j \sum_k p_k \{P^{t-1}(f_k)\}_{ji} = 0. \end{cases}$$

   is an optimal policy for problem (9.6).

Example 9.2

Consider the following MDP: $S = \{1, 2, 3\};\ A(1) = \{1, 2\},\ A(2) = \{1\},\ A(3) = \{1, 2\};$

$r_1(1) = r_1(2) = 0;\ r_2(1) = 1;\ r_3(1) = r_3(2) = 0;$

$p_{11}(1) = 0,\ p_{12}(1) = 1,\ p_{13}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 0,\ p_{13}(2) = 1;$

$p_{21}(1) = 0,\ p_{22}(1) = 1;\ p_{23}(1) = 0;\ p_{31}(1) = 0,\ p_{32}(2) = 0;\ p_{33}(2) = 1;$

$p_{31}(2) = 0,\ p_{32}(2) = 1,\ p_{33}(2) = 0$. Take $\beta_1 = \frac{1}{4},\ \beta_2 = \frac{3}{16},\ \beta_3 = \frac{9}{16}$.

As constraint we have bounds for the value $x_{12}(R)$: $\frac{1}{4} \leq x_{12}(R) \leq \frac{1}{2}$.

If we apply Algorithm 9.1 we obtain the following.

maximize x_{21} *subject to*

$$\begin{array}{rcrcrcrcrcrcrcl}
 x_{11} & + & x_{12} & & & & & & & & & & & = & 0 \\
- x_{11} & & & & & - & x_{32} & & & & & & & = & 0 \\
 & - & x_{12} & & & + & x_{32} & & & & & & & = & 0 \\
 x_{11} & + & x_{12} & & & & & + & y_{11} & + & y_{12} & - & y_{32} & = & \frac{1}{4} \\
 & & & x_{21} & & & & - & y_{11} & & & + & y_{32} & = & \frac{3}{16} \\
 & & & & x_{31} & + & x_{32} & & & - & y_{12} & & & = & \frac{9}{16} \\
 & & & x_{21} & & & & & & & & & & \leq & \frac{1}{2} \\
 & & - & x_{21} & & & & & & & & & & \leq & -\frac{1}{4}
\end{array}$$

$$x_{11},\ x_{12},\ x_{21},\ x_{31},\ x_{32} \geq 0$$

with optimal solution $x_{11} = 0,\ x_{12} = 0,\ x_{21} = \frac{1}{2},\ x_{31} = \frac{1}{2},\ x_{32} = 0;\ y_{11} = 0,\ y_{12} = \frac{1}{4},\ y_{32} = \frac{5}{16}$.

There are four deterministic policies:

$f_1(1) = 1,\ f_1(2) = 1,\ f_1(3) = 1;\ f_2(1) = 1,\ f_2(2) = 1,\ f_2(3) = 2;$

$f_3(1) = 2,\ f_3(2) = 1,\ f_3(3) = 1;\ f_4(1) = 2,\ f_4(2) = 1,\ f_4(3) = 2.$

The corresponding stationary matrices are:

$$P^*(f_1)=\begin{pmatrix}0&1&0\\0&1&0\\0&0&1\end{pmatrix};\ P^*(f_2)=\begin{pmatrix}0&1&0\\0&1&0\\0&1&0\end{pmatrix};\ P^*(f_3)=\begin{pmatrix}0&0&1\\0&1&0\\0&0&1\end{pmatrix};\ P^*(f_4)=\begin{pmatrix}0&1&0\\0&1&0\\0&1&0\end{pmatrix}.$$

The vectors $x^1,\ x^2,\ x^3,\ x^4$ are:

$x^1_{11}=0;\ x^1_{12}=0;\ x^1_{21}=\frac{7}{16};\ x^1_{31}=\frac{9}{16};\ x^1_{32}=0.\ x^2_{11}=0;\ x^2_{12}=0;\ x^2_{21}=1;\ x^2_{31}=0;\ x^2_{32}=0.$

$x^3_{11}=0;\ x^3_{12}=0;\ x^3_{21}=\frac{3}{16};\ x^3_{31}=\frac{13}{16};\ x^3_{32}=0.\ x^4_{11}=0;\ x^4_{12}=0;\ x^4_{21}=1;\ x^4_{31}=0;\ x^4_{32}=0.$

For the numbers $p_1,\ p_2,\ p_3,\ p_4\geq 0$ such that $\sum_{k=1}^4 p_k=1$ and $p_1x^1+p_2x^2+p_3x^3+p_4x^4=4$ we obtain: $p_1=\frac{8}{9},\ p_2=\frac{1}{9},\ p_3=0,\ p_4=0.$

Since $P^t(f_1)=\begin{pmatrix}0&1&0\\0&1&0\\0&0&1\end{pmatrix}$ and $P^t(f_2)=\begin{pmatrix}0&1&0\\0&1&0\\0&1&0\end{pmatrix}$ for all $t\in\mathbb{N}$, we obtain

$$R=(\pi^1,\pi^2,\dots)\text{ with }\pi^t_{11}=1,\ t\in\mathbb{N};\ \pi^t_{21}=1,\ t\in\mathbb{N};\ \pi^t_{31}=\begin{cases}\frac{8}{9} & t=1\\ 1 & t\geq 2\end{cases};\ \pi^t_{32}=\begin{cases}\frac{1}{9} & t=1\\ 1 & t\geq 2\end{cases}.$$

Remark

Algorithm 9.1 is inattractive for practical problems. The number of calculations is prohibitive. Moreover, the use of Markov policies is inefficient in practice. Therefore, in the next pages we discuss the problem of finding an optimal stationary policy, if one exists.

For any feasible solution (x,y) of (9.12) we define a stationary policy π^∞ by

$$\pi_{ia}=\begin{cases}x_i(a)/x_i & i\in S_x\\ y_i(a)/y_i & i\in S_y\\ \text{arbitrary} & \text{if } i\notin S_y\cup S_x\end{cases}\tag{9.13}$$

where $x_i=\sum_a x_i(a),\ y_i=\sum_a y_i(a),\ S_x=\{x\mid x_i>0\}$ and $S_y=\{y\mid x_i=0,\ y_i>0\}$.

Notice that, since $\beta_j=0$ is allowed for one or more $j\in S$, it is possible that $S_x\cup S_y\neq S$.

Lemma 9.2

If (x,y) is an optimal solution of (9.12) and $x_i(a)=\pi_{ia}\cdot\{\beta^TP^*(\pi)\}_i,\ (i,a)\in S\times A$, where π is defined by (9.13), then π^∞ is an optimal solution of (9.6).

Proof

Since $c^k(\pi^\infty)=\beta^TP^*(\pi)c^k(\pi)=\sum_i\{\beta^TP^*(\pi)\}_i\sum_a c^k_i(a)\pi_{ia}=\sum_{(i,a)}c^k_i(a)x_i(a)\leq b_k$ for all $1\leq k\leq m$, the stationary policy π^∞ is a feasible solution of (9.6). Moreover, by Theorem 9.5, part (2), we have $\phi(\pi^\infty)=\beta^TP^*(\pi)r(\pi)=\sum_{(i,a)}r_i(a)x_i(a)=$ optimum (9.12) $=$ optimum (9.6), i.e. π^∞ is an optimal solution of (9.6). □

The next example shows that for an optimal solution (x, y) of (9.12), the pilicy π^∞, where π is defined by (9.13), is not an optimal solution of (9.6), even in the case that (9.6) has a stationary optimal policy.

Example 9.3

Consider the model of Example 9.2, but now with the constraint $x_{21}(R) \leq \frac{1}{4}$. The linear program (9.12) for this constrained problem is

maximize x_{21} *subject to*

$$\begin{array}{rrrrrrrrrrrrrrr}
 & x_{11} & + & x_{12} & & & & & & & & & & = & 0 \\
- & x_{11} & & & & & - & x_{32} & & & & & & = & 0 \\
 & & - & x_{12} & & & + & x_{32} & & & & & & = & 0 \\
 & x_{11} & + & x_{12} & & & & & + & y_{11} & + & y_{12} & - & y_{32} & = & \frac{1}{4} \\
 & & & & x_{21} & & & & - & y_{11} & & & + & y_{32} & = & \frac{3}{16} \\
 & & & & & x_{31} & + & x_{32} & & & - & y_{12} & & & = & \frac{9}{16} \\
 & & & & x_{21} & & & & & & & & & & \leq & \frac{1}{4}
\end{array}$$

$$x_{11},\ x_{12},\ x_{21},\ x_{31},\ x_{32} \geq 0$$

with optimal solution $x_{11} = 0,\ x_{12} = 0,\ x_{21} = \frac{1}{4},\ x_{31} = \frac{3}{4},\ x_{32} = 0;\ y_{11} = 0,\ y_{12} = \frac{1}{4},\ y_{32} = \frac{1}{16}$ and with optimum value $\frac{1}{4}$. The corresponding stationary policy π^∞ satisfies $\pi_{12} = \pi_{21} = \pi_{31} = 1$. This policy is not optimal, because $\phi(\pi^\infty) = \frac{3}{16} < \frac{1}{4}$, the optimum of the linear program.
Consider the stationary policy with $\pi_{11} = \frac{1}{4},\ \pi_{12} = \frac{3}{4},\ \pi_{21} = \pi_{31} = 1$. For this policy we obtain $x_{12}(\pi^\infty) = \frac{1}{4}$ and $\phi(\pi^\infty) = \frac{1}{4}$, the optimum value of the linear program. So, this policy is feasible and optimal.

In order to apply Lemma 9.2 we have to compute the stationary matrix $P^*(\pi)$. The determination of the stationary matrix can be executed in polynomial time (see Algorithm 5.5). However, if $x_i(a)/x_i = y_i(a)/y_i$ for all $a \in A(i),\ i \in \{j \mid x_j > 0,\ y_j > 0\}$, which is for instance the case if $\{j \mid x_j > 0,\ y_j > 0\} = \emptyset$, then the policy π^∞, where π is defined by (9.13), is an optimal policy for problem (9.6) as the next lemma shows.

Lemma 9.3

If $x_i(a)/x_i = y_i(a)/y_i$ for all $a \in A(i),\ i \in \{j \mid x_j > 0,\ y_j > 0\}$, then the stationary policy π^∞, where π is defined by (9.13), is an optimal policy for problem (9.6).

Proof

The conditon $x_i(a)/x_i = y_i(a)/y_i$ for all $a \in A(i),\ i \in \{j \mid x_j > 0,\ y_j > 0\}$ implies that $y_i(a)/y_i = \pi_{ia}$ for all $a \in A(i),\ i \in \{j \mid y_j > 0\}$, i.e. $y_i(a) = \pi_{ia} \cdot y_i$ for all $a \in A(i),\ i \in S$.
Hence, we can write

$$\beta_j = x_j + \textstyle\sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}\pi_{ia} \cdot y_i = x_j + \sum_i y_i\{\delta_{ij} - p_{ij}(\pi)\},\ j \in S.$$

So (x, y) satisfies, in vector notation, $x^T = x^T P(\pi)$ and $x^T + y^T\{I - P(\pi)\} = \beta^T$. Consequently, $x^T = x^T P^*(\pi)$ and $x^T P^*(\pi) = \beta^T P^*(\pi)$. So, $x^T = \beta^T P^*(\pi)$, i.e. x satisfies the conditions of Lemma 9.2.

□

If the conditions of Lemma 9.3 are not satisfied, we can try to find - for the same x - another y, say $\overline{y}$, such that $(x, \overline{y})$ is feasible for (9.12) - and consequently also optimal - and satisfies the conditions of Lemma 9.3. To achieve this, we need $\overline{y}_i(a)/\overline{y}_i = \pi_{ia},\ a \in A(i),\ i \in \{j \mid x_j > 0,\ \overline{y}_j > 0\}$, which is equivalent to $\overline{y}_i(a) = \overline{y}_i \cdot \pi_{ia},\ a \in A(i),\ i \in \{j \mid \overline{y}_j > 0\}$. Hence, $\overline{y}$ has to satisfy the linear system in the y-variables (x is fixed)

$$\begin{cases} \sum_{i \notin S_x} \sum_a \{\delta_{ij} - p_{ij}(a)\}\overline{y}_i(a) + \sum_{i \in S_x} \{\delta_{ij} - p_{ij}(\pi)\}\overline{y}_i = \beta_j - x_j,\ j \in S \\ \overline{y}_i(a) \geq 0,\ i \notin S_x,\ a \in A(i);\ \overline{y}_i \geq 0,\ i \in S_x \end{cases} \tag{9.14}$$

The feasibility of system (9.14) can be checked by the so-called phase I of the simplex method.

Example 9.4

Consider the model of Example 9.3. The optimal solution does not satisfy $x_i(a)/x_i = y_i(a)/y_i$ for all $a \in A(i),\ i \in \{j \mid x_j > 0,\ y_j > 0\}$, because $x_{32}/x_3 = 0$ and $y_{32}/y_3 = 1$.

The system (9.14) becomes $\overline{y}_{11} + \overline{y}_{12} = \frac{4}{16};\ -\overline{y}_{11} = -\frac{1}{16};\ -\overline{y}_{12} = -\frac{3}{16};\ \overline{y}_{11}, \overline{y}_{12} \geq 0$.

This system has the solution $\overline{y}_{11} = \frac{1}{16},\ \overline{y}_{12} = \frac{3}{16}$. Hence, the stationary policy π^∞ with $\pi_{11} = \frac{1}{4},\ \pi_{12} = \frac{3}{4},\ \pi_{21} = \pi_{31} = 1$ is an optimal policy for problem (9.6).

Remark

If the x-part of problem (9.12) is unique and (9.14) is infeasible, then problem (9.6) has no optimal stationary policy, namely:

Suppose that (9.6) has an optimal stationary policy, say π^∞. Then, (x^π, y^π), defined by (5.30) and (5.31), is a feasible solution for problem (9.12) and $\sum_{(i,a)} r_i(a)x_i^\pi(a) = \beta^T P^*(\pi) r(\pi) =$ optimum (9.6). Hence, (x^π, y^π) is an optimal solution of problem (9.12). Consequently, $x^\pi = x$. Then, y^π is a feasible solution of system (9.14), which is contradictory to the assumption that (9.14) is infeasible.

Example 9.5

Consider the model of Example 9.2. We have seen that $x_{11} = 0, x_{12} = 0, x_{21} = \frac{1}{2}, x_{31} = \frac{1}{2}, x_{32} = 0$; $y_{11} = 0,\ y_{12} = \frac{1}{4},\ y_{32} = \frac{5}{16}$ is an optimal solution. It can easily be verified that the x-part of the solution is unique. The system (9.6) is: $\overline{y}_{11} + \overline{y}_{12} = \frac{4}{16};\ -\overline{y}_{11} = -\frac{5}{16};\ -\overline{y}_{12} = \frac{1}{16};\ \overline{y}_{11}, \overline{y}_{12} \geq 0$. The system is infeasible and therefore the problem has no stationary optimal policy.

Unichain case

We will show that $L(S) = Q$, which implies that $L = L(M) = L(C) = L(S) = \overline{L(D)} = Q$. In order to show $L(S) = Q$ we need the following two lemmas.

Lemma 9.4

For every triple (j, a, R), where $j \in S,\ a \in A(j)$ and R a convergent policy, we have

$x_{ja}(R) = lim_{\alpha \uparrow 1}\, (1-\alpha) \sum_{t=1}^{\infty} \alpha^{t-1} \cdot \sum_i \beta_i \cdot \mathbb{P}_R\{X_t = j,\ Y_t = a \mid X_1 = i\}$.

Proof

Let R be a convergent policy and let $x(R) = lim_{T\to\infty}\, x^T(R)$. Take a fixed pair $(j,a) \in S \times A$. Then, $x_{ja}(R) = lim_{T\to\infty}\, \frac{1}{T}\sum_{t=1}^{T} w_t$, where $w_t = \sum_i \beta_i \cdot \mathbb{P}_R\{X_t = j,\ Y_t = a \mid X_1 = i\}$. Since $|w_t|$ is bounded by 1 for all t, the power series $\sum_{t=1}^{\infty} w_t \alpha^{t-1}$ has radius of convergence at least 1. The series $\sum_{t=1}^{\infty} \alpha^{t-1}$ has radius of convergence 1. Hence, we can write

$$(1-\alpha)^{-1} \cdot \textstyle\sum_{t=1}^{\infty} w_t\, \alpha^{t-1} = \{\sum_{t=1}^{\infty} \alpha^{t-1}\} \cdot \{\sum_{t=1}^{\infty} w_t\, \alpha^{t-1}\} = \sum_{t=1}^{\infty} \{\sum_{s=1}^{t} w_s\}\, \alpha^{t-1}.$$

Since $(1-\alpha)^{-2} = \sum_{t=1}^{\infty} t\, \alpha^{t-1}$, we obtain

$$x_{ja}(R) - (1-\alpha) \textstyle\sum_{t=1}^{\infty} \alpha^{t-1} \cdot \sum_i \beta_i \cdot \mathbb{P}\{X_t = j,\ Y_t = a \mid X_1 = i\} =$$
$$(1-\alpha)^2 \textstyle\sum_{t=1}^{\infty} \{x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s\}\, t\, \alpha^{t-1}.$$

Choose $\varepsilon > 0$ arbitrary. Since $x_{ja}(R) = lim_{T\to\infty}\, \frac{1}{T}\sum_{t=1}^{T} w_t$, there exists an integer T_ε such that $|x_{ja}(R) - \frac{1}{T}\sum_{t=1}^{T} w_t| \le \frac{1}{2}\varepsilon$ for all $T \ge T_\varepsilon$. Hence,

$$|(1-\alpha)^2 \textstyle\sum_{t=1}^{T_\varepsilon} \{x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s\}\, t\, \alpha^{t-1}| \le (1-\alpha)^2 M \cdot \sum_{t=1}^{T_\varepsilon} T_\varepsilon\, \alpha^{t-1} \le \tfrac{1}{2}\varepsilon$$

for α sufficiently close to 1 and $M \ge max_{1\le t\le T_\varepsilon} |x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s|$. Furthermore, we have

$$|(1-\alpha)^2 \textstyle\sum_{t=T_\varepsilon+1}^{\infty} \{x_{ja}(R) - \frac{1}{t}\sum_{s=1}^{t} w_s\}\, t\, \alpha^{t-1}| \le (1-\alpha)^2 \sum_{t=T_\varepsilon+1}^{\infty} \tfrac{1}{2}\varepsilon\, t\, \alpha^{t-1} = \tfrac{1}{2}\varepsilon.$$

Hence, $x_{ja}(R) = lim_{\alpha\uparrow 1}\, (1-\alpha) \sum_{t=1}^{\infty} \alpha^{t-1} \cdot \sum_i \beta_i \cdot \mathbb{P}_R\{X_t = j,\ Y_t = a \mid X_1 = i\}$. □

Lemma 9.5

If $x(\pi^\infty)$ is continuous in π, then $L(S) = Q$.

Proof

Since $L(S) \subseteq L(C)$, it is sufficient to show that $L(C) \subseteq L(S)$. Take any $x(R) \in L(C)$. From Theorem 9.1 it follows that for any discount factor $\alpha \in [0,1)$ there exists a stationary policy π_α^∞ such that $x^\alpha(R) = x^\alpha(\pi_\alpha^\infty)$. Choose a fixed pair $(j,a) \in S \times A$ and let the reward function r on $S \times A$ be defined by $r_i(b) = \begin{cases} 1 & \text{if } i = j \text{ and } b = a; \\ 0 & \text{elsewhere.} \end{cases}$

Then, $\beta^T v^\alpha(\pi_\alpha^\infty) = x_{ja}^\alpha(\pi_\alpha^\infty)$ and $\beta^T \phi(\pi_\alpha^\infty) = x_{ja}(\pi_\alpha^\infty)$ for all $\alpha \in [0,1)$. Hence, we can write by Lemma 9.4

$$x_{ja}(R) = lim_{\alpha\uparrow 1}\, (1-\alpha) \cdot x_{ja}^\alpha(R) = lim_{\alpha\uparrow 1}\, (1-\alpha) \cdot x_{ja}^\alpha(\pi_\alpha^\infty) = lim_{\alpha\uparrow 1}\, (1-\alpha) \cdot \beta^T v^\alpha(\pi_\alpha^\infty).$$

Consider a sequence $\{\alpha_k,\ k = 1,2,\dots\}$ such that $\alpha_k \uparrow 1$ and $\pi_{\alpha_k} \to \pi$. Since for any $i \in S$ the sequence $\{(1-\alpha_k) v_i^{\alpha_k}(\pi_{\alpha_k}^\infty),\ k = 1,2,\dots\}$ is dominated by the sequence $\{(1-\alpha_k) v_i^{\alpha_k},\ k = 1,2,\dots\}$ and since $lim_{k\to\infty}\, \{(1-\alpha_k) v_i^{\alpha_k} = \phi_i$, there exists a limit point, say x, of the sequence of vectors $\{(1-\alpha_k) v_i^{\alpha_k}(\pi_{\alpha_k}^\infty),\ k = 1,2,\dots\}$. Therefore, we may assume that

$$x_i = lim_{k\to\infty}\, (1-\alpha_k) v_i^{\alpha_k}(\pi_{\alpha_k}^\infty),\ i \in S, \tag{9.15}$$

implying, by (9.15) and Lemma 9.4,

$$\beta^T x = \sum_i \beta_i \cdot lim_{k\to\infty}\, (1-\alpha_k) v_i^{\alpha_k}(\pi_{\alpha_k}^\infty) = lim_{k\to\infty}\, (1-\alpha_k)\, \beta^T v^{\alpha_k}(\pi_{\alpha_k}^\infty) = x_{ja}(R). \tag{9.16}$$

Since $x(\pi^\infty)$ is continuous in π we can write for $\pi_{\alpha_k} \to \pi$

$$x_{ja}(\pi^\infty) = lim_{k\to\infty}\, x_{ja}(\pi^\infty_{\alpha_k}) \;=\; lim_{k\to\infty}\,(1-\alpha_k)\{\textstyle\sum_{t=1}^{\infty}\alpha^{t-1}\}\cdot\beta^T P^*(\pi^\infty_{\alpha_k})r(\pi_{\alpha_k}).$$

Because $P^* = P^*P^t$ for any stationary matrix P^* and any $t \in \mathbb{N}$, we obtain

$$\begin{aligned} x_{ja}(\pi^\infty) &= lim_{k\to\infty}\,(1-\alpha_k)\{\textstyle\sum_{t=1}^{\infty}\alpha^{t-1}\}\cdot\beta^T P^*(\pi^\infty_{\alpha_k})P^{t-1}(\pi^\infty_{\alpha_k})r(\pi_{\alpha_k}) \\ &= lim_{k\to\infty}\,\beta^T P^*(\pi^\infty_{\alpha_k})(1-\alpha_k)\textstyle\sum_{t=1}^{\infty}\alpha^{t-1}P^{t-1}(\pi^\infty_{\alpha_k})r(\pi_{\alpha_k}) \\ &= lim_{k\to\infty}\,\{x(\pi^\infty_{\alpha_k})\}^T(1-\alpha_k)v^{\alpha_k}(\pi^\infty_{\alpha_k}) \;=\; \{x(\pi^\infty)\}^T x = \beta^T P^*(\pi)x. \end{aligned}$$

Since $v^{\alpha_k}(\pi^\infty_{\alpha_k}) = r(\pi^\infty_{\alpha_k}) \;+\; \alpha P(\pi^\infty_{\alpha_k})v^{\alpha_k}(\pi^\infty_{\alpha_k})$, we can also write

$$(1-\alpha_k)v^{\alpha_k}(\pi^\infty_{\alpha_k}) = (1-\alpha_k)r(\pi^\infty_{\alpha_k}) + \alpha P(\pi^\infty_{\alpha_k})(1-\alpha_k)v^{\alpha_k}(\pi^\infty_{\alpha_k}).$$

Letting $k \to \infty$, then - by (9.15) - we obtain $x = P(\pi)x$ and consequently, $x = P^*(\pi)x$. Finally, we have $x_{ja}(R) = lim_{k\to\infty}\,(1-\alpha_k)\beta^T v^{\alpha_k}(\pi^\infty_{\alpha_k}) = \beta^T x = \beta^T P^*(\pi)x = x_{ja}(\pi)$. Since the stationary policy π is independent of the choice of the pair (j,a), we have shown that $x(R) \in L(S)$. □

Theorem 9.6
$L(S) = Q$.

Proof
By Lemma 9.5 it is sufficient to show that $x(\pi^\infty)$ is continuous in π. Let $\pi^\infty(k),\ k = 1,2,\ldots$ and $\pi^\infty(0)$ be stationary policies such that $\pi(0) = \lim_{k\to\infty}\pi(k)$. By the unichain property the stationary distribution $p^*\big(\pi(k)\big)$ of the Markov chain $P\big(\pi(k)\big)$ is the unique solution of the linear system

$$\left\{\begin{array}{lcl} \sum_i \{\delta_{ij} - p_{ij}\big(\pi(k)\big)\}x_i &=& 0 \\ \sum_i x_i &=& 1 \end{array}\right. \tag{9.17}$$

Since $\pi(k) \to \pi(0)$ for $k \to \infty$, we also have $P\big(\pi(k)\big) \to P\big(\pi(0)\big)$ for $k \to \infty$. Consequently, any limit point of $\{p^*\big(\pi(k)\big),\ k = 1,2,\ldots\}$ is a solution of (9.17) with $k = 0$, i.e. is equal to $p^*\big(\pi(0)\big)$. Hence, $x_{ia}\big(\pi^\infty(k)\big) = p_i^*\big(\pi(k)\big)\cdot\pi_{ia}(k) \to p_i^*\big(\pi(0)\big)\cdot\pi_{ia}(0) = x_{ia}\big(\pi^\infty(0)\big)$, i.e. $x(\pi^\infty)$ is continuous in π. □

By these results an optimal stationary policy for the CMDP in the unichain case can be computed by the following algorithm.

Algorithm 9.2 *Construction of a stationary optimal policy π^∞ for CMDP problem (9.6)*

1. Determine an optimal solution x of linear program

$$max\left\{\sum_{(i,a)} r_i(a)x_i(a)\;\middle|\;\begin{array}{lcl} \sum_{(i,a)}\{\delta_{ij} - p_{ij}(a)\}x_i(a) &=& 0,\ j \in S \\ \sum_{(i,a)} x_i(a) &=& 1 \\ \sum_{(i,a)} c_i^k(a)x_i(a) &\le& b_k,\ k = 1,2,\ldots,m \\ x_i(a) \ge 0,\ (i,a) \in S\times A & & \end{array}\right\}. \tag{9.18}$$

(if (9.18) is infeasible, then problem (9.6) is also infeasible).

2. Take $\pi_{ia} = \begin{cases} x_i(a)/x_i & a \in A(i),\ i \in S_x \\ \text{arbitrary} & \text{otherwise} \end{cases}$ where $x_i = \sum_a x_i(a)$ and $S_x = \{i \mid x_i > 0\}$.

Theorem 9.7

The stationary policy π^∞ obtained by Algorithm 9.2 is an optimal policy for problem (9.6).

Proof

Since $x_i(a) = x_i \cdot \pi_{ia}$ for all $(i,a) \in S \times A$, the constraints of (9.18) imply

$$\begin{cases} \sum_i \{\delta_{ij} - p_{ij}(\pi)\} x_i &= 0 \\ \sum_i x_i &= 1 \end{cases} \tag{9.19}$$

Hence, $x_i = p_i^*(\pi),\ i \in S$, and consequently $x = x(\pi^\infty)$. Therefore, π^∞ is a feasible solution of (9.6). Moreover, $\phi(\pi^\infty) = \sum_{(i,a)} r_i(a) x_i(a) =$ optimum (9.18). From Theorem 9.6 it follows that there exists a stationary optimal policy of problem (9.6), say π_*^∞. Let $x^* = x(\pi_*^\infty)$. Then, x^* is a feasible solution of program (9.18) and consequently,

$$\text{optimum (9.6)} = \phi(\pi_*^\infty) = \sum_{(i,a)} r_i(a) x_i^*(a) \leq \sum_{(i,a)} r_i(a) x_i(a) = \phi(\pi^\infty).$$

Hence, π^∞ obtained by Algorithm 9.2 is an optimal policy for problem (9.6). □

Remark

If the MDP is irreducible, then any solution of (9.19) satisfies $x_i > 0,\ i \in S$. Since any optimal extreme solution of the CMDP has at most $|S| + m$ positive variables, the optimal stationary policy π^∞ chooses actions randomly, i.e. with probability $p \in (0,1)$, in at most m states.

9.2 Multiple objectives

For some problems we may have several sorts of rewards or costs, which we may not be able to optimize simultaneously. Assume that we want to maximize some utility for an m-tuple of immediate rewards, say utilities $u^k(R)$ and immediate rewards $r_i^k(a),\ (i,a) \in S \times A$, for $k = 1, 2, \ldots, m$. For each k one can find an optimal policy R_k, i.e. $u_i^k(R_k) \geq u_i^k(R),\ i \in S$, for all policies R. However, in general, $R_k \neq R_l$ if $k \neq l$, and there does not exist one policy which is optimal for all m rewards simultaneously for all starting states. Therefore, we consider the utility function with respect a given initial distribution β. Given this initial distribution β and a policy R, we denote the utilities by $u^k(\beta, R)$. The goal in multi-objective optimization is to find an β-*efficient solution*, i.e. a policyR_* such that there exists *no other policy* R satisfying

$$u^k(\beta, R) \geq u^k(\beta, R_*) \text{ for all } k \text{ and } u^k(\beta, R) > u^k(\beta, R_*) \text{ for at least one } k.$$

We will consider multiple objectives for both discounted rewards and average rewards.

9.2.1 Discounted rewards

The linear program usually associated with the discounted reward criterion for immediate rewards $r_i(a)$ is

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\}x_i(a) & = & \beta_j,\ j \in S \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\} \tag{9.20}$$

with dual program

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\}v_j \geq r_i(a) \text{for every } (i,a) \in S \times A \right\}. \tag{9.21}$$

Define the utility $v^\alpha(\beta, R)$ by

$$v^\alpha(\beta, R) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{j \in S} \beta_j \cdot \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i(a).$$

A policy R_* is *β-optimal* if $v^\alpha(\beta, R_*) = max_R\, v^\alpha(\beta, R)$. Clearly, a discounted optimal policy is β-optimal, but not conversely.

Theorem 9.8

If x is an optimal solution of the linear program (9.20) and f^∞ is such that $x_i\big(f(i)\big) > 0,\ i \in S_x$, where $S_x = \{j \mid \sum_a x_j(a) > 0\}$, then f^∞ is a β-optimal policy.

Proof

Since v^α is the smallest α-superharmonic vector (see Theorem 3.16), v^α is an optimal solution of program (9.21) (the solution is not necessarily unique because $\beta_j = 0$ is allowed for some $j \in S$). By the complementary property of linear programming, we have

$$\sum_j \{\delta_{ij} - \alpha p_{ij}\big(f(i)\big)\}v_j^\alpha = r_i\big(f(i)\big),\ i \in S_x. \tag{9.22}$$

Next, we show that the set S_x is closed in the Markov chain $P(f)$. Suppose that S_x is not closed, i.e. $p_{kl}(f) > 0$ for some $k \in S_x$ and $l \notin S_x$. Since

$$0 = \sum_a x_l(a) = \beta_l + \alpha \sum_{(k,a)} p_{kl}(a)x_k(a) \geq p_{kl}(f)x_k i\big(f(ki)\big) > 0, \tag{9.23}$$

we have a contradiction. Since S_x is closed and $\beta_j = 0,\ j \notin S_x$ (this follows also from (9.23)), we may consider the process on S_x. Then, by (9.22), $\{I - \alpha P(f)\}v^\alpha = r(f)$, implying $v^\alpha(f^\infty) = \{I - \alpha P(f)\}^{-1}r(f) = v^\alpha$ on S_x. Hence,

$$v^\alpha(\beta, f^\infty) = \textstyle\sum_j \beta_j v_j^\alpha(f^\infty) = \sum_{j \in S_x} \beta_j v_j^\alpha(f^\infty) = \sum_{j \in S_x} \beta_j v_j^\alpha = \sum_j \beta_j v_j^\alpha \geq v^\alpha(\beta, R)$$

for all policies R. □

Define the utilities $v_k^\alpha(\beta, R)$ by

$$v_k^\alpha(\beta, R) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{j \in S} \beta_j \cdot \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i^k(a).$$

Theorem 9.9

Take any $\lambda \in \mathbb{R}^m$ with $\lambda_k > 0,\ k = 1, 2, \ldots, m$ and let x be an optimal solution of the linear program

$$max \left\{ \sum_{(i,a)} \Big\{ \sum_{k=1}^{m} \lambda_k r_i^k(a) \Big\} x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) & = & \beta_j,\ j \in S \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}. \tag{9.24}$$

Take f^∞ such that $x_i\big(f(i)\big) > 0,\ i \in S_x$, then f^∞ a β-efficient policy.

Proof

From Theorem 9.8 it follows that f^∞ is a β-optimal policy with respect to the immediate rewards $r_i(a) = \sum_{k=1}^m \lambda_k r_i^k(a),\ (i,a) \in S \times A$, i.e. $v^\alpha(\beta, f^\infty) \geq v^\alpha(\beta, R)$ for all policies R. Since the discounted rewards are linear in the immediate rewards, we have $v^\alpha(\beta, R) = \sum_{k=1}^m v_k^\alpha(\beta, R)$. Suppose that f^∞ is not β-efficient. Then, there exists a policy R such that

$$\textstyle\sum_{k=1}^m \lambda_k v_k^\alpha(\beta, R) > \sum_{k=1}^m \lambda_k v_k^\alpha(\beta, f^\infty).$$

On the other hand we have

$$\textstyle\sum_{k=1}^m \lambda_k v_k^\alpha(\beta, f^\infty) = v^\alpha(\beta, f^\infty) \geq v^\alpha(\beta, R) = \sum_{k=1}^m \lambda_k v_k^\alpha(\beta, R),$$

implying a contradiction. □

Remark

Suppose that we want to maximixe lexicographically the functions $v_k^\alpha(\beta, R)$ for $k = 1, 2, \ldots, m$. A policy R^* which is lexicographically maximal with respect to $v_1^\alpha(\beta, R),\ v_2^\alpha(\beta, R), \ldots, v_m^\alpha(\beta, R)$ is a *lexicographically efficient* policy.

To determine a lexicographically efficient policy, we compute an optimal solution, say x^1 of the linear program

$$max \left\{ \sum_{(i,a)} r_i^1(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) & = & \beta_j,\ j \in S \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}. \tag{9.25}$$

Next, we solve the following linear program with one additional constraint

$$max \left\{ \sum_{(i,a)} r_i^2(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) & = & \beta_j,\ j \in S \\ \sum_{(i,a)} r_i^1(a) x_i(a) & = & \sum_{(i,a)} r_i^1(a) x_i^1(a) \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}. \tag{9.26}$$

Continuing in this way we stop either when we find a unique optimal solution x^k for some $1 \leq k \leq m$ or when we have solved all m linear programs. Let x^* be the finally obtained solution. Then, a lexicografically efficient solution is the stationary policy π^∞, defined by

$$\pi_{ia} = \begin{cases} x_i^*(a)/x_i^* & \text{if } x_i^* > 0 \\ \text{arbitrary} & \text{if } x_i^* = 0 \end{cases} \quad \text{for all } (i,a) \in S \times A.$$

9.2.2 Average rewards

The average reward case is, as always, more cumbersome that the discounted reward case. The linear program usually associated with the average reward criterion for immediate rewards $r_i(a)$ is

$$max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}x_i(a) & = & 0,\ j \in S \\ \sum_a x_j(a) + \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\}y_i(a) & = & \beta_j,\ j \in S \\ x_i(a), y_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\} \tag{9.27}$$

with dual program

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{rcll} \sum_j \{\delta_{ij} - p_{ij}(a)\}v_j & \geq & 0 & \text{for every } (i,a) \in S \times A \\ v_i + \sum_j \big(\delta_{ij} - p_{ij}(a)\big)u_j & \geq & r_i(a) & \text{for every } (i,a) \in S \times A \end{array} \right\}. \tag{9.28}$$

Define the utility $\phi(\beta, R)$ by

$$\phi(\beta, R) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j \in S} \beta_j \cdot \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i(a).$$

A policy R_* is *β-optimal* if $\phi(\beta, R_*) = max_R\, \phi(\beta, R)$. Clearly, an average optimal policy is β-optimal, but not conversely. For any feasible solution (x, y) of (9.27, we define $x_i,\ y_i,\ i \in S$ and $S_x, S_y \subseteq S$ by $x_i = \sum_a x_i(a),\ y_i = \sum_a y_i(a),\ S_x = \{j \mid \sum_a x_j > 0\},\ S_y = \{j \mid x_j = 0,\ y_j > 0\}$.

Theorem 9.10

If (x, y) is an extreme optimal solution of the linear program (9.27) and f^∞ is such that $x_i\big(f(i)\big) > 0,\ i \in S_x;\ y_i\big(f(i)\big) > 0,\ i \in S_y$, then f^∞ is a β-optimal policy.

Proof

Similarly to the corresponding part in the proof of Theorem 5.16 (the proof is left to the reader as Exercise 9.6) it can be shown that

$$\begin{array}{rcl} \phi_i + \sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\}u_j & = & r_i\big(f(i)\big) \quad ,\ i \in S_x \\ \sum_j \{\delta_{ij} - p_{ij}\big(f(i)\big)\}\phi_j & = & 0 \quad ,\ i \in S_x \cup S_y \end{array} \tag{9.29}$$

We first show that S_x is closed in the Markov chain $P(f)$. Suppose that $p_{kl}\big(f(k)\big) > 0$ for some $k \in S_x,\ l \notin S_x$. From the constraints of (9.27) it follows that

$$0 = \sum_a x_l(a) = \sum_{i,a} p_{il}(a)x_i(a) \geq p_{kl}\big(f(k)\big)x_k\big(f(k)\big) > 0,$$

implying a contradiction.

We now show that $S_x \cup S_y$ is also closed. Suppose that $p_{kl}(f) > 0$ for some $k \in S_x \cup S_y$ and $l \notin S_x \cup S_y$. Then,

if $k \in S_x$: $0 = \sum_a x_l(a) = \sum_{i,a} p_{il}(a)x_i(a) \geq p_{kl}(f)x_k\big(f(k)\big) > 0$;

if $k \in S_y$: $0 = \sum_a x_l(a) + \sum_a y_l(a) = \beta_l + \sum_{i,a} p_{il}(a)y_i(a) \geq p_{kl}(f)y_k\big(f(l)\big) > 0$.

In both cases we have a contradiction: $S_x \cup S_y$ is closed in the Markov chain $P(f)$.

Next, we show that the states of S_y are transient in the Markov chain $P(f)$. Suppose that S_y has an ergodic state. Since S_x and $S_x \cup S_y$ are closed, the set S_y contains an ergodic class, say $J = \{j_1, j_2, \ldots, j_m\}$. Since (x, y) is an extreme solution and $y_j\big(f(j)\big) > 0,\ j \in J$, the corresponding columns in (9.27) are linearly independent. Because these columns have zeros in the first N rows, the second parts of these vectors are also independent vectors. Since for $j \in J$ and $k \notin J$, we have $\delta_{jk} - p_{jk}\big(f(j)\big) = 0 - 0 = 0$, the vectors $b^i,\ 1 \leq i \leq m$, where b^i has components $\delta_{j_i k} - p_{j_i k}\big(f(j_i)\big),\ k \in J$, are also linear independent.

However, $\sum_{k=1}^m b_k^i = \sum_{k=1}^m \{\delta_{j_i j_k} - p_{j_i j_k}\big(f(j_i)\big)\} = 1 - 1 = 0,\ i = 1, 2, \ldots, m$, which contradicts the independency of $b^1, b^2, \ldots, b^m$.

Consider the Markov chain on the closed $S_x \cup S_y$. From (9.29) it follows that $\phi = P(f)\phi$, and consequently we have $\phi = P^*(f)\phi$. Since that states of S_y are transient, the columns of $P^*(f)$ corresponding to S_y are zero-vectors. Hence, also by (9.29),

$$\begin{aligned}
\{P^*(f)r(f)\}_i &= \sum_j p_{ij}^*(f)r_j(f) = \sum_{j \in S_x} p_{ij}^*(f)r_j(f) \\
&= \sum_{j \in S_x} p_{ij}^*(f)\big\{\phi_j + u_j - \{P(f)u\}_j\big\} \\
&= \big\{P^*(f)\{\phi + u - P(f)u\}\big\}_i = \{P^*(f)\phi\}_i = \phi_i,\ i \in S_x \cup S_y.
\end{aligned}$$

Hence,

$$\begin{aligned}
\phi(\beta, f^\infty) &= \sum_i \beta_i\{P^*(f)r(f)\}_i = \sum_{i \in S_x \cup S_y} \beta_i\{P^*(f)r(f)\}_i = \sum_{i \in S_x \cup S_y} \beta_i\phi_i \\
&= \sum_i \beta_i\phi_i = \phi(\beta, \phi) \geq \phi(\beta, R) \text{ for all policies } R.
\end{aligned}$$

□

Let $\phi_(\beta, R)$ be the average reward for immediate rewards $r_i^k(a),\ (i, a) \in S \times A$, given initial distribution β and policy R, i.e.

$$\phi_k(\beta, R) = liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j \in S} \beta_j \cdot \sum_{(i,a)} \mathbb{P}_R\{X_t = i,\ Y_t = a \mid X_1 = j\} \cdot r_i^k(a).$$

Theorem 9.11

Take any $\lambda \in \mathbb{R}^m$ with $\lambda_k > 0,\ k = 1, 2, \ldots, m$ and let (x, y) be an extreme optimal solution of the linear program (9.27). Then, the policy f^∞ satisfying $x_i\big(f(i)\big) > 0,\ i \in S_x;\ y_i\big(f(i)\big) > 0,\ i \in S_y$, is a β-efficient policy.

Proof

From Theorem 9.10 it follows that f^∞ is a β-optimal policy with respect to the immediate rewards $r_i(a) = \sum_{k=1}^m \lambda_k r_i^k(a),\ (i,a) \in S \times A$, i.e. $\phi(\beta, f^\infty) \geq \phi(\beta, R)$ for all policies R. Since the average rewards are linear in the immediate rewards, we have $\phi(\beta, R) = \sum_{k=1}^m \phi_k(\beta, R)$.

Suppose that f^∞ is not β-efficient. Then, there exists a policy R such that

$$\sum_{k=1}^m \lambda_k \phi_k(\beta, R) > \sum_{k=1}^m \lambda_k \phi_k(\beta, f^\infty).$$

On the other hand we have

$$\sum_{k=1}^m \lambda_k \phi_k(\beta, f^\infty) = \phi(\beta, f^\infty) \geq \phi(\beta, R) = \sum_{k=1}^m \lambda_k \phi_k(\beta, R),$$

implying a contradiction. □

Remark

Suppose that we want to maximixe lexicographically the functions $\phi^k(\beta, R)$ for $k = 1, 2, \ldots, m$. A policy R^* which is lexicographically maximal with respect to $\phi^1(\beta, R),\ \phi^2(\beta, R), \ldots, \phi^m(\beta, R)$ is a *lexicographically efficient* policy.

To determine a lexicographically efficient policy, we compute an optimal solution, say (x^1, y^1) of the linear program

$$max \left\{ \sum_{(i,a)} r_i^1(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_a x_j(a) + \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} y_i(a) & = & \beta_j,\ j \in S \\ x_i(a), y_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\} \tag{9.30}$$

Next, we solve the following linear program with one additional constraint

$$max \left\{ \sum_{(i,a)} r_i^2(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_a x_j(a) + \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} y_i(a) & = & \beta_j,\ j \in S \\ \sum_{(i,a)} r_i^1(a) x_i(a) & = & \sum_{(i,a)} r_i^1(a) x_i^1(a) \\ x_i(a), y_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\} \tag{9.31}$$

Continuing in this way we stop either when we find for some $1 \leq k \leq m$ an optimal solution (x^k, y^k) in which x^k is unique or when we have solved all m linear programs. Let (x, y) be the finally obtained solution. Then, as shown in Section 9.1.3, we can construct a convergent Markov policy R such that $x(R) = x$. This policy is obviously a lexicografically efficient solution.

9.3 Mean-variance tradeoffs

9.3.1 Formulations of the problem

In many areas of application, a decision maker may wish to incorporate his attitude toward risk or variability when choosing a policy. One measure of risk is the variance of the rewards generated

by a policy. Frequently one considers tradeoffs between return and risk. Examples of this include a dynamic investment model in which the investor may accept a lower than optimal return to achieve reduced variability in return, and a queueing control model, in which the controller might prefer a policy which results in greater but less variable waiting times. These mean-variance tradeoffs may be analyzed in an MDP using the long-run state-action frequencies.

Given an intial distribution β and a policy R the *long-run variance* $V(\beta, R)$ is defined by

$$\begin{aligned} V(\beta,R) &= limsup_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \sum_i \beta_i \cdot \mathbb{E}_{i,R}\{r_{X_t}(Y_t) - \phi(\beta,R)\}^2 \\ &= limsup_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \sum_i \beta_i \cdot \sum_{j,a} \mathbb{P}_R\{X_t = j,\ Y_t = a \mid X_1 = i\}\{r_j(a) - \phi(\beta,R)\}^2 \end{aligned} \tag{9.32}$$

If $R \in C_1$ the long-run state-action frequencies are unique and the long-run variance can be written as

$$\begin{aligned} V(\beta,R) &= \textstyle\sum_{j,a} x_{ja}(R)\{r_j(a) - \phi(\beta,R)\}^2 \\ &= \textstyle\sum_{j,a} x_{ja}(R) r_j^2(a) - 2\sum_{j,a} x_{ja}(R) r_j(a)\phi(\beta,R) + \sum_{j,a} x_{ja}(R)\phi(\beta,R)^2 \\ &= \textstyle\sum_{j,a} x_{ja}(R) r_j^2(a) - \phi(\beta,R)^2 \\ &= \textstyle\sum_{j,a} x_{ja}(R) r_j^2(a) - \{\sum_{j,a} x_{ja}(R) r_j(a)\}^2. \end{aligned} \tag{9.33}$$

There are several ways to consider the mean-variance tradeoffs. Sobel ([189]) proposed to maximize the mean-standard deviation ratio with upper and lower bounds on the mean. This is equivalent to minimizing the ratio of the variance and the square of the mean under the same constraints. In policy space this concept is

$$min \left\{ \frac{V(\beta,R)}{\phi(\beta,R)^2} \;\middle|\; L \le \phi(\beta,R) \le U \right\}.$$

Using the state action frequencies, problem (9.34) becomes

$$min \left\{ \frac{\sum_{j,a} r_j^2(a) x_j(a) - \{\sum_{j,a} r_j(a) x_j(a)\}^2}{\{\sum_{j,a} r_j(a) x_j(a)\}^2} \;\middle|\; x \in Q;\ L \le \sum_{j,a} r_j(a) x_j(a) \le U \right\},$$

which is equivalent to

$$max \left\{ \frac{-\sum_{j,a} r_j^2(a) x_j(a)}{\{\sum_{j,a} r_j(a) x_j(a)\}^2} \;\middle|\; x \in Q;\ L \le \sum_{j,a} r_j(a) x_j(a) \le U \right\}. \tag{9.34}$$

Kawai ([118]) has considered the problem of minimizing the variance subject a lower bounds on the mean. This problem bevome in the state-action space

$$min \left\{ \sum_{j,a} r_j^2(a) x_j(a) - \Big\{ \sum_{j,a} r_j(a) x_j(a) \Big\}^2 \;\middle|\; x \in Q;\ \sum_{j,a} r_j(a) x_j(a) \ge L \right\},$$

which is equivalent to

$$max \left\{ -\sum_{j,a} r_j^2(a) x_j(a) + \Big\{ \sum_{j,a} r_j(a) x_j(a) \Big\}^2 \;\middle|\; x \in Q;\ \sum_{j,a} r_j(a) x_j(a) \ge L \right\}. \tag{9.35}$$

Filar, Kallenberg and Lee ([71]) proposed a variance-penalized version, i.e.

$$max\{\phi(\beta,R)-\lambda\cdot V(\beta,R),$$

or in the x-space

$$max\left\{\sum_{j,a} r_j(a)x_j(a)-\lambda\Big\{\sum_{j,a} r_j^2(a)x_j(a)-\Big\{\sum_{j,a} r_j(a)x_j(a)\Big\}^2\Big\}\;\Big|\; x\in Q\right\}. \tag{9.36}$$

9.3.2 A unifying framework

In [102] Huang and Kallenberg presented a framework that unifies and extends the approaches posed above. This program is reduced to a parametric linear programming problem. This solution method is at least as good as any known method for the particular problems (9.34), (9.35) and (9.36). This unifying framework is given by the nonlinear program

$$max\left\{\frac{\sum_{j,a} B_j(a)x_j(a)}{D\Big(\sum_{j,a} R_j(a)x_j(a)\Big)}+C\Big(\sum_{j,a} R_j(a)x_j(a)\Big)\;\Big|\; x\in Q;\; L\le\sum_{j,a} R_j(a)x_j(a)\le U\right\}, \tag{9.37}$$

with the following assumptions:

(A1) the functions $D(\cdot)$ and $C(\cdot)$ are convex;

(A2) either $D(\cdot)$ is a positive constant; or $D(\cdot)$ is positive and nondecreasing, $C(\cdot)$ is nondecreasing and $\sum_{j,a} B_j(a)x_j(a)\le 0$ for every $x\in Q$.

We now show that (9.34), (9.35) and (9.36) are special cases of 9.37).

(9.34): take $B_j(a)=-r_j^2(a),\ R_j(a)=r_j(a),\ C(y)\equiv 0,\ D(y)\equiv y^2$.

(9.35): take $B_j(a)=-r_j^2(a),\ R_j(a)=r_j(a),\ C(y)\equiv y^2,\ D(y)\equiv 1$ and $U=\infty$.

(9.36): take $B_j(a)=r_j(a)-\lambda r_j^2(a),\ R_j(a)=r_j(a)\sqrt{\lambda},\ C(y)\equiv y^2,\ D(y)\equiv 1$, $L=-\infty$ and $U=\infty$.

9.3.3 Determination of an optimal solution

In order to solve (9.37)) we consider a parametric version of the linear program for average rewards. The parametric objective function is $\sum_{i,a}\{B_i(a)+\lambda R_i(a)\}x_j(a)$. Hence, the parametric linear program is

$$\begin{aligned}
&maximize\{\textstyle\sum_{i,a}\{B_i(a)+\lambda R_i(a)\}x_i(a)\}\\
&\text{subject to}\\
&\textstyle\sum_{(i,a)}\{\delta_{ij}-p_{ij}(a)\}x_i(a) && = 0,\ j\in S\\
&\textstyle\sum_a x_j(a)\ +\ \sum_{(i,a)}\{\delta_{ij}-p_{ij}(a)\}y_i(a) && = \beta_j,\ j\in S\\
&x_i(a),y_i(a) && \ge 0,\ (i,a)\in S\times A
\end{aligned} \tag{9.38}$$

It is well known (see e.g. Zoutendijk [241], p.165) that the optimum objective function of a parametric linear program is a piecewise linear convex function of the parameter λ, and that on each interval of this piecewise linear convex function an optimal solution exists with is an extreme point of the polytope of the constraints. Thus, there exists $\lambda_0 \equiv -\infty < \lambda_1 < \cdots < \lambda_{m-1} < \lambda_m \equiv +\infty$ and extreme optimal solutions (x^n, y^n) for $n = 1, 2, \ldots, m$. Let $k+1$ and $j+1$ be respectively the smallest integers among $1, 2, \ldots, m$ such that

$$\sum_{i,a} R_i(a) x_i^{k+1}(a) > U \text{ and } \sum_{i,a} R_i(a) x_i^{j+1}(a) \geq L.$$

Furthermore, let $\mu, \nu \in [0, 1]$ be such that

$$x^U = \mu x^k + (1-\mu) x^{k+1} \text{ and } x^L = \nu x^j + (1-\nu) x^{j+1}$$

satisfying

$$\sum_{i,a} R_i(a) x_i^U(a) = U \text{ and } \sum_{i,a} R_i(a) x_i^L(a) = L.$$

Let

$$G(x) = \sum_{i,a} B_i(a) x_i(a),\ g(x) = \sum_{i,a} R_i(a) x_i(a) \text{ and } V(x) = \frac{G(x)}{D\big(g(x)\big)} + C\big(g(x)\big) \text{ for } x \in X,$$

and let $G_n = G(x^n),\ g_n = g(x^n)$ and $V^n = V(x^n)$ for $n = 1, 2, \ldots, m$.

Theorem 9.12

(1) The nonlinear program (9.37) is feasible if and only if $g_m \geq L$ and $g_1 \leq U$.

(2) If program (9.37) is feasible with optimum value V_{opt} and optimal solution x_{opt}, then

$$V_{opt} = max\Big\{max_{j+1 \leq n \leq k} V\big(x^n\big), V\big(x^L\big), V\big(x^U\big)\Big\} \text{ and } x_{opt} = \begin{cases} x^n \text{ if } V\big(x^n\big) = V_{opt} \\ x^L \text{ if } V\big(x^L\big) = V_{opt} \\ x^U \text{ if } V\big(x^U\big) = V_{opt} \end{cases}$$

Proof

Part (1)

Since x^n is optimal for (9.37) for $\lambda_{n-1} \leq \lambda \leq \lambda_n$, we have

$$G_n + \lambda g_n \geq G(x) + \lambda g(x),\ \lambda_{n-1} \leq \lambda \leq \lambda_n,\ x \in X \text{ for } n = 1, 2, \ldots, m. \tag{9.39}$$

For $n = 1$, we obtain $G_1 + \lambda g_1 \geq G(x) + \lambda g(x),\ -\infty < \lambda \leq \lambda_1,\ x \in X$. Hence, $g_1 \leq g(x),\ x \in X$. Similarly, for $n = m$, we have $G_m + \lambda g_m \geq G(x) + \lambda g(x),\ \lambda_{m-1} \leq \lambda < +\infty$, and consequently $g_m \geq g(x),\ x \in X$. Therefore, $g_m < L$ or $g_1 > U$ implies infeasibility of the problem.

Conversely, if program (9.37) is feasible, we have $g_1 \leq U$ and $g_m \geq L$.

Part (2)

We first show that $g_1 \le g_{opt} \le g_m$, where $g_{opt} = g(x_{opt})$. Let $G_{opt} = G(x_{opt})$. Again, specifying (9.39) to the cases $n = 1$ and $n = m$ gives for $x = x_{opt}$:

$$G_1 + \lambda g_1 \ge G_{opt} + \lambda g_{opt},\ -\infty < \lambda \le \lambda_1;\ G_m + \lambda g_m \ge G_{opt} + \lambda g_{opt},\ \lambda_{m-1} \le \lambda < +\infty.$$

Letting $\lambda \to -\infty$ in the first inequality and $\lambda \to +\infty$ in the second estableshes

$$g_1 \le g_{opt} \le g_m. \tag{9.40}$$

Also by (9.39)

$$\begin{cases} G_{n+1} + \lambda_{n+1} g_{n+1} \ge G_n + \lambda_{n+1} g_n \\ G_{n+1} + \lambda_n g_{n+1} = G_n + \lambda_n g_n \end{cases} \to\ (\lambda_{n+1} - \lambda_n)(g_{n+1} - g_n) \ge 0,$$

implying

$$g_{n+1} \ge g_n,\ n = 1, 2, \ldots, m-1. \tag{9.41}$$

From $g_m \ge L,\ g_1 \le U$, (9.40) and (9.41) it follows that there exists an index $1 \le p \le m-1$ such that $g_p \le g_{opt} \le g_{p+1}$, such that exactly one of the following is true:

(a) $L \le g_p \le g_{opt} \le g_{p+1} \le U$; (b) $g_p < g\big(x^L\big) = L \le g_{opt} \le g_{p+1} \le U$;

(c) $L \le g_p \le g_{opt} \le U = g\big(x^U\big) < g_{p+1}$; (d) $g_p < g\big(x^L\big) = L \le g_{opt} \le U = g\big(x^U\big) < g_{p+1}$.

Case d

In this case, we have $j = k = p$, and therefore $V_{opt} = max\Big\{V\big(x^L\big),\ V\big(x^U\big)\Big\}$. By (9.39) for $n = p$, we obtain

$$G_{p+1} + \lambda_p g_{p+1} = G_p + \lambda_p g_p \ge G(x) + \lambda_p g(x),\ x \in X.$$

Since x^L and x^U are convex combinations of x^p and x^{p+1}, we also have

$$G\big(x^L\big) + \lambda_p g\big(x^L\big) = G\big(x^U\big) + \lambda_p g\big(x^U\big) = G_{p+1} + \lambda_p g_{p+1} = G_p + \lambda_p g_p \ge G(x) + \lambda_p g(x),\ x \in X. \tag{9.42}$$

For two distinct real numbers y and z, let $c(y,z) = \frac{C(y)-C(z)}{y-z}$ and $d(y,z) = \frac{D(y)-D(z)}{y-z}$.

We claim that

$$D\big(g(x^U)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^L)}{D\big(g(x^L)\big)} d\big(g(x^L), g(x^U)\big) \ge \lambda_p \tag{9.43}$$

if and only if

$$D\big(g(x^L)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^U)}{D\big(g(x^U)\big)} d\big(g(x^L), g(x^U)\big) \ge \lambda_p. \tag{9.44}$$

From (9.42) it follows that $G(x^U) = G(x^L) + \lambda_p\{g(x^L) - g(x^U)\}$. Since $V(x) = \frac{G(x)}{D\big(g(x)\big)} + C\big(g(x)\big)$, we have

$$\begin{aligned} V(x^U) - V(x^L) &= \frac{G(x^U)}{D\big(g(x^U)\big)} + C\big(g(x^U)\big) - \frac{G(x^L)}{D\big(g(x^L)\big)} - C\big(g(x^L)\big) \\ &= \frac{1}{D\big(g(x^U)\big)} \Big\{ G(x^U) + D\big(g(x^U)\big)C\big(g(x^U)\big) - \frac{G(x^L)D\big(g(x^U)\big)}{D\big(g(x^L)\big)} - D\big(g(x^U)\big)C\big(g(x^L)\big) \Big\} \\ &= \frac{g(x^U)-g(x^L)}{D\big(g(x^U)\big)} \Big\{ \frac{G(x^U)}{g(x^U)-g(x^L)} + D\big(g(x^U)\big) \cdot \frac{C(g(x^U))-C(g(x^L))}{g(x^U)-g(x^L)} - \frac{G(x^L)D(g(x^U))}{D(g(x^L))\{g(x^U)-g(x^L)\}} \Big\}. \end{aligned}$$

Since by (9.42) $G(x^U) = G(x^L) - \lambda_p\{g(x^U) - g(x^L)\}$, we can write $\frac{G(x^U)}{g(x^U)-g(x^L)} = \frac{G(x^L)}{g(x^U)-g(x^L)} - \lambda_p$.
Substituting this expression yields

$$\begin{aligned} V(x^U) - V(x^L) &= \frac{g(x^U)-g(x^L)}{D\big(g(x^U)\big)}\Big\{\frac{G(x^L)}{g(x^U)-g(x^L)} - \lambda_p + D\big(g(x^U)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^L)D(g(x^U))}{D(g(x^L))\{g(x^U)-g(x^L)\}}\Big\} \\ &= \frac{g(x^U)-g(x^L)}{D\big(g(x^U)\big)}\Big\{D\big(g(x^U)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^L)}{D(g(x^L))}d\big(g(x^L), g(x^U)\big) - \lambda_p\Big\}. \end{aligned}$$

Similarly, we can write

$$\begin{aligned} V(x^U) - V(x^L) &= \frac{G(x^U)}{D\big(g(x^U)\big)} + C\big(g(x^U)\big) - \frac{G(x^L)}{D\big(g(x^L)\big)} - C\big(g(x^L)\big) \\ &= \frac{1}{D\big(g(x^L)\big)}\Big\{\frac{G(x^U)D\big(g(x^L)\big)}{D\big(g(x^U)\big)} + D\big(g(x^L)\big)C\big(g(x^U)\big) - G(x^L) - D\big(g(x^L)\big)C\big(g(x^L)\big)\Big\} \\ &= \frac{g(x^U)-g(x^L)}{D\big(g(x^L)\big)}\Big\{\frac{G(x^U)D(g(x^L))}{D(g(x^U))\{g(x^U)-g(x^L)\}} + D\big(g(x^L)\big)\cdot\frac{C(g(x^U))-C(g(x^L))}{g(x^U)-g(x^L)} - \frac{G(x^L)}{g(x^U)-g(x^L)}\Big\}. \end{aligned}$$

Substituting, again by (9.42), $G(x^L) = G(x^U) + \lambda_p\{g(x^U) - g(x^L)\}$, gives

$$\begin{aligned} V(x^U) - V(x^L) &= \frac{g(x^U)-g(x^L)}{D\big(g(x^L)\big)}\Big\{\frac{G(x^U)D(g(x^L))}{D(g(x^U))\{g(x^U)-g(x^L)\}} + D\big(g(x^L)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^U)}{g(x^U)-g(x^L)} - \lambda_p\Big\} \\ &= \frac{g(x^U)-g(x^L)}{D\big(g(x^L)\big)}\Big\{D\big(g(x^L)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^U)}{D(g(x^U))}d\big(g(x^L), g(x^U)\big) - \lambda_p\Big\}. \end{aligned}$$

Hence, (9.43) and (9.44) are equivalent if and only if $D\big(g(x^U)\big)$ and $D\big(g(x^L)\big)$ have the same sign.
By assumption (A2) this is true.

Next, we establish that $V_{opt} = max\{V(x^L), V(x^U)\}$ and $x_{opt} = \begin{cases} x^L \text{ if } V(x^L) = V_{opt}; \\ x^U \text{ if } V(x^U) = V_{opt}. \end{cases}$

We distinguish between two cases:

(1) $D\big(g(x^L)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^U)}{D(g(x^U))}d\big(g(x^L), g(x^U)\big) \geq \lambda_p$;

(2) $D\big(g(x^L)\big)c\big(g(x^L), g(x^U)\big) - \frac{G(x^U)}{D(g(x^U))}d\big(g(x^L), g(x^U)\big) < \lambda_p$.

smallskip Case (1):

We can write, using (9.42),

$$\begin{aligned} 0 \leq V(x_{opt}) - V(x^U) &= \frac{G(x_{opt})}{D\big(g(x_{opt})\big)} + C\big(g(x_{opt})\big) - \frac{G(x^U)}{D\big(g(x^U)\big)} - C\big(g(x^U)\big) \\ &\leq C\big(g(x_{opt})\big) - C\big(g(x^U)\big) + \frac{G(x^U)+\lambda_p\{(g(x^U)-g(x_{opt})\}}{D\big(g(x_{opt})\big)} - \frac{G(x^U)}{D\big(g(x^U)\big)} \\ &= \frac{g(x_{opt})-g(x^U)}{D\big(g(x_{opt})\big)}\Big\{D\big(g(x_{opt})\big)c\big(g(x_{opt}), g(x^U)\big) - \frac{G(x^U)}{D\big(g(x^U)\big)}d\big(g(x_{opt}), g(x^U)\big) - \lambda_p\Big\}. \end{aligned}$$

Case (1a): D is a constant, i.e. $d(\cdot,\cdot)\equiv 0$, and by Case (1), $c\big(g(x^L),g(x^U)\big)\geq\frac{\lambda_p}{D}$.
The above inequality becomes: $0\leq V(x_{opt})-V(x^U)\leq\{g(x_{opt})-g(x^U)\}\big\{c\big(g(x_{opt}),g(x^U)\big)-\frac{\lambda_p}{D}\big\}$.
The convexity of C implies $c(y,z)\leq c(x,z)$ for all x,y,z with $x\leq y\leq z$. Since we consider Case d, we have $g\big(x^L\big)\leq g(x_{opt})\leq g\big(x^U\big)$ and therefore, $c\big(g(x_{opt}),g(x^U)\big)\leq c\big(g(x^L),g(x^U)\big)$.
Consequently

$$0\leq V(x_{opt})-V(x^U)\leq\{g(x_{opt}))-g(x^U)\}\big\{c\big(g(x^L),g(x^U)\big)-\tfrac{\lambda_p}{D}\big\}\leq 0,$$

implying $V(x_{opt})=V(x^U)$ and $x_{opt}=x^U$.
Case (1b): D is not a constant and $D\big(g(x^L)\big)c\big(g(x^L),g(x^U)\big)-\frac{G(x^U)}{D(g(x^U))}d\big(g(x^L),g(x^U)\big)\geq\lambda_p$.
We have seen that

$$0\leq V(x_{opt})-V(x^U)\leq\tfrac{g(x_{opt})-g(x^U)}{D\left(g(x_{opt})\right)}\Big\{D\big(g(x_{opt})\big)c\big(g(x_{opt}),g(x^U)\big)-\tfrac{G(x^U)}{D\left(g(x^U)\right)}d\big(g(x_{opt}),g(x^U)\big)-\lambda_p\Big\}.$$

Because C is convex and D is nondecreasing and convex:

$$c\big(g(x_{opt}),g(x^U)\big)\geq c\big(g(x^L),g(x^U)\big);\ D\big(g(x_{opt})\big)\geq D\big(g(x^L);\ d\big(g(x_{opt}),g(x^U)\big)\geq d\big(g(x^L),g(x^U)\big).$$

Since $G(x^U)\leq 0$, we obtain

$$0\leq V(x_{opt})-V(x^U)\leq\tfrac{g(x_{opt})-g(x^U)}{D\left(g(x_{opt})\right)}\Big\{D\big(g(x^L)\big)c\big(g(x^L),g(x^U)\big)-\tfrac{G(x^U)}{D\left(g(x^U)\right)}d\big(g(x^L),g(x^U)\big)-\lambda_p\Big\}.$$

On the other hand, $g(x_{opt})\leq g(x^U)$ and $D\big(g(x^L)\big)c\big(g(x^L),g(x^U)\big)-\frac{G(x^U)}{D(g(x^U))}d\big(g(x^L),g(x^U)\big)\geq\lambda_p$.
So,

$$\tfrac{g(x_{opt})-g(x^U)}{D\left(g(x_{opt})\right)}\Big\{D\big(g(x^L)\big)c\big(g(x^L),g(x^U)\big)-\tfrac{G(x^U)}{D\left(g(x^U)\right)}d\big(g(x^L),g(x^U)\big)-\lambda_p\Big\}\leq 0.$$

Hence, $V(x_{opt})=V(x^U)$ and $x_{opt}=x^U$.
Case (2):
Similarly as in Case (1) we can write, using (9.42),

$$\begin{aligned}0\ \leq\ V(x_{opt})-V(x^L)\ &=\ \tfrac{G(x_{opt})}{D\left(g(x_{opt})\right)}+C\big(g(x_{opt})\big)-\tfrac{G(x^L)}{D\left(g(x^L)\right)}-C\big(g(x^L)\big)\\ &\leq\ C\big(g(x_{opt})\big)-C\big(g(x^L)\big)+\tfrac{G(x^L)+\lambda_p\{(g(x^L)-g(x_{opt})\}}{D\left(g(x_{opt})\right)}-\tfrac{G(x^L)}{D\left(g(x^L\right)}\\ &=\ \tfrac{g(x_{opt})-g(x^L)}{D\left(g(x_{opt})\right)}\Big\{D\big(g(x_{opt})\big)c\big(g(x_L),g(x_{opt})\big)-\tfrac{G(x^L)}{D\left(g(x^L)\right)}d\big(g(x_L),g(x_{opt})\big)-\lambda_p\Big\}.\end{aligned}$$

Case (2a): D is a constant, i.e. $d(\cdot,\cdot)\equiv 0$, and by Case (2), $c\big(g(x^L),g(x^U)\big)<\frac{\lambda_p}{D}$.
The above inequality becomes: $0\leq V(x_{opt})-V(x^L)\leq\{g(x_{opt})-g(x^L)\}\big\{c\big(g(x_L),g(x_{opt})\big)-\frac{\lambda_p}{D}\big\}$.
The convexity of C and $g\big(x^L\big)\leq g(x_{opt})\leq g\big(x^U\big)$ imply, $c\big(g(x^L),g(x_{opt})\big)\leq c\big(g(x^L),g(x^U)\big)$.
Consequently

$$0\leq V(x_{opt})-V(x^L)\leq\{g(x_{opt}))-g(x^L)\}\big\{c\big(g(x^L),g(x^U)\big)-\tfrac{\lambda_p}{D}\big\}\leq 0,$$

implying $V(x_{opt})=V(x^L)$ and $x_{opt}=x^L$.

Case (2b): D is not a constant and $D\big(g(x^L)\big)c\big(g(x^L),g(x^U)\big) - \frac{G(x^U)}{D(g(x^U)}d\big(g(x^L),g(x^U)\big) < \lambda_p$.

We have seen that

$$0 \le V(x_{opt}) - V(x^U) \le \frac{g(x_{opt})-g(x^L)}{D\big(g(x_{opt}\big)}\Big\{D\big(g(x_{opt})\big)c\big(g(x^L),g(x_{opt})\big) - \frac{G(x^L)}{D\big(g(x^L)\big)}d\big(g(x^L),g(x_{opt})\big) - \lambda_p\Big\}.$$

Since C is convex and D is nondecreasing and convex:

$$c\big(g(x^L),g(x_{opt})\big) \le c\big(g(x^L),g(x^U)\big);\ D\big(g(x_{opt})\big) \le D\big(g(x^U);\ c\big(g(x^L),g(x_{opt})\big) \le c\big(g(x^L),g(x^U)\big).$$

Since $G(x^L) \le 0$, we obtain

$$0 \le V(x_{opt}) - V(x^U) \le \frac{g(x_{opt})-g(x^L)}{D\big(g(x_{opt}\big)}\Big\{D\big(g(x^U)\big)c\big(g(x^L),g(x^U)\big) - \frac{G(x^L)}{D\big(g(x^L)\big)}d\big(g(x^L),g(x^U)\big) - \lambda_p\Big\}.$$

On the other hand, $g(x_{opt}) \ge g(x^L)$ and $D\big(g(x^U)\big)c\big(g(x^L),g(x^U)\big) - \frac{G(x^L)}{D(g(x^L))}d\big(g(x^L),g(x^U)\big) < \lambda_p$, the last inequility by the equivalence of (9.43) and (9.44), So,

$$\frac{g(x_{opt})-g(x^L)}{D\big(g(x_{opt}\big)}\Big\{D\big(g(x^U)\big)c\big(g(x^L),g(x^U)\big) - \frac{G(x^L)}{D\big(g(x^L)\big)}d\big(g(x^L),g(x^U)\big) - \lambda_p\Big\} \le 0.$$

Hence, $V(x_{opt}) = V(x^L)$ and $x_{opt} = x^L$.

The proofs for the cases (a), (b) and (c) can be obtained in a similar way. Instead of x^L and x^U we take: in case (a): x^p and x^{p+1}; in case (b): x^L and x^{p+1}; in case (c): x^p and x^U. □

9.3.4 Determination of an optimal policy

Theorem 9.12 provides an optimal solution for program (9.37), but it does not provide a procedure to construct an optimal policy for the mean-variance problem.

Theorem 9.13

Let (x,y) be an extreme optimal solution for prgram (9.38) for all λ in an open interval I. Then, there exists a policy $f^\infty \in C(D)$ whose limiting state-action frequenties vector $x(f)$ satisfies $\sum_{(i,a)} B_i(a)x_{ia}(f) = \sum_{(i,a)} B_i(a)x_i(a)$, $\sum_{(i,a)} R_i(a)x_{ia}(f) = \sum_{(i,a)} R_i(a)x_i(a)$, $V\big(x(f)\big) = V(x)$.

Proof

Let f^∞ be a policy satisfying $x_i\big(f(i)\big) > 0,\ i \in S_x;\ y_i\big(f(i)\big) > 0,\ i \in S_y$ and $f(i)$ arbitrarily chosen for $i \notin S_x \cup S_y$. From Theorem 9.10 it follows that f^∞ is β-optimal for all $\lambda \in I$. Define the policy f^∞ by (5.30). Then, for $r_i(a) = B_i(a) + \lambda R_i(a),\ (i,a) \in S \times A$, we have

$$\sum_{(i,a)} r_i(a)x_{ia}(f) = \sum_i r_i(f) \cdot \sum_j \beta_j \{P^*(f)\}_{ji} \sum_k \beta_j \cdot \sum_i \{P^*(f)\}_{ji} r_i(f) = \phi(\beta, f^\infty),$$

i.e. $\big(x(f), y(f)\big)$ is also an optimal solution of (9.38). Therefore,

$$\sum_{(i,a)}\{B_i(a) + \lambda R_i(a)\}x_i(a) = \sum_{(i,a)}\{B_i(a) + \lambda R_i(a)\}x_{ia}(f) \text{ for all } \lambda \in I.$$

Therefore, $\sum_{(i,a)}\{B_i(a)x_i(a) = \sum_{(i,a)} B_i(a)x_{ia}(f)$ and $\sum_{(i,a)} R_i(a)\}x_i(a) = \sum_{(i,a)} R_i(a)\}x_{ia}(f)$. Since $g(x) = R_i(a)\}x_i(a),\ G(x) = \sum_{(i,a)}\{B_i(a)x_i(a)$ and $V(x) = \frac{G(x)}{D\big(g(x)\big)} + C\big(g(x)\big)$, we have $g(x) = g\big(x(f)\big),\ G(x) = G\big(x(f)\big)$, implying $V(x) = V\big(x(f)\big)$. □

Theorem 9.14

If program (9.37) is feasible, then either $x_{opt} = x^n$ for some $j+1 \le n \le k$ and there exists an optimal deterministic policy, or $x_{opt} = x^L$ (or x^U) and an initial randomization of two deterministic policies is optimal for the mean-variance tradeoffs problem.

Proof

Suppose that $x_{opt} = x^n$ for some $j+1 \le n \le k$. Since x^n is optimal for all $\lambda \in [\lambda_{n-1}, \lambda_n]$ and $\lambda_{n-1} < \lambda_n$, by Theorem 9.13, there exists a policy f^∞ whose limiting state-action frequencies vector $x(f)$ satisfies $\sum_{(i,a)} B_i(a)x_{ia}(f) = \sum_{(i,a)} B_i(a)x_i^n(a)$, $\sum_{(i,a)} R_i(a)x_{ia}(f) = \sum_{(i,a)} R_i(a)x_i^n(a)$ and $V\big(x(f)\big) = V(x^n)$. Because $x(f)$ also satisfies the constraint $L \le \sum_{(i,a)} R_i(a)x_i(a) \le U$, f^∞ is an optimal policy.

Next, suppose that $x_{opt} = x^L$, where $x^L = \nu x^j + (1-\nu)x^{j+1}$ and $\sum_{(i,a)} R_i(a)x_i^L(a) = L$ (the case $x_{opt} = x^U$ can be shown similarly). By Theorem 9.13, corresponding to x^j and x^{j+1}, there are policies $f_j^\infty, f_{j+1}^\infty \in C(D)$ whose limiting state-action frequenties vectors $x(f_j)$ and $x(f_{j+1})$ satisfy $\sum_{(i,a)} B_i(a)x_{ia}(f_j) = \sum_{(i,a)} B_i(a)x_i^j(a)$, $\sum_{(i,a)} R_i(a)x_{ia}(f_j) = \sum_{(i,a)} R_i(a)x_i^j(a)$ and $\sum_{(i,a)} B_i(a)x_{ia}(f_{j+1}) = \sum_{(i,a)} B_i(a)x_i^{j+1}(a)$, $\sum_{(i,a)} R_i(a)x_{ia}(f_{j+1}) = \sum_{(i,a)} R_i(a)x_i^{j+1}(a)$, respectively. Then, setting $x^* = \nu x(f_j) + (1-\nu)x(f_{j+1})$, we obtain

$$\begin{aligned}\sum_{(i,a)} R_i(a)x_i^*(a) &= \sum_{(i,a)} R_i(a)\{\nu x_i^j(a) + (1-\nu)x_i^{j+1}(a)\} \\ &= \sum_{(i,a)} R_i(a)\{\nu x_i^j(a) + (1-\nu)x_i^{j+1}(a)\} = \sum_{(i,a)} R_i(a)x_i^L(a) = L\end{aligned}$$

and

$$\begin{aligned}\sum_{(i,a)} B_i(a)x_i^*(a) &= \sum_{(i,a)} B_i(a)\{\nu x_i^j(a) + (1-\nu)x_i^{j+1}(a)\} \\ &= \sum_{(i,a)} B_i(a)\{\nu x_i^j(a) + (1-\nu)x_i^{j+1}(a)\} = \sum_{(i,a)} B_i(a)x_i^L(a).\end{aligned}$$

Hence, $V(x_{opt}) = V(x^L) = V(x^*)$. From Theorem 1.1 it follows that the policy R_* which initially randomizes between f_j^∞ and f_{j+1}^∞ with coefficients ν and $1-\nu$ yields as state-action frequencies vector $x(R_*) = \nu x(f_j) + (1-\nu)x(f_{j+1}) = x^*$. Therefore, R_* is an optimal policy for the mean-variance tradeoffs problem. □

Corollary 9.1

For an unconstrained problem, i.e. without the constraint $L \le \sum_{(i,a)} R_i(a)x_i(a) \le U$, there exists a deterministic optimal policy.

Remark

1. The average-unichain case can be treated in a similar way, but is more simple: the initial distribution β has no influence and the parametric linear progam has only the x-variables.
2. The optimal policy R_* is also *Pareto optimal* with respect to the pair $\big(\phi(R), -V(R)\big)$, i.e. there does not exist a policy R such that $\phi(R) \ge \phi(R_*)$ and $V(R) \le V(R_*)$, with a strict inequality holding for at least one of the two inequalities.

9.4 Bibliographic notes

The first reference on MDPs with additional constraints is the paper of Derman and Klein ([56]). Derman was the first who presented a comprehensive treatment to analyze a constrained MDP ([55], chapter 7). He introduced the state-action frequency approach for the analysis of these problems, and developed its relationship to linear programming. Derman and Veinott ([59]) analyzed CMDPs by applying the Dantzig-Wolfe decomposition principle. Kallenberg ([108]) and Hordijk and Kallenberg ([98] and [99]) developed further properties of the sets of limiting state action frequencies, and extended the linear programming approach to include constrained multichain models. Altman, Hordijk and Kallenberg studied the value function for constrained MDPs ([2]).

The sensitivity of CMDPs was considered by Altman and Shwartz ([3]). White ([231]) and Beutler and Ross ([17]) used Lagrange multipliers to analyse constrained models. A more recent comprehensive survey of constrained MDPs with an emphasis on the Lagrange approach is Altman's book ([1]). We also mension some papers on CMDPs written by Ross and Varadarajan ([165], [166], [167]). Contrubutions on multi-objective MDPs were given by White ([232]) for the discounted case and by Durinovic, Lee, Kathehakis and Filar ([62]) for the undiscounted case.

Sobel ([189]) and Chung ([34], [35], [36]) considered the mean-variance ratio with a lower bound on the mean. Kawai ([118]) investigated the minimization of the variance with a lower bound on the mean. White ([233]) surveyed various models with mean-variance criteria amd reviewed the importance of and relationship between the limiting state action frequencies in different classes of models. Filar, Kallenberg and Lee ([71]) and White ([234], [235]) analyzed the variance penalized model. Other contributions to the literature on mean-variance tradeoffs are Kawai and Katoh ([119]), Bayal-Gursoy and Ross ([9]) and Sobel ([190]). Huang and Kallenberg ([102]) presented a framework that unifies and extends most of these approaches.

9.5 Exercises

Exercise 9.1
Consider the following MDP model: $S = \{1, 2\}$; $A(1) = \{1, 2\}$, $A(2) = \{1\}$;
$p_{11}(1) = 1$, $p_{12}(1) = 0$; $p_{11}(2) = 0$, $p_{12}(2); = 1$; $p_{21}(1) = 0$, $p_{22}(1) = 1$; $\beta_1 = \beta_2 = \frac{1}{2}$.
Determine the set Q of the long-run average state-action frequenties.

Exercise 9.2
Show by a counterexample that in the multichain case $x(\pi^\infty)$ is in general not continuous in π.

Exercise 9.3
Consider the inventory model with backlogging of Example 1.1. The state represent the inventory on hand and negative states represent backlogged orders. Suppose that we are interested in maximizing the long-run average profit, subject to the requirement that the average probability that there is out of stock is at most γ. Formulate the constraint of this optimization problem.

Exercise 9.4

Consider the following irreducible MDP model: $S = \{1, 2\}$; $A(1) = \{1\}$, $A(2) = \{1, 2, 3\}$;
$p_{11}(1) = 0.4$, $p_{12}(1) = 0.6$; $p_{21}(1) = 1$, $p_{22}(1) = 0$; $p_{21}(2) = 0.8$, $p_{22}(2) = 0.2$;
$p_{21}(3) = 0.3$, $p_{22}(3) = 0.7$.

a. Determine an average optimal deterministic policy f^∞ by linear programming.

b. Add the constraint that the limiting state-action frequencies in state 2 is no more than 0.4 and solve the constrained model, i.e. determine an optimal stationary policy π^∞.

Exercise 9.5

Consider a unichain multi-objective MDP with immediate rewards $r_i^k(a)$, $k = 1, 2, \ldots, m$.
Let x be an optimal solution of the linear program

$$max \left\{ \sum_{(i,a)} r_i(a) x_i(a) \;\middle|\; \begin{array}{rcl} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = & 0,\ j \in S \\ \sum_{(i,a)} x_i(a) & = & 1 \\ x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\},$$

where $r_i(a) = \sum_{k=1}^m \lambda_k r_i^k(a)$ for some $\lambda \in \mathbb{R}^m$ with $\lambda_k > 0$, $k = 1, 2, \ldots, m$.

Define the stationary policy π^∞ by $\pi_{ia} = \begin{cases} x_i(a)/x_i & \text{if } x_i > 0; \\ \text{arbitrary} & \text{if } x_i = 0. \end{cases}$

Show that policy π^∞ is a β-efficient solution for any initial distribution β.

Exercise 9.6

Prove Lemma 9.14.

Exercise 9.7

Consider the following irreducible MDP model: $S = \{1, 2\}$; $A(1) = \{1, 2\}$, $A(2) = \{1\}$.
$p_{11}(1) = 0.8$, $p_{12}(1) = 0.2$; $p_{11}(2) = 0$, $p_{12}(2) = 1$; $p_{21}(1) = 1$, $p_{22}(1) = 0$.
$r_1(1) = 1$, $r_1(2) = 0$, $r_2(1) = 3$.

a. Determine for the two deterministic stationary policies the average reward and the variance.

b. Consider the mean-variance tradeoffs problem $min\{V(R) \mid \phi(R) \geq \frac{17}{12}\}$.

   (1) Formulate the parametric linear program for this problem and solve it.

   (2) Determine the optimal solution x_{opt} of problem (9.37) and the optimum value $V(x_{opt})$.

   (3) Determine an optimal policy according to the proof of Theorem 9.14.

   (4) Show that if π^∞ is the stationary policy which randomizes in state 1 betweeen the two actions with the same randomization as the optimal policy uses between the two deterministic policies, then π^∞ is not an optimal policy for the mean-variance tradeoffs problem.

   (5) Try to find a stationary policy π^∞ with the same average reward and variance as the policy of part (3).

Chapter 10

Stochastic Games

10.1 Introduction

10.1.1 The model

In this chapter we consider *two-person zero-sum stochastic games*. As in MDPs a stochastic game is a dynamic system that evolves along discrete time points. The state of the system at every time point is assumed to be one of a finite set $S = \{1, 2, \ldots, N\}$. At these discrete time points each of the two players has the possibility to earn rewards and to influence the course of the system by choosing, independently of the choice of the other player, an action out of a finite action set. Let $A(i)$ and $B(i)$ be the action sets of player 1 and player 2, respectively, in state $i,\ i \in S$. If in state i player 1 chooses action $a \in A(i)$ and player 2 action $b \in B(i)$ then two things happen:

(1) Player 1 earns an immediate reward $r_i(a, b)$ from player 2 (*zero-sum game*);

(2) The next state is determined by a transitions which depend on the actions a and b, i.e. the state of the next decision time point is state j with probability $p_{ij}(a, b),\ j \in S$, where $\sum_j p_{ij}(a, b) = 1$ for every $i \in S,\ a \in A(i)$ and $b \in B(i)$.

Consider the Cartesian product

$$S \times A \times B = \{(i, a, b) \mid i \in S,\ a \in A(i),\ b \in B(i)\}$$

and let H_t denote the set of the possible *histories* of the system up to time point t, i.e.

$$H_t := \{h_t = (i_1, a_1, b_1, \ldots, i_{t-1}, a_{t-1}, b_{t-1}, i_t \mid (i_k, a_k, b_k) \in S \times A \times B,\ 1 \le k \le t-1;\ i_t \in S\}.$$

A decision rule π^t at time point t for player 1 is a function on H_t which prescribes the action to be taken at time t as a transition probability from H_t into A, i.e.

$$\pi^t_{h_t a_t} \ge 0 \text{ for every } a_t \in A(i_t) \text{ and } \sum_{a_t} \pi^t_{h_t a_t} = 1 \text{ for every } h_t \in H_t.$$

A policy R_1 for player 1 is a sequence of decision rules: $R_1 = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$, where π^t is the decision rule at time point $t,\ t = 1, 2, \ldots$. Similarly, the concept of a decision rule and a policy for player 2 is defined. As in the MDP model we distinguish between Markov, stationary and deterministic policies.

For stationary policies π^∞ and ρ^∞ for player 1 and 2, respectively, the transition matrix $P(\pi,\rho)$ and the reward vector $r(\pi,\rho)$ are defined by

$$p_{ij}(\pi,\rho) = \sum_{a,b} p_{ij}(a,b)\pi_{ia}\rho_{ib} \text{ for every } (i,j) \in S \times S; \tag{10.1}$$

$$r_i(\pi,\rho) = \sum_{a,b} r_i(a,b)\pi_{ia}\rho_{ib} \text{ for every } i \in S. \tag{10.2}$$

Furthermore, we define

$$p_{ij}(a,\rho) = \sum_b p_{ij}(a,b)\rho_{ib},\ i,j \in S,\ a \in A(i);\ p_{ij}(\pi,b) = \sum_a p_{ij}(a,b)\pi_{ia},\ i,j \in S,\ b \in B(i);$$

$$r_i(a,\rho) = \sum_b r_i(a,b)\rho_{ib},\ i \in S,\ a \in A(i);\ r_i(\pi,b) = \sum_a r_i(a,b)\pi_{ia},\ i \in S\ ,\ b \in B(i).$$

10.1.2 Optimality criteria

Let $X_t,\ Y_t,\ Z_t$ be random variables denoting the observed state, the action chosen by player 1 and the action chosen by player 2, respectively, at time point t. For any two policies R_1 and R_2 for player 1 and player 2, respectively, and initial state i, we denote the *total expected discounted reward* and the *average expected reward* by $v_i^\alpha(R_1,R_2)$ and $\phi_i(R_1,R_2)$, defined by

$$v_i^\alpha(R_1,R_2) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{j,a,b} \mathbb{P}_{i,R_1,R_2}\{X_t = j, Y_t = a, Z_t = b\} \cdot r_j(a,b). \tag{10.3}$$

and

$$\phi_i(R_1,R_2) = liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j,a,b} \mathbb{P}_{i,R_1,R_2}\{X_t = j, Y_t = a, Z_t = b\} \cdot r_j(a,b). \tag{10.4}$$

The *total expected reward*, given initial state i and the policies R_1 and R_2 is denoted by $v_i(R_1,R_2)$ and defined by

$$v_i(R_1,R_2) = \sum_{t=1}^{\infty} \sum_{j,a,b} \mathbb{P}_{i,R_1,R_2}\{X_t = j, Y_t = a, Z_t = b\} \cdot r_j(a,b), \tag{10.5}$$

under the following assumptions:

(1) The model is *substochastic*, i.e. $\sum_j p_{ij}(a,b) \le 1$ for all $(i,a,b) \in S \times A \times B$.

(2) For any initial state i and any two policies R_1, R_2 the expected total reward $v_i(R_1,R_2)$ is well-defined (possibly $\pm\infty$).

Under the assumption that the model is *transient*, i.e. $\sum_{t=1}^{\infty} \mathbb{P}_{i,R_1,R_2}\{X_t = j, Y_t = a, Z_t = b\} < \infty$ for all i,j and a,b, it can be shown that, with $\alpha = 1$, most properties of the discounted model are valid for the total reward.

10.1.3 Matrix games

For the solution of stochastic games we sometimes make use of properties of matrix games. Therefore we present in this section a number of concepts and results in the theory of matrix games.[1] A *two-person zero-sum matrix game* can be represented by an $m \times n$-matrix $A = (a_{ij})$, the *game matrix* or *payoff matrix*. The actions of player 1 correspond to the rows and the actions of player 2 to the columns of A. When player 1 chooses row i and player 2 column j, player 2 has to pay the amount a_{ij} to player 1. If player 1 chooses row i, he will get at least $min_j\, a_{ij}$. Hence, by an optimal choice of row i, he can achieve $\underline{w}(A) = max_i min_j\, a_{ij}$. Similarly, player 2 can obtain a payoff of at most $\overline{w}(A) = min_j max_i\, a_{ij}$. It is well known that

$$\overline{w}(A) = min_i max_j\, a_{ij} \geq max_i min_j\, a_{ij} = \underline{w}(A).$$

Let us now allow the choice of a strategy by a player to be random. The set of *mixed strategies* of player 1 is the simplex $X = \{(x_1, x_2, \ldots, x_m) \mid x_i \geq 0,\ 1 \leq i \leq m;\ \sum_{i=1}^m\ x_i = 1\}$. An element $x \in X$ is the probability on the set of rows of A. Similarly, the set of mixed strategies of player 2 is the simplex $Y = \{(y_1, y_2, \ldots, y_n) \mid y_j \geq 0,\ 1 \leq j \leq n;\ \sum_{j=1}^n\ y_j = 1\}$. If player 1 uses $x \in X$ and player 2 $y \in Y$, the (average) payoff is $x^T A y = \sum_{i=1}^m \sum_{j=1}^n\ x_i a_{ij} y_j$.

Note that the *pure strategy* for player 1 of choosing row i may be represented as the mixed strategy e_i, the unit vector with a 1 in the i-th position and 0's elsewhere. Similarly, the pure strategy for player 2 of choosing column j may be represented as the mixed strategy e_j.

It is natural to consider the mixed *maxmin* and *minmax*, namely $\underline{v}(A) = max_{x \in X} min_{y \in Y}\, x^T A y$ and $\overline{v}(A) = min_{y \in Y} max_{x \in X}\, x^T A y$. Since $min_{y \in Y}\, x^T A y = min_j\, x^T A e_j$, we can write

$$\underline{v}(A) = max_{x \in X} min_{y \in Y}\, x^T A y = max_{x \in X} min_j\, x^T A e_j \geq max_i min_j\, a_{ij} = \underline{w}(A).$$

Similarly, $\overline{w}(A) \geq \overline{v}(A)$, implying

$$\overline{v}(A) - \underline{v}(A) \leq \overline{w}(A) - \underline{w}(A),$$

i.e. mixed strategies reduce the 'duality gap'. Since $max_{x \in X}\, x^T A y \geq max_{x \in X} min_{y \in Y}\, x^T A y$ for all $y \in Y$, we obtain

$$\overline{v}(A) = min_{y \in Y} max_{x \in X}\, x^T A y \geq max_{x \in X} min_{y \in Y}\, x^T A y = \underline{v}(A).$$

The matrix game with payoff matrix A has a *value* $val(A)$ if $val(A) = \overline{v}(A) = \underline{v}(A)$. The policy $x^* \in X$ is an *optimal policy for player 1* if

$$(x^*)^T A y \geq \overline{v}(A) \text{ for all } y \in Y.$$

The policy $y^* \in Y$ is an *optimal policy for player 2* if

$$x^T A y^* \leq \underline{v}(A) \text{ for all } x \in X.$$

The basic Minmax Theorem for two-person zero-sum matrix games proves that the game has a value and that both players have optimal mixed strategies.

[1] For a comprehensive survey of matrix games we refer to Owen, G.: *Game Theory*, Academic Press, 1982.

Theorem 10.1 *Minmax theorem*

Two-person zero-sum matrix games have a value and both players have optimal mixed strategies.

Proof

Consider the linear programming problem

$$min \left\{ y_0 \;\middle|\; y_0 \geq \sum_{j=1}^{n} a_{ij} y_j,\ 1 \leq i \leq m;\ \sum_{j=1}^{n} y_j = 1;\ y_j \geq 0,\ 1 \leq j \leq n \right\} \tag{10.6}$$

with corresponding dual program

$$max \left\{ x_0 \;\middle|\; x_0 \leq \sum_{i=1}^{m} a_{ij} x_i, 1 \leq j \leq n;\ \sum_{i=1}^{m} x_i = 1;\ x_i \geq 0, 1 \leq i \leq m \right\}. \tag{10.7}$$

Let (y_0^*, y^*) and (x_0^*, x^*) be optimal solutions of (10.6) and (10.7), respectively. Take any $x \in X$ and $y \in Y$. Then, we can write

$$y_0^* = \textstyle\sum_{i=1}^{m} x_i y_0^* \geq \sum_{i=1}^{m} x_i \sum_{j=1}^{n} a_{ij} y_j^* = x^T A y^*$$

and

$$x_0^* = \textstyle\sum_{j=1}^{n} y_j x_0^* \leq \sum_{j=1}^{n} y_j \sum_{i=1}^{m} a_{ij} x_i^* = (x^*)^T A y.$$

Hence, $x^T A y^* \leq y_0^* = x_0^* \leq (x^*)^T A y$ for all $x \in X$ and $y \in Y$. Therefore,

$$x_0^* = y_0^* = (x^*)^T A y^* \text{ and } (x^*)^T A y^* = max_{x \in X}\, x^T A y^* \text{ and } (x^*)^T A y^* = min_{y \in Y}\, (x^*)^T A y,$$

implying

$$\begin{aligned} \underline{v}(A) &= max_{x \in X} min_{y \in Y}\, x^T A y \;\geq\; min_{y \in Y}\, (x^*)^T A y \;=\; (x^*)^T A y^* \;=\; max_{x \in X}\, x^T A y^* \\ &\geq min_{y \in Y} max_{x \in X}\, x^T A y = \overline{v}(A). \end{aligned}$$

Since we also have $\overline{v}(A) \geq \underline{v}(A)$, we have shown that $\overline{v}(A) = \underline{v}(A) = val(A)$ and $(x^*)^T A y \geq val(A)$ for all $y \in Y$ and $x^T A y^* \leq val(A)$, i.e. x^* and y^* are optimal policies for player 1 and 2, respectively. □

The simplest case of all occurs if a *saddle point* exists, i.e. there exists an entry a_{kl} which is both the maximum entry in its column and the minimum entry in its row. In this case the pure strategies row k for player 1 and column l for player 2 are optimal strategies as the following lemma shows.

Lemma 10.1

If $a_{kl} \geq a_{il}$ for all i and $a_{kl} \leq a_{kj}$ for all j, then $x = e_k$ and $y = e_l$ are optimal pure strategies for player 1 and 2, respectively, and a_{kl} is the value of the game.

Proof

The result follows immediately from the following observation.

$a_{kl} = max_i\, a_{il} \geq min_j max_i\, a_{ij} = \overline{w}(A) \geq val(A) \geq \underline{w}(A) = max_i min_j\, a_{ij} \geq min_j\, a_{kj} = a_{kl}.$ □

Suppose that player 1 has the pure optimal strategy e_k. From $(e_k)^T Ay \geq val(A) = (e_k)^T Ay^*$ for all $y \in Y$ and some $y^* \in Y$ it follows that $a_{kj} \geq val(A) = (e_k)^T Ay^*$ for $j = 1, 2, \ldots, n$. Therefore, e_l is an optimal pure strategy for player 2, where l satisfies $a_{kl} = min_j\, a_{kj}$. Since e_l is an optimal pure strategy for player 2, we also have $x^T Ae_l \leq val(A) = a_{kl}$ for all $x \in X$, implying $a_{il} \leq a_{kl}$ for $i = 1, 2, \ldots, m$. Hence, A has a saddle point a_{kl} and we obtain the following result.

Lemma 10.2

If one of the players has a pure optimal strategy, both players have optimal pure strategies and the game has a saddle point.

Lemma 10.3

(1) For any $c \in \mathbb{R}$ and any $m \times n$-matrix A, $val(A + cJ) = val(A) + c$, where J is the $m \times n$-matrix with each entry equal to 1.

(2) For any two $m \times n$-matrices A and B with $a_{ij} \leq b_{ij}$ for all (i, j), we have $val(A) \leq val(B)$.

(3) For any two $m \times n$-matrices A and B, $|val(A) - val(B)| \leq max_{(k,l)} |a_{kl} - b_{kl}|$.

Proof

(1) and (2): Since $x^T(A + cJ)y = x^T Ay + c$ and $x^T Ay \leq x^T By$, it is straightforward that $val(A + cJ) = val(A) + c$ and $val(A) \leq val(B)$.

(3) Notice that $a_{ij} - max_{(k,l)} |a_{kl} - b_{kl}| \leq b_{ij} \leq a_{ij} + max_{(k,l)} |a_{kl} - b_{kl}|$ for all (i, j). Hence, by (1) and (2), $val(A) \leq val(B) + max_{(k,l)} |a_{kl} - b_{kl}|$ and $val(B) \leq val(A) + max_{(k,l)} |a_{kl} - b_{kl}|$, implying $|val(A) - val(B)| \leq max_{(k,l)} |a_{kl} - b_{kl}|$. □

2×2 games

Suppose we are given the 2×2 matrix game $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$. It may be that this game has a saddle point; if so, this entry is the value and provides the optimal strategies which are pure. Suppose that the game has no saddle point. Then, by Lemma 10.2 both players have completely mixed optimal strategies x and y, i.e. $x_1 > 0,\ x_2 > 0,\ y_1 > 0$ and $y_2 > 0$. For the value of the game we have $val(A) = x_1\{a_{11}y_1 + a_{12}y_2\} + x_2\{a_{21}y_1 + a_{22}y_2\}$. The two terms between brackets are at most $val(A)$ (see the linear program (10.6)), we have $val(A) = a_{11}y_1 + a_{12}y_2 = a_{21}y_1 + a_{22}y_2$. Similarly, it can been seen that $val(A) = a_{11}x_1 + a_{21}x_2 = a_{12}x_1 + a_{22}x_2$. In vector notation, $v = Ay$ and $v = A^T x$, where $v = \binom{val(A)}{val(A)}$.

If A is nonsingular, we can write

$$
\begin{aligned}
v = A^T x \quad &\rightarrow \quad (A^{-1})^T v = x \ \rightarrow\ v^T A^{-1} e = \{(A^{-1})^T v\}^T e = x^T e = 1 \ \rightarrow\ val(A) \cdot e^T A^{-1} e = 1 \\
&\rightarrow \quad val(A) = \tfrac{1}{e^T A^{-1} e}. \\
v = Ay \quad &\rightarrow \quad y = A^{-1} v = val(A) \cdot A^{-1} e = \tfrac{A^{-1} e}{e^T A^{-1} e}. \\
v = A^T x \quad &\rightarrow \quad x = (A^{-1})^T v = val(A) \cdot (A^{-1})^T e = \tfrac{(A^{-1})^T e}{e^T A^{-1} e}.
\end{aligned}
$$

If A is singular, this is of course meaningless. Then, it can be shown that

$$val(A) = \frac{|A|}{e^T A^* e} = \frac{a_{11}a_{22}-a_{12}a_{21}}{a_{11}+a_{22}-a_{12}-a_{21}},\ y = \frac{A^* e}{e^T A^* e} = \left(\frac{a_{22}-a_{21}}{a_{11}+a_{22}-a_{12}-a_{21}}, \frac{a_{11}-a_{12}}{a_{11}+a_{22}-a_{12}-a_{21}}\right),$$

$$x = \frac{A^* e}{e^T A^* e} = \left(\frac{a_{22}-a_{12}}{a_{11}+a_{22}-a_{12}-a_{21}}, \frac{a_{11}-a_{21}}{a_{11}+a_{22}-a_{12}-a_{21}}\right).$$

where $|A|$ the determinant of A and A^* is the adjoint of A. Note that the formulas in the nonsingular case coincide with the above formulas, because $A^*A = AA^* = |A| \cdot I$. For the details on the adjoint of A and the property $A^*A = AA^* = |A| \cdot I$ we refer to text books on linear algebra.

10.2 Discounted rewards

10.2.1 Value and optimal policies

A policy R_1^* is *optimal for player 1* if $v^\alpha(R_1^*, R_2) \geq inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2)$ for all policies R_2.

A policy R_2^* is *optimal for player 2* if $v^\alpha(R_1, R_2^*) \leq sup_{R_1}\, inf_{R_2}\, v^\alpha(R_1, R_2)$ for all policies R_1.

The stochastic discounted game has a *value* if $inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2) = sup_{R_1}\, inf_{R_2}\, v^\alpha(R_1, R_2)$.

A policy R_1^* is *ε-optimal for player 1* if $v^\alpha(R_1^*, R_2) \geq inf_{R_2} sup_{R_1} v^\alpha(R_1, R_2) - \varepsilon$ for all policies R_2.

A policy R_2^* is *ε-optimal for player 2* if $v^\alpha(R_1, R_2^*) \leq sup_{R_1} inf_{R_2} v^\alpha(R_1, R_2) + \varepsilon$ for all policies R_1.

Theorem 10.2
If the policies R_1^* and R_2^* satisfy $v^\alpha(R_1, R_2^*) \leq v^\alpha(R_1^*, R_2^*) \leq v^\alpha(R_1^*, R_2)$ for all policies R_1 and R_2, the game has a value and R_1^* and R_2^* are optimal policies.

Proof
We can write

$$\begin{aligned} sup_{R_1}\, inf_{R_2}\, v^\alpha(R_1, R_2) &\geq inf_{R_2}\, v^\alpha(R_1^*, R_2) \geq v^\alpha(R_1^*, R_2^*) \\ &\geq sup_{R_1}\, v^\alpha(R_1, R_2^*) \geq inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2). \end{aligned}$$

On the other hand, $inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2) \geq inf_{R_2}\, v^\alpha(R_1, R_2)$ for all policies R_1, implying $inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2) \geq sup_{R_1} inf_{R_2}\, v^\alpha(R_1, R_2)$. Hence, we have shown that $inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2) = sup_{R_1} inf_{R_2}\, v^\alpha(R_1, R_2) = v^\alpha(R_1^*, R_2^*)$ i.e. the game has a value.
Since $v^\alpha(R_1^*, R_2) \geq inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2)$ for all R_2 and $v^\alpha(R_1, R_2^*) \leq sup_{R_1}\, inf_{R_2}\, v^\alpha(R_1, R_2)$ for all R_1, i.e. R_1^* and R_2^* are optimal policies. □

We will show in this section that the game has a value and that there exist stationary optimal policies for both players. Furthermore, we present algorithms to approximate the value and stationary optimal policies arbitrarily close. Let Π and Γ be the set of stationary policies for player 1 and 2, respectively. Define for any $x \in \mathbb{R}^N$ the mapping $T : \mathbb{R}^N \to \mathbb{R}^N$ by

$$(Tx)_i = inf_{\rho^\infty \in \Gamma}\, sup_{\pi^\infty \in \Pi}\, \{r_i(\pi, \rho) + \alpha \sum_j p_{ij}(\pi, \rho) x_j\},\ i \in S. \tag{10.8}$$

$(Tx)_i$ is the value of a matrix game with matrix $M_x[i]$. The matrix $M_x[i]$ has $m = \#A(i)$ rows and $n = \#B(i)$ columns and the payoff, if player 1 chooses row a and player 2 column b, is $r_i(a,b) + \alpha \sum_j p_{ij}(a,b)x_j$. We will show in the next theorem that T is a monotone contraction.

Theorem 10.3

The mapping T, defined in (10.8), is a monotone contraction with respect to the supremum norm $\|\cdot\|_\infty$ with contraction factor α and fixed point $v^\alpha = inf_{R_1}\, sup_{R_2}\, v^\alpha(R_1,R_2)$.

Proof

Let $x, y \in \mathbb{R}^N$ with $x \leq y$. Take any $i \in S$. Then, $\{M_x[i]\}_{ab} \leq \{M_y[i]\}_{ab}$ for all (a,b). By Lemma 10.3 part (2),

$$(Tx)_i = val(M_x[i]) \leq val(M_y[i]) = (Ty)_i,$$

proving the monotonicity.

$\|Tx - Ty\|_\infty = max_i|(Tx)_i - (Ty)_i|$. Notice that $|(Tx)_i - (Ty)_i| = |val(M_x[i]) - val(M_y[i])|$.

By Lemma 10.3 part (3), we can write,

$$\begin{aligned} |val(M_x[i]) - val(M_y)[i]| &\leq max_{(a,b)}\, |\{r_i(a,b) + \alpha \textstyle\sum_j p_{ij}(a,b)x_j\{-\{r_i(a,b) + \alpha \sum_j p_{ij}(a,b)y_j\}| \\ &= \alpha \cdot max_{(a,b)}\, |\textstyle\sum_j p_{ij}(a,b)(x_j - y_j)| \leq \alpha \cdot \|x - y\|_\infty, \end{aligned}$$

implying that T is a contraction with contraction factor α.

Hence, T has a unique fixed point, say v^α. We now show that there exist stationary policies $(\pi^*)^\infty$ and $(\rho^*)^\infty$ such that $v^\alpha\big(\pi^\infty, (\rho^*)^\infty\big) \leq v^\alpha \leq v^\alpha\big((\pi^*)^\infty, \rho^\infty\big)$ for every $\pi^\infty \in \Pi$ and $\rho^\infty \in \Gamma$.

Let π^* be such that $\pi^*_{ia},\ a \in A(i)$, is an optimal mixed strategy in the matrix game with matrix $\{r_i(a,b) + \alpha \sum_j p_{ij}(a,b)v^\alpha_j\}$, which - because of the fixed point property - has value $v^\alpha_i,\ i \in S$.

So, $r(\pi^*, \rho) + \alpha P(\pi^*, \rho)v^\alpha \geq v^\alpha$ for all $\rho^\infty \in \Gamma$, implying $v^\alpha\big((\pi^*)^\infty, \rho^\infty\big) \geq v^\alpha$ for every $\rho^\infty \in \Gamma$.

Similarly, it can be shown that $v^\alpha\big(\pi^\infty, (\rho^*)^\infty\big) \leq v^\alpha$ for every $\pi^\infty \in \Pi$. Therefore

$$v^\alpha\big(\pi^\infty, (\rho^*)^\infty\big) \leq v^\alpha \leq v^\alpha\big((\pi^*)^\infty, \rho^\infty\big) \text{ for every } \pi^\infty \in \Pi \text{ and } \rho^\infty \in \Gamma. \tag{10.9}$$

As in the proof of Theorem 10.2, we obtain from these inequalities

$$v^\alpha = v^\alpha\big((\pi^*)^\infty, (\rho^*)^\infty\big) = inf_{\rho^\infty \in \Gamma}\, sup_{\pi^\infty \in \Pi}\, v^\alpha(\pi^\infty, \rho^\infty) = sup_{\pi^\infty \in \Pi}\, inf_{\rho^\infty \in \Gamma}\, v^\alpha(\pi^\infty, \rho^\infty). \tag{10.10}$$

Finally we show that $v^\alpha = inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1,R_2) = sup_{R_1}\, min_{R_2}\, v^\alpha(R_1,R_2)$.

Since $sup_{R_1}\, v^\alpha(R_1,R_2) \geq sup_{R_1}\, min_{R_2}\, v^\alpha(R_1,R_2)$ for all policies R_2, we have

$$inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1,R_2) \geq sup_{R_1}\, min_{R_2}\, v^\alpha(R_1,R_2).$$

Take any fixed policy ρ^∞ for player 2. This induces an MDP, so we obtain

$$sup_{R_1}\, v^\alpha(R_1,\rho^\infty) = max_{\pi^\infty \in \Pi}\, v^\alpha(\pi^\infty, \rho^\infty) \text{ for any fixed } \rho^\infty \in \Gamma$$

and similarly

$$inf_{R_2}\, v^\alpha(\pi^\infty, R_2) = min_{\rho^\infty\in\Gamma}\, v^\alpha(\pi^\infty, \rho^\infty) \text{ for any fixed } \pi^\infty \in \Pi.$$

Because

$$\begin{aligned} sup_{R_1}\, inf_{R_2}\, v^\alpha(R_1, R_2) &\geq sup_{\pi^\infty\in\Pi}\, inf_{R_2}\, v^\alpha(\pi^\infty, R_2) = sup_{\pi^\infty\in\Pi}\, inf_{\rho^\infty\in\Gamma}\, v^\alpha(\pi^\infty, \rho^\infty) \\ &= v^\alpha = inf_{\rho^\infty\in\Gamma}\, sup_{\pi^\infty\in\Pi}\, v^\alpha(\pi^\infty, \rho^\infty) = inf_{\rho^\infty\in\Gamma}\, sup_{R_1}\, v^\alpha(R_1, \rho^\infty) \\ &\geq inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2), \end{aligned}$$

we have shown that $v^\alpha = inf_{R_2}\, sup_{R_1}\, v^\alpha(R_1, R_2) = sup_{R_1}\, min_{R_2}\, v^\alpha(R_1, R_2)$. □

Corollary 10.1

The game has a value v^α, which satisfies $v_i^\alpha = val(M_{v^\alpha}[i])$, $i \in S$. Furthermore, there are stationary optimal policies for both players.

Proof

From the last line of the proof of Theorem 10.3 we obtain that v^α is value of the game. Since v^α is the unique fixed point of T, we have $v_i^\alpha = val(M_{v^\alpha}[i])$, $i \in S$. Furthermore, we can write,

$$v^\alpha\big((\pi^*)^\infty, R_2\big) \geq inf_{R_2}\, v^\alpha\big((\pi^*)^\infty, R_2\big) = inf_{\rho^\infty\in\Gamma}\, v^\alpha\big((\pi^*)^\infty, \rho^\infty\big) = v^\alpha,$$

the last equality by (10.9), i.e. $(\pi^*)^\infty$ is an optimal policy for player 1. Similarly, we have

$$v^\alpha\big(R_1, (\rho^*)^\infty\big) \leq sup_{R_1}\, v^\alpha\big(R_1, (\rho^*)^\infty\big) = sup_{\pi^\infty\in\Pi}\, v^\alpha\big((\pi^\infty, (\rho^*)^\infty\big) = v^\alpha,$$

i.e. $(\rho^*)^\infty$ is an optimal policy for player 2. □

Example 10.1

$S = \{1, 2\};\ A(1) = B(1) = \{1, 2\},\ A(2) = B(2) = \{1\};\ \alpha = \frac{1}{2}.$

$r_1(1,1) = \frac{1}{2},\ r_1(1,2) = 1,\ r_1(2,1) = 3,\ r_1(2,2) = \frac{3}{2},\ r_2(1,1) = 1.$

$p_{11}(1,1) = \frac{1}{3},\ p_{12}(1,1) = \frac{2}{3};\ p_{11}(1,2) = 0,\ p_{12}(1,2) = 1;\ p_{11}(2,1) = 0,\ p_{12}(2,1) = 1;$

$p_{11}(2,2) = \frac{1}{2},\ p_{12}(2,2) = \frac{1}{2};\ p_{21}(1,1) = 0,\ p_{22}(1,1) = 1.$

Consider the fixed point equation $x = Tx$, i.e.

$$x_1 = val\begin{pmatrix} \frac{1}{2} + \frac{1}{6}x_1 + \frac{1}{3}x_2 & 1 + \frac{1}{2}x_2 \\ 3 + \frac{1}{2}x_2 & \frac{3}{2} + \frac{1}{4}x_1 + \frac{1}{4}x_2 \end{pmatrix};\ x_2 = val\big(1 + \tfrac{1}{2}x_2\big).$$

Hence, $v_2^\alpha = x_2 = 2$ and $x_1 = val\begin{pmatrix} \frac{5}{6} + \frac{1}{6}x_1 & 2 \\ 4 & 2 + \frac{1}{4}x_1 \end{pmatrix}$. Since the maximum reward is $\frac{3}{2}$, the total expected discounted reward is at most $\frac{3/2}{1-\alpha} = 3$. Therefore the second row of the matrix dominates the first one and player 1 and 2 will both choose the second action:

$x_1 = 2 + \frac{1}{4}x_1 \ \rightarrow\ v_1^\alpha = x_1 = \frac{8}{3}.$

Perfect information

A stochastic game is said to be a game of *perfect information* if the state space S can be divided into two disjoint sets S_1 and S_2 such that $|A(i)| = 1$ for $i \in S_1$ and $|B(i)| = 1$ for $i \in S_2$. Then, the matrices in the matrix game with matrix M_x has either one row (if $i \in S_1$) or one column (if $i \in S_2$). Hence, the optimal policies are pure, i.e. nonrandomized, and we obtain the following result.

Corollary 10.2

In a discounted stochastic game with perfect information, both players possess optimal deterministic policies.

Remark

It is an open problem to find an efficient finite algorithm for this class of games.

10.2.2 Mathematical programming

A vector $v \in \mathbb{R}^N$ is called *superharmonic* if there exists a policy $\rho^\infty \in \Gamma$ such that

$$v_i \geq r_i(a, \rho) + \alpha \sum_j p_{ij}(a, \rho) v_j, \ a \in A(i), \ i \in S.$$

A vector $v \in \mathbb{R}^N$ is called *subharmonic* if there exists a policy $\pi^\infty \in \Pi$ such that

$$v_i \leq r_i(\pi, b) + \alpha \sum_j p_{ij}(\pi, \rho) v_j, \ b \in B(i), \ i \in S.$$

Theorem 10.4

(1) The value vector v^α is the smallest superharmonic vector.

(2) The value vector v^α is the largest subharmonic vector.

Proof

Let $(\pi^*)^\infty$ and $(\rho^*)^\infty$ be the policies mentioned in Theorem 10.3. If player 2 uses policy $(\rho^*)^\infty$, then the game becomes an MDP. We know from Theorem 3.16 that $x = sup_{R_1} v^\alpha(R_1, (\rho^*)^\infty)$ is the smallest superharmonic vector. Since $v^\alpha = v^\alpha((\pi^*)^\infty, (\rho^*)^\infty)$, we have $x \geq v^\alpha$. On the other hand, it follows from the proof of Corollary 10.1 that $x = sup_{R_1} v^\alpha(R_1, (\rho^*)^\infty) \leq v^\alpha$.

The proof of part (2) is analogous to the proof of part (1). □

Consider the two nonlinear programs

$$min \left\{ \sum_i v_i \ \middle| \ \begin{array}{rcl} \sum_j \{\delta_{ij} - \alpha \sum_b p_{ij}(a,b)\rho_{ib}\} v_j & - & \sum_b r_i(a,b)\rho_{ib} \geq 0, \ a \in A(i), \ i \in S \\ & & \sum_b \rho_{ib} = 1, \ i \in S \\ & & \rho_{ib} \geq 0, \ b \in B(i), \ i \in S \end{array} \right\} \tag{10.11}$$

and

$$max \left\{ \sum_i w_i \ \middle| \ \begin{array}{rcl} \sum_j \{\delta_{ij} - \alpha \sum_a p_{ij}(a,b)\pi_{ia}\} w_j & - & \sum_a r_i(a,b)\pi_{ia} \leq 0, \ b \in B(i), \ i \in S \\ & & \sum_a \pi_{ia} = 1, \ i \in S \\ & & \pi_{ia} \geq 0, \ a \in A(i), \ i \in S \end{array} \right\}. \tag{10.12}$$

Theorem 10.5

The nonlinear programs (10.11) and (10.12) have both optimal solutions, say (v^*, ρ^*) *and* (w^*, π^*). *Furthermore,* $v^* = w^* = v^\alpha$ *and* $(\pi^*)^\infty$ *and* $(\rho^*)^\infty$ *are optimal policies for player 1 and player 2.*

Proof

From Theorem 10.4 it follows that both nonlinear programs have optimal solutions and that $v^* = w^* = v^\alpha$. The constraints of the programs imply

$$r(\pi, \rho^*) + \alpha P(\pi, \rho^*)v^\alpha \leq v^\alpha \leq r(\pi^*, \rho) + \alpha P(\pi^*, \rho)v^\alpha \text{ for all } \pi \text{ and } \rho.$$

Therefore, $\{I - \alpha P(\pi, \rho^*)\}v^\alpha \geq r(\pi, \rho^*)$ and $\{I - \alpha P(\pi^*, \rho)\}v^\alpha \leq r(\pi^*, \rho)$ for all π and ρ. Hence, $v^\alpha(\pi^\infty, (\rho^*)^\infty) = \{I - \alpha P(\pi, \rho^*)\}^{-1}r(\pi, \rho^*) \leq v^\alpha \leq \{I - \alpha P(\pi^*, \rho)\}^{-1}r(\pi^*, \rho) = v^\alpha((\pi^*)^\infty, \rho^\infty)$ for all $\pi^\infty \in \Pi$ and $\rho^\infty \in \Gamma$. Then, by the proof of Corollary 10.1, it follows that $(\pi^*)^\infty$ and $(\rho^*)^\infty)$ are optimal policies for player 1 and player 2. □

10.2.3 Iterative methods

Since T is a contraction with fixed point the value vector v^α, it follows that the value iteration algorithm stated below approximately computes v^α.

Algorithm 10.1 *Value Iteration for discounted games*

1. Choose $\varepsilon > 0$ and $x \in \mathbb{R}^N$ arbitrary.
2. Compute for each $i \in S$: $y_i = val(M_x[i])$, where $M_x[i]$ is the matrix with entries $r_i(a, b) + \alpha \sum_j p_{ij}(a, b)x_j,\ a \in A(i),\ b \in B(i)$.
3. If $\|y - x\|_\infty \leq (1 - \alpha)\alpha^{-1}\varepsilon$:
   a. Determine for each $i \in S$ an optimal strategy $\pi^*_{ia},\ a \in A(i)$, for player 1 in the matrix game $M_x[i]$;
   b. Determine for each $i \in S$ an optimal strategy $\rho^*_{ib},\ b \in B(i)$ for player 2 in the matrix game $M_x[i]$;
   c. y is an ε-approximation of the value vector v^α and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are 2ε-optimal policies for player 1 and 2, respectively (STOP);

   Otherwise: $x := y$ and return to step 2.

Theorem 10.6
Algorithm 10.1 is correct.

Proof
Since T is a monotone contraction with contraction factor α and fixed point v^α, it follows from Corollary 3.1 that $\|v^\alpha - y\|_\infty \le \alpha(1-\alpha)^{-1}\|y-x\|_\infty \le \varepsilon$, i.e. y is a ε-approximation of the value vector v^α.
For any two policies $\pi^\infty \in \Pi$ and $\rho^\infty \in \Gamma$, we define the operator $L_{\pi,\rho} : \mathbb{R}^N \to \mathbb{R}^N$ by

$$L_{\pi,\rho}x = r(\pi,\rho) + \alpha P(\pi,\rho)x.$$

It is straightforward to show that $L_{\pi,\rho}$ is a monotone contraction with contraction factor α and fixed point $v^\alpha(\pi^\infty,\rho^\infty)$. Because $(\pi^*)^\infty$ is an optimal policy in the matrix games of step 2 of Algorithm 10.1, which have values $y_i,\ i \in S$, we can write

$$L_{\pi^*,\rho}x = r(\pi^*,\rho) + \alpha P(\pi^*,\rho)x \ge y = x + (y-x) \ge x - \|y-x\|_\infty \cdot e \ge x - \frac{1-\alpha}{\alpha}\varepsilon \cdot e \quad (10.13)$$

Hence, applying $L_{\pi^*,\rho}$ to (10.13), we also have

$$L^2_{\pi^*,\rho}x \ge L_{\pi^*,\rho}\{x - \tfrac{1-\alpha}{\alpha}\varepsilon \cdot e\} = L_{\pi^*,\rho}x - (1-\alpha)\varepsilon \cdot e \ge y - (1-\alpha)\varepsilon \cdot e.$$

By iterating (10.13), we obtain $L^n_{\pi^*,\rho}x \ge y - (1-\alpha)\{1+\alpha+\cdots+\alpha^{n-2}\}\varepsilon \cdot e$. Taking the limit for $n \to \infty$ yields $v^\alpha\big((\pi^*)^\infty,\rho\big) \ge y - \varepsilon \cdot e \ge v^\alpha - 2\varepsilon \cdot e$. Since the fixed stationary policy $(\pi^*)^\infty$ induces an MDP, we also have $v^\alpha\big((\pi^*)^\infty,R_2\big) \ge v^\alpha - 2\varepsilon \cdot e$, i.e. $(\pi^*)^\infty$ is an 2ε-optimal policy for player 1. Similarly, it can be shown that $(\rho^*)^\infty$ is an 2ε-optimal policy for player 2. □

Example 10.1 (continued)
We apply Algorithm 10.1 with $\varepsilon = 0.2$ and starting value $x = (2,2)$.
Iteration 1:

$$i=1:\ y_1 = val\begin{pmatrix} \frac{3}{2} & 2 \\ 4 & \frac{5}{2} \end{pmatrix} = \tfrac{5}{2};\ i=2:\ y_2 = val(2) = 2;\ x = \big(\tfrac{5}{2},2\big).$$

Iteration 2:

$$i=1:\ y_1 = val\begin{pmatrix} \frac{19}{12} & 2 \\ 4 & \frac{21}{8} \end{pmatrix} = \tfrac{21}{8};\ i=2:\ y_2 = val(2) = 2;\ \|y-x\| = 18 < (1-\alpha)\alpha^{-1}\varepsilon = 0.2.$$

$\big(\frac{21}{8},2\big)$ is a 0.2-approximation of the value vector v^α; f^∞_* with $f(1)=2,\ f(2)=1$ is a 0.4-optimal policy for player 1 and g^∞_* with $g(1)=2,\ g(2)=1$ is a 0.4-optimal policy for player 2.

Algorithm 10.1 does not utilize the information contained in the optimal strategies of the matrix games at each iteration. The next algorithm attemps to improve the basis scheme of Algorithm 10.1 by using these optimal strategies. This algorithm iterates in both value space and policy space.

Algorithm 10.2 *Value Iteration for discounted games (Modification 1)*

1. Choose $\varepsilon > 0$ and choose a stationary policy $(\rho^*)^\infty$ for player 2.
2. Solve the MDP induced by the policy $(\rho^*)^\infty$: $x = max_{f^\infty \in C(D)}\, v^\alpha(f^\infty, (\rho^*)^\infty)$.
3. Compute for each $i \in S$:
   a. $y_i = val(M_x[i])$, where $M_x[i]$ is the matrix with entries
   $r_i(a,b) + \alpha \sum_j p_{ij}(a,b)x_j,\ a \in A(i),\ b \in B(i)$;
   b. $\rho^*_{ib},\ b \in B(i)$, an optimal mixed policy for player 2 in the matrix game of 3a.
4. If $\|y - x\|_\infty \leq (1-\alpha)\alpha^{-1}\varepsilon$:
   a. Determine for each $i \in S$ an optimal strategy $\pi^*_{ia},\ a \in A(i)$, for player 1 in the matrix game $M_x[i]$;
   b. y is an ε-approximation of the value vector v^α and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are 2ε-optimal policies for player 1 and 2, respectively (STOP);

   Otherwise: return to step 2.

Example 10.1 (continued)

We apply Algorithm 10.2 with $\varepsilon = 0.2$ and starting policy $\rho^*_{11} = \rho^*_{12} = \frac{1}{2};\ \rho^*_{21} = 1$.

$r_1(1,\rho^*) = \frac{3}{4},\ r_1(2,\rho^*) = \frac{5}{4},\ r_2(1,\rho^*) = 1;\ p_{11}(1,\rho^*) = \frac{1}{6},\ p_{12}(1,\rho^*) = \frac{5}{6}$;

$p_{11}(2,\rho^*) = \frac{1}{4},\ p_{12}(2,\rho^*) = \frac{3}{4};\ p_{21}(2,\rho^*) = 0,\ p_{22}(2,\rho^*) = 1$.

Iteration 1:

$x = \left(\frac{16}{7}, 2\right)$.

$i = 1$: $y_1 = val\begin{pmatrix} \frac{50}{21} & 2 \\ 4 & \frac{18}{7} \end{pmatrix} = \frac{18}{7};\ \rho^*_{11} = 0,\ \rho^*_{12} = 1.\ i = 2:\ x_2 = val(2) = 2;\ \rho^*_{21} = 1.$

Iteration 2:

$r_1(1,\rho^*) = 1,\ r_1(2,\rho^*) = \frac{3}{2},\ r_2(1,\rho^*) = 1;\ p_{11}(1,\rho^*) = \frac{1}{6},\ p_{12}(1,\rho^*) = \frac{5}{6}$;

$p_{11}(2,\rho^*) = \frac{1}{2},\ p_{12}(2,\rho^*) = \frac{1}{2};\ p_{21}(2,\rho^*) = 0,\ p_{22}(2,\rho^*) = 1$.

$x = \left(\frac{8}{3}, 2\right)$.

$i = 1$: $y_1 = val\begin{pmatrix} \frac{29}{18} & 2 \\ 4 & \frac{8}{3} \end{pmatrix} = \frac{8}{3};\ \ \rho^*_{11} = 0,\ \rho^*_{12} = 1.\ i = 2:\ y_2 = val(2) = 2;\ \ \rho^*_{21} = 1.$

$\|y - x\| = 0 < (1-\alpha)\alpha^{-1}\varepsilon = 0.2$.

$\left(\frac{8}{3}, 2\right)$ is a 0.2-approximation of the value vector v^α; f^∞_* with $f(1) = 2,\ f(2) = 1$ is a 0.4-optimal policy for player 1 and g^∞_* with $g(1) = 2,\ g(2) = 1$ is a 0.4-optimal policy for player 2.

Theorem 10.7
Algorithm 10.2 is correct.

Proof
Let x^n and y^n be the values of x and y in iteration n; let f_n^∞ be the optimal policy obtained in step 2 in iteration n; let π^n and ρ^n be the optimal mixed strategies of the two players obtained in step 3 in iteration n. Then,

$$x^n = r(f_n, \rho^{n-1} + \alpha P(f_n, \rho^{n-1})x^n \geq r(\pi, \rho^{n-1}) + \alpha P(\pi, \rho^{n-1})x^n,\ \pi^\infty \in \Pi. \tag{10.14}$$

and

$$r(\pi^n, \rho) + \alpha P(\pi^n, \rho)x^n \geq r(\pi^n, \rho^n) + \alpha P(\pi^n, \rho^n)x^n = y^n \geq r(\pi, \rho^n) + \alpha P(\pi, \rho^n)x^n,\ \pi^\infty \in \Pi,\ \rho^\infty \in \Gamma. \tag{10.15}$$

Hence, $y^n \leq r(\pi^n, \rho^{n-1}) + \alpha P(\pi^n, \rho^{n-1})x^n \leq x^n$. From (10.14) and the monotonicity of $L_{\pi,\rho}$ it follows that $y^n \geq L_{f_{n+1},\rho^n} x^n \geq L_{f_{n+1},\rho^n} x^n$, implying $y^n \geq v^\alpha\big(f_{n+1}^\infty, (\rho^n)^\infty\big)x^n = x^{n+1}$.
So we obtain the sequence $x^0 \geq y^0 \geq x^1 \geq y^1 \geq \cdots \geq x^n \geq y^n \geq \cdots$, bounded below by $\frac{-1}{1-\alpha} \cdot max_{i,a,b} |r_i(a,b)| \cdot e$. Therefore, $lim_{n\to\infty} x^n = lim_{n\to\infty} y^n = x^*$ for some $x^* \in \mathbb{R}^N$.
Since the sets Π and Γ are compact, there are subsequences $\{n_k\}_{k=1}^\infty$ such that $\pi^{n_k} \to \pi^*$ and $\rho^{n_k} \to \rho^*$ for some $(\pi^*)^\infty \in \Pi$ and $(\rho^*)^\infty \in \Gamma$. From (10.14) it follows that

$$r(\pi^*, \rho) + \alpha P(\pi^*, \rho)x^* \geq x^* \geq r(\pi, \rho^*) + \alpha P(\pi, \rho^*)x^*,\ \pi^\infty \in \Pi,\ \rho^\infty \in \Gamma,$$

implying $v^\alpha\big((\pi^*)^\infty, \rho^\infty\big) \geq x^* \geq v^\alpha\big(\pi^\infty, (\rho^*)^\infty\big),\ \pi^\infty \in \Pi,\ \rho^\infty \in \Gamma$.
Hence, x^* is the value, and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal policies and the algorithm terminates.
Let x and y be the vectors at termination of the algorithm. Then, we can write

$$\|y - v^\alpha\|_\infty = \|Tx - Tv^\alpha\|_\infty \leq \alpha\|x - v^\alpha\|_\infty \leq \alpha\|x - y\|_\infty + \alpha\|y - v^\alpha\|_\infty.$$

Therefore, $\|y - v^\alpha\|_\infty \leq \alpha(1-\alpha)^{-1}\|x - y\|_\infty < \varepsilon$ at termination, i.e. y is an ε-approximation of the value vector v^α. Similarly as in the proof of Theorem 10.6 we can show that the policies $(\pi^*)^\infty$ and $(\rho^*)^\infty$, defined in the steps 4a and 3b, respectively, are 2ε-optimal policy for the players. □

In the next algorithm the optimal mixed strategies of the two players obtained by the matrix game $M_x[i]$ are used in another way.

Algorithm 10.3 *Value Iteration for discounted games (Modification 2)*

1. Choose $\varepsilon > 0,\ x \in \mathbb{R}^N$ and let $\beta = \frac{\alpha}{1-\alpha} \cdot max_i\big\{\sum_j \{max_{(a,b)} p_{ij}(a,b) - min_{(a,b)} p_{ij}(a,b)\}\big\}$.
2. Compute for each $i \in S$:
   a. $y_i = val(M_x[i])$;
   b. $\pi_{ia},\ a \in A(i)$, an optimal mixed policy for player 1 in the matrix game of 2a.
   c. $\rho_{ib},\ b \in B(i)$, an optimal mixed policy for player 2 in the matrix game of 2a.

3. Compute $z = v^{\alpha}(\pi^{\infty}, \rho^{\infty})$.

4. If $\|z - x\|_{\infty} \leq \frac{1-\alpha}{(1+\alpha)\beta}\varepsilon$:

   z is an ε-approximation of the value vector v^{α} and π^{∞} and ρ^{∞} are $\{1 + \frac{2}{\beta(1+\alpha)}\}\varepsilon$-optimal policies for player 1 and 2, respectively (STOP).

   Otherwise: $x := z$ and return to step 2.

The next example shows that Algorithm 10.3 does not converge in general.

Example 10.2

$S = \{1, 2\};\ A(1) = B(1) = \{1, 2\},\ A(2) = B(2) = \{1\};\ \alpha = \frac{3}{4}.$

$r_1(1,1) = 3,\ r_1(1,2) = 6,\ r_1(2,1) = 2,\ r_1(2,2) = 1,\ r_2(1,1) = 0.$

$p_{11}(1,1) = 1,\ p_{12}(1,1) = 0,\ p_{11}(1,2) = \frac{1}{3};\ p_{12}(1,2) = \frac{2}{3},\ p_{11}(2,1) = 1,\ p_{12}(2,1) = 0;$

$p_{11}(2,2) = 1,\ p_{12}(2,2) = 0;\ p_{21}(1,1) = 0,\ p_{22}(1,1) = 1.$

Take $\varepsilon = 0.2$ and $x = (0, 0);\ \beta = 5.$

Iteration 1:

$$y_1 = val\begin{pmatrix} 3 & 6 \\ 2 & 1 \end{pmatrix} = 3;\ \pi_{11} = \rho_{11} = 1,\ \pi_{12} = \rho_{12} = 0;\ z = v^{\alpha}(\pi^{\infty}, \rho^{\infty}) = (12, 0);\ x = (12, 0).$$

Iteration 2:

$$y_1 = val\begin{pmatrix} 12 & 9 \\ 11 & 10 \end{pmatrix} = 10;\ \pi_{11} = \rho_{11} = 0,\ \pi_{12} = \rho_{12} = 1;\ z = v^{\alpha}(\pi^{\infty}, \rho^{\infty}) = (4, 0);\ x = (4, 0).$$

Iteration 3:

$$y_1 = val\begin{pmatrix} 6 & 7 \\ 5 & 4 \end{pmatrix} = 6;\ \pi_{11} = \rho_{11} = 1,\ \pi_{12} = \rho_{12} = 0;\ z = v^{\alpha}(\pi^{\infty}, \rho^{\infty}) = (12, 0);\ x = (12, 0).$$

Hence, we are in the same situation as at the start of iteration 2 and there is no convergence.

Since the mapping T is a contraction, it is a continuous mapping. In case $\frac{\partial (Tx)_i}{\partial x_j}$, the partial derivatives in x, exist, then $\frac{\partial (Tx)_i}{\partial x_j} = \alpha p_{ij}(\pi, \rho)$, because $(Tx)_i = r_i(\pi, \rho) + \alpha \sum_j p_{ij}(\pi.\rho)x_j$, where π and ρ are optimal mixed strategies in the matrix game $M_x[i]$. Let $F : \mathbb{R}^N \to \mathbb{R}^N$ be defined

by $Fx = Tx - x$. Then, the problem of finding the value vector of the stochastic game is the same as solving the nonlinear equation $Fx = 0$. We will show that Algorithm 10.3 is equivalent to Newton's method for solving $Fx = 0$. From Algorithm 10.3 we obtain

$$\begin{aligned} x^{n+1} &= v^{\alpha}\big((\pi^n)^{\infty}, (\rho^n)^{\infty}\big) = x^n + v^{\alpha}\big((\pi^n)^{\infty}, (\rho^n)^{\infty}\big) - x^n \\ &= x^n + \big\{I - \alpha P(\pi^n, \rho^n)\big\}^{-1} r(\pi^n, \rho^n) - x^n \\ &= x^n + \big\{I - \alpha P(\pi^n, \rho^n)\big\}^{-1} r(\pi^n, \rho^n) - \big\{I - \alpha P(\pi^n, \rho^n)\big\}^{-1}\big\{I - \alpha P(\pi^n, \rho^n)\big\}x^n \\ &= x^n - \big\{\alpha P(\pi^n, \rho^n) - I\big\}^{-1}\{r(\pi^n, \rho^n) + \alpha P(\pi^n, \rho^n)x^n - x^n\}. \end{aligned}$$

Because $\left\{\frac{\partial (Fx)_i}{\partial x_j}\right\}_{x=x^n} = \alpha p_{ij}(\pi^n,\rho^n) - \delta_{ij}$ and $r(\pi^n,\rho^n) + \alpha P(\pi^n,\rho^n)x^n - x^n = Tx^n - x^n = Fx^n$, we have

$$x^{n+1} = x^n - \{\nabla F x^n\}^{-1} F x^n. \tag{10.16}$$

i.e. Algorithm 10.3 is Newton's method for solving $Fx = 0$.

Let $\Delta x_n = x^{n+1} - x^n$, $\Delta T_n = Tx^{n+1} - Tx^n$ and $\Delta F_n = Fx^{n+1} - Fx^n$.

$\Delta F_n = (Tx^{n+1} - x^{n+1}) - (Tx^n - x^n) = \Delta T_n - \Delta x_n$; $Fx^{n+1} = Fx^n + \Delta F_n = Fx^n + \Delta T_n - \Delta x_n$.

Similarly as in the proof of Lemma 10.3 it can be shown that

$$\alpha \cdot \textstyle\sum_j \{min_{(a,b)} p_{ij}(a,b)\}(\Delta x^n)_j \le (\Delta T_n)_i \le \alpha \cdot \sum_j \{max_{(a,b)} p_{ij}(a,b)\}(\Delta x^n)_j.$$

Then, $(\Delta T_n)_i$ is a convex combination of the upper and lower bound, i.e. $(\Delta T_n)_i = \sum_j q_{ij}(n)(\Delta x_n)_j$, where $q_{ij}(n) = \alpha\{\lambda \cdot max_{(a,b)}\, p_{ij}(a,b) + (1-\lambda)\cdot min_{(a,b)}\, p_{ij}(a,b)\}$ for some $\lambda \in [0,1]$.

Hence,

$$Fx^{n+1} = Fx^n - \{I - Q(n)\}\Delta x_n. \tag{10.17}$$

From (10.17) and (10.16) it follows that

$$\begin{aligned}
Fx^{n+1} &= Fx^n + \{I - Q(n)\}\{\nabla F x^n\}^{-1} F x^n \\
&= Fx^n - \{I - Q(n)\}\{I - \alpha P(\pi^n,\rho^n)\}^{-1} Fx^n \\
&= \big\{I - \{I - Q(n)\}\{I - \alpha P(\pi^n,\rho^n)\}^{-1}\big\} Fx^n \\
&= \big\{I - \{I - \alpha P(\pi^n,\rho^n)\}^{-1} + Q(n)\{I - \alpha P(\pi^n,\rho^n)\}^{-1}\big\} Fx^n \\
&= \big\{ - \alpha P(\pi^n,\rho^n)\{I - \alpha P(\pi^n,\rho^n)\}^{-1} + Q(n)\{I - \alpha P(\pi^n,\rho^n)\}^{-1}\big\} Fx^n \\
&= \{Q(n) - \alpha P(\pi^n,\rho^n)\}\{I - \alpha P(\pi^n,\rho^n)\}^{-1} Fx^n.
\end{aligned}$$

Therefore,

$$\begin{aligned}
(Fx^{n+1})_i &= \textstyle\sum_j \big\{\{Q(n) - \alpha P(\pi^n,\rho^n)\}\{I - \alpha P(\pi^n,\rho^n)\}^{-1}\big\}_{ij} (Fx^n)_j \\
&= \textstyle\sum_j \big\{ \sum_k \{Q(n) - \alpha P(\pi^n,\rho^n)\}_{ik} \{\{I - \alpha P(\pi^n,\rho^n)\}^{-1}\}_{kj}\big\}(Fx^n)_j \\
&= \textstyle\sum_k \{Q(n) - \alpha P(\pi^n,\rho^n)\}_{ik} \cdot \sum_j \{\{I - \alpha P(\pi^n,\rho^n)\}^{-1}\}_{kj} (Fx^n)_j \\
&= \textstyle\sum_k \{Q(n) - \alpha P(\pi^n,\rho^n)\}_{ik} \cdot \frac{1}{1-\alpha} \cdot \|Fx^n\|.
\end{aligned}$$

Notice that

$$\begin{aligned}
|\textstyle\sum_k \{Q(n) - \alpha P(\pi^n,\rho^n)\}_{ik}| &= \alpha \textstyle\sum_k |\lambda \cdot max_{(a,b)}\, p_{ik}(a,b) + (1-\lambda)\cdot min_{(a,b)}\, p_{ik}(a,b) - p_{ik}(\pi^n,\rho^n)| \\
&\le \alpha \cdot max_i \big\{ \textstyle\sum_k \{max_{(a,b)}\, p_{ik}(a,b) - min_{(a,b)}\, p_{ik}(a,b)\}\big\} = (1-\alpha)\beta.
\end{aligned}$$

Hence, $\|Fx^{n+1}\|_\infty \le \beta \cdot \|Fx^n\|_\infty$, i.e. the process converges if $\beta < 1$.

Remark

The condition $\beta < 1$ is very restrictive. However, for problems that do not satisfy $\beta < 1$ the algorithm terminates in most cases.

Theorem 10.8

Assume that $\beta < 1$. Then, Algorithm 10.3 is correct.

Proof

For $\beta < 1$, we have shown that $Fx^n \to 0$ for $n \to \infty$, implying that $\|x^{n+1} - x^n\| \to 0$ for $n \to \infty$, i.e. the algorithm terminates. At termination with $z = x^{n+1}$ and $x = x^n$, we have

$$\begin{aligned} \|x^{n+1} - v^\alpha\|_\infty &= \|Tx^{n+1} - Fx^{n+1} - Tv^\alpha\|_\infty \leq \|Tx^{n+1} - Tv^\alpha\|_\infty + \|Fx^{n+1}\|_\infty \\ &\leq \alpha \cdot \|x^{n+1} - v^\alpha\|_\infty + \beta \cdot \|Fx^n\|_\infty. \end{aligned}$$

Hence,

$$\begin{aligned} \|x^{n+1} - v^\alpha\|_\infty &\leq \tfrac{\beta}{1-\alpha} \cdot \|Fx^n|_\infty = \tfrac{\beta}{1-\alpha} \cdot \|\nabla Fx^n(x^{n+1} - x^n)\|_\infty \\ &\leq \tfrac{\beta}{1-\alpha} \cdot \|I - \alpha P(\pi^n, \rho^n)\|_\infty \cdot \|x^{n+1} - x^n)\|_\infty \\ &\leq \tfrac{1+\alpha}{1-\alpha} \cdot \beta \cdot \|x^{n+1} - x^n)\|_\infty \leq \varepsilon. \end{aligned}$$

Let $\gamma = \frac{\varepsilon}{\beta} \cdot \frac{1-\alpha}{1+\alpha}$, then $-\gamma \cdot e \leq x^{n+1} - x^n \leq \gamma \cdot e$. We can also write for any $\rho^\infty \in \Gamma$,

$$\begin{aligned} L_{\pi^n,\rho} x^n &= r(\pi^n, \rho) + \alpha P(\pi^n, \rho) x^n \geq r(\pi^n, \rho^n) + \alpha P(\pi^n, \rho^n) x^n \\ &= r(\pi^n, \rho^n) + \alpha P(\pi^n, \rho^n) x^{n+1} + \alpha P(\pi^n, \rho^n)(x^n - x^{n+1}) \\ &= x^{n+1} + \alpha P(\pi^n, \rho^n)(x^n - x^{n+1}) \\ &\geq x^{n+1} - \alpha\gamma P(\pi^n, \rho^n) e = x^{n+1} - \alpha\gamma \cdot e \geq x^n - (1+\alpha)\gamma \cdot e = x^n - \delta \cdot e, \end{aligned}$$

with $\delta = (1+\alpha)\gamma$. The monotonicity of $L_{\pi^n,\rho}$ yields $L^k_{\pi^n,\rho} x^n \geq x^n - \delta(1 + \alpha + \cdots + \alpha^k) \cdot e$, $k \in \mathbb{N}$, implying

$$\begin{aligned} v^\alpha\big((\pi^n)^\infty, \rho^\infty\big) &\geq x^n - \delta(1-\alpha)^{-1} \cdot e = x^{n+1} + (x^n - x^{n+1}) - (1-\alpha)^{-1}\delta \cdot e \\ &\geq v^\alpha - \varepsilon \cdot e - \tfrac{1+\alpha}{1-\alpha}\gamma \cdot e = v^\alpha - \{1 + \tfrac{2}{\beta(1+\alpha)}\varepsilon \cdot e. \end{aligned}$$

From this result it follows that $(\pi^n)^\infty$ is a $\{1 + \frac{2}{\beta(1+\alpha)}\}\varepsilon$-optimal policy for player 1. Similarly it can be shown that $(\rho^n)^\infty$ is a $\{1 + \frac{2}{\beta(1+\alpha)}\}\varepsilon$-optimal policy for player 2. □

The last method in this section uses an integer k, where $1 \leq k \leq \infty$. For $k = 1$ we obtain Algorithm 10.1 and for $k = \infty$ Algorithm 10.3. So, this algorithm is of the type of modified policy iteration as analysed in Section 3.7 for the MDP model.

Algorithm 10.4 *Modified policy iteration for discounted games*

1. Choose $\varepsilon > 0$, an integer $1 \leq k \leq \infty$ and a vector $x \in \mathbb{R}^N$ such that $Tx \leq x$.
2. Compute for each $i \in S$:
   a. $y_i = val(M_x[i])$;
   b. π_{ia}, $a \in A(i)$, an optimal mixed policy for player 1 in the matrix game of 2a.
   c. ρ_{ib}, $b \in B(i)$, an optimal mixed policy for player 2 in the matrix game of 2a.

3. Compute $z = U^k(\rho)x$, where $U(\rho)$ is defined by $\{U(\rho)x\}_i = max_a \{r_i(a,\rho)+\alpha \sum_j p_{ij}(a,\rho)x_j\}$
4. If $\|z - x\|_\infty \leq \frac{1-\alpha}{\alpha}\varepsilon$:

   z is an ε-approximation of the value vector v^α and π^∞ and ρ^∞ are $\frac{1}{\alpha}\varepsilon$-optimal policies for player 1 and 2, respectively (STOP).

   Otherwise: $x := z$ and return to step 2.

We denote the vectors x, y, z, the strategies π and ρ and the operator $U(\rho)$ in the n-th iteration by $x^n, y^n, z^n, \pi^n, \rho^n$ and U_n, respectively. For any fixed $\rho^\infty \in \Gamma$ and $x \in \mathbb{R}^N$, we have the property

$$U(\rho)x = max_\pi \{r(\pi,\rho) + \alpha P(\pi,\rho)x\} \geq max_\pi min_\rho \{r(\pi,\rho) + \alpha P(\pi,\rho)x\} = Tx, \qquad (10.18)$$

implying that $U^m(\rho)x \geq T^m x$ for all $\rho^\infty \in \Gamma$, $x \in \mathbb{R}^N$ and $m \in \mathbb{N}$. Furthermore, notice that $y^n = Tx^n = U_n x^n$ for all n.

Lemma 10.4
$x^n \geq Tx^n \geq x^{n+1} \geq v^\alpha$ for $n = 0, 1, \ldots$.

Proof
We apply induction on n.
For $n = 0$, we have $x^0 \geq Tx^0 = y^0 = U_0 x^0$ (the first inequality by step 1 of the algorithm).
Since U_0 is monotone and $x^0 \geq U_0 x^0$, we obtain $x^1 = U_0^k x^0 \leq U_0 x^0 = Tx^0 \leq x^0$.
From $x^1 \leq Tx^0 \leq x^0$ and the monotonicity of T it follows that $T^m x^1 \leq x^0$ for $m = 0, 1, 2, \ldots$.
Hence, $v^\alpha = lim_{m\to\infty} T^m x^1 \leq x^0$. Therefore, $x^1 = U_0^k x^0 \geq T^k x^0 \geq Tv^\alpha = v^\alpha$, and we have shown that $x^n \geq Tx^n \geq x^{n+1} \geq v^\alpha$ for $n = 0$.
Suppose that $x^n \geq Tx^n \geq x^{n+1} \geq v^\alpha$. Now, we will show that $x^{n+1} \geq Tx^{n+1} \geq x^{n+2} \geq v^\alpha$.
We have, $U_n x^{n+1} = Tx^{n+1} = T\{U_n^k x^n\} \leq U_n^{k+1} x^n \leq U_n^k x^n = x^{n+1}$, the last inequality since $U_n x^n = Tx^n \leq x^n$ and the monotonicity of U_n. From $U_n x^{n+1} \leq x^{n+1}$ it follows that $x^{n+2} = U_{n+1}^k x^{n+1} \leq U_{n+1}^{k-1} x^{n+1} \leq \cdots \leq U_{n+1} x^{n+1} = Tx^{n+1}$. Since $x^{n+1} \geq v^\alpha$, we obtain $x^{n+2} = U_{n+1}^k x^{n+1} \geq T^k x^{n+1} \geq T^k v^\alpha = v^\alpha$. □

Corollary 10.3
$lim_{n\to\infty} x^n = v^\alpha$.

Proof
From Lemma 10.4 it follows that $v^\alpha \leq x^n \leq Tx^{n-1} \leq T^2 x^{n-2} \leq \cdots \leq T^{n-1} x^1 \leq T^n x^0$ for $n = 0, 1, 2, \ldots$. Since $lim_{n\to\infty} T^n x^0 = v^\alpha$, we also have $lim_{n\to\infty} x^n = v^\alpha$. □

Theorem 10.9
Algorithm 10.4 is correct.

Proof

Because $lim_{n\to\infty} x^n = v^\alpha$, the algorithm terminates. Let x^n and z^n be the vectors x and z in the final iteration. Since $0 \le x^{n+1} - v^\alpha \le Tx^n - v^\alpha$, we obtain

$$\begin{aligned} \|x^{n+1} - v^\alpha\|_\infty &\le \|Tx^n - v^\alpha\|_\infty = \|Tx^n - Tv^\alpha\|_\infty \le \alpha \cdot \|x^n - v^\alpha\|_\infty \\ &\le \alpha \cdot \|x^n - x^{n+1}\|_\infty + \alpha \cdot \|x^{n+1} - v^\alpha\|_\infty. \end{aligned}$$

Hence, $\|z^n - v^\alpha\|_\infty = \|x^{n+1} - v^\alpha\|_\infty \le \frac{\alpha}{1-\alpha} \cdot \|x^n - x^{n+1}\|_\infty = \frac{\alpha}{1-\alpha} \cdot \|z^n - x^n\|_\infty < \varepsilon$,

i.e. z^n is an ε-approximation of the value vector.

Furthermore, we have for any $\rho^\infty \in \Gamma$,

$L_{\pi^n,\rho} x^n = r(\pi^n, \rho) + \alpha P(\pi^n, \rho) x^n \ge Tx^n \ge x^{n+1} \ge x^n - \|x^n - x^{n+1}\|_\infty \cdot e \ge x^n - \frac{1-\alpha}{\alpha}\varepsilon \cdot e.$

Hence, $L^m_{\pi^n,\rho} x^n \ge x^n - \{1 + \alpha + \cdots + \alpha^{m-1}\}\frac{1-\alpha}{\alpha}\varepsilon \cdot e$ for $m = 1, 2, \ldots$.

Therefore, we obtain

$$v^\alpha\big((\pi^n)^\infty, \rho^\infty\big) = lim_{m\to\infty} L^m_{\pi^n,\rho} x^n \ge x^n - \frac{1}{\alpha}\varepsilon \cdot e \ge v^\alpha - \frac{1}{\alpha}\varepsilon \cdot e.$$

From this result it follows that $(\pi^n)^\infty$ is a $\frac{1}{\alpha}\varepsilon$-optimal policy for player 1. Similarly it can be shown that $(\rho^n)^\infty$ is a $\frac{1}{\alpha}\varepsilon$-optimal policy for player 2. □

10.2.4 Finite methods

In general, solutions to stochastic games lack an important algebraic property., which suggests that effectively solving is essentially more difficult than solving matrix games. This is illustrated by the following example.

Example 10.3

$S = \{1, 2\};\ A(1) = B(1) = \{1, 2\},\ A(2) = B(2) = \{1\};\ \alpha = \frac{1}{2}.$

$r_1(1,1) = 1,\ r_1(1,2) = 0,\ r_1(2,1) = 0,\ r_1(2,2) = 3,\ r_2(1,1) = 0.$

$p_{11}(1,1) = 1,\ p_{12}(1,1) = 0;\ p_{11}(1,2) = 0,\ p_{12}(1,2) = 1;\ p_{11}(2,1) = 0,\ p_{12}(2,1) = 1;$

$p_{11}(2,2) = 1,\ p_{12}(2,2) = 0;\ p_{21}(1,1) = 0,\ p_{22}(1,1) = 1.$

Consider the fixed point equation $x = Tx$, i.e.

$$x_1 = val\begin{pmatrix} 1 + \frac{1}{2}x_1 & 0 + \frac{1}{2}x_2 \\ 0 + \frac{1}{2}x_2 & 3 + \frac{1}{4}x_1 \end{pmatrix};\ x_2 = val\big(0 + \tfrac{1}{2}x_2\big) \ \to\ v_2^\alpha = x_2 = 0.$$

$$x_1 = val\begin{pmatrix} 1 + \frac{1}{2}x_1 & 0 \\ 0 & 3 + \frac{1}{2}x_1 \end{pmatrix} \ \to\ x_1 = \tfrac{(1+\frac{1}{2}x_1)(3+\frac{1}{2}x_1)}{(1+\frac{1}{2}x_1)+(3+\frac{1}{2}x_1)} \ \to\ v_1^\alpha = x_1 = \tfrac{2}{3}\{-2 + \sqrt{13}\}.$$

The optimal policies are for both players $\left(\frac{7+\sqrt{13}}{8+2\sqrt{13}}, \frac{1+\sqrt{13}}{8+2\sqrt{13}}\right)$.

The above example shows that while all the data defining the stochastic game (the rewards, the transition probabilities and the discount factor) are rational, the value vector has irrational entries. Thus the data and the solution are not in the same ordered Archimedean field. This phenomenon is called lack of the *ordered field property*. It essentially eliminates the possibility of solving discounted stochastic games by performing only finitely many arithmetic operations. Note that since linear programs solve a general matrix game, and since an optimal basis of that program can be found via finitely many pivots of the simplex method, matrix games possess the ordered field property.

One line of research that has evolved from the preceding considerations is focussed on identifying those natural classes of stochastic games for which the ordered field property holds, and on developing algorithms for their solution. We will consider the following special games:
(1) The single-controller stochastic game.
(2) The switching-controller stochastic game.
(3) The separabel reward - state independent transitions (SER-SIT) stochastic game.
(4) The additive reward - additive transitions (ARAT) stochastic game.

Single-controller stochastic game

In the single-controller stochastic game is player 1 the 'single-controller'. This means that the transition probabilities $p_{ij}(a,b)$ are independent of b. Therefore, we denote these probabilities as $p_{ij}(a)$. Under this assumption the nonlinear program (10.11) becomes the following linear program

$$min \left\{ \sum_i v_i \;\middle|\; \begin{array}{rcrl} \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j & - & \sum_b r_i(a,b)\rho_{ib} & \geq 0,\ a \in A(i),\ i \in S \\ & & \sum_b \rho_{ib} & = 1,\ i \in S \\ & & \rho_{ib} & \geq 0,\ b \in B(i),\ i \in S \end{array} \right\}. \tag{10.19}$$

The dual program is

$$max \left\{ \sum_i z_i \;\middle|\; \begin{array}{rcrl} \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) & & = & 1,\ j \in S \\ -\sum_a r_i(a,b) x_i(a) & + \quad z_i & \leq & 0,\ (i,b) \in S \times B \\ & x_i(a) & \geq & 0,\ (i,a) \in S \times A \end{array} \right\}. \tag{10.20}$$

The following theorem shows that the value vector and optimal stationary policies for both players can be obtained from the optimal solutions of the dual pair of linear programs.

Theorem 10.10

Let (v^*, ρ^*) and (x^*, z^*) be optimal solutions of the linear programs (10.19) and (10.20), respectively. Define the stationary policy $(\pi^*)^\infty$ by $\pi^*_{ia} = \frac{x_i^*(a)}{\sum_a x_i^*(a)}$, $(i,a) \in S \times A$.
Then, v^* is the value vector and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and 2, respectively.

Proof

Theorem 10.4 implies that v^* is the value vector of the stochastic game. Since

$$\sum_a x_i^*(a) = 1 + \alpha \sum_{(i,a)} p_{ij}(a) x_i(a) > 0, \ j \in S,$$

the stationary policy $(\pi^*)^\infty$ is well defined. From the constraints of program (10.19) it follows that $\{I - \alpha P(\pi)\}v^* \geq r(\pi, \rho^*)$ for every $\pi^\infty \in \Pi$. Therefore,

$$v^* \geq \{I - \alpha P(\pi)\}^{-1} r(\pi, \rho^*) = v^\alpha\big(\pi^\infty, (\rho^*)^\infty\big) \text{ for every } \pi^\infty \in \Pi. \tag{10.21}$$

From the complentary slackness property of linear programming it follows that

$$x_i^*(a) \cdot \Big\{ \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j^* - \sum_b r_i(a,b) \rho_{ib}^* \Big\} = 0 \text{ for all } (i,a) \in S \times A.$$

Since $x_i^*(a) > 0$ if and only if $\pi_{ia}^* > 0$, we also have

$$\pi_{ia}^* \cdot \Big\{ \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j^* - \sum_b r_i(a,b) \rho_{ib}^* \Big\} = 0 \text{ for all } (i,a) \in S \times A.$$

Therefore

$$\sum_a \pi_{ia}^* \cdot \Big\{ \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j^* - \sum_b r_i(a,b) \rho_{ib}^* \Big\} = 0 \text{ for all } i \in S,$$

implying $\sum_j \{\delta_{ij} - \alpha p_{ij}(\pi^*)\} v_j^* = r_i(\pi^*, \rho^*)$ for all $i \in S$, i.e. $\{I - \alpha P(\pi^*)\}v^* = r(\pi^*, \rho^*)$. So,

$$v^* = \{I - \alpha P(\pi^*)\}^{-1} r(\pi^*, \rho^*) = v^\alpha\big((\pi^*)^\infty, (\rho^*)^\infty\big).$$

Since the optimum values of (10.19) and (10.20) are equal, we can write

$$\sum_j v_j^\alpha\big((\pi^*)^\infty, (\rho^*)^\infty\big) = \sum_i z_i^*. \tag{10.22}$$

Since $z_i^* \leq \sum_a r_i(a,b) x_i^*(a)$ for all $b \in B(i)$, we also have $z_i^* \leq \sum_a r_i(a,\rho) x_i^*(a)$ for all $\rho^\infty \in \Gamma$. From the constraints of (10.20) it follows that, with $x_i^* = \sum_a x_i^*(a), \ i \in S$,

$$1 = \sum_{(i,a)} \{\delta_{ij} - \alpha p_{ij}(a)\} \pi_{ia}^* \cdot x_i^* = \sum_i \{\delta_{ij} - \alpha p_{ij}(\pi^*)\} \cdot x_i^*,$$

or, in vector notation, $e^T = (x^*)^T \{I - \alpha P(\pi^*)\}$, implying $(x^*)^T = e^T \{I - \alpha P(\pi^*)\}^{-1}$. Then, because $z_i^* \leq \sum_a r_i(a,\rho) x_i^*(a)$,

$$\begin{aligned} \sum_i z_i^* &\leq \sum_{(i,a)} r_i(a,\rho) x_i^*(a) = \sum_{(i,a)} r_i(a,\rho) \pi_{ia}^* x_i^* \\ &= \sum_i r_i(\pi^*, \rho) x_i^* = (x^*)^T r(\pi^*, \rho) = e^T \{I \ \alpha P(\pi^*)\}^{-1} r(\pi^*, \rho) \\ &= e^T v^\alpha\big((\pi^*)^\infty, \rho^\infty\big) = \sum_j v_j^\alpha\big((\pi^*)^\infty, \rho^\infty\big). \end{aligned}$$

With (10.22) we obtain $\sum_j v_j^\alpha\big((\pi^*)^\infty, \rho^\infty\big) \geq \sum_j v_j^\alpha\big((\pi^*)^\infty, (\rho^*)^\infty\big)$ for all $\rho^\infty \in \Gamma$. Hence, $(\rho^*)^\infty$ is an optimal policy for player 2 in the MDP induced by policy $(\pi^*)^\infty$. Therefore,

$$v^\alpha\big((\pi^*)^\infty, \rho^\infty\big) \geq v^* = v^\alpha\big((\pi^*)^\infty, (\rho^*)^\infty\big) \text{ for all } \rho^\infty \in \Gamma. \tag{10.23}$$

Hence, by (10.21 and (10.23, we have

$$v^\alpha\big((\pi^*)^\infty, \rho^\infty\big) \geq v^* = v^\alpha\big((\pi^*)^\infty, (\rho^*)^\infty\big) \geq v^\alpha\big(\pi^\infty, (\rho^*)^\infty\big), \ \pi^\infty \in \Pi, \rho^\infty \in \Gamma,$$

i.e. $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and 2, respectively. □

Algorithm 10.5 *Single-controller game with discounting*

1. Compute optimal solutions (v^*, ρ^*) and (x^*, z^*) of the linear programs (10.19) and (10.20).
2. Define the stationary policy $(\pi^*)^\infty$ by $\pi^*_{ia} = \frac{x^*_i(a)}{\sum_a x^*_i(a)}$, $(i,a) \in S \times A$.
3. v^* is the value vector and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ optimal stationary policies for player 1 and 2.

Example 10.4

$S = \{1,2\}$; $A(1) = \{1,2\}$, $B(1) = \{1,2,3\}$, $A(2) = \{1,2,3\}$, $B(2) = \{1,2\}$; $\alpha = \frac{1}{2}$.

$r_1(1,1) = 5$, $r_1(1,2) = 1$, $r_1(1,3) = 6$, $r_1(2,1) = 4$, $r_1(2,2) = 6$, $r_1(2,3) = 2$;

$r_2(1,1) = 6$, $r_2(1,2) = 0$, $r_2(2,1) = 3$, $r_2(2,2) = 4$, $r_2(3,1) = 0$, $r_2(3,2) = 6$.

$p_{11}(1) = 1$, $p_{12}(1) = 0$; $p_{11}(2) = 0$, $p_{12}(2) = 1$; $p_{21}(1) = 1$, $p_{22}(1) = 0$;

$p_{21}(2) = 0$, $p_{22}(2) = 1$; $p_{21}(3) = 1$, $p_{22}(3) = 0$.

The linear programs (10.19) and (10.20) are

$$min \left\{ v_1 + v_2 \;\middle|\; \begin{array}{rcrcrcrcrcrcl}
\frac{1}{2}v_1 & & & - & 5\rho_{11} & - & \rho_{12} & - & 6\rho_{13} & & & \geq 0 \\
v_1 & - & \frac{1}{2}v_2 & - & 4\rho_{11} & - & 6\rho_{12} & - & 2\rho_{13} & & & \geq 0 \\
- \frac{1}{2}v_1 & + & v_2 & & & & & & & - & 6\rho_{21} & \geq 0 \\
& & \frac{1}{2}v_2 & & & & & & & - & 3\rho_{21} - 4\rho_{22} & \geq 0 \\
- \frac{1}{2}v_1 & + & v_2 & & & & & & & & - 6\rho_{22} & \geq 0 \\
\multicolumn{12}{c}{\rho_{11} + \rho_{12} + \rho_{13} = 1;\ \rho_{21} + \rho_{22} = 1;\ \rho_{11}, \rho_{12}, \rho_{13}, \rho_{21}, \rho_{22} \geq 0}
\end{array} \right\}$$

and

$$max \left\{ z_1 + z_2 \;\middle|\; \begin{array}{rcrcrcrcrcrcl}
\frac{1}{2}x_{11} & + & x_{12} & - & \frac{1}{2}x_{21} & & & - & \frac{1}{2}x_{23} & & & = 1 \\
& - & \frac{1}{2}x_{12} & + & x_{21} & + & \frac{1}{2}x_{22} & + & x_{23} & & & = 1 \\
- 5x_{11} & - & 4x_{12} & & & & & & & + & z_1 & \leq 0 \\
- x_{11} & - & 6x_{12} & & & & & & & + & z_1 & \leq 0 \\
- 6x_{11} & - & 2x_{12} & & & & & & & + & z_1 & \leq 0 \\
& & & - & 6x_{21} & - & 3x_{22} & & & + & z_2 & \leq 0 \\
& & & & & - & 4x_{22} & - & 6x_{23} & + & z_2 & \leq 0 \\
\multicolumn{12}{c}{x_{11}, x_{12}, x_{21}, x_{22}, x_{23} \geq 0}
\end{array} \right\}$$

The optimal solutions are:

$v^*_1 = 7.327$, $v^*_2 = 6.916$; $\rho^*_{11} = 0$, $\rho^*_{12} = 0.467$, $\rho^*_{13} = 0.533$, $\rho^*_{21} = 0.542$, $\rho^*_{22} = 0.458$ and

$z^*_1 = 5.720$, $z^*_2 = 8.523$; $x^*_{11} = 0.673$, $x^*_{12} = 0.841$, $x^*_{21} = 0.355$, $x^*_{22} = 2.131$, $x^*_{23} = 0$.

The optimal policy for player 1 is: $\pi^*_{11} = 0.446$, $\pi^*_{12} = 0.554$, $\pi^*_{21} = 0.856$, $\pi^*_{22} = 0.144$, $\pi^*_{23} = 0$.

Switching-controller stochastic game

In a switching-controller stochastic game we assume that the set of states is the union of two disjoint sets S_1 and S_2 such that player 1 controls the transitions in S_1 and player 2 in S_2. Notice that a game with perfect information and the single-controller stochastic game are special cases of the switching-controller stochastic game.

Denote the transitions by $p_{ij}(a,b) = \begin{cases} p_{ij}(a),\ i \in S_1,\ a \in A(i),\ b \in B(i),\ j \in S; \\ p_{ij}(b),\ i \in S_2,\ a \in A(i),\ b \in B(i),\ j \in S. \end{cases}$

It appears that to solve such a game by a finite algorithm, a finite sequence of linear programs and matrix games needs to be solved instead of only a single one. The linear programs are the programs of the type of linear programs for single-controller stochastic games.

Suppose that player 2 fixes his strategy ρ^∞ in the states of S_2. Then we denote the corresponding single-controller stochastic game by $SCSG(\rho)$ with data

$$r_i(a,b) = \begin{cases} r_i(a,b) & ,\ i \in S_1,\ a \in A(i),\ b \in B(i) \\ r_i(a,\rho) = \sum_b r_i(a,b)\rho_{ib} & ,\ i \in S_2,\ a \in A(i),\ b \in B(i) \end{cases}$$

and

$$p_{ij}(a,b) = \begin{cases} p_{ij}(a) & ,\ i \in S_1,\ a \in A(i),\ b \in B(i),\ j \in S \\ p_{ij}(\rho) = \sum_b p_{ij}(b)\rho_{ib} & ,\ i \in S_2,\ a \in A(i),\ b \in B(i),\ j \in S. \end{cases}$$

Notice that the transitions in the states of S_2 are independent of any choice of the players and the rewards depend only on the action taken by player 1. So, in the states of S_2 player 2 is a dummy and the first player will choose the action which maximizes $r_i(a,\rho)$ over the action set $A(i)$. Let $a[i,\rho]$ be that action, i.e. $a[i,\rho] = argmax_{a \in A(i)}\, r_i(a,\rho),\ i \in S_2$. The linear program for the single-controller stochastic game by $SCSG(\rho)$ is:

$$min \left\{ \sum_i v_i \;\middle|\; \begin{array}{rcrl} \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j & - & \sum_b r_i(a,b)\rho_{ib} & \geq 0,\ a \in A(i),\ i \in S_1 \\ \sum_j \{\delta_{ij} - \alpha p_{ij}(\rho)\} v_j & - & r_i(a,\rho) & \geq 0,\ a \in A(i),\ i \in S_2 \\ & & \sum_b \rho_{ib} & = 1,\ i \in S_1 \\ & & \rho_{ib} & \geq 0,\ b \in B(i),\ i \in S_1 \end{array} \right\}. \tag{10.24}$$

The inequalities for $i \in S_2$ can be written as $v_i \geq \alpha \sum_j p_{ij}(\rho) v_j + r_i(a,\rho),\ a \in A(i),\ i \in S_2$, and are equivalent to a single inequality for each $i \in S_2$, namely $v_i \geq \alpha \sum_j p_{ij}(\rho) v_j + r_i(a[i,\rho])$. Therefore, program (10.24) is equivalent to the program

$$min \left\{ \sum_i v_i \;\middle|\; \begin{array}{rcrl} \sum_j \{\delta_{ij} - \alpha p_{ij}(a)\} v_j & - & \sum_b r_i(a,b)\rho_{ib} & \geq 0,\ a \in A(i),\ i \in S_1 \\ \sum_j \{\delta_{ij} - \alpha p_{ij}(\rho)\} v_j & - & r_i(a[i,\rho]) & \geq 0,\ i \in S_2 \\ & & \sum_b \rho_{ib} & = 1,\ i \in S_1 \\ & & \rho_{ib} & \geq 0,\ b \in B(i),\ i \in S_1 \end{array} \right\}. \tag{10.25}$$

Note

For different choices of ρ in S_2, the linear programs (10.25) only differ in the inequalities for the states S_2. This property can be used, e.g. by using the dual simplex method for the solution of subsequent programs (10.25).

Example 10.5

$S = \{1,2\};\ S_1 = \{1\},\ S_2 = \{2\};\ A(1) = B(1) = A(2) = B(2) = \{1,2\};\ \alpha = \frac{1}{2}$.

$r_1(1,1) = 3,\ r_1(1,2) = 1,\ r_1(2,1) = 1,\ r_1(2,2) = 4;$

$r_2(1,1) = 4,\ r_2(1,2) = 6,\ r_2(2,1) = 7,\ r_2(2,2) = 5.$

$p_{11}(1) = 0,\ p_{12}(1) = 1;\ p_{11}(2) = 1,\ p_{12}(2) = 0;\ p_{21}(1) = 1,\ p_{22}(1) = 0;\ p_{21}(2) = 0,\ p_{22}(2) = 1.$

Let player 2 choose in state 2 both action 1 and 2 with probability $\frac{1}{2}$.

The rewards and probabilities in state 2 are $r_2(1,\rho) = \frac{1}{2}\cdot 4 + \frac{1}{2}\cdot 6 = 5;\ r_2(2,\rho) = \frac{1}{2}\cdot 7 + \frac{1}{2}\cdot 5 = 6.$

$p_{21}(\rho) = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 0 = \frac{1}{2};\ p_{22}(\rho) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2}.$

Program (10.24) becomes

$$min \left\{ v_1 + v_2 \;\middle|\; \begin{array}{rcrcrcrl} v_1 & - & \frac{1}{2}v_2 & - & 3\rho_{11} & - & \rho_{12} & \geq 0 \\ \frac{1}{2}v_1 & & & - & \rho_{11} & - & 4\rho_{12} & \geq 0 \\ -\frac{1}{4}v_1 & + & \frac{3}{4}v_2 & & & & & \geq 5 \\ -\frac{1}{4}v_1 & + & \frac{3}{4}v_2 & & & & & \geq 6 \\ & & & & \rho_{11} & + & \rho_{12} & = 1 \\ & & & & & & \rho_{11},\ \rho_{12} & \geq 0 \end{array} \right\},$$

which is equivalent to

$$min \left\{ v_1 + v_2 \;\middle|\; \begin{array}{rcrcrcrl} v_1 & - & \frac{1}{2}v_2 & - & 3\rho_{11} & - & \rho_{12} & \geq 0 \\ \frac{1}{2}v_1 & & & - & \rho_{11} & - & 4\rho_{12} & \geq 0 \\ -\frac{1}{4}v_1 & + & \frac{3}{4}v_2 & & & & & \geq 6 \\ & & & & \rho_{11} & + & \rho_{12} & = 1 \\ & & & & & & \rho_{11},\ \rho_{12} & \geq 0 \end{array} \right\}.$$

The solution of this program is: $v_1 = 6.57,\ v_2 = 10.19,\ \rho_{11} = 0.24,\ \rho_{12} = 0.76.$

Denote the value vector of the single controller stochastic game $SCSG(\rho)$ by v^ρ. This value vector satisfies the fixed point equation $x = T^\rho x$, i.e. $x_i = val\big(M_x^\rho[i]\big),\ i \in S$, where $M_x^\rho[i]$ has the elements $\begin{cases} r_i(a,b) + \alpha \sum_j p_{ij}(a)x_j & ,i \in S_1; \\ r_i(a,\rho) + \alpha \sum_j p_{ij}(\rho)x_j & ,i \in S_2. \end{cases}$

If it turns out that $x = Tx$, with $(Tx)_i = val\big(M_x[i]\big),\ i \in S$, where $M_x[i]$ has the elements

$\begin{cases} r_i(a,b) + \alpha \sum_j p_{ij}(a)x_j & ,i \in S_1 \\ r_i(a,b) + \alpha \sum_j p_{ij}(b)x_j & ,i \in S_2 \end{cases}$, then x is the value vector of the original game.

Therefore, we compute $val\big(M_{v^\rho}[i]\big),\ i \in S_2$, and check whether $v_i^\rho = val\big(M_{v^\rho}[i]\big),\ i \in S_2$.

If this is the case, we have found the value vector and the corresponding optimal stationary policies of the two players; if not, our 'guess' for $\rho_{ib},\ i \in S_2,\ b \in B(i)$, was not optimal and we try another ρ for the states in S_2, namely the ρ's we found in the matrix games $M_{v^\rho}[i],\ i \in S_2$.

For a matrix game it is well known that the optimal stategy spaces are polytopes. We need for our algorithm extreme optimal strategies. If we use linear programming to compute the value and optimal strategies of the game we find extreme optimal strategies. The algorithm for the switching-controller game with discounting is as follows.

Algorithm 10.6 *Switching-controller game with discounting*

1. Set $n = 0$ and choose an arbitrary extreme strategy for player 2 on S_2, say:
   $\rho^0_{ib},\ i \in S_2,\ b \in B(i)$.
2. Set $n := n+1$, solve the single-controller stochastic game $SCSG(\rho^{n-1})$, i.e. solve the linear program (10.25) and denote the value vector by x^n.
3. Determine, for each $i \in S_2$, $val(M_{x^n}[i]$ and the corresponding extreme stationary strategy $\rho^n_{ib},\ b \in B(i)$.
4. If $x^n_i = val(M_{x^n}[i])$ for all $i \in S_2$, then x^n is the value vector and $\rho^n_{ib},\ i \in S_2,\ b \in B(i)$ is part of an optimal policy for player 2; the optimal actions for player 2 in the states $i \in S_1$ and the optimal strategy for player 1 follow from the linear program (10.25) and its dual, respectively.
   Otherwise: return to step 2.

Example 10.5 (continued)
Iteration 1:
$n = 0$ and start with $\rho^0_{21} = 0,\ \rho^0_{22} = 1$. Then, $r_2(1,\rho) = 6,\ r_2(2,\rho) = 5,\ p_{21}(\rho) = 0,\ p_{22}(\rho) = 1$.
$n = 1$ and the linear program for $SCSG(\rho^0)$ is

$$min \left\{ v_1 + v_2 \;\middle|\; \begin{array}{rcrcrcrl} v_1 & - & \frac{1}{2}v_2 & - & 3\rho_{11} & - & \rho_{12} & \geq 0 \\ \frac{1}{2}v_1 & & & - & \rho_{11} & - & 4\rho_{12} & \geq 0 \\ & & \frac{1}{2}v_2 & & & & & \geq 6 \\ & & & & \rho_{11} & + & \rho_{12} & = 1 \\ & & & & & & \rho_{11},\ \rho_{12} & \geq 0 \end{array} \right\}.$$

The solution of this program is: $v_1 = \frac{29}{7},\ v_2 = 12,\ \rho_{11} = \frac{1}{8},\ \rho_{12} = \frac{7}{8}$ and $x^1 = (7.25, 12)$.
$i = 2$: $M_{x^1}[2] = \begin{pmatrix} 7\frac{5}{8} & 12 \\ 10\frac{5}{8} & 11 \end{pmatrix}$ with $val\big(M_{x^1}[2]\big) = 10\frac{5}{8}$ and $\rho^1_{21} = 1,\ \rho^1_{22} = 0$.

Iteration 2:

$n = 2$ and $\rho^1_{21} = 1,\ \rho^1_{22} = 0$. Then, $r_2(1,\rho) = 4,\ r_2(2,\rho) = 7,\ p_{21}(\rho) = 1,\ p_{22}(\rho) = 0$.

The linear program for $SCSG(\rho^1)$ is

$$min\left\{ v_1 + v_2 \;\middle|\; \begin{array}{rcrcrcrl} v_1 & - & \frac{1}{2}v_2 & - & 3\rho_{11} & - & \rho_{12} & \geq 0 \\ \frac{1}{2}v_1 & & & - & \rho_{11} & - & 4\rho_{12} & \geq 0 \\ -\frac{1}{2}v_1 & + & v_2 & & & & & \geq 7 \\ & & & & \rho_{11} & + & \rho_{12} & = 1 \\ & & & & & & \rho_{11},\ \rho_{12} \geq 0 & \end{array} \right\}$$

with solution: $v_1 = 6.62,\ v_2 = 10.31,\ \rho_{11} = 0.23,\ \rho_{12} = 0.77$ and $x^2 = (6.62, 10.31)$.

$i = 2$: $M_{x^2}[2] = \begin{pmatrix} 7.31 & 11.15 \\ 10.31 & 10.15 \end{pmatrix}$ with $val\big(M_{x^2}[2]\big) = 10.19$ and $\rho^2_{21} = \frac{1}{4},\ \rho^2_{22} = \frac{3}{4}$.

Iteration 3:

$n = 3$ and $\rho^2_{21} = \frac{1}{4},\ \rho^2_{22} = \frac{3}{4}$. Then, $r_2(1,\rho) = 5\frac{1}{2},\ r_2(2,\rho) = 5\frac{1}{2},\ p_{21}(\rho) = \frac{1}{4},\ p_{22}(\rho) = \frac{3}{4}$.

The linear program for $SCSG(\rho^1)$ is

$$min\left\{ v_1 + v_2 \;\middle|\; \begin{array}{rcrcrcrl} v_1 & - & \frac{1}{2}v_2 & - & 3\rho_{11} & - & \rho_{12} & \geq 0 \\ \frac{1}{2}v_1 & & & - & \rho_{11} & - & 4\rho_{12} & \geq 0 \\ -\frac{1}{8}v_1 & + & \frac{1}{8}v_2 & & & & & \geq 5\frac{1}{2} \\ & & & & \rho_{11} & + & \rho_{12} & = 1 \\ & & & & & & \rho_{11},\ \rho_{12} \geq 0 & \end{array} \right\}$$

with solution: $v_1 = 6.54,\ v_2 = 10.11,\ \rho_{11} = \frac{1}{4},\ \rho_{12} = \frac{3}{4}$ and $x^3 = (6.54, 10.11)$.

$i = 2$: $M_{x^3}[2] = \begin{pmatrix} 7.27 & 11.05 \\ 10.27 & 10.05 \end{pmatrix}$ with $val\big(M_{x^3}[2]\big) = 10.11$ and $\rho^3_{21} = \frac{1}{4},\ \rho^3_{22} = \frac{3}{4}$.

Since $x^3_2 = val\big(M_{x^3}[2]\big) = 10.11$, we have found the optimal solution: $v^\alpha = (6.54, 10.11)$, $\pi_{11} = \frac{3}{5},\ \pi_{12} = \frac{2}{5},\ \pi_{21} = 0.055,\ \pi_{22} = 0.945;\ \rho_{11} = \frac{1}{4};\ \rho_{12} = \frac{3}{4};\ \rho_{21} = \frac{1}{4};\ \rho_{22} = \frac{3}{4}$.

Lemma 10.5

For $n = 1, 2, \ldots$, we have $x^{n+1} \leq x^n$. Furthermore, if $val\big(M_{x^n}[i]\big) \neq x^n_i$ for some $i \in S_2$, then $x^{n+1} < x^n$, i.e. $x^{n+1}_i \leq x^n_i,\ i \in S$, with at least one strict inequality.

Proof

x^n is the value vector of the single-controller stochastic game $SCSG(\rho^{n-1})$, implying $x^n_i = val\Big(M^{\rho^{n-1}}_{x^n}[i]\Big),\ i \in S$. Therefore,

$$x^n_i = val\big(M_{x^n}[i]\big),\ i \in S_1 \text{ and } x^n_i = max_a \left\{ r_i(a, \rho^n) + \alpha \sum_j p_{ij}(\rho^n) x^n_j \right\},\ i \in S_2. \tag{10.26}$$

Let $\{\rho^n_{ib},\ i \in S_1,\ b \in B(i)\}$ be the optimal strategy for player 2 in the matrix games $M_{x^n}[i],\ i \in S_1$.

From (10.26) and the definition of ρ^n it follows that

$$x_i^n \geq r_i(a, \rho^n) + \alpha \sum_j p_{ij}(\rho^n) x_j^n, \ i \in S_1, \ a \in A(i). \tag{10.27}$$

By (10.26) and (10.27), we obtain

$$x^n \geq r(f, \rho^n) + \alpha P(f, \rho^n) x^n \text{ for all } f^\infty \in C(D), \tag{10.28}$$

from which it follows that $x^n \geq v^\alpha\big(f^\infty, (\rho^n)^\infty\big)$ for all $f^\infty \in C(D)$. Hence, we can write

$$\begin{aligned} x^n &\geq max_{f^\infty \in C(D)}\, v^\alpha\big(f^\infty, (\rho^n)^\infty\big) = max_{\pi^\infty \in \Pi}\, v^\alpha\big(\pi^\infty, (\rho^n)^\infty\big) \\ &\geq max_{\pi^\infty \in \Pi}\, inf_{\{\rho^\infty \in \Gamma \,|\, \rho_{ib} = \rho_{ib}^n, \ i \in S_2, \ b \in B(i)\}}\, v^\alpha\big(\pi^\infty, (\rho^n)^\infty\big) \\ &= \text{value vector of } SCSG(\rho^n) = x^{n+1}, \end{aligned}$$

which proves the first part of the lemma. If $val\big(M_{x^n}[i]\big) \neq x_i^n$ for some $i \in S_2$, then (10.28) holds with a strict inequality for at least one $i \in S_2$, i.e. $x^n > r(f, \rho^n) + \alpha P(f, \rho^n) x^n$ for all $f^\infty \in C(D)$, implying $x^n > x^{n+1}$. □

Theorem 10.11

Algorithm 10.6 is finite and correct.

Proof (outline)

If the algorithm is not finite, then by Lemma 10.5 $x^1 > x^2 > \cdots > x^n > \cdots$. Hence, the subsequent extreme strategies ρ^n, $n = 1, 2, \ldots$ are different. Now, by results of Parthasarathy and Raghavan ([145]) and Shapley and Snow ([184]), it can be shown that for any $i \in S$ and any $n \in \mathbb{N}$, the actions ρ_{ib}^n, $b \in B(i)$, are chosen from a same finite set. This yields a contradiction: the algorithm is finite.

Let the algorithm terminate at the n-th iteration, i.e. $x_i^n = val\big(M_{x^n}[i]\big)$ for all $i \in S_2$. Since we always have $x_i^n = val\big(M_{x^n}[i]\big)$ for all $i \in S_1$ (see (10.26)), we have $x^n = val\big(M_{x^n}\big)$: x^n is the value vector v^α of the game. The optimal stationary strategies in the matrix games $M_{x^n}[i]$, $i \in S$, say π_{ia}^n, $a \in A(i)$, and ρ_{ib}^n, $b \in B(i)$, are optimal policies in the stochastic game, namely: $r(\pi, \rho^n) + \alpha P(\pi, \rho^n) v^\alpha \leq v^\alpha$, $\pi^\infty \in \Pi$ implies $v^\alpha\big(\pi^\infty, (\rho^n)^\infty\big) = \{I - \alpha P(\pi, \rho^n)\}^{-1} r(\pi, \rho^n) \leq v^\alpha$ for all $\pi^\infty \in \Pi$. Similarly, we derive $v^\alpha\big((\pi^n)^\infty, \rho^\infty\big) \geq v^\alpha$ for all $\rho^\infty \in \Gamma$. □

Remark

The correctness of Algorithm 10.6 provides also the proof that the value vector and the optimal policies lie in the same ordered field as the data: linear programming is used and the dat of the stochastic games $SCSG(\rho^n)$ are also in the same field. Hence, the ordered field property holds also for switching control staochastic games.

SER-SIT stochastic game

In this game we assume that the rewards are *separable*, i.e. $r_i(a,b) = s_i + t(a,b)$ for all $i.a,b$ (SER property), and the transitions are *state independent*, i.e. $p_{ij}(a,b) = p_j(a,b),\ j \in S$ for all $i.a,b$ (SIT property). Note that the above is meaningful if the set $\{(a,b)\}$ is independent of the states. Therefore, we assume that $|A(i)| = m$ and $|B(i)| = n$ for all $i \in S$. Thus a fixed pair of actions (a,b) determines the same transition law, $p_j(a,b),\ j \in S$, in every $i \in S$. In addition, the SER property implies that all rewards are a sum of a contribution due the current state (s_i) and a contribution due to the action pair selected ($t(a,b)$).

Let $s = (s_1, s_2, \ldots, s_N)^T$ and define the $m \times n$ matrix $M = (m_{ab})$ by $m_{ab} = t(a,b) + \alpha \sum_j p_j(a,b)s_j$, $1 \le a \le m,\ 1 \le b \le n$, which is, unlike $M_x[i]$ in the previous section, independent of the state i.

Lemma 10.6

Let $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$ and $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ be an arbitrary pair of mixed strategies of the matrix game with matrix M. Then, $v^\alpha(\pi^\infty, \rho^\infty) = s + (1-\alpha)^{-1}\pi^T M \rho \cdot e$.

Proof

Since

$$v^\alpha(\pi^\infty, \rho^\infty) = r(\pi,\rho) + \alpha P(\pi,\rho)v^\alpha(\pi^\infty,\rho^\infty) = s + t(\pi,\rho)\cdot e + \alpha P(\pi,\rho)v^\alpha(\pi^\infty,\rho^\infty),$$

we also have

$$v_i^\alpha(\pi^\infty,\rho^\infty) - s_i = t(\pi,\rho) + \alpha \textstyle\sum_j p_j(\pi,\rho)s_j + \alpha \sum_j p_j(\pi,\rho)\{v_j^\alpha(\pi^\infty,\rho^\infty) - s_j\},\ i \in S.$$

In vector notation: $\{I - \alpha P(\pi,\rho)\}\{v^\alpha(\pi^\infty,\rho^\infty) - s\} = t(\pi,\rho)\cdot e + \alpha P(\pi,\rho)s = \pi^T M\rho \cdot e$.

Hence,

$v^\alpha(\pi^\infty,\rho^\infty) - s = \pi^T M\rho \cdot \{I - \alpha P(\pi,\rho)\}^{-1}\cdot e = \pi^T M\rho \cdot \sum_{t=0}^\infty \{\alpha P(\pi,\rho)\}^t \cdot e = \pi^T M\rho \cdot (1-\alpha)^{-1}\cdot e$,

i.e. $v^\alpha(\pi^\infty,\rho^\infty) = s + (1-\alpha)^{-1}\pi^T M\rho \cdot e$. □

Corollary 10.4

Let $v^* = val(M)$ and let $\pi^* = (\pi_1^*, \pi_2^*, \ldots, \pi_m^*)$ and $\rho^* = (\rho_1^*, \rho_2^*, \ldots, \rho_n^*)$ be a pair of optimal mixed strategies of the matrix game with matrix M. Then the value vector of the stochastic game $v^\alpha = s + \frac{1}{1-\alpha}v^* \cdot e$ and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal policies for player 1 and player 2, respectively.

Proof

Since π^* and ρ^* are optimal strategies for the matrix game with matrix M, we have for all strategies π and ρ: $\pi^T M\rho^* \le (\pi^*)^T M\rho^* \le (\pi^*)^T M\rho$. Therefore, by Lemma 10.6, for all $\pi^\infty \in \Pi$ and all $\rho^\infty \in \Gamma$: $v^\alpha\big(\pi^\infty, (\rho^*)^\infty\big) \le v^\alpha\big((\pi^*)^\infty, (\rho^*)^\infty\big) \le v^\alpha\big((\pi^*)^\infty, \rho^\infty\big)$. Hence, by Theorem 10.2, the value vector of the stochastic game is $s + \frac{1}{1-\alpha}v^* \cdot e$ and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal policies for player 1 and player 2, respectively. □

Algorithm 10.7 *SER-SIT game with discounting*

1. Determine the value v^* and optimal stationary strategies π^* and ρ^* of the matrix game with matrix M, where M is the $m \times n$ matrix defined by $m_{ab} = t(a,b) + \alpha \sum_j p_j(a,b)s_j$.
2. $v^\alpha = s + \frac{1}{1-\alpha} v^* \cdot e$ is the value vector, and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal policies for player 1 and player 2, respectively.

Remark

Since the value v^* and the optimal stationary strategies π^* and ρ^* can be computed by linear programming, $SER - SIT$ games possess the ordered field property.

Example 10.6

Consider the following example, which is a SIT game but no SER game.

$S = \{1,2,3\};\ A(1) = A(2) = A(3) = \{1,2\};\ B(1) = B(2) = B(3) = \{1,2\}.$

$r_1(1,1) = 0,\ r_1(1,2) = 0,\ r_1(2,1) = 0,\ r_1(2,2) = 1;\ r_2(1,1) = -1,\ r_2(1,2) = -1;$

$r_2(2,1) = -1,\ r_2(2,2) = -1,\ r_3(1,1) = -2,\ r_3(1,2) = -2;\ r_3(2,1) = 2,\ r_3(2,2) = 1.$

$p_1(1,1) = 1,\ p_2(1,1) = 0,\ p_3(1,1) = 0;\ p_1(1,2) = 0,\ p_2(1,2) = 0,\ p_3(1,2) = 1;$

$p_1(2,1) = 0,\ p_2(2,1) = 1,\ p_3(2,1) = 0;\ p_1(2,2) = 1,\ p_2(2,2) = 0,\ p_3(2,2) = 0.$

For the value vector v^α, we have

$$v_1^\alpha = val \begin{pmatrix} \alpha v_1^\alpha & \alpha v_3^\alpha \\ \alpha v_2^\alpha & \alpha v_1^\alpha \end{pmatrix};\ v_2^\alpha = val \begin{pmatrix} -1+\alpha v_1^\alpha & -1+\alpha v_3^\alpha \\ -1+\alpha v_2^\alpha & -1+\alpha v_1^\alpha \end{pmatrix};\ v_3^\alpha = val \begin{pmatrix} -2+\alpha v_1^\alpha & -2+\alpha v_3^\alpha \\ -1+\alpha v_2^\alpha & -1+\alpha v_1^\alpha \end{pmatrix}.$$

The matrix of v_1^α has entries which are all 1 larger than the entries of the matrix of v_2^α: $v_1^\alpha = v_2^\alpha + 1$. Furthermore, we see in the matrix of v_3^α that the entry at position (2,1), i.e. $-1 + \alpha v_2^\alpha$, is the largest in the first column and the smallest in the second row. So, this entry is a saddle point, i.e. $v_3^\alpha = -1 + \alpha v_2^\alpha$. Hence, we obtain from the equation for v_1^α,

$$\tfrac{1}{\alpha} v_1^\alpha = val \begin{pmatrix} v_1^\alpha & v_3^\alpha \\ v_2^\alpha & v_1^\alpha \end{pmatrix} = val \begin{pmatrix} v_1^\alpha & \alpha v_1^\alpha - 1 - \alpha \\ v_1^\alpha - 1 & v_1^\alpha \end{pmatrix}.$$

Hence, $\frac{1}{\alpha} v_1^\alpha = \frac{v_1^\alpha \cdot v_1^\alpha - (v_1^\alpha - 1)\cdot(\alpha v_1^\alpha - 1 - \alpha)}{v_1^\alpha + v_1^\alpha - (v_1^\alpha - 1) - (\alpha v_1^\alpha - 1 - \alpha)} = \frac{(1-\alpha)\left(v_1^\alpha\right)^2 + v_1^\alpha(1+2\alpha) - (1+\alpha)}{(1-\alpha)v_1^\alpha + 2 + \alpha}$.

The solution of this quadratic equation yields $v_1^\alpha = \frac{-(1+\alpha) + \sqrt{(1+\alpha)}}{1-\alpha}$. Let $\alpha = \frac{1}{2}$, then $v_1^\alpha = -3 + \sqrt{6}$. So, we conclude that the SIT game without the SER property does not possess the ordered field property.

Remark

It can also be shown (see Exercise 10.9) that a SER game without the SIT property does not possess the ordered field property.

ARAT stochastic game

An additive reward and additive transition ($ARAT$) stochastic game is defined by the property that the rewards as well as the transitions can be written as the sum of a term determined by player 1 and a term determined by player 2: $r_i(a,b) = r_i^1(a) + r_i^2(b),\ i \in S,\ a \in A(i),\ b \in B(i)$ and $p_{ij}(a,b) = p_{ij}^1(a) + p_{ij}^2(b),\ i,j \in S,\ a \in A(i),\ b \in B(i)$.

Theorem 10.12

(1) Both players have optimal deterministic and stationary policies.

(2) The ordered field property holds.

Proof

(1) By the additivity of $r_i(a,b)$ and $p_{ij}(a,b)$, the matrix $M_x[i]$, with entries $r_i(a,b)+\alpha\sum_j p_{ij}(a,b)x_j$ can be written as the sum of two matrices: $M_x[i] = A_x[i] + B_x[i]$, where $A_x[i]$ and $B_x[i]$ have elements $r_i^1(a) + \alpha\sum_j p_{ij}^1(a)x_j$ and $r_i^2(b) + \alpha\sum_j p_{ij}^2(b)x_j$. The matrix $A_x[i]$ has identical columns and the matrix $B_x[i]$ has identical rows. Consider the equation $x_i = val(M_x[i])$, which has as unique solution v_i^α. Effectively, this means that player 1 is only interested in the matrix $A_{v^\alpha}[i]$ with identical columns, and player 2 is only interested in the matrix $B_{v^\alpha}[i]$ with identical rows. Hence, in each state i, both players possess deterministic optimal strategies. Hence, the stochastic game has optimal deterministic and stationary policies.

(2) Let f_*^∞ and g_*^∞ be optimal deterministic and stationary optimal policies for player 1 and 2, respectively. Then, the value vector $v^\alpha = v^\alpha(f_*^\infty, g_*^\infty) = \{I - \alpha P(f,g)\}^{-1} r(f,g)$, which shows the ordered field property. □

Since there are only a finite number of deterministic and stationary policies, there is a finite algorithm. The next algorithm is a special version of Algorithm 10.2 with $\varepsilon = 0$.

Algorithm 10.8 *ARAT game with discounting*

1. Choose a deterministic and stationary policy g^∞ for player 2.
2. Solve the MDP induced by the policy g^∞: $x = max_{f^\infty \in C(D)}\, v^\alpha(f^\infty, g^\infty)$.
3. Compute for each $i \in S$: $y_i = val(M_x[i])$ and a deterministic and stationary optimal strategy $g(i) \in B(i)$.
4. If $\|y - x\|_\infty = 0$:
   a. Determine for each $i \in S$ an optimal deterministic and stationary strategy $f(i) \in A(i)$, for player 1 in the matrix game $M_x[i]$;
   b. y is the value vector v^α and f^∞ and g^∞ are optimal policies for player 1 and 2, respectively (STOP);

   Otherwise: return to step 2.

10.3 Average rewards

10.3.1 Value and optimal policies

A policy R_1^* is *optimal for player 1* if $\phi(R_1^*, R_2) \geq inf_{R_2}\, sup_{R_1}\, \phi(R_1, R_2)$ for all policies R_2.

A policy R_2^* is *optimal for player 2* if $\phi(R_1, R_2^*) \leq sup_{R_1}\, inf_{R_2}\, \phi(R_1, R_2)$ for all policies R_1.

The stochastic undiscounted game has a *value* if $inf_{R_2}\, sup_{R_1}\, \phi(R_1, R_2) = sup_{R_1}\, inf_{R_2}\, \phi(R_1, R_2)$.

The value vector of an undiscounted stochastic game is denoted by ϕ.

A policy R_1^* is ε*-optimal for player 1* if $\phi(R_1^*, R_2) \geq inf_{R_2} sup_{R_1} \phi(R_1, R_2) - \varepsilon$ for all policies R_2.

A policy R_2^* is ε*-optimal for player 2* if $\phi(R_1, R_2^*) \leq sup_{R_1} inf_{R_2} \phi(R_1, R_2) + \varepsilon$ for all policies R_1.

Theorem 10.13
If the policies R_1^* and R_2^* satisfy $\phi(R_1, R_2^*) \leq \phi(R_1^*, R_2^*) \leq \phi(R_1^*, R_2)$ for all policies R_1 and R_2, the game has a value and R_1^* and R_2^* are optimal policies.

Proof
The proof is analogous to the proof of Theorem 10.2. □

We have seen in Chapter 5 that, for Markov decision processes, the average reward criterion is considerably more difficult to analyze than the discounted reward criterion. Nonetheless, after overcoming a number of technical difficulties, results of qualitative strength and generality were established for both the discounted and the average reward criterium. For instance, we have seen that in both cases there are optimal deterministic and stationary optimal policies and that they can be found by policy iteration, linear programming and value iteration methods. Consequently, one might think that in the case of stochastic games (perhaps at the cost of extra analysis) it might be possible to obtain qualitatively the same results in the average and the discounted case. Unfortunately, this is not the case. In the next section we shall illustrate some of the problems that arise.

10.3.2 The Big Match

The seemingly simple example described below can be used to illustrate many of the difficulties arising in the analysis of stochastic games under the average reward criterion.

Example 10.7 *The Big Match*
$S = \{1, 2, 3\};\ A(1) = B(1) = \{1, 2\},\ A(2) = B(2) = A(3) = B(3) = \{1\}.$
$r_1(1,1) = 1,\ r_1(1,2) = 0,\ r_1(2,1) = 0,\ r_1(2,2) = 1;\ r_2(1,1) = 0;\ r_3(1,1) = 1.$
$p_{11}(1,1) = 1,\ p_{12}(1,1) = 0,\ p_{13}(1,1) = 0;\ p_{11}(1,2) = 1,\ p_{12}(1,2) = 0,\ p_{13}(1,2) = 0;$
$p_{11}(2,1) = 0,\ p_{12}(2,1) = 1,\ p_{13}(2,1) = 0;\ p_{11}(2,2) = 0,\ p_{12}(2,2) = 0,\ p_{13}(2,2) = 1.$
$p_{21}(1,1) = 0,\ p_{22}(1,1) = 1,\ p_{23}(1,1) = 0;\ p_{31}(1,1) = 0,\ p_{32}(1,1) = 0,\ p_{33}(1,1) = 1.$
The states 2 and 3 are absorbing: $\phi_2(R_1, R_2) = 0$ and $\phi_3(R_1, R_2) = 1$ for all policies R_1 and R_2.

However, it seems that the structure of the transition data makes the choice for player 1 in state 1 extremely difficult. While the choice of the first action leads to a repetition of the same game, the choice of the second action absorbs the game either in state 2 or state 3, depending on the choice of player 2. Thus the consequence of the second choice is so permanent and with such different payoffs that it is a risky action.

To make the above point more precise, suppose that player 1 uses a stationary policy π^∞ with probability p for action 1 and probability $1-p$ for action 2 in state 1, and that player 2 uses a stationary policy ρ^∞ with probability q for action 1 and probability $1-q$ for action 2 in state 1. Let $P[p,q]$ and $r[p,q]$ be the corresponding transition matrix and reward vector.

There are now two cases.

Case 1: $p = 1$

$$P[p,q] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix};\ r[p,q] = (q,0,1)^T \ \rightarrow\ P^*[p,q] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ \rightarrow\ \phi_1\big(\pi^\infty,\rho^\infty\big) = q.$$

Case 2: $0 \le p < 1$

$$P[p,q] = \begin{pmatrix} p & (1-p)q & (1-p)(1-q) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix};\ r[p,q] = (pq + (1-p)(1-q), 0, 1)^T.$$

$$P^*[p,q] = \begin{pmatrix} 0 & q & 1-q \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ \rightarrow\ \phi_1\big(\pi^\infty,\rho^\infty\big) = 1-q.$$

Hence,

$$max_{0\le p\le 1}\,\phi_1(\pi^\infty,\rho^\infty) = max_{0\le q\le 1}\,(q,1-q) \ \rightarrow\ min_{0\le q\le 1}\,\max_{0\le p\le 1}\,\phi_1(\pi^\infty,\rho^\infty) = \tfrac{1}{2}$$

and

$$min_{0\le q\le 1}\,\phi_1(\pi^\infty,\rho^\infty) = min_{0\le q\le 1}\,(q,1-q) = 0 \ \rightarrow\ max_{0\le p\le 1}\,\min_{0\le q\le 1}\,\phi_1(\pi^\infty,\rho^\infty) = 0.$$

Therefore,

$$max_{\pi^\infty\in\Pi}\,min_{\rho^\infty\in\Gamma}\,\phi_1(\pi^\infty,\rho^\infty) = 0 < \frac{1}{2} = min_{\rho^\infty\in\Gamma}\,max_{\pi^\infty\in\Pi}\,\phi_1(\pi^\infty,\rho^\infty). \tag{10.29}$$

Of course, the above strict inequality implies that optimal stationary policies do not exist in the Big Match. Since we alway have $sup_{R_1}\,inf_{R_2}\,\phi(R_1,R_2) \le inf_{R_2}\,sup_{R_1}\,\phi(R_1,R_2)$, it is sufficient for the existence of the value of this stochastic game to show that there exitst a policy R_1^* for player 1 such that $\phi_1(R_1^*,R_2) \ge \frac{1}{2} - \varepsilon$ for every $\varepsilon > 0$ and every policy R_2 for player 2, namely: In that case we have:

$$\begin{aligned} sup_{R_1}\,inf_{R_2}\,\phi_1(R_1,R_2) &\ge inf_{R_2}\,\phi_1(R_1^*,R_2) \\ &\ge \tfrac{1}{2} = min_{\rho^\infty\in\Gamma}\,max_{\pi^\infty\in\Pi}\,\phi_1(\pi^\infty,\rho^\infty) \\ &= min_{\rho^\infty\in\Gamma}\,sup_{R_1}\,\phi_1(R_1,\rho^\infty) \\ &\ge inf_{R_2}\,sup_{R_1}\,\phi_1(R_1,R_2). \end{aligned}$$

Thus, in order to show that the Big Match has a value vector, it is sufficient to show that for any $M \in \mathbb{N}$ there exists a policy R_1^M for player 1 such that, for any realization $Z = (Z_1, Z_2, \dots)$ of the actions of player 2, player 1 has an average reward of at least $\frac{1}{2} \cdot \frac{M}{M+1}$.
At decision point $t+1$ player 1 knows the realizations of $Z_1, Z_2, \dots, Z_t$, say $b_1, b_2, \dots, b_t$, where $b_i \in \{1,2\},\ 1 \leq i \leq t$. Let k_t^1 be the number of 1's and k_t^2 be the number of 2's in $\{b_1, b_2, \dots, b_t\}$, and let $k_t = k_t^1 - k_t^2$ for $t = 1, 2, \dots$. The policy R_1^M is history-dependent, but depends only on the numbers k_t. Let $\pi^{t+1}(k_t)$ be the probability that player 1 chooses action 2 in state 1 at time point $t+1$, given k_t. Then, in policy R_1^M, we take

$$\pi^{t+1}(k_t) = \frac{1}{(k_t + M + 1)^2} \text{ for } t = 0, 1, 2, \dots, \text{ where } k_0 \equiv 0.$$

Intuitively, when k_t is positive and large, i.e. player 2 seems more willing for action 1, which leads - when player 1 chooses action 2 - to the for player 1 bad state 2, then the probability that player 1 chooses action 2 is very small; when k_t is negative and tends to $-M$, i.e. player 2 seems more willing for action 2, which leads - when player 1 chooses action 2 - to the for player 1 good state 3, then the probability that player 1 chooses action 2 is increasing to 1.

Lemma 10.7
Define the events E_m by $E_m = \{Y_1 = Y_2 = \dots = Y_m = 1$ or $Y_n = Z_n = 2$ for some $1 \leq n \leq m\}$ for $m = 1, 2, \dots$. Then, $\mathbb{P}_{R_1^M}\{E_m\} \geq \frac{1}{2} \cdot \frac{M}{M+1}$ for all $m, M \in \mathbb{N}$.

Proof
The proof is inductively on m.
If $m = 1$, then $E_m = \{Y_1 = 1 \text{ or } Y_1 = Z_1 = 2\}$.
Hence, $\mathbb{P}_{R_1^M}\{E_m\} \geq \mathbb{P}_{R_1^M}\{Y_1 = 1\} = 1 - \frac{1}{(M+1)^2} \geq \frac{1}{2} \cdot \frac{M}{M+1}$ for all $M \in \mathbb{N}$.
Assume that $\mathbb{P}_{R_1^M}\{E_m\} \geq \frac{1}{2} \cdot \frac{M}{M+1}$ for all $m, M \in \mathbb{N}$ and consider E_{m+1}.
We distinguish between the two possibilities for Z_1.

Case 1: $Z_1 = 1$, i.e. $k_1 = 1$

$$\begin{aligned} E_{m+1} &= \{Y_1 = Y_2 = \dots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 2 \leq n \leq m+1\} \\ &= \Big\{Y_1 = 1 \text{ or } \{Y_2 = \dots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 2 \leq n \leq m+1\}\Big\}. \end{aligned}$$

If $Y_1 = 2$, then the next state is state 2 and $Y_n = Z_n = 2$ for some $2 \leq n \leq m+1$ is impossible.
Therefore,

$$\mathbb{P}_{R_1^M}\{E_{m+1}\} = \mathbb{P}_{R_1^M}\{Y_1 = 1\} \cdot \mathbb{P}_{R_1^M}\{Y_2 = \dots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 2 \leq n \leq m+1\}.$$

Because $k_1 = 1$,

$$\mathbb{P}_{R_1^M}\{Y_2 = \dots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 2 \leq n \leq m+1\} = \mathbb{P}_{R_1^{M+1}}\{E_m\}.$$

Hence,

$$\mathbb{P}_{R_1^M}\{E_{m+1}\} = \Big\{1 - \tfrac{1}{(M+1)^2}\Big\} \cdot \mathbb{P}_{R_1^{M+1}}\{E_m\} = \tfrac{M(M+2)}{(M+1)^2} \cdot \mathbb{P}_{R_1^{M+1}}\{E_m\} \geq \tfrac{M(M+2)}{(M+1)^2} \cdot \tfrac{1}{2} \cdot \tfrac{M+1}{M+2} = \tfrac{1}{2} \cdot \tfrac{M}{M+1}.$$

Case 2: $Z_1 = 2$, *i.e.* $k_1 = -1$

$$\begin{aligned} E_{m+1} &= \{Y_1 = Y_2 = \cdots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 1 \le n \le m+1\} \\ &= \Big\{Y_1 = Z_1 = 2 \text{ or } \{Y_1 = Y_2 = \cdots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 2 \le n \le m+1\}\Big\}. \end{aligned}$$

Since $Z_1 = 2$, we have $\mathbb{P}_{R_1^M}\{Y_1 = Z_1 = 2\} = \mathbb{P}_{R_1^M}\{Y_1 = 2\} = \frac{1}{(M+1)^2}$.

If $Y_1 = 2$, then the next state is state 3 and $Y_n = Z_n = 2$ for some $2 \le n \le m+1$ is impossible.

Therefore,

$$\mathbb{P}_{R_1^M}\{E_{m+1}\} = \tfrac{1}{(M+1)^2} + \mathbb{P}_{R_1^M}\{Y_1 = 1\}\,\mathbb{P}_{R_1^M}\{Y_2 = \cdots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 2 \le n \le m+1\}.$$

Because $k_1 = -1$,

$$\mathbb{P}_{R_1^M}\{Y_2 = \cdots = Y_{m+1} = 1 \text{ or } Y_n = Z_n = 2 \text{ for some } 2 \le n \le m+1\} = \mathbb{P}_{R_1^{M-1}}\{E_m\}.$$

Hence,

$$\begin{aligned} \mathbb{P}_{R_1^M}\{E_{m+1}\} &= \tfrac{1}{(M+1)^2} + \Big\{1 - \tfrac{1}{(M+1)^2}\Big\} \cdot \mathbb{P}_{R_1^{M-1}}\{E_m\} \ge \tfrac{1}{(M+1)^2} + \tfrac{M(M+2)}{(M+1)^2} \cdot \tfrac{1}{2} \cdot \tfrac{M-1}{M} \\ &= \tfrac{1}{(M+1)^2} \cdot \{1 + \tfrac{1}{2} \cdot (M-1)(M+2)\} = \tfrac{1}{2} \cdot \tfrac{M}{M+1}. \end{aligned}$$

So, we have shown that $\mathbb{P}_{R_1^M}\{E_m\} \ge \frac{1}{2} \cdot \frac{M}{M+1}$ for all $m, M \in \mathbb{N}$. □

Lemma 10.8

$\phi_1(R_1^M, R_2) \ge \frac{1}{2} \cdot \frac{M}{M+1}$ for all $M \in \mathbb{N}$ and all policies R_2.

Proof

Take any $M \in \mathbb{N}$ and any policy R_2 with realization $Z_1, Z_2, \ldots$.

Consider the cases $k_t > -M$ for all t and $k_t = -M$ for some t separately.

Case 1: $k_t = -M$ *for some* t

In this case player 1 chooses at time $t+1$ (or earlier) action 2. Let m be the smallest time point for which $Y_m = 2$. Then, we have $\phi_1(R_1^M, R_2) = \mathbb{P}\{Z_m = 2\} = \mathbb{P}_{R_1^M}\{E_m\} \ge \frac{1}{2} \cdot \frac{M}{M+1}$ for any policy R_2 of player 2.

Case 2: $k_t > -M$ *for all* t

Let t be the smallest time point for which $Y_{t+1} = 2$ (if $Y_n = 1$ for all n, then $t = \infty$). Define for $m \ge 2$: $\lambda(m) = \mathbb{P}\{t < m \text{ and } Z_{t+1} = 1\}$ and $\mu(m) = \mathbb{P}\{t < m \text{ and } Z_{t+1} = 2\}$. Then, the sequences $\{\lambda(m)\}$ and $\{\mu(m)\}$ are nondecreasing. Let $\lambda = lim_{m\to\infty}\,\lambda(m)$ and $\mu = lim_{m\to\infty}\,\mu(m)$: λ is the probability that the game ends in state 2, μ the probability that the game ends in state 3, and $1 - \lambda - \mu$ is the probability that the game never leaves state 1. Since $k_t^1 + k_t^2 = t$ and $k_t = k_t^1 - k_t^2 > -M$ for all t, we have $k_t^1 > \frac{1}{2}(t - M)$: $\frac{k_t^1}{t} > \frac{1}{2}\big(1 - \frac{M}{t}\big)$ for all t, implying $liminf_{t\to\infty}\,\frac{k_t^1}{t} \ge \frac{1}{2}$. Hence, $\phi_1(R_1^M, R_2) \ge \mu + (1 - \lambda - \mu) \cdot \frac{1}{2} = \frac{1}{2}(1 - \lambda + \mu)$.

Finally, we have to show that $\frac{1}{2}(1 - \lambda + \mu) \ge \frac{M}{M+1}$. Therefore, consider the following policy for player 2: first he plays according to $Z_1, Z_2, \ldots, Z_m$ and thereafter he uses a fain coin, i.e. with probability $\frac{1}{2}$ he chooses action 1 and action 2. Then, the expected average reward for player 1 is: the probability to move from state 1 to state 3 during the first m time point plus the probability to be at time point $m+1$ in state 1 multiplied with the average reward from time point $m+1$. This yields $\mu(m) + \{1 - \lambda(m) - \mu(m)\} \cdot \frac{1}{2}$.

On the other hand, any realization of this policy will, with probability 1, reach $k_t = -M$ for some t. Hence, by case 1 of the lemma, $\mu(m) + \{1 - \lambda(m) - \mu(m)\} \cdot \frac{1}{2} \geq \frac{M}{M+1}$ for all m. Letting $m \to \infty$ completes the proof that $\frac{1}{2}(1 - \lambda + \mu) \geq \frac{M}{M+1}$. □

Theorem 10.14

The Big Match has the following properties:

(1) There exists a value vector ϕ and $\phi = \left(\frac{1}{2}, 0, 1\right)^T$.

(2) Player 2 has an optimal stationary policy ρ^∞: $\rho_{11} = \rho_{12} = \frac{1}{2}$.

(3) For any $\varepsilon > 0$ player 1 has a ε-optimal policy: R_1^M with $M = \frac{1}{2}\{\frac{1}{\varepsilon} - 2\}$.

(4) Player 1 has no optimal policy.

Proof

(1) We have shown above that this game has a value vector ϕ and that $\phi = \left(\frac{1}{2}, 0, 1\right)^T$.

(2) Take for player 2 the stationary policy with $\rho_{11} = \rho_{12} = \frac{1}{2}$. Then, in state 1 in each period player 1 earns $\frac{1}{2}$ independent of his strategy: $\phi_1(R_1, \rho^\infty) = \frac{1}{2} = \phi_1$ for all policies R_1, i.e. ρ^∞ is an optimal policy for player 2.

(3) We have shown in Lemma 10.8 that $\phi_1(R_1^M, R_2) \geq \frac{1}{2} \cdot \frac{M}{M+1}$ for all $M \in \mathbb{N}$ and all policies R_2. Hence, with $M = \frac{1}{2}\{\frac{1}{\varepsilon} - 2\}$, we obtain $\phi_1(R_1^M, R_2) \geq \frac{1}{2} - \varepsilon$, i.e. R_1^M with $M = \frac{1}{2}\{\frac{1}{\varepsilon} - 2\}$ is an optimal policy for player 1.

(4) Suppose that player 1 has an optimal policy, say $R_1^* = (\pi^1, \pi^2, \dots)$, i.e. $\phi_1(R_1^*, R_2) \geq \frac{1}{2}$ for all R_2. The game is only interesting in state 1, i.e. as long as player 1 uses action 1.

We distinguish between two cases.

Case 1: $R_1^* = f_*^\infty$ with $f_*(1) = 1$

Take $R_2 = g^\infty$ with $g(1) = 2$. Then, $\phi_1(R_1^*, R_2) = 0 < \frac{1}{2} = \phi_1$: R_1^* is not optimal for player 1.

Case 2: $R_1^* \neq f_*^\infty$ with $f_*(1) = 1$

Suppose that $\pi^t_{h_t 2} = \varepsilon > 0$ for some t and some history h_t. Let t be the smallest time point for which this case holds and suppose that $b_1, b_2, \cdots, b_{t-1}$ is the sequence of actions for player 2 in h_t. Take $R_2 = (\rho^1, \rho^2, \dots)$ such that $\rho^n_{h_n b} = \begin{cases} 1 & 1 \leq n \leq t-1 \quad b = b_n \\ 0 & 1 \leq n \leq t-1 \quad b \neq b_n \\ 1 & n = t \quad b = 1 \\ 0 & n = t \quad b = 2 \\ \frac{1}{2} & n \geq t \quad b = 1, 2 \end{cases}$

Then, $\phi_1(R_1^*, R_2) = \varepsilon \cdot 0 + (1 - \varepsilon) \cdot \frac{1}{2} < \frac{1}{2} = \phi_1$: R_1^* is not optimal for player 1. □

The above lack of a solution in the space of stationary policies naturally gives reason for the following questions:

(1) Have stochastic games under the average reward criterion a value vector?

(2) Are there optimal (nonstationary) policies?

(3) For which subclasses do exist stationary optimal policies?

The answer to the first question remained open for over twenty years and was answered in the affirmative by Mertens and Neyman ([137]). This is a deep result, based on ingenious analysis in a series of three papers by Bewley and Kohlberg ([18],[19],[20]), who expressed the value vector of the discounted stochastic game in a Puiseux series, the so-called *limit discount equation*. We will not present the proof of the existence of the value vector in these lecture notes.
The Big Match shows that in general there are no optimal policies. The existence of ε-optimal policies follows from the existence of the value vector. Let ϕ be the value vector. Then, for any $\varepsilon > 0$ and any state i, we obtain from $sup_{R_1}\, inf_{R_2}\, \phi_i(R_1, R_2) = \phi_i$ that there exists a policy R_1^ε such that $inf_{R_2}\, \phi_i(R_1^\varepsilon, R_2) \geq \phi_i - \varepsilon$, implying $\phi_i(R_1^\varepsilon, R_2) \geq \phi_i - \varepsilon$ for all policies R_2 for player 2. Therefore, policy R_1^ε is a ε-optimal policy for player 1. Similarly it can be shown that player 2 has an ε-optimal policy for any $\varepsilon > 0$.

In view of the fact that in general undiscounted games need not possess optimal stationary policies, the algorithmic development for computing such policies centered around 'natural' classes that possess optimal stationary policies and on supplying algorithms for their computation. These classes of games can be roughly divided into two groups:
(1) Those that make assumptions on the ergodic properties of the underlying Markov chains.
(2) Those that make assumptions on the structure of the game data (transitions and/or rewards).
In the sequel we will encounter several special stochastic games which have stationary or even deterministic optimal policies.

10.3.3 Mathematical programming

Inspired by the concepts of super- and subharmonicity for both MDPs (cf. Theorem 5.15) and discounted stochastic games (cf. Theorem 10.4) we define for undiscounted stochastic games super- and subharmonicity as follows:

A vector $v \in \mathbb{R}^N$ is *superharmonic* if there exists a vector $t \in \mathbb{R}^N$ and a policy $\rho^\infty \in \Gamma$ such that the triple (v, t, ρ) satisfies the following system of inequalities

$$\begin{cases} v_i & \geq & \sum_j p_{ij}(a,\rho)v_j & \text{for every } (i,a) \in S \times A \\ v_i + t_i & \geq & r_i(a,\rho) + \sum_j p_{ij}(a,\rho)t_j & \text{for every } (i,a) \in S \times A \end{cases} \tag{10.30}$$

A vector $v \in \mathbb{R}^N$ is *subharmonic* if there exists a vector $u \in \mathbb{R}^N$ and a policy $\pi^\infty \in \Pi$ such that the triple (v, u, π) satisfies the following system of inequalities

$$\begin{cases} v_i & \leq & \sum_j p_{ij}(\pi,b)v_j & \text{for every } (i,b) \in S \times B \\ v_i + u_i & \leq & r_i(\pi,b) + \sum_j p_{ij}(\pi,b)u_j & \text{for every } (i,b) \in S \times B \end{cases} \tag{10.31}$$

Theorem 10.15
An undiscounted stochastic game has stationary optimal policies $(\pi^*)^\infty$ and $(\rho^*)^\infty$ for player 1 and 2, respectively, if and only if $(v, t, \rho = \rho^*)$ and $(v, u, \pi = \pi^*)$ are feasible solutions of (10.30) and (10.31), respectively.

Proof

Assume that (v, t, ρ^*) and (v, u, π^*) are feasible solutions of (10.30) and (10.31). Then, for any $\pi^\infty \in \Pi$, $v \geq P(\pi, \rho^*)v$ and $v + t \geq r(\pi, \rho^*) + P(\pi, \rho^*)t$. The first inequality yields $v \geq P^*(\pi, \rho^*)v$ and consequently,

$$v \geq P^*(\pi, \rho^*)v \geq P^*(\pi, \rho^*)\{r(\pi, \rho^*) + P(\pi, \rho^*)t\} = P^*(\pi, \rho^*)r(\pi, \rho^*) = \phi\big(\pi^\infty, (\rho^*)^\infty\big).$$

Hence,

$$v \geq \phi\big(\pi^\infty, (\rho^*)^\infty\big) \text{ for all } \pi^\infty \in \Pi. \tag{10.32}$$

Similarly, we derive

$$v \leq \phi\big((\pi^*)^\infty, \rho^\infty\big) \text{ for all } \rho^\infty \in \Gamma. \tag{10.33}$$

So,

$$\phi\big(\pi^\infty, (\rho^*)^\infty\big) \leq v \leq \phi\big((\pi^*)^\infty, \rho^\infty\big) \text{ for all } \pi^\infty \in \Pi \text{ and } \rho^\infty \in \Gamma,$$

implying that the stochastic game has value vector $\phi = v$ and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and 2, respectively.

Now assume that $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are stationary optimal policies for player 1 and 2, respectively. Then,

$$\phi\big((\pi^*)^\infty, R_2\big) \geq inf_{R_2}\, sup_{R_1}\, \phi(R_1, R_2) \geq sup_{R_1}\, inf_{R_2}\, \phi(R_1, R_2) \geq \phi\big(R_1, (\rho^*)^\infty\big) \text{ for all } R_1 \text{ and } R_2,$$

implying

$$\phi\big((\pi^*)^\infty, R_2\big) \geq \phi = \phi\big((\pi^*)^\infty, (\rho^*)^\infty\big) \geq \phi\big(R_1, (\rho^*)^\infty\big) \text{ for all } R_1 \text{ and } R_2.$$

Hence, $(\pi^*)^\infty$ is an optimal policy in the MDP induced by the stationary policy $(\rho^*)^\infty$. Consequently ϕ is the smallest superharmonic vector in the sense of an undiscounted MDP problem, i.e. $(v = \phi, t, \rho = \rho^*)$ is a feasible solution of (10.30) for some t.

Similarly, it follows that $(\rho^*)^\infty$ is an optimal policy in the MDP induced by the stationary policy $(\pi^*)^\infty$ with respect to minimizing the average rewards. Therefore, ϕ is the largest subharmonic vector in the sense of an undiscounted MDP problem, i.e. $(v = \phi, u, \pi = \pi^*)$ is a feasible solution of (10.31) for some u. □

The systems (10.30) and (10.31) contain a mixture of linear and nonlinear terms. A method to solve these systems is to transform the systems to a nonlinear program. In the next corollary we have exhibit this idea. The nonlinear parts are moved to the objective function and the constraints are linear. We add some variables: $w_i(a), x_i(a)$ in (10.30) and $y_i(b), z_i(b)$ in (10.31) and obtain the following result.

Corollary 10.5

An undiscounted stochastic game has a value vector and optimal stationary policies if and only if the following nonlinear program has a global minimum value zero.

$$min \left\{\sum_{(i,a)} \left\{w_i(a) - \sum_j \sum_b p_{ij}(a,b)\rho_{ib}v_j\right\}^2 + \sum_{(i,a)} \left\{x_i(a) - \sum_j \sum_b p_{ij}(a,b)\rho_{ib}t_j\right\}^2 +\right.$$

$$\sum_{(i,b)} \left\{ y_i(b) - \sum_i \sum_a p_{ij}(a,b)\pi_{ia} v_j \right\}^2 + \sum_{(i,b)} \left\{ z_i(b) - \sum_j \sum_b p_{ij}(a,b)\pi_{ia} u_j \right\}^2 \Bigg\}$$

subject to

(1) $v_i - w_i(a) \geq 0,\ (i,a) \in S \times A;$

(2) $v_i + t_i - x_i(a) - \sum_b r_i(a,b)\rho_{ib} \geq 0,\ (i,a) \in S \times A;$

(3) $-v_i + y_i(b) \geq 0,\ (i,b) \in S \times B;$

(4) $-v_i - t_i + z_i(b) + \sum_a r_i(a,b)\pi_{ia} \geq 0,\ (i,b) \in S \times B;$

(5) $\pi_{ia} \geq 0,\ (i,a) \in S \times A;\ \sum_a \pi_{ia} = 1,\ i \in S;$

(6) $\rho_{ib} \geq 0,\ (i,b) \in S \times B;\ \sum_b \rho_{ib} = 1,\ i \in S.$

(7) $w_i(a),\ x_i(a) \geq 0,\ (i,a) \in S \times A;$

(8) $y_i(b),\ z_i(b) \geq 0,\ (i,b) \in S \times B.$

We can also formulate another, strongly related, nonlinear program in which the objective function is linear and the constraints are (partly) nonlinear. In this formulation we use different vectors for the superharmonicity (v^1, t^1) and the subharmonicity (v^2, t^2). It turns out that there are optimal stationary policies if and only if the program is feasible with optimum objective value 0.

Theorem 10.16

An undiscounted stochastic game has stationary optimal policies $(\pi^*)^\infty$ and $(\rho^*)^\infty$ for player 1 and 2, respectively, if and only if $(v^1, v^2, t^1, t^2, \pi = \pi^*, \rho = \rho^*)$ is an optimal solution of the nonlinear program

min $\{\sum_i (v_i^1 - v_i^2)\}$

subject to

(1) $v_i^1 \geq \sum_j \sum_b p_{ij}(a,b)\rho_{ib} v_j^1,\ (i,a) \in S \times A;$

(2) $v_i^1 + t_i^1 \geq \sum_b r_i(a,b)\rho_{ib} + \sum_j \sum_b p_{ij}(a,b)\rho_{ib} t_j^1,\ (i,a) \in S \times A;$

(3) $-v_j^2 \geq -\sum_i \sum_a p_{ij}(a,b)\pi_{ia} v_j^2,\ (i,b) \in S \times B;$

(4) $-v_j^2 - t_j^2 \geq -\sum_a r_i(a,b)\pi_{ia} - \sum_i \sum_a p_{ij}(a,b)\pi_{ia} t_j^2,\ (i,b) \in S \times B;$

(5) $\pi_{ia} \geq 0,\ (i,a) \in S \times A;\ \sum_a \pi_{ia} = 1,\ i \in S;$

(6) $\rho_{ib} \geq 0,\ (i,b) \in S \times B;\ \sum_b \rho_{ib} = 1,\ i \in S.$

with optimum value 0.

Proof

Assume that $(v^1, v^2, t^1, t^2, \pi = \pi^*, \rho = \rho^*)$ is an optimal solution of the nonlinear program with optimum value 0. Then, (v^1, t^1, ρ^*) and (v^2, t^2, π^*) are feasible solutions of (10.30) and (10.31), respectively. Then, for any $\pi^\infty \in \Pi$, $v^1 \geq P(\pi, \rho^*)v^1$ and $v^1 + t^1 \geq r(\pi, \rho^*) + P(\pi, \rho^*)t^1$, which implies $v^1 \geq \phi\big(\pi^\infty, (\rho^*)^\infty\big)$ for all $\pi^\infty \in \Pi$. Similarly, we derive $v^2 \leq \phi\big((\pi^*)^\infty, \rho^\infty\big)$ for all $\rho^\infty \in \Gamma$. So, $v^1 \geq \phi\big((\pi^*)^\infty, (\rho^*)^\infty\big) \geq v^2$, i.e. $v^1 - v^2 \geq 0$. Since $\sum_i (v_i^1 - v_i^2) = 0$, we obtain $v^1 = v^2$. Furthermore, $\phi\big(\pi^\infty, (\rho^*)^\infty\big) \leq v^1 = v^2 \leq \phi\big((\pi^*)^\infty, \rho^\infty\big)$ for all $\pi^\infty \in \Pi$ and $\rho^\infty \in \Gamma$, implying that the stochastic game has value vector $\phi = v^1 = v^2$ and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and 2, respectively.

Now assume that $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and 2. Then, $\phi\big((\pi^*)^\infty, R_2\big) \geq inf_{R_2}\, sup_{R_1}\, \phi(R_1, R_2) \geq sup_{R_1}\, inf_{R_2}\, \phi(R_1, R_2) \geq \phi\big(R_1, (\rho^*)^\infty\big)$ for all R_1 and R_2, implying $\phi\big((\pi^*)^\infty, R_2\big) \geq \phi = \phi\big((\pi^*)^\infty, (\rho^*)^\infty\big) \geq \phi\big(R_1, (\rho^*)^\infty\big)$ for all R_1 and R_2. Hence, $(\pi^*)^\infty$ is an optimal policy in the MDP induced by the stationary policy $(\rho^*)^\infty$. Consequently ϕ is the smallest superharmonic vector in the sense of an undiscounted MDP problem, i.e. $(v^1 = \phi, t^1, \rho = \rho^*)$ is a feasible solution of (10.30) for some t^1.
Similarly, it follows that $(\rho^*)^\infty$ is an optimal policy in the MDP induced by the stationary policy $(\pi^*)^\infty$ with respect to minimizing the average rewards. Therefore, ϕ is the largest subharmonic vector in the sense of an undiscounted MDP problem, i.e. $(v^2 = \phi, t^2, \pi = \pi^*)$ is a feasible solution of (10.31) for some t^2. Hence, $(v^1 = \phi, v^2 = \phi, t^1, t^2, \pi = \pi^*, \rho = \rho^*)$ is an optimal solution of the nonlinear program with optimum value 0. □

10.3.4 Perfect information and irreducible games

Perfect information

We have seen in Corollary 10.2 that a discounted stochastic game with perfect information has optimal deterministic policies. For undiscounted stochastic games we have the same result, but the proof is more complicated.

Theorem 10.17
In an undiscounted stochastic game with perfect information, both players possess optimal deterministic policies.

Proof
For any $\alpha \in (0,1)$ there exists deterministic stationary policies f_α^∞ and g_α^∞ such that

$$v^\alpha(f^\infty, g_\alpha^\infty) \leq v^\alpha(f_\alpha^\infty, g_\alpha^\infty) \leq v^\alpha(f_\alpha^\infty, g^\infty) \text{ for all } f^\infty \in F \text{ and } g^\infty \in G,$$

where F and G are the sets of deterministic stationary policies for player 1 and 2, respectively. Since $F \times G$ is a finite set, we can therefore find a pair $f_*^\infty \in F$ and $g_*^\infty \in G$ and a sequence $\{\alpha_n\}_{n=1}^\infty$ such that

$$v_n^\alpha(f^\infty, g_*^\infty) \leq v_n^\alpha(f_*^\infty, g_*^\infty) \leq v_n^\alpha(f_*^\infty, g^\infty) \text{ for all } f^\infty \in F,\ g^\infty \in G \text{ and } n = 1, 2, \ldots,$$

For any $f^\infty \in F,\ g^\infty \in G$, the vector $v^\alpha(f^\infty, g^\infty)$ is the unique solution of the linear system $x = r(f,g) + \alpha P(f,g)x$. Since this linear system can be solved by Cramer's rule, the numbers $v_i^\alpha(f^\infty, g^\infty),\ i \in S$, are rational functions in α. Therefore, also - for all $i \in S$ - the functions $h_i^1(\alpha) = v_i^\alpha(f^\infty, g_*^\infty) - v_i^\alpha(f_*^\infty, g_*^\infty)$ and $h_i^2(\alpha) = v_i^\alpha(f_*^\infty, g^\infty) - v_i^\alpha(f_*^\infty, g_*^\infty)$ are rational in α for all $f^\infty \in F,\ g^\infty \in G$ and $i \in S$. Hence, for $k = 1, 2$ and all $i \in S$, either $h_i^k(\alpha) \equiv 0$ or $h_i^k(\alpha)$ has a finite number of zero's in $(0,1)$. In the last case let $\alpha_* \in (0,1)$ be the largest zero of the finite number of functions $h_i^k(\alpha),\ k = 1, 2,\ i \in S$. With this α_*, we have for all $\alpha \geq \alpha_*$:

$$v^\alpha(f^\infty, g_*^\infty) \leq v^\alpha(f_*^\infty, g_*^\infty) \leq v^\alpha(f_*^\infty, g^\infty) \text{ for all } f^\infty \in F,\ g^\infty \in G \text{ and all } \alpha \geq \alpha_*,$$

and consequently

$(1-\alpha)v^{\alpha}(f^{\infty},g_{*}^{\infty}) \leq (1-\alpha)v^{\alpha}(f_{*}^{\infty},g_{*}^{\infty}) \leq (1-\alpha)v^{\alpha}(f_{*}^{\infty},g^{\infty})$ for all $f^{\infty} \in F,\ g^{\infty} \in G,\ \alpha \geq \alpha_{*}$.

Then, using the Laurent series expansion, which implies $\lim_{\alpha\to\infty}(1-\alpha)v^{\alpha}(f^{\infty},g^{\infty}) = \phi(f^{\infty},g^{\infty})$ for all $f^{\infty} \in F,\ g^{\infty} \in G$, we obtain

$$\phi(f^{\infty},g_{*}^{\infty}) \leq \phi(f_{*}^{\infty},g_{*}^{\infty}) \leq \phi(f_{*}^{\infty},g^{\infty}) \text{ for all } f^{\infty} \in F \text{ and } g^{\infty} \in G,$$

also implying (by MDP) $\phi(R_1,g_{*}^{\infty}) \leq \phi(f_{*}^{\infty},g_{*}^{\infty}) \leq \phi(f_{*}^{\infty},R_2)$ for all policies R_1 and R_2, i.e. f_{*}^{∞} and $g^{\infty} \in G$ are optimal deterministic policies. □

Remark

Like discounted stochastic games, it is for undiscounted stochastic games with perfect information also an open problem to find an efficient finite algorithm.

Irreducible games

The class of *irreducible stochastic games* are characterized by the propery that for each pair of stationary policies, say $(\pi^{\infty},\rho^{\infty})$, the Markov chain $P(\pi,\rho)$ is an irreducible Markov chain.

We first show a lemma on the relation between the linear program (6.3) and the optimality equation (6.1).

Lemma 10.9

Any optimal solution (x^*,y^*) of the linear program (6.3) is a solution of the optimality equation (6.1), i.e. $x^* + y_i^* = max_{a\in A(i)}\{r_i(a) + \sum_j p_{ij}(a)y_j^*\},\ i \in S$.

Proof

From the constraints of the linear program (6.3) it follows that

$$x^* + y_i^* \geq max_{a\in A(i)}\{r_i(a) + \textstyle\sum_j p_{ij}(a)y_j^*\},\ i \in S.$$

From the theory of irreducible MDPs we know that any feasible solution of the set

$$\begin{cases} \sum_{i,a}\{\delta_{ij} - p_{ij}(a,\rho)\}x_i(a) = 0, j \in S \\ \sum_{i,a} x_i(a) = 1;\ x_i(a) \geq 0,\ i \in S,\ a \in A(i) \end{cases} \quad \text{has } \textstyle\sum_a x_j(a) > 0 \text{ for all } j \in S.$$

The complementary slackness property of linear programming implies that any $(i,a) \in S \times A$ with $x_i^*(a) > 0$, satisfies $x^* + \sum_j \{\delta_{ij} - p_{ij}(a)\}y_j^* = r_i(a)$. Hence,

$$x^* + y_i^* = max_{a\in A(i)}\{r_i(a) + \textstyle\sum_j p_{ij}(a)y_j^*\},\ i \in S.$$ □

Suppose that player 2 plays a fixed stationary policy ρ^{∞}. Then, the game becomes an MDP for player 1 and let $\phi(\rho) = max_{R_1}\,\phi(R_1,\rho^{\infty})$. Because of the property of irreducibility, $\phi(\rho)$ has identical components. So, we may view $\phi(\rho)$ as a real function of ρ. We will first show that $\phi(\rho)$

is a continuous function of ρ. Therefore, we consider the following set of linear (in)equalities, which are a combination of the linear programs (6.3) and (6.4).

$$\begin{cases} z + \sum_j \{\delta_{ij} - p_{ij}(a,\rho)\} y_j & \geq & r_i(a,\rho), & i \in S,\ a \in A(i) \\ \sum_{i,a} \{\delta_{ij} - p_{ij}(a,\rho)\} x_i(a) & = & 0, & j \in S \\ \sum_{i,a} x_i(a) & = & 1 & \\ \sum_{i,a} r_i(a,\rho) x_i(a) - z & \geq & 0 & \\ x_i(a) & \geq & 0, & i \in S,\ a \in A(i) \\ y_1 & = & 0 & \end{cases} \tag{10.34}$$

Since, without $y_1 = 0$, for any solution (z, y, x) also $(z, y + c \cdot e, x)$ is a solution, this additional constraint may be imposed. From the theory of linear programming we know that for any pair of feasible solutions (z, y) and x of (6.3) and (6.4), respectively, the value of the objective function z is at least the value of the objective function $\sum_{i,a} r_i(a) x_i(a)$. Hence, the inequality $\sum_{i,a} r_i(a,\rho) x_i(a) - z \geq 0$ implies that only optimal solutions are feasible for (10.34) and that $z = \phi(\rho)$.

Consider a sequence of $\rho^n,\ n = 1, 2, \ldots$, and the corresponding feasible solutions (z^n, y^n, x^n) of (10.34). Then, these elements are bounded, namely:

(1) $\rho^n_{ib} \geq 0$ for all $(i, b) \in S \times B$ and $\sum_b \rho^n_{ib} = 1$ for all $i \in S$: the set $\{\rho^n\}_{n=1}^{\infty}$ is bounded.

(2) $z^n = \phi(\rho^n)$, which is bounded because $\phi(\rho^n) \leq max_{(i,a,b)} |r_i(a, b)|$: the set $\{z^n\}_{n=1}^{\infty}$ is bounded.

(3) From Lemma 10.9 we obtain $z^n + y^n_i = max_{a \in A(i)} \left\{ r_i(a\rho^n) + \sum_j p_{ij}(a, \rho^n) y^n_j \right\},\ i \in S$. Then, with $y^n_1 = 0$, Theorem 6.1 yields that $y^n = u^0\big(f_0(\rho^n)\big) - u^0_1\big(f_0(\rho^n)\big) \cdot e$, where $\big(f_0(\rho^n)\big)$ is a Blackwell optimal deterministic stationary policy in the MDP induced by ρ^n. Since there are only a finite number of deterministic stationary policies there are only a finite number of different y^n: the set $\{y^n\}_{n=1}^{\infty}$ is bounded.

(4) $x_i(a) \geq 0$ for all $(i, a) \in S \times A$ and $\sum_{i,a} x_i(a) = 1$: the set $\{x^n\}_{n=1}^{\infty}$ is bounded.

Consider a limit point (ρ^*, z^*, y^*, x^*) of the sequence $\{(\rho^n, z^n, y^n, x^n)\}_{n=1}^{\infty}$. For convenience, let $(\rho^*, z^*, y^*, x^*) = lim_{n\to\infty} (\rho^n, z^n, y^n, x^n)$. For $z^* = lim_{n\to\infty} z^n = lim_{n\to\infty} \phi(\rho^n)$ we have to show $z^* = \phi(\rho^*)$, i.e. $z^* \geq \phi\big(\pi^\infty, (\rho^*)^\infty\big)$ for all $\pi^\infty \in \Pi$ and $z^* = \phi\big((\pi^*)^\infty, (\rho^*)^\infty\big)$ for some $(\pi^*)^\infty \in \Pi$. From the first set of the constraints of (10.34), we obtain $z^* \cdot e + \{I - P(\pi, \rho^*\} y^* \geq r(\pi, \rho^*)$, implying, by multiplication with $P^*(\pi, \rho^*)$, that $z^* \geq \phi\big(\pi^\infty, (\rho^*)^\infty\big)$ for all $\pi^\infty \in \Pi$.

Let $(\pi^n)^\infty$ be the stationary policy that corresponds to x^n (see Theorem 6.5). The linear function $\sum_{i,a} r_i(a, \rho^n) x^n_i(a) = \phi\big((\pi^n)^\infty, (\rho^n)^\infty\big) \ \to\ \sum_{i,a} r_i(a, \rho^*) x^*_i(a) = \phi\big((\pi^*)^\infty, (\rho^*)^\infty\big)$, where $(\pi^*)^\infty$ is the stationary policy that corresponds to x^*. From the fourth constraint of (10.34) it follows that $\phi\big((\pi^*)^\infty, (\rho^*)^\infty\big) \geq z^*$. Hence, we have shown that $\phi\big((\pi^*)^\infty, (\rho^*)^\infty\big) \geq z^* \geq \phi\big(\pi^\infty, (\rho^*)^\infty\big)$ for all $\pi^\infty \in \Pi$, i.e. $z^* = \phi(\rho^*)$, completing the proof that $\phi(\rho)$ is a continuous function of ρ.

The function $\phi(\rho)$ is continuous on the compact set Γ of all stationary policies. Therefore, there exists a stationary policy, say $(\rho^*)^\infty \in \Gamma$ such that $\phi(\rho^*) = min_{\rho^\infty \in \Gamma}\, max_{R_1}\, \phi(R_1, \rho^\infty)$. We will show that $(\rho^*)^\infty$ is an optimal policy for player 2. Therefore, we consider the associate MDP with rewards $r_i(a, \rho^*),\ (i, a) \in S \times A$ and transition probabilities $p_{ij}(a, \rho^*),\ j \in S,\ (i, a) \in S \times A$. For this model, let (x^*, y^*) be an optimal solution of the linear program (6.3).

Let $M_x[i]$ be a payoff matrix with $m = \#A(i)$ rows and $n = \#B(i)$ columns and with payoff $r_i(a, b) + \sum_j p_{ij}(a, b)x_j$, if player 1 chooses row a and player 2 column b.

Theorem 10.18

Let (x^*, y^*) be an optimal solution of the linear program (6.3) associated with policy $(\rho^*)^\infty$ for player 2. Then, $x^* + y_i^* = val\big(M_{y^*}[i]\big)$ for all $i \in S$.

Proof

We have to show that $max_a\, min_b \{r_i(a, b) + \sum_j p_{ij}(a, b)y_j^*\} = x + y_i^*,\ i \in S$. From Lemma 10.9 it follows that

$$x^* + y_i^* = max_{a \in A(i)} \{r_i(a, \rho^*) + \sum_j p_{ij}(a, \rho^*)y_j^*\},\ i \in S, \tag{10.35}$$

implying

$$max_a\, min_b \{r_i(a, b) + \textstyle\sum_j p_{ij}(a, b)y_j^*\} \le max_a \{r_i(a, \rho^*) + \textstyle\sum_j p_{ij}(a, \rho^*)y_j^*\} = x + y_i^*,\ i \in S.$$

Finally, we have to show that $max_a\, min_b \{r_i(a, b) + \sum_j p_{ij}(a, b)y_j^*\} \ge x + y_i^*,\ i \in S$.

Suppose the contrary, i.e. there is a state $k \in S$ and a mixed strategy $\{\rho_{kb},\ b \in B(k)\}$ such that

$$x^* + y_k^* > max_{a \in A(k)} \{r_k(a, \rho) + \sum_j p_{kj}(a, \rho)y_j^*\}. \tag{10.36}$$

Consider the policy $\overline{\rho}^\infty$ defined by $\overline{\rho}_{ib} = \begin{cases} \rho_{ib}^* & \text{if } i \ne k,\ b \in B(i); \\ \rho_{kb} & \text{if } i = k,\ b \in B(k). \end{cases}$

Then, $\phi(\overline{\rho})$ is the optimum value of the linear program associated with policy $\overline{\rho}^\infty$. Since (x^*, y^*) is also feasible for this linear program (because of (10.35) and (10.36)) and Lemma 10.9 is not satisfied (because of (10.19)), we obtain $\phi(\overline{\rho}) < \phi(\rho^*)$. However, this contradicts the property of ρ^*, namely that $\phi(\rho^*) = min_{\rho^\infty \in \Gamma}\, \phi(\rho)$. □

Theorem 10.19

If $x + y_i = val(M_y[i]),\ i \in S$ and $x^* + y_i^* = val(M_{y^*}[i]),\ i \in S$, then $x = x^*$ and $y = y^* + c \cdot e$ for some scalar c, i.e. x is unique and y is unique up to an additional constant.

Proof

Let $\{\pi_{ia}^*,\ a \in A(i)\}$ be an optimal mixed strategy for player 1 in the matrix game $M_y^*[i]$, and let $\{\rho_{ib},\ b \in B(i)\}$ be an optimal mixed strategy for player 2 in the matrix game $M_y[i]$. Therefore,

$$val(M_y[i]) = x + y_i \ge r_i(\pi^*, \rho) + \textstyle\sum_j p_{ij}(\pi^*, \rho)y_j.$$

and

$$val(M_{y^*}[i]) = x^* + y_i^* \leq r_i(\pi^*, \rho) + \sum_j p_{ij}(\pi^*, \rho) y_j^*.$$

Subtracking the second inequelity from the first one obtains

$$(x - x^*) \cdot e + (y - y^*) \geq P(\pi^*, \rho)(y - y^*).$$

Multiplying this equation by $P^*(\pi^*, \rho)$ yields $x \geq x^*$. Interchanging the roles of the solutions (x, y) and (x^*, y^*), we may establish similarly that $x^* \geq x$. Therefore, $x = x^*$, and, setting $x - x^* = 0$ and $z = y - y^*$, we have $z - P(\pi^*, \rho)z \geq 0$. Since $P^*(\pi^*, \rho)\{z - P(\pi^*, \rho)z\} = 0$ and $P^*(\pi^*, \rho)$ is a matrix with strictly positive elements, we obtain $z = P(\pi^*, \rho)$, implying $z = P^*(\pi^*, \rho)z$. Because the matrix $P^*(\pi^*, \rho)$ has identical rows, all components of $z = y - y^*$ are equal. Hence, $y = y^* + c \cdot e$ for some scalar c. □

Corollary 10.6

The equation $x + y_i = val(M_y[i]),\ i \in S$, has a solution (x^*, y^*) in which x^* is unique. Furthermore, x^* is the value of the stochastic game and optimal strategies in the matrix games $M_{y^*}[i],\ i \in S$, are optimal stationary policies for the stochastic game.

Proof

From the Theorems 10.18 and 10.19 it follows that the equation $x + y_i = val(M_y[i]),\ i \in S$, has a solution (x^*, y^*) in which x^* is unique. Let $\{\pi_{ia}^*,\ a \in A(i)\}$ and $\{\rho_{ia}^*,\ b \in B(i)\}$ be optimal mixed strategies for player 1 and 2, respectively, in the matrix game $M_y^*[i],\ i \in S$. Then,

$$r(\pi, \rho^*) + P(\pi, \rho^*)y^* \leq x^* \cdot e + y^* \leq r(\pi^*, \rho) + P(\pi^*, \rho)y^* \text{ for all } \pi^\infty \in F \text{ and } \rho^\infty \in \Gamma.$$

Hence, by multiplying the first inequality by $P^*(\pi, \rho^*)$ we obtain $\phi\big(\pi^\infty, (\rho^*)^\infty\big) \leq x^*$. Similarly, by multiplying the second inequality by $P^*(\pi^*, \rho)$ we obtain $x^* \leq \phi\big((\pi^*)^\infty, \rho^\infty\big)$. Therefore, $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies and x^* is the value. □

Algorithm 10.9 *Value iteration for undiscounted games (irreducible case)*

1. Let $t = 0$ and choose a stationary policy $(\rho^t)^\infty$ for player 2.

2. Solve the MDP induced by policy $(\rho^t)^\infty$: compute (x^t, y^t) such that $y_1^t = 0$ and $x^t + y_i^t = max_a \{r_i(a, \rho^t) + \sum_j p_{ij}(a, \rho^t) y_j^t\},\ i \in S$.

3. Compute for each $i \in S$ in the matrix game $M_{y^t}[i]$, where $M_{y^t}[i]$ is the matrix with entries $r_i(a, b) + \sum_j p_{ij}(a, b) y_j^t,\ a \in A(i),\ b \in B(i)$, stationary strategies $\pi_{ia}^{t+1},\ a \in A(i)$, for player 1, and $\rho_{ib}^{t+1},\ b \in B(i)$, for player 2.

4. If $val\big(M_{y^t}[i]\big) = x^t + y_i^t,\ i \in S$, x^t is the value and $(\pi^{t+1})^\infty$ and $(\rho^{t+1})^\infty$ are optimal stationary policies for player 1 and 2, respectively (STOP).

   Otherwise: $t := t + 1$ and return to step 2.

Remark

Step 2 of this algorithm can be solved by the linear program (6.3). We will show that the sequence $\{x^t,\ t = 0, 1, \dots\}$ converges to the value of the stochastic game.

Theorem 10.20

(1) The sequences $\{x^t,\ t = 0, 1, \dots\}$ and $\{y^t,\ t = 0, 1, \dots\}$ are convergent.

(2) Let $x^* = lim_{t\to\infty} x^t$ and $y^* = lim_{t\to\infty} y^t$. Then $x^* + y_i^* = val(M_{y^*}[i]),\ i \in S$.

(3) If $x^t + y_i^t = val(M_{y^t}[i]),\ i \in S$, then x^t is the value and $(\pi^{t+1})^\infty$ and $(\rho^{t+1})^\infty$ are optimal stationary policies for player 1 and 2, respectively.

Proof

Since π^{t+1} and ρ^{t+1} are optimal strategies in the matrix games $M_{y^t}[i],\ i \in S$, we have

$$val(M_{y^t}) \geq r(\pi, \rho^{t+1}) + P(\pi, \rho^{t+1})y^t \text{ for all } \pi^\infty \in \Pi,$$

and

$$val(M_{y^t}) \leq r(\pi^{t+1}, \rho) + P(\pi^{t+1}, \rho)y^t \text{ for all } \rho^\infty \in \Gamma,$$

implying $val(M_{y^t}) \leq r(\pi^{t+1}, \rho^t) + P(\pi^{t+1}, \rho^t)y^t$. By step 2 of the algorithm we have

$$y^t = r(\pi^{t+1}, \rho^t) + P(\pi^{t+1}, \rho^t)y^t - x^t \cdot e \geq val(M_{y^t}) - x^t \cdot e.$$

Therefore,

$$val(M_{y^t}) \geq r(\pi, \rho^{t+1}) + P(\pi, \rho^{t+1})\{val(M_{y^t}) - x^t \cdot e\} \text{ for all } \pi^\infty \in \Pi.$$

Multiplication with $P^*(\pi, \rho^{t+1})$ gives

$$x^t \geq \phi\big(\pi^\infty, (\rho^{t+1})^\infty\big) \text{ for all } \pi^\infty \in \Pi, \text{ i.e. } x^t \geq max_{\pi^\infty\in\Pi}\, \phi\big(\pi^\infty, (\rho^{t+1})^\infty\big) = x^{t+1}.$$

Hence, the sequence $\{x^t,\ t = 0, 1, \dots\}$ is nonincreasing and bounded below by $-max|r_i(a, b)|$: $\{x^t,\ t = 0, 1, \dots\}$ is convergent. In the proof that $\phi(\rho)$, defined as $\phi(\rho) = max_{R_1}\, \phi(R_1, \rho^\infty)$, is a continuous function of ρ, we have seen that $\{y^t\}_{t=1}^\infty$ is a bounded sequence. Therefore, we may choose a convergent subsequence of vectors $\{(x^t, y^t)\}_{t=1}^\infty$, and let us denote the vector to which they converge by (x^+, y^+). Let the corresponding stationary policies ρ^t for player 2 converge to ρ^+. Since ρ^{t+1} is an optimal solution for player 2 in the matrix games $M_{y^t}[i],\ i \in S$, it follows by continuity that ρ^+ is an optimal policy in the matrix games $M_{y^+}[i],\ i \in S$.

Since $x^t + y_i^t \geq val\big(M_{y^t}\big)[i]$, it follows - also by continuity - that $x^+ + y_i^+ \geq val\big(M_{y^+}\big)[i]$. If, for some k, $x^+ + y_k^+ > val\big(M_{y^+}\big)[k]$, then this implies $x^+ + y_k^+ > max_a\, \{r_i(a, \rho^+) + \sum_j\, p_{kj}(a, \rho^+)y_j^+\}$. Similarly as in the proof of Theorem 10.18 we obtain $\phi(\rho^+) < x^+$. But $x^+ \leq x^t = \phi(\rho^t)$ and the continuity of $\phi(\rho)$ imply that $x^+ \leq \phi(\rho^+)$, which yields a contradiction. This contradiction establishes that $x^+ + y_i^+ = val\big(M_{y^+}\big)[i],\ i \in S$. Because the solution of this functional equation is unique, every convergent subsequence has the same limit, and it follows that the seqeuence produced in the algorithm converges to this functional equation. Part (3) follows directly from Corollary 10.6.

□

10.3.5 Finite methods

When the value and the optimal policies lie in the same ordered field as the data one can hope to arrive at a solution by a finite number of operations. If the ordered field property is not valid then one can only try iterative procedures for solving these stochastic games. As is the case for discounted stochastic games, also in undiscounted stochastic games the ordered field property does not hold, in general. This is illustrated by the following example.

Example 10.6 (continued)

In Example 10.6 we have derived that $v_1^\alpha = \frac{-(1+\alpha)+\sqrt{(1+\alpha)}}{1-\alpha}$. It can be shown [2] that ϕ, the value vector of the undiscounted game, satisfies $\phi = lim_{\alpha\uparrow 1}\,(1-\alpha)v^\alpha$, where v^α is the value vector of the α-discounted stochastic game. Hence, $\phi_1 = lim_{\alpha\uparrow 1}\,\{-(1+\alpha)+\sqrt{1+\alpha}\} = -2+\sqrt{2}$, which lies not in the ordered field of the rational numbers.

We consider the following special games, which have the ordered field property, as we will show:

(1) The single-controller stochastic game.

(2) The switching-controller stochastic game.

(3) The separabel reward - state independent transitions (SER-SIT) stochastic game.

(4) The additive reward - additive transitions (ARAT) stochastic game.

Single-controller stochastic game

In the single-controller stochastic game, where player 1 is the 'single-controller', the transition probabilities $p_{ij}(a,b)$ are independent of b denoted by $p_{ij}(a)$. Under this assumption the concept of superharmonicity for a vector $v \in \mathbb{R}^N$ means that there exists a vector $t \in \mathbb{R}^N$ and a policy $\rho^\infty \in \Gamma$ such that the triple (v,t,ρ) satisfies

$$\begin{cases} v_i & \geq & \sum_j p_{ij}(a)v_j & \text{for every } (i,a) \in S \times A; \\ v_i + t_i & \geq & r_i(a,\rho) + \sum_j p_{ij}(a)t_j & \text{for every } (i,a) \in S \times A. \end{cases} \tag{10.37}$$

Therefore, the problem to find the smallest superharmonic vector is the following linear program

$$min \left\{ \sum_i v_i \;\middle|\; \begin{array}{rrrrrrl} \sum_j\{\delta_{ij} - p_{ij}(a)\}v_j & & & & & \geq 0, & a \in A(i),\ i \in S \\ v_i & + & \sum_j\{\delta_{ij} - p_{ij}(a)\}t_j & - & \sum_b r_i(a,b)\rho_{ib} & \geq 0, & a \in A(i),\ i \in S \\ & & & & \sum_b \rho_{ib} & = 1, & i \in S \\ & & & & \rho_{ib} & \geq 0, & b \in B(i),\ i \in S \end{array} \right\}. \tag{10.38}$$

[2] see Corollary 5.2.7 in [76]

The dual program is

$$max\left\{\sum_i z_i \;\middle|\; \begin{array}{rcrcl} \sum_{(i,a)}\{\delta_{ij}-p_{ij}(a)\}x_i(a) & & & = & 0,\ j\in S \\ \sum_a x_j(a) & + & \sum_{(i,a)}\{\delta_{ij}-p_{ij}(a)\}y_i(a) & = & 1,\ j\in S \\ -\sum_a r_i(a,b)x_i(a) & + & z_i & \le & 0,\ (i,b)\in S\times B \\ & & x_i(a),y_i(a) & \ge & 0,\ (i,a)\in S\times A \end{array}\right\}. \tag{10.39}$$

Lemma 10.10

The linear programs (10.38) and (10.39) have finite optimal solutions.

Proof

Take an arbitrary stationary policy ρ^∞ for player 2, and let $t=0$ and $v_i = max_{i,a,b}\, r_i(a,b),\ i\in S$. Then, (v,t,ρ) is a feasible solution of (10.38). For the existence of finite optimal solutions it is sufficient to show, by the duality theorem of linear programming, that the optimum of (10.38) is bounded below. Let (v,t,ρ) be any feasible solution of (10.38). Then, for any $\pi^\infty\in F$, we have from the first equations of (10.38) $v\ge P(\pi)v$, implying $v\ge P^*(\pi)v$. From the second set of equation, we obtain $v+\{I-P(\pi)\}t\ge r(\pi,\rho)$. Therefore, we can write,

$$v\ge P^*(\pi)v\ge P^*(\pi)\big\{\{r(\pi,\rho)-\{I-P(\pi)\}t\big\}=P^*(\pi)\{r(\pi,\rho)=\phi(\pi^\infty,\rho^\infty).$$

Now we have $v_i\ge\phi_i(\pi^\infty,\rho^\infty)\ge min_{i,a,b}\, r_i(a,b),\ i\in S$, which shows that the optimum of (10.38) is bounded below. □

The following theorem shows that the value vector and optimal stationary policies for both players can be obtained from the optimal solutions of the dual pair of linear programs.

Theorem 10.21

Let (v^*,t^*,ρ^*) and (x^*,y^*,z^*) be optimal solutions of the linear programs (10.38) and (10.39). Define the policy $(\pi^*)^\infty$ by $\pi^*_{ia}=\begin{cases}\frac{x_i^*(a)}{\sum_a x_i^*(a)},\ i\in S_{x^*},\ a\in A(i),\ \text{where } S_{x^*}=\{i\mid\sum_a x_i^*(a)>0\};\\ \frac{y_i^*(a)}{\sum_a y_i^*(a)},\ i\notin S_{x^*},\ a\in A(i).\end{cases}$

Then, v^* is the value vector and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and 2, respectively.

Proof

The constraints of program (10.39) imply $\sum_a x_j^*(a)+\sum_a y_j^*(a)=1+\sum_{(i,a)} p_{ij}(a)y_i^*(a)>0,\ j\in S$. Hence, the policy $(\pi^*)^\infty$ is well-defined. From the constaints of program (10.38) we obtain

$$v^*\ge P(\pi)v^* \text{ and } v^*\ge r(\pi,\rho^*)-\{I-P(\pi)\}t^* \text{ for all } \pi^\infty\in\Pi.$$

Therefore, we obtain

$$v^*\ge P^*(\pi)v^*\ge P^*(\pi)\{r(\pi,\rho^*)-\{I-P(\pi)\}t^*\}=\phi\big(\pi^\infty,(\rho^*)^\infty\big) \text{ for all } \pi^\infty\in\Pi. \tag{10.40}$$

Since $\pi^*_{ia}>0$ if and only if $\begin{cases}x_i^*(a)>0 \text{ for } i\in S_{x^*}\\ y_i^*(a)>0 \text{ for } i\notin S_{x^*}\end{cases}$ it follows from the complementary

slackness property of linear programming that

$$\begin{cases} \sum_a \pi_i^*(a) \cdot \{v_i^* + \sum_j \{\delta_{ij} - p_{ij}(a)\} t_j^* - \sum_b r_i(a,b)\rho_{ib}^*\} &= 0, \ i \in S_{x^*} \\ \sum_a \pi_i^*(a) \cdot \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j^* &= 0, \ i \notin S_{x^*} \end{cases}$$

Suppose that $\pi_k^*(a_k) \cdot \sum_j \{\delta_{kj} - p_{kj}(a_k)\} v_j^* \neq 0$ for some $k \in S_{x^*}, \ a_k \in A(k)$. Then, the definition of π^* and the constraints of (10.38) imply that $x_k^*(a_k) \cdot \sum_j \{\delta_{kj} - p_{kj}(a_k)\} v_j^* > 0$. Hence, we get $\sum_{(i,a)} x_i^*(a) \cdot \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j^* > 0$, which is contradictory to

$$\sum_{(i,a)} x_i^*(a) \cdot \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j^* = \sum_j \left\{ \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i^*(a) \right\} v_j^* = 0.$$

Therefore, we have

$$\begin{cases} \sum_a \pi_i^*(a) \cdot \{v_i^* + \sum_j \{\delta_{ij} - p_{ij}(a)\} t_j^* - \sum_b r_i(a,b)\rho_{ib}^*\} &= 0, \ i \in S_{x^*}; \\ \sum_a \pi_i^*(a) \cdot \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j^* &= 0, \ i \in S. \end{cases}$$

Hence,

$$\begin{cases} v_i^* + \{\{I - P(\pi^*)\} t^*\}_i &= r_i(\pi^*, \rho^*), \quad i \in S_{x^*}; \\ \{\{I - P(\pi^*)\} v^*\}_i &= 0, \quad i \in S. \end{cases}$$

The second equation implies $v^* = P^*(\pi^*)v^*$. Since S_{x^*} is the set of recurrent states in the Markov chain induced by $P(\pi^*)$ (see the proof of Theorem 5.18), we obtain

$$v^* = P^*(\pi^*)v^* = P^*(\pi^*)\big\{r(\pi^*, \rho^*) - \{I - P(\pi^*)\}t^*\big\} = P^*(\pi^*)r(\pi^*, \rho^*) = \phi\big((\pi^*)^\infty, (\rho^*)^\infty\big),$$

implying, using (10.40),

$$\phi\big(\pi^\infty, (\rho^*)^\infty\big) \leq v^* = \phi\big((\pi^*)^\infty, (\rho^*)^\infty\big) \text{ for all } \pi^\infty \in \Pi. \tag{10.41}$$

Let $x_i^* = \sum_a x_i^*(a), \ i \in S$. Suppose that $S_1, S_2, \ldots, S_m$ are the ergodic sets and let T be the set of transient states in the Markov chain induced by $P(\pi^*)$. Let $n_k = |S_k|, \ k = 1, 2, \ldots, m$. Then, we shall show that $x^* = \{P^*(\pi^*)\}^T \gamma$, where γ is a strictly positive vector with elements

$$\gamma_l = \begin{cases} \frac{1}{n} & l \in T; \\ \frac{1}{n_k} \cdot \sum_{j \in S_k} \{x_j^* - \frac{1}{n} \sum_{i \in T} p_{ij}^*(\pi^*)\} & l \in S_k, \ k = 1, 2, \ldots, m. \end{cases}, \text{ where } n \text{ is sufficienlty large}$$

such that $\gamma_l > 0$, i.e. $n > max_{j \in S_{x^*}} \left\{ \frac{1}{x_j^*} \cdot \sum_{i \in T} p_{ij}^*(\pi^*) \right\}$. Now, we have

$$\begin{aligned} \sum_l \gamma_l \cdot \sum_{j \in S_k} p_{lj}^*(\pi^*) &= \sum_{l \in T} \gamma_l \cdot \sum_{j \in S_k} p_{lj}^*(\pi^*) + \sum_{l \in S_k} \gamma_l \cdot \sum_{j \in S_k} p_{lj}^*(\pi^*) \\ &= \frac{1}{n} \sum_{l \in T} \sum_{j \in S_k} p_{lj}^*(\pi^*) + \sum_{l \in S_k} \gamma_l \\ &= \frac{1}{n} \sum_{l \in T} \sum_{j \in S_k} p_{lj}^*(\pi^*) + \sum_{l \in S_k} \left\{ \frac{1}{n_k} \cdot \sum_{j \in S_k} \{x_j^* - \frac{1}{n} \sum_{i \in T} p_{ij}^*(\pi^*)\} \right\} \\ &= \frac{1}{n} \sum_{l \in T} \sum_{j \in S_k} p_{lj}^*(\pi^*) + \sum_{j \in S_k} \{x_j^* - \frac{1}{n} \sum_{i \in T} p_{ij}^*(\pi^*)\} \\ &= \sum_{j \in S_k} x_j^*, \ k = 1, 2, \ldots, m. \end{aligned}$$

From program (10.39) and the definition of π^* it follows that $x^* = \{P(\pi^*)\}^T x^*$ and, consequently, $x^* = \{P^*(\pi^*)\}^T x^*$. Since $S \backslash S_{x^*}$ is the set of transient states T in the Markov chain induced by $P(\pi^*)$ (see the proof of Theorem 5.18), we have $p_{li}^* = 0, \ l \in S$. Therefore, we obtain

$$0 = x_i^* = \sum_l p_{li}^*(\pi^*)\gamma_l = \left\{\{P^*(\pi^*)\}^T \gamma\right\}_i, \ i \notin S_{x^*}. \tag{10.42}$$

For $i \in S_k$, it follows that

$$\begin{aligned} x_i^* &= \textstyle\sum_j p_{ji}^*(\pi^*)x_j^* = \sum_{j\in S_k} p_{ji}^*(\pi^*)x_j^* + \sum_{j\in T} p_{ji}^*(\pi^*)x_j^* \\ &= p_{ii}^*(\pi^*)\cdot\textstyle\sum_{j\in S_k} x_j^* + \sum_{j\in S\setminus S_{x^*}} p_{ji}^*(\pi^*)x_j^* \;=\; p_{ii}^*(\pi^*)\cdot\big\{\sum_l \gamma_l\cdot\sum_{j\in S_k} p_{lj}^*(\pi^*)\big\} + 0 \\ &= \textstyle\sum_l \gamma_l\cdot\sum_{j\in S_k} p_{lj}^*(\pi^*)p_{ji}^*(\pi^*) \;=\; \sum_l \gamma_l\cdot p_{li}^*(\pi^*), \end{aligned}$$

implying

$$x_i^* = \big\{\{P^*(\pi^*)\}^T\gamma\big\}_i,\ i\in S_k,\ k=1,2,\ldots,m. \tag{10.43}$$

Combining (10.42) and (10.42) yields $x^* = \{P^*(\pi^*)\}^T\gamma$.
Using agian the complementary slackness property of linear programming yields

$$\textstyle\sum_i \sum_b \rho_{ib}\cdot\{z_i^* - \sum_a r_i(a,b)x_i^*(a)\} = 0.$$

Therefore,

$$\begin{aligned} \textstyle\sum_i z_i^* &= \textstyle\sum_i\sum_b\sum_a r_i(a,b)\rho_{ib}^*x_i^*(a) \;=\; \sum_i\big\{\sum_b\sum_a r_i(a,b)\rho_{ib}^*\pi_{ia}^*\cdot x_i^*\big\} \\ &= \textstyle\sum_i\big\{r_i(\pi^*,\rho^*)\cdot x_i^*\big\} \;=\; \sum_i\big\{r_i(\pi^*,\rho^*)\cdot\sum_l\gamma_l\cdot p_{li}^*(\pi^*)\big\} \\ &= \textstyle\sum_l\gamma_l\cdot\big\{\sum_i p_{li}^*(\pi^*)r_i(\pi^*,\rho^*)\big\}, \end{aligned}$$

implying

$$\sum_i z_i^* = \gamma^T\phi\big((\pi^*)^\infty,(\rho^*)^\infty\big). \tag{10.44}$$

For any stationary policy $\rho^\infty\in\Gamma$, we have in view of the constraints of linear program (10.39)

$$\sum_i z_i^* = \sum_i\sum_b \rho_{ib}z_i^* \le \sum_i\sum_b\sum_a r_i(a,b)\rho_{ib}\pi_{ia}^*\cdot x_i^* = \gamma^T\phi\big((\pi^*)^\infty,\rho^\infty\big). \tag{10.45}$$

Since γ is strictly positive, (10.44) and (10.45) yields

$$\phi\big((\pi^*)^\infty,(\rho^*)^\infty\big) \le \phi\big((\pi^*)^\infty,\rho^\infty\big) \text{ for every } \rho^\infty\in\Gamma. \tag{10.46}$$

From (10.41) and (10.46) we obtain

$$\phi\big(\pi^\infty,(\rho^*)^\infty\big) \le v^* = \phi\big((\pi^*)^\infty,(\rho^*)^\infty\big) \le \phi\big((\pi^*)^\infty,\rho^\infty\big) \text{ for all } \pi^\infty\in\Pi \text{ and } \rho^\infty\in\Gamma, \tag{10.47}$$

showing that v^* is the value vector and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and 2, respectively. □

Algorithm 10.10 *Single-controller game with no discounting*

1. Compute optimal solutions (v^*,t^*,ρ^*) and (x^*,y^*,z^*) of the linear programs (10.38) and (10.39).
2. Define the stationary policy $(\pi^*)^\infty$ by $\pi_{ia}^* = \begin{cases} \frac{x_i^*(a)}{\sum_a x_i^*(a)}, & i\in S_{x^*},\ a\in A(i); \\ \frac{y_i^*(a)}{\sum_a y_i^*(a)}, & i\notin S_{x^*},\ a\in A(i), \end{cases}$
   where $S_{x^*} = \{i \mid \sum_a x_i^*(a) > 0\}$.
3. v^* is the value vector and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ optimal stationary policies for player 1 and 2.

Example 10.4 (continued)

For this example the linear programs (10.38) and (10.39) become

minimize $v_1 + v_2$

subject to

$$
\begin{array}{rcl}
v_1 - v_2 & \geq & 0 \\
-v_1 + v_2 & \geq & 0 \\
-v_1 + v_2 & \geq & 0 \\
v_1 - 5\rho_{11} - \rho_{12} - 6\rho_{13} & \geq & 0 \\
v_1 + t_1 - t_2 - 4\rho_{11} - 6\rho_{12} - 2\rho_{13} & \geq & 0 \\
v_2 - t_1 + t_2 - 6\rho_{21} & \geq & 0 \\
v_2 - 3\rho_{21} - 4\rho_{22} & \geq & 0 \\
v_2 - t_1 + t_2 - 6\rho_{22} & \geq & 0 \\
\rho_{11} + \rho_{12} + \rho_{13} = 1;\ \rho_{21} + \rho_{22} = 1;\ \rho_{11}, \rho_{12}, \rho_{13}, \rho_{21}, \rho_{22} & \geq & 0
\end{array}
$$

and

maximize $z_1 + z_2$

subject to

$$
\begin{array}{rcl}
x_{12} - x_{21} - x_{23} & = & 1 \\
-x_{12} + x_{21} + x_{23} & = & 1 \\
x_{11} + x_{12} + y_{12} - y_{21} - y_{23} & = & 1 \\
x_{21} + x_{22} - y_{12} + y_{21} + y_{23} & = & 1 \\
-5x_{11} - 4x_{12} + z_1 & \leq & 0 \\
-x_{11} - 6x_{12} + z_1 & \leq & 0 \\
-6x_{11} - 2x_{12} + z_1 & \leq & 0 \\
-6x_{21} - 3x_{22} + z_2 & \leq & 0 \\
-4x_{22} - 6x_{23} + z_2 & \leq & 0 \\
x_{11}, x_{12}, x_{21}, x_{22}, x_{23}, y_{11}, y_{12}, y_{21}, y_{22}, y_{23} & \geq & 0
\end{array}
$$

The optimal solutions are:

$v_1^* = 3.5,\ v_2^* = 3.5;\ t_1^* = 0.5;\ t_2 = 0;\ \rho_{11}^* = 0,\ \rho_{12}^* = 0.5,\ \rho_{13}^* = 0.5,\ \rho_{21}^* = 0.5,\ \rho_{22}^* = 0.5$ and

$z_1^* = 1.546,\ z_2^* = 5.454;\ x_{11}^* = 0.182,\ x_{12}^* = 0.227,\ x_{21}^* = 0.227,\ x_{22}^* = 1.364,\ x_{23}^* = 0;$

$y_{12}^* = 0.591,\ y_{21}^* = 0,\ y_{23}^* = 0;$

The optimal policy for player 1 is: $\pi_{11}^* = 0.444,\ \pi_{12}^* = 0.556,\ \pi_{21}^* = 0.143,\ \pi_{22}^* = 0.857,\ \pi_{23}^* = 0.$

Switching-controller stochastic game

In a switching-controller stochastic game we assume that the set of states is the union of two disjoint sets S_1 and S_2 such that player 1 controls the transitions in S_1 and player 2 in S_2. Also the switching-controller stochastic game has the ordered field property, stationary optimal policies

and a finite algorithm can be developed. However, this algorithm has a considerable complexity. It involves the solutions of a sequence of single-controller stochastic games and a combinatorial problem. We will not give the algorithm here, but refer to [222].

SER-SIT games

In this subsection we consider a stochastic game with *separable* rewards SER) and *state independent* transitions (SIT), i.e. $r_i(a,b) = s_i + t(a,b)$ and $p_{ij}(a,b) = p_j(a,b),\ j \in S$, for all $i.a,b$, under the average reward criterion. Let $|A(i)| = m$ and $|B(i)| = n$ for all $i \in S$ (notice that the SIT-property makes only sense when in all states the number of actions for player 1 (player 2) is the same). Consider the matrix games with $m \times n$ matrix $M = (m_{ab})$, where $m_{ab} = t(a,b) + \sum_j p_j(a,b)s_j,\ 1 \le a \le m,\ 1 \le b \le n$.

Theorem 10.22
Let $\pi^* = (\pi_1, \pi_2, \ldots, \pi_m)$ and $\rho^* = (\rho_1, \rho_2, \ldots, \rho_n)$ be optimal mixed strategies of the matrix game with matrix M. Then, $\phi = val(M)\cdot e$ is the value vector and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal policies for player 1 and player 2, respectively.

Proof
Since $val(M) \le t(\pi^*, \rho) + \sum_j p_j(\pi^*, \rho)s_j$ for all ρ, we also have, in vector notation, where the matrix $P(\pi^*, \rho)$ has identical rows, $s + val(M)\cdot e \le s + t(\pi^*, \rho)\cdot e + P(\pi^*, \rho)s$ for all ρ. By applying $P^*(\pi^*, \rho)$ on both sides, we obtain

$$val(M)\cdot e \le P^*(\pi^*, \rho)\{s + t(\pi^*, \rho)\cdot e\} = \phi\big((\pi^*)^\infty, \rho^\infty\big) \text{ for all } \rho^\infty \in \Gamma.$$

Similarly, one can prove

$$val(M)\cdot e \ge P^*(\pi, \rho^*)\{s + t(\pi, \rho^*)\cdot e\} = \phi\big(\pi^\infty, (\rho^*)^\infty\big) \text{ for all } \pi^\infty \in \Pi.$$

Hence, $\phi = val(M)\cdot e$ is the value vector and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal policies for player 1 and player 2, respectively. □

Algorithm 10.11 *SER-SIT game with no discounting*

1. Determine the value ϕ and optimal mixed strategies π^* and ρ^* in the matrix game with matrix M, where M is the $m \times n$ matrix defined by $m_{ab} = t(a,b) + \sum_j p_j(a,b)s_j$.

2. $\phi \cdot e$ is the value vector, and $(\pi^*)^\infty$ and $(\rho^*)^\infty$ are optimal stationary policies for player 1 and player 2, respectively.

Remark
Since ϕ and the optimal stationary strategies π^* and ρ^* can be computed by linear programming, $SER - SIT$ games possess the ordered field property.

ARAT games

An additive reward and additive transition ($ARAT$) stochastic game is defined by the property that the rewards as well as the transitions can be written as the sum of a term determined by player 1 and a term determined by player 2: $r_i(a,b) = r_i^1(a) + r_i^2(b),\ i \in S,\ a \in A(i),\ b \in B(i)$ and $p_{ij}(a,b) = p_{ij}^1(a) + p_{ij}^2(b),\ i,j \in S,\ a \in A(i),\ b \in B(i)$. We will argue the result that both players have optimal deterministic and stationary policies and that the ordered field property holds. For the details we refer to [162] and [76].

We have seen in Theorem 10.12 that, in the case of discounted rewards, both players have optimal deterministic and stationary policies and that the ordered field property holds. As usual, since there are only a finite number of deterministic and stationary policies, taking a sequence of discount factors tending to 1, some optimal deterministic and stationary pair of policies appears infinitely often, giving rise to a uniform discount optimal policy. But then, such pair is also average reward optimal.

A finite algorithm to compute the value vector and optimal deterministic and stationary policies resembles the algorithm of Vrieze, Raghavan, Tijs and Filar ([222]). there are some simplifications: no partition of the state space is needed, so $S_1 = S$ and $S_2 = \emptyset$. Furthermore, the policies can be taken deterministic and stationary.

10.4 Bibliographic notes

The name *stochastic game* stems from the seminal paper by Shapley ([183]). Some authors use the name *Markov game*, which expresses the relation with Markov decision processes. For books and surveys on stochastic games we refer to [146], [221], [26], [161] and [76]. The book of Von Neumann and Morgenstern ([219]) generally is seen as the starting point of game theory. A standard book, including much material on matrix games, is Owen ([144]).

The fixed point result, i.e. the value vector v^α is the unique solution of $x = Tx$, and the method of value iteration (Algorithm 10.1) for discounted games are due to Shapley ([183]). The mathematical programming formulations of section 10.2.2 were presented by Rothblum ([171]), and Hordijk and Kallenberg ([96]).

The iterative algorithms 10.2, 10.3 and 10.4 were proposed by Hoffman and Karp ([88]), Pollatschek and Avi-Itzhak ([151]), and Van der Wal ([200]), respectively. Example 10.2 that shows that Algorithm 10.3 does not converge in general is also due to Van der Wal ([200]). For a survey on (modified) value iteration methods we refer to Van der Wal and Wessels ([205]).

The notion that the ordered field property holds for discounted stochastic games in which one player controls the transitions, which yields a finite algorithm for such games, is due to Parthasarathy and Raghavan ([145]), a paper that also contains Example 10.3. See also Hordijk

and Kallenberg ([96]). The switching-controller stochastic game first was studied by Filar ([69]). Algorithm 10.6 is due to Vrieze ([221]). See also Vrieze, Tijs, Raghavan and Filar ([222]). The $SER-SIT$ game was introduced by Sobel ([188]) and later studied by Parthasarathy, Tijs and Vrieze ([147]). Raghavan, Tijs and Vrieze ([162]) have solved the $ARAT$ stochastic game.

The average reward stochastic games were introduced in 1957 by Gillette ([78]) who studied two special classes: games with perfect information and irreducible games. Gillette's proof that the above classes possess stationary optimal policies were later completed by Liggett and Lippman ([127]). Gillette's paper contains also the example of the Big Match, showing that undiscounted games were inherently more complex than discounted games. The complete analysis of the Big Match was made by Blackwell and Ferguson ([24]).

For the results that an undiscounted stochastic game possesses optimal stationary policies if and only if a global minimum with objective value zero can be found to an appropriate nonlinear program we refer to Filar and Schultz ([72]) and to Filar, Schultz, Thuijsman and Vrieze ([73]). The proof that in a game with perfect information both players possess optimal deterministic policies is due to Federgruen ([65]). The value iteration method, described in Algorithm 10.9, can be found in the paper by Hoffman and Karp ([88]). Van der Wal ([201]) developed a value iteration algorithm for the unichain case.

Stern, in his PhD thesis ([191]) proved that in the undiscounted single-controller stochastic game both players possess optimal stationary policies. Hordijk and Kallenberg ([97] and independently Vrieze ([220]) discovered the linear programmig solution this class of games. Also Filar ([70]) and Filar and Raghavan ([74]) made contributions to the undiscounted single-controller stochastic game.

The existence of optimal stationary policies for the switching-controller undiscounted stochastic game is due to Filar ([69]). Vrieze, Raghavan, Tijs and Filar ([222]) have develoed a finite, but complicated, algorithm for this model. The solution of the undiscounted $SER-SIT$ game was presented in Parthasarathy, Tijs and Vrieze ([147]). It is not known whether stochastic games with additive transitions have stationary optimal policies. When also the rewards are additive ($ARAT$ games), then Raghavan, Tijs and Vrieze ([162]) have shown that the undiscounted $ARAT$ stochastic game possesses the ordered field property and that both players have deterministic and stationary optimal policies.

10.5 Exercises

Exercise 10.1

Consider the following discounted stochastic game:

$S = \{1,2\};\ A(1) = \{1,2,3\},\ A(2) = \{1,2\};\ B(1) = \{1,2\},\ B(2) = \{1,2,3\};\ \alpha = \frac{1}{2}.$

$r_1(1,1) = 1;\ r_1(1,2) = 2;\ r_1(2,1) = 5;\ r_1(2,2) = 0;\ r_1(3,1) = 0;\ r_1(3,2) = 4;$

$r_2(1,1) = 0;\ r_2(1,2) = 3;\ r_2(1,3) = 6;\ r_2(2,1) = 6;\ r_2(2,2) = 2;\ r_2(2,3) = 0.$

$p_{11}(1,1) = 1,\ p_{12}(1,1) = 0;\ p_{11}(1,2) = 0,\ p_{12}(1,2) = 1;\ p_{11}(2,1) = 1,\ p_{12}(2,1) = 0;$

$p_{11}(2,2) = 0,\ p_{12}(2,2) = 1;\ p_{11}(3,1) = 1,\ p_{12}(3,1) = 0;\ p_{11}(3,2) = 0,\ p_{12}(3,2) = 1;$

$p_{21}(1,1) = 1,\ p_{22}(1,1) = 0;\ p_{21}(1,2) = 0,\ p_{22}(1,2) = 1;\ p_{21}(1,3) = 1,\ p_{22}(1,3) = 0;$

$p_{21}(2,1) = 1,\ p_{22}(2,1) = 0;\ p_{21}(2,2) = 0,\ p_{22}(2,2) = 1;\ p_{21}(2,3) = 1,\ p_{22}(2,3) = 0.$

Apply Algorithm 10.1 to compute x^2, starting with $x^0 = (0,0)$.

Exercise 10.2

Execute one iteration of Algorithm 10.2 on the model of Exercise 10.1.
Start with $\rho^*_{11} = \rho^*_{12} = \frac{1}{2};\ \rho^*_{21} = \rho^*_{22} = \rho^*_{23} = \frac{1}{3}$.

Exercise 10.3

Execute one iteration of Algorithm 10.3 on the model of Exercise 10.1. Start with $x = (0,0)$.

Exercise 10.4

Execute one iteration of Algorithm 10.4 on the model of Exercise 10.1. Start with $x = (6,6)$ and take $k = 2$.

Exercise 10.5

Consider the single-controller stochastic game in which player 2 controls the transitions.

a. Formulate the dual pair of linear progams for this stochastic game analogous to the programs (10.19) and (10.20).

b. Give the analogon of Theorem 10.10.

Exercise 10.6

The stochastic game of Exercise 10.1 is a single-controller stochastic game in which player 2 controls the transitions. Determine the value vector and optimal policies for the two players by linear programming as indicated in Exercise 10.5.

Exercise 10.7

Apply Algorithm 10.6 to the following switching control stochastic game.
$S = \{1,2\};\ S_1 = \{1\},\ S_2 = \{2\};\ A(1) = B(1) = A(2) = B(2) = \{1,2\};\ \alpha = \frac{1}{2}$.
$r_1(1,1) = 4,\ r_1(1,2) = 0,\ r_1(2,1) = 0,\ r_1(2,2) = 6;$
$r_2(1,1) = 3,\ r_2(1,2) = 5,\ r_2(2,1) = 6,\ r_2(2,2) = 4.$
$p_{11}(1) = 1,\ p_{12}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = 1;\ p_{21}(1) = 1,\ p_{22}(1) = 0;\ p_{21}(2) = 0,\ p_{22}(2) = 1.$
Start with $\rho_{21}^0 = 1,\ \rho_{22}^0 = 0$.

Exercise 10.8

Apply Algorithm 10.7 to the following $SER - SIT$ stochastic game.
$S = \{1,2\};\ A(1) = B(1) = A(2) = B(2) = \{1,2\};\ \alpha = \frac{1}{2}$.
$s_1 = 0,\ s_2 = 1;\ t(1,1) = 0,\ t(1,2) = 2,\ t(2,1) = 1,\ t(2,2) = 3.$
$p_1(1,1) = \frac{1}{2},\ p_2(1,1) = \frac{1}{2};\ p_1(1,2) = 1,\ p_2(1,2) = 0;$
$p_1(2,1) = 0,\ p_2(2,1) = 1;\ p_1(2,2) = \frac{1}{2},\ p_2(2,2) = \frac{1}{2}.$

Exercise 10.9

Consider the following model which has the SER property but not the SIT property.
$S = \{1,2,3\};\ A(i) = B(i) = \{1,2\}$ for $i = 1,2,3.\ s_1 = 1,\ s_2 = 1,\ s_3 = 2;\ t(a,b) = 0$ for all (a,b).
$p_{11}(1,1) = 1,\ p_{12}(1,1) = 0,\ p_{13}(1,1) = 0;\ p_{11}(1,2) = 0,\ p_{12}(1,2) = 1,\ p_{13}(1,2) = 0;$
$p_{11}(2,1) = 0,\ p_{12}(2,1) = 1,\ p_{13}(2,1) = 0;\ p_{11}(2,2) = 1,\ p_{12}(2,2) = 0,\ p_{13}(2,2) = 0;$
$p_{21}(1,1) = 0,\ p_{22}(1,1) = 1,\ p_{23}(1,1) = 0;\ p_{21}(1,2) = 0,\ p_{22}(1,2) = 0,\ p_{23}(1,2) = 1;$
$p_{21}(2,1) = 1,\ p_{22}(2,1) = 0,\ p_{23}(2,1) = 0;\ p_{21}(2,2) = 0,\ p_{22}(2,2) = 1,\ p_{23}(2,2) = 0;$
$p_{31}(1,1) = 0,\ p_{32}(1,1) = 1,\ p_{33}(1,1) = 0;\ p_{31}(1,2) = 0,\ p_{32}(1,2) = 1,\ p_{33}(1,2) = 0;$
$p_{31}(2,1) = 0,\ p_{32}(2,1) = 1,\ p_{33}(2,1) = 0;\ p_{31}(2,2) = 0,\ p_{32}(2,2) = 1,\ p_{33}(2,2) = 0.$

a. Show that $v_1^\alpha = 1 + \frac{1}{2}\alpha(v_1^\alpha + v_2^\alpha)$ and $v_3^\alpha = 2 + \alpha v_2^\alpha$.

b. Show that $v_2^\alpha = \frac{(6+\alpha) - \sqrt{(\alpha^2 - 20\alpha + 36)}}{8(1-\alpha)}$.

c. Show that this game does not possess the ordered field property.

Exercise 10.10

Show that, without using Theorem 10.15, the Big Match does not satisfy both (10.30) and (10.31).

Exercise 10.11

Execute Algorithm 10.10 to compute the value vector and optimal stationary policies for both players for the following undiscounted stochastic game:
$S = \{1,2,3\};\ A(1) = B(1) = \{1,2\};\ A(2) = \{1\},\ B(2) = \{1,2\};\ A(3) = B(3) = \{1\}.$
$p_{11}(1) = 1,\ p_{12}(1) = 0,\ p_{13}(1) = 0;\ p_{11}(2) = 0,\ p_{12}(2) = \frac{1}{2},\ p_{13}(2) = \frac{1}{2};$
$p_{21}(1) = 0,\ p_{22}(1) = 1,\ p_{23}(1) = 0;\ p_{31}(1) = 0,\ p_{32}(2) = 0,\ p_{33}(1) = 1.$
$r_1(1,1) = 1,\ r_1(1,2) = 0,\ r_1(2,1) = 0,\ r_1(2,2) = 1;\ r_2(1,1) = 4,\ r_2(1,2) = 2,\ r_3(1,1) = -1.$

Bibliography

[1] Altman, E.: *Constrained Markov decision processes*, Chapman & Hall/CRC, 1999.

[2] Altman, E., A. Hordijk and L.C.M. Kallenberg: *On the value function in constrained control of Markov chains*, Mathematical Methods of Operations Research 44 (1996) 389–399.

[3] Altman, E. and A. Shwartz: *Sensitivity of constrained Markov decision processes*, Annals of Operations Research 33 (1991) 1–22.

[4] Avrachenkov, K.E. and E. Altman: *Sensitive discount optimality via nested linear programs for ergodic Markov decision processes*, IDC'99 Proceedings (1999) 53–58.

[5] Baras, J.S., D.J. Ma and A.M. Makowsky: *K competing queues with linear costs and geometric service requirements: the μc-rule is always optimal*, Systems Control Letters 6 (1985) 173–180.

[6] Bather, J.: *Optimal decision procedures for finite Markov chains. Part I: Examples*, Advances in Applied Probability 5 (1973) 328–339.

[7] Bather, J.: *Optimal decision procedures for finite Markov chains. Part II: Communicating systems*, Advances in Applied Probability 5 (1973) 521–540.

[8] Bauer, H.: *Probability theory and elements of measure theory*, Second English Edition, Academic Press, London, 1981.

[9] Bayal-Gursoy, M. and K.W. Ross: *Viariability-sensitive Markov decision processes*, Mathematics of Operations Research 17 (1992) 558–571.

[10] Beckmann, M.: *An inventory model for arbitrary interval and quantity distributions of demands*, Management Science 8 (1961) 35–57.

[11] Bellman, R.: *Dynamic programming*, Princeton University Press, Princeton, 1957.

[12] Bellman, R., I. Glicksberg and O. Gross: *On the optimal inventory equation*, Management Science 2 (1955) 83–104.

[13] Ben-Israel, A. and S.D. Flåm: *A bisection/successive approximation method for computing Gittins indices*, Zeitschrift für Operations Research 34 (1990) 411–422.

[14] Bertsekas, D.P. and S.E. Shreve: *Stochastic optimal control: the discrete time case*, Academic Press, New York, 1978.

[15] Bertsekas, D.P.: *Dynamic programming: deterministic and stochastic models*, Prentice-Hall, 1987.

[16] Bertsimas, D. and J. Nino-Mora: *Conservation laws, extended polymatroids and multi-armed bandit problems: a unified approach to indexable systems*, Mathematics of Operations Research 21 (1996) 257–306.

[17] Beutler, F.J. and K.W. Ross: *Optimal policies for controlled Markov chains with a constraint*, Journal of Mathematical Analysis and Applications 112 (1985) 236–252.

[18] Bewley, T. and E. Kohlberg: *The asymptotic theory of stochastic games*, Mathematics of Operations Research 1 (1976) 197–208.

[19] Bewley, T. and E. Kohlberg: *The asymptotic solution of a recursive equation arising in stochastic games*, Mathematics of Operations Research 1 (1976) 321–336.

[20] Bewley, T. and E. Kohlberg: *On stochastic games with stationary optimal solutions*, Mathematics of Operations Research 3 (1978) 104–127.

[21] Bierth, K.-J.: *An expected average reward criterion*, Stochastic Processes and Applications 26 (1987) 133–140.

[22] Blackwell, D.: *Discrete dynamic programming*, Annals of Mathematical Statistics 33 (1962) 719–726.

[23] Blackwell, D.: *Positive dynamic programming*, Proceedings Fifth Berkeley Symposium Mathematical Statistics and Probability, Volume 1 (1967) 415–418.

[24] Blackwell, D. and T. Ferguson: *The Big Match*, Annals of Mathematical Statistics 39 (1968) 159–163.

[25] Breiman, L.: *Stopping-rule problems*, in: E.F. Beckenbach (ed.), *Applied Combinatorial Mathematics*, Wiley, New York, 1964, 284–319.

[26] Breton, M., J.A. Filar, A. Haurie and T.A. Shultz: *On the computation of equilibria in discounted stochastic games*, in: T.Basar (ed.), *Dynamic games and applications in economics*, Lecture Notes in Economics and Mathematical Systems no. 265 (1985), Springer-Verlag.

[27] Brown, B.W.: *On the iterative method of dynamic programming on a finite space discrete Markov process*, Annals of Mathematical Statistics 36 (1965) 1279–1285.

[28] Bruno, J., P. Downey and G. Frederickson: *Sequencing tasks with exponential service times to minimize the expected flowtime or makespan* Journal of the ACM 28 (1981) 100–113.

[29] Buyukkoc, C., P. Varaiya and J. Walrand: *The μc-rule revisited*, Advances in Applied Probability 17 (1985) 237–238.

[30] Cesaro, E.: *Sur la multiplication des séries*, Bulletin des Sciences Mathématiques 14 (1890) 114–120.

[31] Chen, Y.-R. and M.N. Katehakis: *Linear programming for finite state bandit problems*, Mathematics of Operations Research 11 (1986) 180–183.

[32] Cheng, M.C.: *New criteria for the simplex method*, Mathematical Programming 19 (1980) 230–236.

[33] Chung, K.L.: *Markov chains with stationary transition probabilities*, Springer, 1960.

[34] Chung, K.-J.: *A note on maximal mean/standard deviation ratio in an undiscounted MDP*, OR Letters 8 (1989) 201–204.

[35] Chung, K.-J.: *Remarks on maximal mean/standard deviation ratio in undiscounted MDPs*, Optimization 26 (1992) 385–392.

[36] Chung, K.-J.: *Mean-variance tradeoffs in an undiscounted MDP: the unichain case*, Operations Research 42 (1994) 184–188.

[37] Cox, D.R. and W.L. Smith: *Queues*, Methuen, London, 1961.

[38] Dantzig, G.B.: *Linear programming and extensions*, Princeton University Presss, 1963.

[39] De Ghellinck, G.T.: *Les problèmes de décisions séquentielles*, Cahiers du Centre de Recherche Opérationelle 2 (1960) 161–179.

[40] De Ghellinck, G.T. and G.D. Eppen: *Linear programming solutions for separable Markovian decision problems*, Management Science 13 (1967) 371–394.

[41] Dekker, R.: *Denumerable Markov decision chains: Optimal policies for small interest rate*, Ph.D. Dissertation, Leiden University, 1985.

[42] Dekker, R. and A. Hordijk: *Average, sensitive and Blackwell optimality in denumerable Markov decision chains with unbounded rewards*, Mathematics of Operations Research 13 (1988) 395–421.

[43] Dembo, R.S. and M. Haviv: *Truncated policy iteration methods*, OR Letters 3 (1984) 243–246.

[44] Denardo, E.V.: *Contraction mappings in the theory underlying dynamic programming*, SIAM Review 9 (1967) 165–177.

[45] Denardo, E.V.: *Separable Markov decision problem*, Management Science 14 (1968) 451–462.

[46] Denardo, E.V.: *On linear programming in a Markov decision problem*, Management Science 16 (1970) 281–288.

[47] Denardo, E.V.: *A Markov decision problem*, in: T.C.Hu and S.M.Robinson (eds.) *Mathematical Programming*, Academic Press (1973) 33–68.

[48] Denardo, E.V.: *Stopping and regeneration*, Draft of Chapter 7, problem 4 (1975).

[49] Denardo, E.V.: *Dynamic programming: Models and Applications*, Prentice-Hall, 1982.

[50] Denardo, E.V. and B.L. Fox: *Multichain Markov renewal programs*, SIAM Journal on Applied Mathematics 16 (1968) 468–487.

[51] Denardo, E.V. and B.L. Miller: *An optimality condition for discrete dynamic programming with no discounting*, Annals of Mathematical Statistics 39 (1968) 1220–1227.

[52] Denardo, E.V. and U.G. Rothum: *Overtaking optimality for Markov decision chains*, Mathematics of Operations Research 4 (1979) 144–152.

[53] D'Epenoux, F.: *Sur un problème de production et de stockage dans l'aléatoire*, Revue Française de Recherche Opérationelle 14 (1960) 3–16 .

[54] Derman, C.: *On optimal replacement rules when changes of state are Markovian*, in: R.Bellman (ed.) *Mathematical Optimization Techniques*, University of California Press, Berkeley (1963) 201–210.

[55] Derman, C.: *Finite state Markovian decision processes*, Academic Press, New York, 1970.

[56] Derman, C. and M. Klein: *Some remarks on finite horizon Markovian decision models*, Operations Research 13 (1965) 272–278.

[57] Derman, C. and R. Strauch: *A note on memoryless rules for controlling sequential control problems*, Annals of Mathematical Statistics 37 (1966) 276–278.

[58] Derman, C., G.J. Lieberman and S.M. Ross: *A sequential stochastic assignment model*, Management Science 18 (1972) 349–355.

[59] Derman, C. and A.F. Veinott Jr.: *Constrained Markov decision chains*, Management Science 19 (1972) 389–390.

[60] Doob, J.L.: Stochastic processes, Wiley, 1953.

[61] Dubins, L.E. and L.J. Savage: *How to gamble if you must: inequalities for stochastic processes*, McGraw-Hill, New York, 1965.

[62] Durinovic, S., H.M. Lee, M.N. Kathehakis and J.A. Filar: *Multiobjective Markov decision processes with average reward criterion* Large Scale Systems 10 (1986) 215–226.

[63] Eaves, B.C. and A.F. Veinott, Jr.: *Maximum-stopping-value policies in finite Markov population decision chains*, Report, Stanford University, 2007.

[64] Ephremides, A., P. Varaiya and J. Walrand: *A simple dynamic routing problem*, IEEE Transactions on Automatic Control AC-25 (1980) 690–693.

[65] Federgruen, A.: *Markovian control problems: functional equations and algorithms*, Mathematical Centre Tracts no.97, Amsterdam, 1984.

[66] Federgruen, A., P.J. Schweitzer and H.C. Tijms: *Contraction mappings underlying undiscounted Markov decision problems*, Journal of Mathematical Analyss and Applications 65 (1978) 711–730.

[67] Feinberg, E.A. and F. Yang: *On Polynomial Classification Problems for Markov Decision Processes* Proceedings of the 2008 NSF Engineering Research and Innovation Conference, Knoxville, TN.

[68] Feller, W.: *An introduction to probability theory and its aplications*, Volume I, third edition, Wiley, 1970.

[69] Filar, J.A.: *Ordered field property for stochastic games when the player who controls transitions changes from state to state*, Journal on Optimization Theory and Applications 34 (1981) 503–513.

[70] Filar, J.A.: *The completely mixed single-controller stochastic game*, Proceedings of the American Mathematical Society 95 (1985) 585–594.

[71] Filar, J.A., L.C.M. Kallenberg and H.M. Lee: *Variance-penalized Markov decision processes*, Mathematics of Operations Research 14 (1989) 147–161.

[72] Filar, J.A. and T. Schultz: *Nonlinear programming and stationary strategies in stochastic games*, Mathematical Programming 35 (1988) 243–247.

[73] Filar, J.A., Schultz, F. Thuijsman and O.J. Vrieze: *Nonlinear programming and stationary equilibria in stochastic games*, Mathematical Programming 50 (1991) 227–237.

[74] Filar, J.A. and T.E.S. Raghavan: *A matrix game solution of the single-controller stochastic game*, Mathematics of Operations Research 9 (1984) 356–362.

[75] Filar, J.A. and T. Schultz: *Communicating MDPs: Equivalence and LP properties*, Operations Research Letters 7 (1988) 303–307.

[76] Filar, J.A. and O.J. Vrieze: *Competitive Markov decision processes*, Springer-Verlag, 1997.

[77] Gal, S.: *A $\mathcal{O}(N^3)$ algorithm for optimal replacement problems*, SIAM Journal of Control and Optimization 22 (1984) 902–910.

[78] Gillette, D: *Stochastic games with zero stop probabilities* in: Dresher, M., A.W. Tucker and P. Wolfe (eds.), *Contributions to the theory of games*, vol. III, Princeton University Press, Annals of Mathematics Studies 39 (1957) 179–187.

[79] Gittins, J.C.: *Bandit processes and dynamic allocation indices*, Journal of the Royal Statistic Society Series B 14 (1979) 148–177.

[80] Gittins, J.C. and D.M. Jones: *A dynamic allocation index for the sequential design of experiments*, in: J. Gani (ed.) *Progress in Statistics* North Holland, Amsterdam (1974) 241–266.

[81] Glazebrook, K.D.: *Scheduling tasks with exponential service times on parallel processors* Journal of Applied Probability 16 (1979) 685–689.

[82] Glazebrook, K.D. and R.W. Owen: *New results for generalized bandit problems* International Journal of System Science22 (1991) 479–494.

[83] Hastings, N.A.J.: *Some notes on dynamic programming and replacement*, Operational Research Quarterly 19 (1968) 453–464.

[84] Hastings, N.A.J.: *Optimization of discounted Markov decision problems*, Operations Research Quarterly 20 (1969) 499–500.

[85] Hastings, N.A.J.: *A test for nonoptimal actions in undiscounted finite Markov decision chains*, Management Science 23 (1976) 87–92.

[86] Haviv, M. and M.L. Puterman: *An improved algorithm for solving communicating average reward Markov decision processes*, Annals of Operations Research 28 (1991) 229–242.

[87] Heyman, D.P. and M.J. Sobel: *Stochastic models in Operations Research, Volume II: Stochastic optimization*, MacGraw-Hill, 1984.

[88] Hoffman, A.J. and R.M. Karp: *On non-terminating stochastic games*, Management Science 12 (1966) 359–370.

[89] Hordijk, A.: *A sufficient condition for the existence of an optimal policy with respect to the average cost criterion in Markovian decision processes*, Transactions of the Sixth Conference on Information Theory, Statistical Decision Functions, Random Processes (1971) 263–274.

[90] Hordijk, A.: *Dynamic programming and Markov potential theory*, Mathematical Centre, Amsterdam, 1974.

[91] Hordijk, A.: *Convergent dynamic programming*, Report BW 47/75, Mathematical Centre, Amsterdam, 1975.

[92] Hordijk, A.: *Stochastic dynaming programming*, Course notes, University of Leiden (in Dutch), 1976.

[93] Hordijk, A.: *From linear to dynamic programming via shortest paths*, Mathematical Centre Tract no. 100, Amsterdam, 1978.

[94] Hordijk, A., R. Dekker and L.C.M. Kallenberg: *Sensitivity analysis in discounted Markov decision problems*, OR Spektrum 7 (1985) 143–151.

[95] Hordijk, A. and L.C.M. Kallenberg: *Linear programming and Markov decision chains*, Management Science 25 (1979) 352–362.

[96] Hordijk, A. and L.C.M.Kallenberg: *Linear programming and Markov games I*, in: O. Moeschlin and D. Pallaschke (eds.), *Game theory and mathematical economics*, North Holland (1981) 291–305.

[97] Hordijk, A. and L.C.M.Kallenberg: *Linear programming and Markov games II*, in: O. Moeschlin and D. Pallaschke (eds.), *Game theory and mathematical economics*, North Holland (1981) 307–320.

[98] Hordijk, A. and L.C.M. Kallenberg: *Transient policies in discrete dynamic programming: linear programming including suboptimality and additional constraints*, Mathematical Programming 30 (1984) 46–70.

[99] Hordijk, A. and L.C.M. Kallenberg: *Constrained undiscounted stochastic dynamic programming*, Mathematics of Operations Research 9 (1984) 276–289.

[100] Hordijk, A. and H.C. Tijms: *Colloquium Markov programming*, Mathematical Centre Report BC 1/70, Mathematical Centre, Amsterdam (in Dutch).

[101] Howard, R.A.: *Dynamic programming and Markov processes*, MIT Press, Cambridge, 1960.

[102] Huang, Y and L.C.M. Kallenberg: *On finding optimal policies for Markov decision chains: A unifying framework for mean-variance tradeoffs*, Mathematics of Operations Research 19 (1994) 434–448.

[103] Iglehart, D.: *Optimality of (s, S)-policies in the infinite horizon dynamic inventory problem*, Management Science 9 (1963) 259–267.

[104] Iglehart, D.: *Dynamic programming and stationary analysis of inventory problems*, Chapter 1 in: H. Scarf, D. Gilford and M. Shelly (eds.), *Multistage inventory models and techniques*, Stanford University Press, Stanford, 1963.

[105] Johnson, S.M.: *Optimal two- and three-stages production schedules with setup times included*, Naval Research Logistics Quarterly 1 (1954) 61–68.

[106] Kallenberg, L.C.M.: *Finite horizon dynamic programming and linear programming*, Methods of Operations Research 43 (1981) 105–112.

[107] Kallenberg, L.C.M.: *Unconstrained and constrained dynamic programming over a finite horizon*, Report, University of Leiden, 1981.

[108] Kallenberg, L.C.M.: *Linear programming and finite Markovian control problems*, Mathematical Centre Tract no.148, Amsterdam, 1983.

[109] Kallenberg, L.C.M.: *A note on M.N.Katehakis and Y.-R.Chen's computation of the Gittins index*, Mathematics of Operations Research 11 (1986) 184–186.

[110] Kallenberg, L.C.M.: *Separable Markov decision problems*, OR Spektrum 14 (1992) 43–52.

[111] Kallenberg, L.C.M.: *Classification problems in MDPs*, in: Z. How, J.A. Filar and A. Chen (ed.) *Markov processes and controlled Markov chains*, Kluwer Boston (2002) 151–165.

[112] Kao, E.P.C.: *Optimal replacement rules when changes of state are semi-Markovian*, Operations Research 21 (1973) 1231–1249.

[113] Karlin, S.: *Mathematical methods and theory in games, programming and economics*, Volume I, Addison-Wesley, 1959.

[114] Karlin, S.: *Dynamic inventory policy with varying stochastic demands*, Management Science 6 (1960) 231–258.

[115] Katehakis, M.N. and C. Derman: *Optimal repair allocation in a series system*, Mathematics of Operations Research 9 (1984) 615–623.

[116] Katehakis M. N. and U. Rothblum: *Finite state multi-armed bandit sensitive-discount, average-reward and average-overtaking optimality*, Annals of Applied Probability 6 (1996) 1024–1034.

[117] Katehakis, M.N. and A.F. Veinott Jr.: *The multi-armed bandit problem: decomposition and computation*, Mathematics of Operations Research 12 (1987) 262–268.

[118] Kawai, H.: *A variance minimization problem for a Markov decision process*, European Journal of Operations Research 31 (1987) 140–145.

[119] Kawai, H. and N. Katoh: *Variance constrained Markov decision process*, Journal of the Operations Research Society of Japan 30 (1987) 88–100.

[120] Kemeny, J. and L. Snell: *Finite Markov chains*, Van Nostrand, 1960.

[121] Kolesar, P.: *Minimum cost replacement under Markovian deterioration*, Management Science 12 (1966) 694–706.

[122] Koole, G.M.: *Stochastic scheduling and dynamic programming*, CWI Tract 113, CWI, Amsterdag, 1995.

[123] Kushner, H.J. and A.J. Kleinman: *Mathematical programming and the control of Markov chains*, IEEE Transactions on Automatic Control AC-13 (1968) 801–820.

[124] Kushner, H.J. and A.J. Kleinman: *Accelerated procedures for the solution of discrete Markov control problems*, IEEE Transactions on Automatic Control AC-16 (1971) 147–152.

[125] Lasserre, J.B.: *Detecting optimal and non-optimal actions in average-cost Markov decision processes* Journal of Applied Probability 31 (1994) 979–990.

[126] Lasserre, J.B.: *A new policy iteration scheme for Markov decision processes using Schweitzer's formula*, Journal of Applied Probability 31 (1994) 268–273.

[127] Liggett, T.M. and S.A. Lippman: *Stochastic games with perfect information and time averqage payoff*, SIAM Review 11 (1969) 604–607.

[128] Lin, W. and P.R. Kumar: *Optimal control of a queueing system with two heterogeneous servers*, IEEE Tansactions on Automatic Control AC-29 (1984) 696–705.

[129] Lippman, S.A.: *Applying a new device in the optimization of exponential queueing systems*, Operations Research 23 (1975) 687–710.

[130] Liu, J.Y. and K. Liu: *An algorithm on the Gittins index*, Systems Science and Mathematical Science 7 (1994) 106–114.

[131] MacQueen, J.: *A modified programming method for Markovian decision problems*, Journal of Mathematical Analysis and Applications 14 (1966) 38–43.

[132] MacQueen, J.: *A test for suboptimal actions in Markov decision problems*, Operations Research 15 (1967) 559–561.

[133] Manne, A.S.: *Programming of economic lot sizes*, Management Science 4 (1958) 115–135.

[134] Manne, A.S.: *Linear programming and sequential decisions*, Management Science 6 (1960) 259–267.

[135] Manne, A.S. and A.F. Veinott, Jr.: Chapter 11 in A.S. Manne (ed.), *Investments for capacity expansion: size, location and time-phasing*, MIT Press, 1967.

[136] McCuaig, W.: *Intercyclic digraphs*, in: N. Robertson and P. Seymour (eds.) *Graph structure theory*, Contemporary Mathematics 147, American Mathematical Society (1993) 203–245.

[137] Mertens, J.F. and A. Neyman: *Stochastis games*, International Journal of Game Theory 10 (1981) 53–56.

[138] Miller, B.L. and A.F. Veinott Jr.: *Discrete dynamic programming with a small interest rate*, Annals of Mathematical Statistics 40 (1969) 366–370.

[139] Mine, H. and S. Osaki: *Markovian decision processes*, Elsevier, New York, 1970.

[140] Morton, T.E.: *On the asymptotic convergence rate of cost differences for Markovian decision processes*, Operations Research 19 (1971) 244–248.

[141] Nazareth, J.L. and R.B. Kulkarni: *Linear programming formulations of Markov decision processes* Operations Research Letters 5 (1986) 13–16.

[142] Nino-Mora, J.: *A* $(2/3)n^3$ *fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain* INFORMS Journal of Computing 19 (2007) 596–606.

[143] Norman, J.M.: *Dynamic programming in tennis - when to use a fast serve*, Journal of the Operational Research Society 36 (1987) 75–77.

[144] Owen, G.: *Game theory*, Academic Press, 1982.

[145] Parthasarathy, T. and T.E.S. Raghavan: *An orderfield property for stochastic games when one player controls the transitions*, Journal of Optimization Theory and Applications 33 (1981) 375–392.

[146] Parthasarathy, T. and T.E.S. Raghavan: *Some topics in two-person games*, Elsevier, 1971.

[147] Parthasarathy, T., S.H. Tijs and O.J. Vrieze: *Stochastic games with state independent transitions and separable rewards*, in: Hammer, G. and D. Pallschke (eds.): *Selected topics in Operations Research and Mathematical Economics*, Springer (1984) 262–271.

[148] Pinedo, M. and L. Schrage: *Stochastic shop scheduling: a survey*, in: Dempster, M.A.H., J.K. Lenstra and A.H.G. Rinnooy Kan (eds.), *Deterministic and stochastic scheduling*, Reidel, Dordrecht, Holland (1982) 181–196.

[149] Pinedo, M. and G. Weiss: *Scheduling of stochastic tasks on two parallel processors* Naval Research Logistics Quarterly 26 (1979) 527–535.

[150] Platzman, L.K.: *Improved conditions for convergence in undiscounted Markov renewal programming*, Operations Research 25 (1977) 529–533.

[151] Pollatschek, M. and Avi-Itzhak: *Algorithms for stochastic games with geometric interpretation*, Manangement Science 15 (1969) 399–415.

[152] Porteus, E.L.: *Some bounds for discounted sequential decision processes*, Management Science 18 (1971) 7–11.

[153] Porteus, E.L.: *On the optimality of generalized* (s, S) *policies*, Management Science 17 (1971) 411–426.

[154] Porteus, E.L.: *Bounds and transformations for discounted finite Markov decision chains*, Operations Research 23 (1975) 761–784.

[155] Powell, R.E. and S.M. Shah: *Summability theory and applications*, Van Nostrand Reinhold, London (1972).

[156] Prussing, J.E.: *How to serve in tennis*, The Mathematical Gazette 61 (1977) 294–296 .

[157] Puterman, M.L.: *Markov decision processes*, Wiley, New York, 1994.

[158] Puterman, M.L. and S.L. Brumelle: *On the convergence of policy iteration in stationary dynamic programming*, Mathematics of Operations Research 4 (1979) 60–69.

[159] Puterman, M.L. and M.C. Shin: *Modified policy iteration algorithms for discounted Markov decision chains*, Management Science 24 (1978) 1127–1137.

[160] Puterman, M.L. and M.C. Shin: *Action elimination procedures for modified policy iteration algorithms*, Operations Research 30 (1982) 301–318.

[161] Raghavan, T.E.S. and J.A. Filar: *Algorithms for stochastic games - a survey*, Zeitschrift für Operations Research 35 (1991) 437–472.

[162] Raghavan, T.E.S., S.H. Tijs and O.J. Vrieze: *On stochastic games with additive reward and trasition structure*, Journal of Optimization Theory and Applications 47 (1985) 451–464.

[163] Reetz, D.: *Solution of a Markovian decision problem by successive overrelaxation*, Zeitschrift für Operations Research 17 (1973) 29–32.

[164] Righter, R.: *Scheduling*, in: Shaked, M. and J.G. Shanthikumar (eds.), *Stochastic orders and their applications*, Academic Press, 1994, 381–432.

[165] Ross, K.W.: *Randomized and past=dependent policies for Markov decision processes with multiple constraints*, Operations Research 37 (1989) 474–477.

[166] Ross, K.W. and R. Varadarajan: *Markov decision processes with sample path constraints: the communicating case*, Operations Research 37 (1989) 780–790.

[167] Ross, K.W. and R. Varadarajan: *Multichain Markov decision processes with a sample path constraint: a decomposition approach*, Mathematics of Operations Research 16 (1991) 195–207.

[168] Ross, S.M.: *Applied probability models with optimization applications*, Holden-Day, San Francisco, 1970.

[169] Ross, S.M.: *Dynamic programming and gambling models*, Advances in Applied Probability 6 (1974) 593–606.

[170] Ross, S.M.: *Introduction to stochastic dynamic programming*, Academic Press, New York, 1983.

[171] Rothblum, U.G.: *Solving stopping stochastic games by maximizing a linear function subject to quadratic constraints*, O. Moeschlin and D. Pallaschke (eds.), *Game theory and mathematical economics*, North Holland (1978) 103–105.

[172] Scarf, H.: *The optimality of (s, S)-policies in the dynamic inventory problem*, Chapter 13 in: K.J. Arrow, S. Karlin and P. Suppes (eds.) *Mathematical methods in the social sciences*, Stanford University Press, Stanford, 1960.

[173] Scarf, H.: *A survey of analytic techniques in inventory theory*, Chapter 7 in: H. Scarf, D. Gilford and M. Shelly (eds.), *Multistage inventory models and techniques*, Stanford University Press, Stanford, 1963.

[174] Schweitzer, P.J.: *Multiple policy improvements in undiscounted Markov renewal programming*, Operations Research 19 (1971) 784–793.

[175] Schweitzer, P.J.: *Iterative solution of the functional equations of undiscounted Markov renewal programming*, Journal of Mathematical Analysis and Applications 34 (1971) 495–501.

[176] Schweitzer, P.J. and A. Federgruen: *The asymptotic behavior of undiscounted value iteration in a Markov decision problem*, Mathematics of Operations Research 2 (1977) 360–381.

[177] Schweitzer, P.J. and A. Federgruen: *The functional equation of undiscounted Markov renewal programming*, Mathematics of Operations Research 3 (1978) 308–321.

[178] Schweitzer, P.J. and A. Federgruen: *Geometric convergence of value-iteration in multichain Markovian renewal programming*, Advances in Applied Probability 11 (1979) 188–217.

[179] Sennott, L.I.: *Stocastic dynamic programming and the control of queueing systems*, Wiley, 1999.

[180] Serfozo, R.: *Monotone optimal policies for Markov decision processes*, Mathematical Programming Study 6 (1976) 202–215.

[181] Serfozo, R.: *An equivalence between continuous and discrete time Markov decision processes*, Operations Research 27 (1979) 616–620.

[182] Shapiro, J.F.: *Brouwer's fixed-point theorem and finite state space Markovian decision theory*, Journal of Mathematical Analysis and Applications 49 (1975) 710–712.

[183] Shapley, L.S.: *Stochastic games*, Proceedings of the National Academy of Sciences 39 (1953) 1095–1100.

[184] Shapley, L.S. and R.N. Snow: *Basic solutions of discrete games*, in: H.W. Kuhn & A.W. Tucker (eds.), *Contributions to the theory of games*, Vol. I, Annals of Mathematical Studies no. 24, pp. 27–35, Princeton University Press (1950).

[185] Sherif, Y.S. and M.L. Smith: *Optimal maintenance policies for systems subject to failure - A review*, Naval Research Logistics Quarterly 28 (1981) 47–74.

[186] Sladky, K.: *On the set of optimal controls for Markov chains with rewards*, Kybernetica 10 (1974) 350–367.

[187] Smith, D.R.: *Optimal repair of a series system*, Operations Research 26 (1978) 653–662.

[188] Sobel, M.J.: *Myopic solutions of Markov decisin processes and stochastic games*, Operations Research 29 (1981) 995–1009.

[189] Sobel, M.J.: *Maximal mean/standard deviation ratio in an undiscounted MDP*, OR Letters 4 (1985) 157–159.

[190] Sobel, M.J.: *Mean-variance tradeoffs in undiscounted MDP*, Operations Research 42 (1994) 175–183.

[191] Stern, M.: *On stochastic games with limiting average payoff*, PhD thesis, University of Illinois at Chicago, Chicago, 1975.

[192] Stoer, J. and R. Bulirsch: *Introduction to numerical analysis*, Springer,1980.

[193] Strauch, R.: *Negative dynamic programming*, Annals Mathematical Statistics 37 (1966) 871–889.

[194] Strauch, R. and A.F. Veinott Jr.: *A property of sequential control processes*, Report, Rand McNally, Chicago, 1966.

[195] Tarjan, R.E.: *Depth-first search and linear graph algorithms*, SIAM Journal of Computing 1 (1972) 146–160.

[196] Topkis, D.: *Minimizing a submodular function on a lattice*, Operations Research 26 (1978) 305–321.

[197] Tsitsiklis, J.N.: *A lemma on the multi-armed bandit problem*, IEEE Transactions on Automatic Control 31 (1986) 576–577.

[198] Tsitsiklis, J.N.: *A short proof of the Gittins index theorem*, Annals of Applied Probability 4 (1994) 194–199.

[199] Tsitsiklis, J.N.: *NP-hardness of checking the unichain condition in average cost MDPs*, Operations Research Letters 35 (2007) 319–323.

[200] Van der Wal, J.: *Discounted Markov games: successive approximations and stopping times*, International Journal of Game Theory 6 (1977) 11–22.

[201] Van der Wal, J.: *Successive approximations for average reward Markov games*, International Journal of Game Theory 9 (1980) 13–24.

[202] Van der Wal, J.: *The method of value oriented successive approximation for the average reward Markov decision processes*, OR Spektrum 1 (1980) 233–242.

[203] Van der Wal, J.: *Stochastic dynamic programming*, Mathematical Centre, Amsterdam, 1981.

[204] Van der Wal, J. and J.A.E.E. Van Nunen: *A note on the convergence of value oriented successive approximations method*, Report, Eindhoven University of Technology, 1977.

[205] Van der Wal, J. and J. Wessels: *Successive approximations for Markov games*, in: H. Tijms and J. Wessels (eds.), *Markov decision theory*, Mathematical Centre Tract no. 93, 1977, Amsterdam.

[206] Van Hee, K.M., A. Hordijk and J. Van der Wal: *Successive approximations for convergent dynamic programming*, in: H.C. Tijms and J. Wessels (eds.) *Markov decision theory*, Mathematical Centre Tract no.93, Mathematical Centre, Amsterdam, 1977, 183–211.

[207] Van Nunen, J.A.E.E.: *A set of successive approximation method for discounted Markovian decision problems*, Zeitschrift für Operations Research 20 (1976) 203–208.

[208] Van Nunen, J.A.E.E.: *Contracting Markov decision processes* Mathematical Centre Tract 71, Mathematical Centre, Amsterdam, 1976.

[209] Van Nunen, J.A.E.E. and J. Wessels: *A principle for generating optimization procedures for discounted Markov decision processes*, Colloquia Mathematica Societatis Bolyai Janos, Vol. 12, North Holland, Amsterdam, 1976, 683–695.

[210] Van Nunen, J.A.E.E. and J. Wessels: *The generation of successive approximations for Markov decision processes using stopping times*, in: H. Tijms and J. Wessels (eds.) *Markov decision theory*, Mathematical Centre Tract no.93, Mathematical Centre, Amsterdam, 1977, 25–37.

[211] Van Nunen, J.A.E.E. and J. Wessels: *Markov decision processes with unbounded rewards*, in: H.C. Tijms and J. Wessels (eds.) *Markov decision theory*, Mathematical Centre Tract no.93, Mathematical Centre, Amsterdam, 1977, 1–24.

[212] Varaiya, P.P., J.C. Walrand and C. Buyukkoc: *Extensions of the multi-armed bandit problem: the discounted case*, IEEE Transactions on Automatic Control 30 (1985) 426–439.

[213] Veinott, A.F. Jr.: *Optimal policy for a multi-product, dynamic nonstationary inventory problem*, Management Science 12 (1965) 206–222.

[214] Veinott, A.F. Jr.: *On finding optimal policies in discrete dynamic progamming with no discounting*, Annals of Mathematical Statistics 37 (1966) 1284–1294.

[215] Veinott, A.F. Jr.: *On the optimality of (s, S) inventory policies: new conditions and a new proof*, SIAM Journal on Applied Mathematics 14 (1966) 1067–1083.

[216] Veinott, A.F. Jr.: *Discrete dynamic programming with sensitive discount optimality criteria (preliminary report)*, Annals of Mathematical Statistics 39 (1968) 1372.

[217] Veinott, A.F. Jr.: *Discrete dynamic programming with sensitive discount optimality criteria*, Annals of Mathematical Statistics 40 (1969) 1635–1660.

[218] Veinott, A.F. Jr.: *Markov decision chains*, in: G.B.Dantzig and B.C.Eaves (eds.) *Studies in Mathematics, vol. 10: studies in optimization*, The Mathematical Association of America (1974) 124–159.

[219] Von Neumann, J. and O. Morgenstern: *The theory of games and economic behaviour*, Princeton University Press, 1950.

[220] Vrieze, O.J.: *Linear programming and undiscounted stochastic games*, OR Spektrum 3 (1981) 29–35.

[221] Vrieze, O.J.: *Stochastic games with finite state and action spaces*, CWI Tracts 33, 1987.

[222] Vrieze, O.J., S.H. Tijs, T.E.S. Raghavan and J.A. Filar: *A finite algorithm for the switching controller stochastic game*, OR Spektrum 5 (1983) 15–83.

[223] Wagner, H.M. and T. Whithin: *Dynamic problems in the theory of the firm*, T. Whithin (ed.): *Theory of inventory management*, App. 6, 2nd ed., Princeton University Press, 1957.

[224] Walrand, J.: *An introduction to queueing networks*, Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[225] Weber, R.R.: *Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime*, Journal of Applied Probabitily 19 (1982) 167–182.

[226] Weber, R.R.: *On the Gittins index for multi-armed bandits*, Annals of Applied Probabitily 2 (1992) 1024–1033.

[227] Weiss, G.: *Multiserver stochastic scheduling*, in: Dempster, M.A.H., J.K. Lenstra and A.H.G. Rinnooy Kan (eds.), *Deterministic and stochastic scheduling*, Reidel, Dordrecht, Holland (1982) 157–179.

[228] Weiss, G.: *Braching bandit processes*, in: Probability in the Engineering and Informational Sciences 2 (1988) 269–278.

[229] Wessels, J.: *Stopping times and Markov programming*, in: Transactions of the 7-th Prague conference on information theory, statistical decision functions and random processes, Academia, Prague (1977) 575–585.

[230] White, D.J.: *Dynamic programming, Markov chains and the method of successive approximations*, Journal of Mathematical Analysis and Applications 6 (1963) 373–376.

[231] White, D.J.: *Dynamic programming and probabilistic constraints*, Operations Research 22 (1974) 654–664.

[232] White, D.J.: *Multi-objective infinite-horizon discounted Markove decision processes*, Journal of Mathematical Analysis and Applications 89 (1982) 639–647.

[233] White, D.J.: *Mean, variance and probabilistic criteria in finite decision processes: a review*, Journal of Optimization Theory and Applications 56 (1988) 1–30.

[234] White, D.J.: *Computational approaches to variance-penalized Markov decision processes*, OR Spektrum 14 (1992) 79–83.

[235] White, D.J.: *A mathematical programming approach to a problem in variance penalised Markov decision processes*, OR Spektrum 15 (1994) 225–230.

[236] Whittle, P.: *Multi-armed bandits and the Gittins index*, Journal of the Royal Statistical Society, Series B 42 (1980) 143–149.

[237] Whittle, P.: *Optimization over time*, Wiley, 1982.

[238] Winston, W.: *Optimality of the shorstest line discipline*, Journal of Applied Probability 14 (1977) 181–189.

[239] Zangwill, W.I.: *A deterministic multi-period production scheduling model with backlogging*, Management Science 13 (1966) 105–119.

[240] Zangwill, W.I.: *A backlogging model and a multi-echelon model of a dynamic economic lot size production system - a network approach*, Management Science 15 (1969) 506–527.

[241] Zoutendijk, G.: *Mathematical programming methods*, North Holland, 1976.