# Stochastic games

Janssen, P.A.M.

**Pascal Janssen**

# Stochastic Games

**Master thesis, defended on August 22, 2006**

**Thesis advisor: Prof.dr. L.C.M. Kallenberg**


**Mathematisch Instituut, Universiteit Leiden**

# Contents

# Introduction

In this thesis we will take a look at stochastic games. The first two chapters are based on chapter 2 and 3 of the book 'Markov Decision Processes', by Filar and Vrieze [6]. Some parts are condensed, whereas other parts get a more extensive treatment. Solutions to some of the exercises posed in the book are integrated in the text.

In chapter 1 we start by taking a look at Markov Decision Processes, which can be regarded as one-person games. Some basic definitions and theorems are introduced, on which we will expand in chapter 3. Chapter 1 covers a couple of summable Markov decision processes, also called MDPs. In order to compare different strategies, four methods of evaluation are introduced here, of which the $\beta$-discounted model and the limiting average model are our main objects of study. We discuss methods of finding an optimal strategy for these models, and, if a strategy is already given, how to improve upon it. It turns out we have an important tool in our possession in the form of Linear Programming. Furthermore the theory of Markov decision processes is applied to derive a formulation of the Hamilton cycle problem. Besides stationary strategies, we also take a brief look at Markov and behaviour strategies. An explanation is given as to why we can restrict ourselves to stationary strategies in this chapter, without loss of generality.

In chapter 2 we try taking these ideas to a higher level, and we try to apply them to 2-player games. Unfortunately, linearity is lost in most of the cases, and consequently we cannot always use Linear Programming to find optimal strategies. For $\beta$-discounted games, we show that subclasses exist in which the linearity is retained, and the LP approach still holds. In order to solve more general stochastic games it is necessary to study nonlinear programming methods. We take a look at a modified version of Newton's method, and use this to solve general $\beta$-discounted models. Problems arise in the analysis of the limiting average games. These are shown by way of the example of the Big Match. As it turns out, it is not always possible to formulate an optimal stationary strategy in this model. A natural question which now arises is how well we can perform with stationary strategies. This leads to the introduction of $\varepsilon$-optimal stationary strategies. The nonlinear programming approach is also tried on general-sum, $K$-player games, both discounted and limiting average.

Chapter 3 deals with an article by Mertens and Neyman [12] about the existence of the value of a limiting average game. They provided an answer for the question whether the value of such a game always exists. In this thesis we try to clarify their result for the case with finite actions and finite states. In order to do this we briefly discuss summation theory, in particular Abel and Cesaro summation. Our main line of approach is the consideration of a sequence of $\beta$-discounted games with $\beta$ approaching 1. This brings us to the subject of Puiseux series and results published by Bewley and Kohlberg [1]. After these prelimaries we present the main result of Mertens and Neyman.

Finally, I would like to thank Prof.Dr.Kallenberg for all his time, effort and patience in coaching me. Whenever I hit a dead end, a meeting with him would see me through. Furthermore, I would like to thank Dr.Spieksma and Dr.Kooman for being part of the exam committee. And finally, Claire Coombes for correcting my English, and encouraging me along the entire way. I couldn't have done it without her.

# Chapter 1

# Markov Decision Processes

## 1.1 Introduction

In this chapter we approach the subject of Stochastic Games from the perspective of Markov Decision Processes with a finite state/action space. These can be viewed as a special case of stochastic games, namely where there is only one player. This means we can consider it to be an optimization problem. Our main approach will be to reformulate the problem in terms of an optimal control problem using mathematical programming, resembling the following:

Find a control that:
maximizes (objective function)
subject to:
satisfaction of feasibility constraints.

This approach will prove very fruitful and many of the techniques used in this chapter will be extended to the multiplayer case in the next chapter.

## 1.2 The Summable Markov Decision Processes

A Markov Decision Process can be viewed as an extension of Markov Processes. Whereas Markov Processes have only one action to choose from, which consequently defines the transition probabilities to the other stages, a Markov Decision Process can have more than one action, each of which defines its own transition probabilities. As in Markov Processes, the process is observed at discrete time intervals $t = 0, 1, 2, 3, \ldots$, which are referred to as *stages*. The *state* of the process at time point $t$ will be denoted by the random variable $S_t$. The values $S_t$ can take can be finite or infinite. We will restrict ourselves however to the finite case. This means $S_t \in \mathbf{S} = \{1, 2, \ldots, N\}$, where $\mathbf{S}$ is called the *state space*. When the process is in state $i$ at time $t$ we will write $\{S_t = i\}$.

The process is controlled by a *controller* (also called *player* or *decision maker*), who chooses an *action* $a_t \in \mathbf{A}(i) = \{1, 2, \ldots, m(i)\}$ at time $t$ if the process is in state $i$ at that time. As with the state space, we will take this action space to be finite. The action taken at time point $t$ will be denoted by the random variable $A_t$. Each decision $a_t \in \mathbf{A}(i)$ results in an immediate *reward* or *output* $r(i, a_t)$ and a transition to another state $j \in \mathbf{S}$, which depends on the transition probabilities connected to this choice. This brings us to another restriction we will impose. We will only consider cases where the actions chosen are independent of time and any previous states and actions. This means that *stationary transition probabilities* exist, such that

$$p_{ij}(a) := \mathbb{P}\{S_{t+1} = j \mid S_t = i, A_t = a\} \tag{1.1}$$

for all $t = 0, 1, 2, \ldots$.

In every state $i$ we can impose a probability distribution on the choices we have. We can write this as a nonnegative row vector

$$\mathbf{f}(i) = (f(i,1),\ f(i,2), \ldots, f(i, m(i))),\ \text{with} \sum_{a=1}^{m(i)} f(i,a) = 1.$$

We can now define a *strategy* as the block row vector

$$\mathbf{f} = (\mathbf{f}(1), \ldots, \mathbf{f}(i), \ldots, \mathbf{f}(N)).$$

A strategy will be called *pure* or *deterministic* if for all $i \in \mathbf{S}$ there is an $a_i \in \mathbf{A}(i)$ such that $f(i, a_i) = 1$ and $f(i, a) = 0$ for all $a \neq a_i$. In such a case we will often shorten $f(s, a)$ to $f(i) = a_i$. Other strategies we will call *mixed*. If the controller's decision in state $i$ is not dependent on the history of the process, i.e. if the controller's decision in state $i$ is invariant with respect to the time of visit to $i$, a strategy is called *stationary*.

Every strategy $\mathbf{f}$ defines a *probability transition matrix*

$$P(\mathbf{f}) = (p_{ij}(\mathbf{f}))_{i,j=1}^{N} \tag{1.2}$$

with entries given by

$$p_{ij}(\mathbf{f}) = \sum_{a=1}^{m(i)} p_{ij}(a) f(i,a). \tag{1.3}$$

Under the most natural assumption the process has to move into one of the states of $\mathbf{S}$ at every transition, that is

$$\sum_{j=1}^{N} p_{ij}(a) = 1 \qquad \text{for all } a \in \mathbf{A}(i),\ i \in \mathbf{S}. \tag{1.4}$$

This means every stochastic matrix $P(\mathbf{f})$ uniquely defines a Markov Chain.

Now that we have established the definition of a strategy, a natural question is how to compare two different strategies. To do this however we first need to have a way to evaluate them. For every strategy we have a sequence of random rewards/outputs $\{R_t\}_{t=0}^{\infty}$, with $R_t$ being the reward for the period $[t, t+1)$. Once the initial state $i$ and a strategy $\mathbf{f}$ are specified, then so is the probability distribution of $R_t$ for every $t = 0, 1, 2, \ldots$. The expectation of $R_t$ is well defined and will be denoted by

$$\mathbb{E}_{i\mathbf{f}}[R_t] := \mathbb{E}_{\mathbf{f}}[R_t | S_0 = i]. \tag{1.5}$$

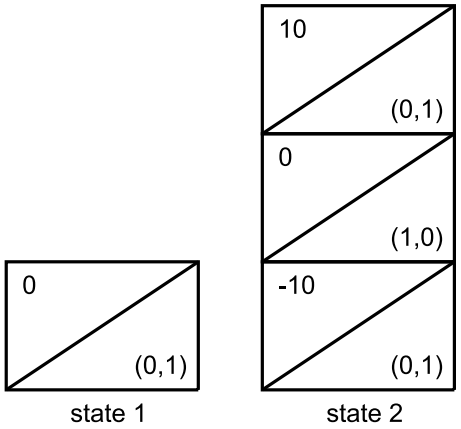
We will now discuss two ways of evaluating a strategy, the $\beta$-discounted Markov decision model and the terminating Markov decision model.

### 1.2.1 Total Reward Markov Decision Model $\Gamma_\sigma$

The most obvious way of trying to evaluate a strategy is to sum up the rewards in every stage. Consider a Markov decision process over an infinite time horizon in which the value of a stationary strategy $\mathbf{f} \in \mathbf{F}_S$ from a starting state $i \in \mathbf{S}$ is defined by

$$v_\sigma(i, \mathbf{f}) := \sum_{t=0}^{\infty} \mathbb{E}_{i\mathbf{f}}(R_t). \tag{1.6}$$

The above is called the *total reward criterion*, and we will denote the corresponding model by $\Gamma_\sigma$. The problem with this model however is that summability is not guaranteed. For example, let $S = \{1, 2\}$, $A(1) = 1$, $A(2) = 3$. The transitional probabilities and rewards are represented by the following picture:




In this representation a box corresponds to an action in a state and its reward/transitions, with the rewards above the diagonal divider, and the

transitions below it. Taking $\mathbf{f}_1 = (1,(0,0,1))$, $\mathbf{f}_2 = (1,(0,1,0))$ and $\mathbf{f}_3 = (1,(1,0,0))$, it can now easily be seen that $\mathbf{f}_1 = -\infty$, $\mathbf{f}_2 = 0$ and $\mathbf{f}_3 = \infty$.

In order to evaluate the stream of expected rewards resulting from the use of strategy $\mathbf{f}$, define the *immediate reward vector* by

$$\mathbf{r}(\mathbf{f}) = (r(1,\mathbf{f}), r(2,\mathbf{f}), \ldots, r(N,\mathbf{f}))^T \tag{1.7}$$

where, for each $i \in S$,

$$r(i,\mathbf{f}) := \sum_{a \in A(i)} r(i,a) f(i,a). \tag{1.8}$$

We now have for arbitrary $i \in S$

$$\begin{aligned}
\mathbb{E}_{i\mathbf{f}}[R_0] &= r(i,\mathbf{f}) = [\mathbf{r}(\mathbf{f})]_i \\
\mathbb{E}_{i\mathbf{f}}[R_1] &= \sum\nolimits_{j=1}^{N} p_{ij}(\mathbf{f}) r(j,\mathbf{f}) = [P(\mathbf{f})\mathbf{r}(\mathbf{f})]_i \\
\mathbb{E}_{i\mathbf{f}}[R_2] &= \sum\nolimits_{j=1}^{N} p_{ij}^{(2)}(\mathbf{f}) r(j,\mathbf{f}) = [P^2(\mathbf{f})\mathbf{r}(\mathbf{f})]_i \\
&\vdots \\
\mathbb{E}_{i\mathbf{f}}[R_t] &= \sum\nolimits_{j=1}^{N} p_{ij}^{(t)}(\mathbf{f}) r(j,\mathbf{f}) = [P^t(\mathbf{f})\mathbf{r}(\mathbf{f})]_i,
\end{aligned} \tag{1.9}$$

where $[\mathbf{u}]_i$ denotes the $i$-th entry of a vector $\mathbf{u}$, and $p_{ij}^{(t)}(\mathbf{f})$ is the $t$-step transition probability from $i$ to $j$ in the Markov chain defined by $\mathbf{f}$. Using (1.6) we now have

$$\mathbf{v}_\sigma(i,\mathbf{f}) = \sum_{t=0}^{\infty} \left( P^{(t)}\mathbf{r}(\mathbf{f}) \right)_i. \tag{1.10}$$

From Markov chain theory we know that the $t$-th power of $P(\mathbf{f})$ contains all such $t$-step transition probabilities, that is,

$$P^t(\mathbf{f}) = (p_{ij}^{(t)}(\mathbf{f}))_{i,j=1}^{N}.$$

To ensure summability we have to pose some restrictions on the stochastic game. Suppose that for every strategy $\mathbf{f}$ and every pair of states $(i,j) \in \mathbf{S}\times\mathbf{S}$ we have that

$$\sum_{t=1}^{\infty} p_{ij}^{(t)}(\mathbf{f}) < \infty.$$

Then the model $\Gamma_\sigma$ will be called a *transient total reward MDP*. In the transient model $\Gamma_\sigma$ for every $\mathbf{f} \in \mathbf{F}_S$ we have that

$$\mathbf{v}_\sigma(\mathbf{f}) = [I - P(\mathbf{f})]^{-1}\, \mathbf{r}(\mathbf{f}).$$

We will use the following arguments to show this. If we can show that $[I - P(\mathbf{f})]$ is an invertible matrix, and that

$$[I - P(\mathbf{f})]^{-1} = I + P(\mathbf{f}) + P^2(\mathbf{f}) + \ldots$$

we are done. We know that $\sum_{t=1}^{\infty} p_{ij}^{(t)}(\mathbf{f}) < \infty$, so $\lim_{t\to\infty} p_{ij}^{(t)}(\mathbf{f}) = 0$. Furthermore,

$$\lim_{t\to\infty}[I - P(\mathbf{f})][I + P(\mathbf{f}) + P^2(\mathbf{f}) + \ldots + P^{t-1}(\mathbf{f})] = \lim_{t\to\infty}[I - P^t(\mathbf{f})] = I.$$

Since $\det(\mathrm{I}) \neq 0$, this means $\det[I - P^t(\mathbf{f})] \neq 0$ for $t$ sufficiently large. From this we can conclude that $\det[I - P(\mathbf{f})] \neq 0$, so $[I - P(\mathbf{f})]$ is nonsingular, and

$$[I - P(\mathbf{f})]^{-1} = \sum_{t=0}^{\infty} P^t(\mathbf{f}). \tag{1.11}$$

Since we have (1.10) we know that

$$\mathbf{v}(\mathbf{f})_\sigma = \sum_{t=0}^{\infty} P^t(\mathbf{f})\mathbf{r}(\mathbf{f}) \tag{1.12}$$

This gives us:

$$\mathbf{v}_\sigma(\mathbf{f}) = [I - P(\mathbf{f})]^{-1}\mathbf{r}(\mathbf{f})$$

Another possible restriction is that the transition probabilities are such that scalars $\mu_1, \mu_2, \ldots, \mu_N > 0$ and $\gamma \in [0, 1)$ exist satisfying

$$\sum_{j=1}^{N} p_{ij}(a)\mu_j \leq \gamma\mu_i$$

for all $a \in \mathbf{A}(i)$, $i, j \in \mathbf{S}$. Then the model $\Gamma_\sigma$ will be called a *contracting total reward MDP*. If $\Gamma_\sigma$ is contracting, then it is also transient. Let $\mathbf{m} = (\mu_1, \mu_2, \ldots, \mu_N)$. Multiplying over $f(i, a)$ and summing over $a$ gives us:

$$\sum_{a\in A(i)} \sum_{j=1}^{N} p_{ij}(a)f(i, a)\mu_j \leq \gamma\mu_i.$$

(Remember that $\sum_{a\in A(i)} f(i, a) = 1$). Written in matrix notation this becomes:

$$P(\mathbf{f})\mathbf{m} \leq \gamma\mathbf{m}.$$

From this we can conclude that for every $t > 0$

$$P^{(t)}(\mathbf{f})\mathbf{m} \leq \gamma^t\mathbf{m}$$

and

$$\sum_{t=0}^{\infty} P^{(t)}(\mathbf{f})\mathbf{m} \leq \sum_{t=0}^{\infty} \gamma^t\mathbf{m} = \frac{\mathbf{m}}{1-\gamma}.$$

This implies

$$\sum_{t=0}^{\infty} p_{ij}^{(t)}(\mathbf{f}) < \infty, \quad \text{for all } i, j \in S.$$

### 1.2.2 $\beta$-Discounted Markov Decision Model $\Gamma_\beta$

A very important way to ensure summability is to use the overall "discounted value" of strategy $\mathbf{f}$ from the initial state $i$. This will be defined by

$$v_\beta(i, f) := \sum_{t=0}^{\infty} \beta^t \, \mathbb{E}_{i\mathbf{f}}[R_t], \tag{1.13}$$

where $\beta \in [0, 1)$ is called the *discount factor*. The model that uses (1.13) as its performance criterion will be called the *discounted Markov decision process* (or DMD, for short). In this model $\beta$ ensures the summability of the series. In economical terms, $\beta$ can be thought of as some sort of inflation correction. That this model yields a result for all $\beta \in [0, 1)$ can be see by the following. Take $D = \max_{i \in \mathbf{S}, a \in \mathbf{A}(i)} |r(i, a)|$. This results in

$$|v_\beta(i, f)| \leq D \sum_{t=0}^{\infty} \beta^t = \frac{D}{1 - \beta}$$

for all $i \in \mathbf{S}$ and $\mathbf{f} \in \mathbf{F}_S$. If $\mathbf{v}_\beta(\mathbf{f}) := (v_\beta(1, \mathbf{f}), \ldots, v_\beta(N, \mathbf{f}))^T$ we now have analogous to the Total Reward MDM

$$\mathbf{v}_\beta(\mathbf{f}) = \sum_{t=0}^{\infty} \beta^t P^t(\mathbf{f}) \mathbf{r}(\mathbf{f}), \tag{1.14}$$

where $P^0(\mathbf{f}) := I_N$, the $N \times N$ identity matrix. Similar to the derivation of (1.11)) we obtain that

$$[I - \beta P(\mathbf{f})]^{-1} := I + \beta P(\mathbf{f}) + \beta^2 P^2(\mathbf{f}) + \ldots \tag{1.15}$$

Substituting the above into (1.14) we obtain the following compact matrix expression for the (discounted) *value vector of* $\mathbf{f}$ (which will also be referred to as the *value of* $\mathbf{f}$):

$$\mathbf{v}_\beta(\mathbf{f}) = [I - \beta P(\mathbf{f})]^{-1} \mathbf{r}(\mathbf{f}). \tag{1.16}$$

The discounted model $\Gamma_\beta$ can be regarded as a special case of contracting $\Gamma_\sigma$. Redefine the transition probabilities by

$$\overline{p}_{ij}(a) := \beta p_{ij}(a)$$

for all $a \in \mathbf{A}(i), i \in \mathbf{S}, j \in \mathbf{S}$, and $\mu_i = 1$ for $i = 1, 2, \ldots, N$ and $\gamma = \beta$. This gives us

$$\sum_{j=1}^{N} \overline{p}_{ij}(a) = \sum_{j=1}^{N} \beta p_{ij}(a) = \beta \sum_{j=1}^{N} p_{ij}(a) \leq \beta = \gamma \mu_i.$$

Now that we have a way to evaluate an arbitrary stationary strategy/control $\mathbf{f}$, we can define the corresponding "optimal control problem"

*Find, if possible, a strategy* $\mathbf{f}^0$ *that "maximizes"* $\mathbf{v}(\mathbf{f})$.

In order to make the above optimization problem precise, we first must define its feasible region, that is, the *space of stationary strategies or controls* that will from now on be denoted by $\mathbf{F}_S$. We can view this space as the polyhedron

$$\mathbf{F}_S := \left\{ \mathbf{f} = (\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(N) \mid f(i,a) \geq 0 \right. \\ \left. \text{and } \sum_{a=1}^{m(i)} f(i,a) = 1, \text{ for all } a \in A(i),\ i \in \mathbf{S} \right\}.$$

If $\mathbf{u}$,$\mathbf{v} \in \mathbb{R}^n$, then $\mathbf{u} \geq \mathbf{v}$ if and only if $[\mathbf{u}]_i \geq [\mathbf{v}]_i$ for every $i$th entry. The corresponding component-wise strict inequality and vector maximum/minimum will have analogous interpretations. We can now state the *discounted optimal Markov control problem* as

$$\max \mathbf{v}_\beta(\mathbf{f})$$

subject to:

$$\mathbf{f} \in \mathbf{F}_S$$

It is far from clear that this component-wise maximum exists. In later sections we shall prove the existence and demonstrate two well-known algorithms for computing the component-wise maximum and a corresponding optimal control.

### 1.2.3 Terminating Markov Decision Model $\Gamma_\tau$

Finally we consider a model in which

$$\sum_{j=1}^{N} p_{ij}(a) < 1 \text{ for all } a \in A(i), i \in \mathbf{S}. \tag{1.17}$$

This is a relaxation of the assumption made in the previous model. This assumption has the interpretation that with every action $a \in A(i)$ selected in every state $i$, there is a *positive stopping probability* of

$$1 - \sum_{j=1}^{N} p_{ij}(a) > 0$$

that signifies the "termination" of the process. The transition matrix $P(\mathbf{f})$ has the property that

$$\sum_{j=1}^{N} p_{ij}(\mathbf{f}) < 1 \text{ for all } i \in \mathbf{S}. \tag{1.18}$$

It follows in a manner analogous to that in the DMD model that the (terminating) *value vector of* $\mathbf{f}$ is

$$\begin{aligned} \mathbf{v}_\tau &:= \sum_{t=0}^{\infty} P^t(\mathbf{f})\mathbf{r}(\mathbf{f}) \\ &= [I - P(\mathbf{f})]^{-1}\mathbf{r}(\mathbf{f}). \end{aligned}$$

The corresponding *terminating optimal Markov control problem* is the optimization problem

$$\max \mathbf{v}_\tau(\mathbf{f})$$

subject to

$$\mathbf{f} \in \mathbf{F}_S.$$

## 1.3 The Finite Horizon Markov Decision Process

In the previous section we assumed that the time horizon of the Markov decision models was infinite. A logical question is what happens when the time horizon is finite? In this section we will discuss the *finite horizon Markov process* $\Gamma_T$ in which we assume that the stages are indexed by the time variable $t \in \{0, 1, 2, \ldots, T\}$. In the development of stochastic games the Finite Horizon model has not received much attention. There are two practical reasons for this:

1. When the horizon $T$ is "short" there is an elegant solution, which will presented later on.

2. When the horizon $T$ is "long" the computational complexity of the problem becomes too big.

In the finite case we will have to extend the notion of a strategy beyond the class of stationary strategies considered so far. A decision might be unfavourable in the early stages because it might lead to an unfavourable state later on, but favourable in the later stages, because there is not enough time left to reach the unfavourable state. This means the worth of an action is now a function of the time left before termination, and hence an optimal control also should be time dependent.

This leads us to extend the notion of a strategy/control to that of a finite sequence

$$\pi = (f_0, f_1, \ldots, f_T)$$

such that $f_t \in \mathbf{F}_T$. We shall denote the set of all such strategies by $\mathbf{F}_M^T$ and call it the set of *Markov strategies* of the $T$-horizon Markov decision process.

Now, for every $\pi \in \mathbf{F}_M^T$ the expectation $\mathbb{E}_{i\pi}[R_t] := \mathbb{E}_\pi[R_t|S_0 = i]$ is well defined for each $t = 0, 1, 2, \ldots, T$, and hence so is the *T-stage value of* $\pi$

$$v_T(i, \pi) := \sum_{t=0}^{T} \mathbb{E}_{i\pi}[R_t] \tag{1.19}$$

for every initial state $i \in S$. The corresponding $T$-stage *value vector of* $\pi$ is the vector $\mathbf{v}_T(\pi)$ whose entries are $v_T(i, \pi)$ for $i = 1, 2, \ldots, N$, and the related $T$-stage Markov decision process will be denoted by $\Gamma_T$. The optimization problem in this case is

$$\max \ \mathbf{v}_T(\pi)$$

subject to:

$$\pi \in \mathbf{F}_M^T.$$

If we now assume that an optimal control and optimal payoff with $(n-1)$ stages to go are known, then with $n$ stages to go all that we needs to do is to maximize the sum of the immediate reward and the maximal expected payoff for the remainder of the process with $(n-1)$ stages to go. This idea is called the principle of dynamic programming and is reflected in the following algorithm.

In the following $\text{argmax}_{z\in\mathbb{Z}}\ h(z)$ denotes the value $z$ for which a real valued function $h(z)$ over $\mathbb{Z}$ attains its maximum.

**Algorithm 1.3.1.** The Backward Recursion of Dynamic Programming.

**Step 1.** (Initiation) Set $V_{-1}(i) = 0$ for all $i \in \mathbf{S}$ and define

$$f_T^*(i) := a_i^T = \underset{a\in A(i)}{\text{argmax}} \left\{ r(i, a) + \sum_{j=1}^{N} p_{ij}(a) V_{-1}(j) \right\}$$

and

$$V_0(i) := r(i, a_i^T) = \max_{a\in A(i)} \{r(i, a) + 0\}.$$

**Step 2.** (Recursion) For each $n = 1, 2, \ldots, T$ calculate for each $i \in \mathbf{S}$

$$f_{T-n}^*(i) := a_i^{T-n} = \text{argmax}_{A(i)} \left\{ r(i, a) + \sum_{j=1}^{N} p_{ij}(a) V_{n-1}(j) \right\}$$

and

$$V_n(i) := r(i, a_i^{T-n}) + \sum_{j=1}^{N} p_{ij}(a_i^{T-n}) V_{n-1}(j).$$

**Step 3.** Construct the strategy

$$\pi^* = (f_0^*, f_0^*, \ldots, f_T^*) \in \mathbf{F}_M^T.$$

For each $i \in S$, every $\pi \in \mathbf{F}_M^T$ and $n = 0, 1, \ldots, T$ define

$$V_n(i, \pi) := \sum_{t=T-n}^{T} \mathbb{E}_\pi[R_t | S_{T-n} = i].$$

$V_n(i, \pi)$ represents the expected reward over the last $n$ stages given that the state at time $(T - n)$ is $i$. Note that when $n = T$, this is simply the total expected reward, that is

$$V_T(i, \pi) = v_T(i, \pi)$$

for all $i \in \mathbf{S}$ and $\pi \in \mathbf{F}_M^T$.

**Theorem 1.3.1.** *Consider the $T$-horizon Markov decision process $\Gamma_T$, and let $\pi^* \in \boldsymbol{F}_M^T$ be a strategy constructed by the dynamic programming algorithm 2.1. Then $\pi^*$ is an optimal strategy over $\boldsymbol{F}_M^T$, and for all $n = 0, 1, \ldots, T$ and $i \in S$*

$$V_n(i) = \max_{A(i)} \left\{ r(i, a) + \sum_{i=1}^{N} p_{ij}(a) V_{n-1}(j) \right\}. \tag{1.20}$$

For proofs of theorems in this paper we refer from now on to [6], unless stated otherwise. The equation (1.20) is sometimes referred to as *optimality equation of dynamic programming* and is regarded as the most fundamental mathematical tool for the analysis of Markov decision processes. In section 1.7 and 1.8 we will see that we can restrict the search for optimal strategies to this class without loss of generality.

## 1.4 Linear Programming and the Summable Markov Decision Models

In the previous section we introduced the $\beta$-discounted Markov decision process $\Gamma_\beta$ and formulated the related optimal control problem:

$$\max \ \mathbf{v}_\beta(\mathbf{f})$$

subject to:

$$\mathbf{f} \in \mathbf{F}_S.$$

A control/strategy $f^0 \in \mathbf{F}_S$ that achieves the maximum will be called an *optimal control/strategy*, and the corresponding value vector of $\mathbf{f}^0$

$$\mathbf{v}_\beta := \mathbf{v}_\beta(\mathbf{f}^0) = \max_{\mathbf{f}} \ \mathbf{v}_\beta(\mathbf{f}) \tag{1.21}$$

will be called the (discounted) *value vector of the process* $\Gamma_\beta$.

The question now remains whether an optimal $\mathbf{f}^0$ and corresponding $\mathbf{v}_\beta$ exist. The following will show that this is indeed the case, and that they correspond to optimal solutions of suitably constructed linear programs. We shall begin by characterizing some of the properties that an optimal strategy and the discounted value vector might be expected to possess.

Multiplying (1.16) by $[I - \beta P(\mathbf{f})]$ gives us

$$\mathbf{v}_\beta(\mathbf{f}) = \mathbf{r}(\mathbf{f}) + \beta P(\mathbf{f})\mathbf{v}_\beta(\mathbf{f}). \tag{1.22}$$

This means that any optimal strategy $\mathbf{f}^0 \in \mathbf{F}_S$ and $\mathbf{v}_\beta$ that may exist, must also satisfy this equation:

$$\mathbf{v}_\beta = \mathbf{r}(\mathbf{f}^0) + \beta P(\mathbf{f}^0)\mathbf{v}_\beta. \tag{1.23}$$

Furthermore, if $\mathbf{v}_\beta$, the discounted value vector, exists, then for each $i \in \mathbf{S}$ it satisfies

$$v_\beta(i) = \max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j=1}^{N} p_{ij}(a) v_\beta(j) \right\}, \tag{1.24}$$

where $\mathbf{v}_\beta(i)$ is the $i$th entry of $\mathbf{v}_\beta$.

This suggests that the value vector should satisfy the following set of linear inequalities expressed in terms of some arbitrary variable vector $\mathbf{v} = (v(1), \ldots, v(N))^T$:

$$v(i) \geq r(i,a) + \beta \sum_{j=1}^{N} p_{ij}(a) v(j) \tag{1.25}$$

for all $a \in A(i)$, $i \in \mathbf{S}$. However, if for an arbitrary $\mathbf{f} \in \mathbf{F}_S$ we multiply each of the above inequalities by the corresponding entry $f(i,a)$ of $\mathbf{f}$ and sum over all the $a \in A(i)$, then we shall obtain for each $i \in \mathbf{S}$

$$v(i) \geq r(i,\mathbf{f}) + \beta \sum_{j=1}^{N} p_{ij}(\mathbf{f}) v(j), \tag{1.26}$$

or in matrix form

$$\mathbf{v} \geq \mathbf{r}(\mathbf{f}) + \beta P(\mathbf{f})\mathbf{v}. \tag{1.27}$$

When we substitute the inequality above into itself $k$ times we obtain

$$\mathbf{v} \geq \mathbf{r}(\mathbf{f}) + \beta P(\mathbf{f})\mathbf{r}(\mathbf{f}) + \cdots + \beta^{k-1} P^{k-1}(\mathbf{f})\mathbf{r}(\mathbf{f}) + \beta^k P^k(\mathbf{f})\mathbf{v} \tag{1.28}$$

which for $k \to \infty$ yields

$$\mathbf{v} \geq [I - \beta P(\mathbf{f})]^{-1}\mathbf{r}(\mathbf{f}) = \mathbf{v}_\beta(\mathbf{f}). \tag{1.29}$$

We see that an arbitrary vector $\mathbf{v}$ satisfying the system of linear inequalities (1.25) is an upper bound on the discounted value vector due to any stationary strategy $\mathbf{f}$. This suggests that the discounted value vector of the process $\Gamma_\beta$ might be the optimal solution of the linear program

$$\min \sum_{i=1}^{N} \frac{1}{N} v(i)$$

subject to

$$v(i) \geq r(i,a) + \beta \sum_{i=1}^{N} p_{ij}(a) v(j), \text{ for } a \in A(i), \ i \in \mathbf{S}. \tag{1.30}$$

The coefficients $\frac{1}{N}$ in the objective function of (1.30) can be interpreted as the equal probabilities that the process $\Gamma_\beta$ begins in any given state.

If we regard the problem (1.30) as a primal linear program, and associate with each constraint a dual variable $x_{ia}$, then the dual linear program will be of the form:

$$\max \sum_{i=1}^{N} \sum_{a=1}^{m(i)} r(i,a) x_{ia}$$

subject to

$$\sum_{i=1}^{N} \sum_{a=1}^{m(i)} [\delta(i,j) - \beta p_{ij}(a)] x_{ia} = \frac{1}{N}, \ j \in \mathbf{S} \tag{1.31}$$

$$x_{ia} \geq 0, \ a \in A(i), \ i \in S,$$

where $\delta(i,j)$ is the Kronecker delta. Here $x_i = \left(x_{i1}, \ldots, x_{im(i)}\right)$ can be seen as some sort of frequency vector. As we can see in the following theorem, normalizing this will give us a strategy.

We will now state the main result connecting linear programming with the discounted Markov decision model.

**Theorem 1.4.1.**

1. *The primal-dual linear programs (1.30) and (1.31) possess finite optimal solutions.*
2. *Let $\mathbf{v}^0 = (v^0(1), v^0(2), \ldots, v^0(N))^T$ be an optimal solution of (1.30); then $\mathbf{v}^0 = \mathbf{v}_\beta$, the value vector of the process $\Gamma_\beta$.*
3. *Let $\boldsymbol{x}^0 = \left\{x_{ia}^0 | a \in A(i), i \in \mathbf{S}\right\}$ be an optimal solution of (1.31) and define $x_i^0 := \sum_{a=1}^{m(i)} x_{ia}^0$ for each $i \in \mathbf{S}$; then $x_i^0 > 0$, and the strategy $\mathbf{f}^0 \in \mathbf{F}_S$ defined by*

$$f^0(i,a) := \frac{x_{ia}^0}{x_i^0} \text{ for all } a \in A(i), \ i \in \mathbf{S} \tag{1.32}$$

*is an optimal stationary strategy in the process $\Gamma_\beta$.*

**Corollary 1.4.1.** *(Validity of Optimality Equation)*

1. *The value vector* $\mathbf{v}_\beta$ *is the unique solution of the optimality equation (1.24).*
2. *For each* $i \in \mathbf{S}$ *select any one action* $a_i \in A(i)$ *that achieves the maximum in (1.24), that is,*

$$\begin{aligned} v(i) &= r(i,a_i) + \beta \sum_{j=1}^{N} p_{ij}(a_i) v(j) \\ &= \max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j=1}^{N} p_{ij}(a) v(j) \right\}, \end{aligned}$$

*where* $\mathbf{v}$ *is the solution of (1.24). Define* $\mathbf{f}^* \in \mathbf{F}_S$ *by*

$$f^*(i,a) = \begin{cases} 1 & \text{if } a = a_i \\ 0 & \text{otherwise} \end{cases}$$

*for each* $i \in \mathbf{S}$. *Then* $\mathbf{f}^*$ *is an optimal deterministic strategy in* $\Gamma_\beta$,

**Remark 1.4.1.** From the corollary follows that the problem of finding an optimal strategy is straightforward once the value vector $\mathbf{v}_\beta$ is known since it requires only the computation of $N$ maxima specified in part (ii). This leads to a family of algorithms, for the calculation of the value vector, that has come to be known as the "methods of successive approximations".

**Remark 1.4.2.** The analysis for the discounted model is also valid for the terminating model. Notationally, the main change is the omission of the discount factor $\beta$ from all of the relevant equations. For instance, the primal linear program $(P_\beta)$ is replaced by the analogous primal problem

$$\min \sum_{i=1}^{N} \frac{1}{N} v(i) \tag{1.33}$$

subject to:

$$v(i) \geq r(i,a) + \sum_{j=1}^{N} p_{ij}(a) v(j), \;\; a \in A(i), i \in \mathbf{S}. \tag{1.34}$$

## 1.5 The Irreducible Limiting Average Process

A more longterm performance criterion is the *limiting average Markov decision process*. We define the *limiting average value* of the stationary strategy $\mathbf{f}$ from the initial state $i$ as

$$v_\alpha(i,\mathbf{f}) := \lim_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{i\mathbf{f}}[R_t]. \tag{1.35}$$

The associated model will be called the *limiting average Markov decision process* (or AMD for short) and will be denoted by $\Gamma_\alpha$.

Analogous to the $\beta$-discounted MDP we have the (limiting average) *value vector of* $\mathbf{f}$ defined as

$$\mathbf{v}_\alpha := (\mathbf{v}_\alpha(1,\mathbf{f}), \mathbf{v}_\alpha(2,\mathbf{f}), \ldots, \mathbf{v}_\alpha(N,\mathbf{f}))^T, \tag{1.36}$$

and an associated optimal control problem:

$$\max \ \mathbf{v}_\alpha(\mathbf{f})$$

subject to:

$$\mathbf{f} \in \mathbf{F}_S. \tag{1.37}$$

An optimal control/strategy $\mathbf{f}^0$ will achieve the maximum in (1.37), and the vector

$$\mathbf{v}_\alpha := \mathbf{v}_\alpha(\mathbf{f}^0) = \max_f \mathbf{v}_\alpha(f) \tag{1.38}$$

will be called the (limiting average) *value vector of the process* $\Gamma_\alpha$.

Due to $\mathbb{E}_{i\mathbf{f}}[R_t] = [P^t(\mathbf{f})\mathbf{r}(\mathbf{f})]_i$ for each $i \in \mathbf{S},\ \mathbf{f} \in \mathbf{F}_S$ (see (1.9)), and the existence of a Markov matrix $Q(\mathbf{f})$ such that

$$Q(\mathbf{f}) := \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f})$$

we can also write (1.35) in vector form as

$$\begin{aligned} \mathbf{v}_\alpha(\mathbf{f}) &= \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f})\mathbf{r}(\mathbf{f}) \\ &= \left[\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f})\right] \mathbf{r}(\mathbf{f}) \\ &= Q(\mathbf{f})\mathbf{r}(\mathbf{f}). \end{aligned} \tag{1.39}$$

The matrix $Q(\mathbf{f})$ is sometimes called the *Cesaro-limit matrix* of $P(\mathbf{f})$, or a *stationary distribution matrix* of the Markov chain determined by $P(\mathbf{f})$. From this follows that the (limiting average) value vector of any $\mathbf{f} \in \mathbf{F}_S$ is given by (1.39).

A problem that may arise in this model is the existence of one or more absorbing states. Depending on the stragegy it is possible that the process might get trapped in one of these states. This means we have to somehow balance these absorption probabilities when evaluating the model. To avoid these problems will restrict ourselves to irreducible Markov chains for now. We will assume that for every $\mathbf{f} \in \mathbf{F}_S$ the probability transition matrix $P(\mathbf{f})$ determines an *irreducible* (or *completely ergodic*) Markov chain. Irreducibility means the process will visit every state infinitely often, regardless of

the choice of $\mathbf{f}$. Or, in mathematical terms, it means that for every pair of states $(i, j)$ there exists some positive integer $t$ such that $[P^t(\mathbf{f})]_{ij} > 0$. The following lemma states a property crucial to the arguments used in this section.

**Lemma 1.5.1.** *Let $P$ be the probability transition matrix of an irreducible Markov chain and $Q$ be the corresponding Cesaro-limit matrix. Then*

*(i) $Q$ has identical rows.*

*(ii) Let $\mathbf{q} = (q_1, \ldots, q_N)$ be a row of $Q$. Then every entry of $\mathbf{q}$ is strictly positive, and $\mathbf{q}$ is the unique solution of the linear system of equations:*

$$\mathbf{q}P = \mathbf{q}$$
$$\mathbf{q}\mathbf{1} = 1,$$

*where $\mathbf{1}$ is an $N$-dimensional column vector with unity in every entry. The vector $\mathbf{q}$ is called the "stationary distribution" of the irreducible Markov chain.*

*Proof.* We will use the Perron-Frobenius theorem [10]:

**Theorem 1.5.1** (Perron-Frobenius)**.** *Let $A$ be an irreducible $n \times n$ matrix with nonnegative entries $a_{ij}$. Then the following statements hold:*

*(i) there is a real eigenvalue $r$ of $A$ such that any other eigenvalue $\lambda$ satisfies $|\lambda| < r$.*

*(ii) the eigenvalue $r$ is simple: $r$ is a simple root of the characteristic polynomial of $A$. In particular both the right and left eigenspace associated to $r$ are 1-dimensional.*

*(iii) there is a left (respectively right) eigenvector associated with $r$ having positive entries. This means that a row-vector $v = (v_1, \ldots, v_n)$ and a column-vector $w = (w_1, \ldots, w_n)^t$ exist with positive entries $v_i > 0, w_i > 0$ such that $vA = rv, Aw = rw$. The vector $v$ (resp. $w$) is then called a left (resp. right) eigenvector associated with $r$. In particular two uniquely determined left (resp. right) positive eigenvectors exist associated with $r$ (sometimes also called "stochastic" eigenvectors) vnorm and wnorm such that $\sum_i v_i = \sum_i w_i = 1$.*

*(iv) one has the eigenvalue estimate $\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}$*

Since $Pe = e$, we know an eigenvalue with $\lambda = 1$ exists. Suppose now an eigenvalue $\lambda > 1$ exists. This means:

$$\lim_{t \to \infty} P^t x = \lambda^t x \to \infty.$$

Furthermore $||P^t x|| \leq ||P||^t ||x|| = ||x|| < \infty$. This leads to a contradiction, so $\lambda = 1$ is maximal and unique. Applying part (iii) Perron-Frobenius now gives the required result. □

Now consider the irreducible Markov chain determined by some fixed $\mathbf{f} \in \mathbf{F}_S$ via the transition matrix $P(\mathbf{f})$. Let $\mathbf{q}(\mathbf{f})$ be its stationary distribution as defined in the previous lemma. For each $a \in A(i)$, $i \in \mathbf{S}$, define

$$x_{ia}(\mathbf{f}) := q_s(\mathbf{f}) f(i,a) \tag{1.40}$$

and

$$x_i(\mathbf{f}) := \sum_{a \in A(i)} x_{ia}(\mathbf{f}) = q_i(\mathbf{f}), \tag{1.41}$$

where the last equality follows from the fact that $\sum_{a \in A(i)} f(i,a) = 1$. Since $q_i(\mathbf{f})$ normally is interpreted as the long-run proportion of visits to state $i$, we shall call $x_i(\mathbf{f})$ the *long-run frequency of visits to state* $s$ and $x_{ia}(\mathbf{f})$ the *long-run frequency of the state-action pair* $(i,a)$, induced by the control $\mathbf{f}$. Furthermore, define the *long-run (state-action) frequency vector* $\mathbf{x}(\mathbf{f})$ induced by $\mathbf{f}$ as the block-column vector whose $i$th block is

$$\mathbf{x}_i(\mathbf{f}) = (x_{i1}(\mathbf{f}), x_{i2}(\mathbf{f}), \ldots, x_{im(i)}(\mathbf{f}))^T.$$

Analogously, the *long-run state frequency vector* induced by $\mathbf{f}$ will be the row $N$-vector

$$\overline{\mathbf{x}}(\mathbf{f}) = (x_1(\mathbf{f}), x_2(\mathbf{f}), \ldots, x_N(\mathbf{f})).$$

Since lemma 1.5.1 states that $\mathbf{q}(\mathbf{f})P(\mathbf{f}) = \mathbf{q}(\mathbf{f})$, which implies that $\mathbf{q}(\mathbf{f})\,[I - P(\mathbf{f})] = 0$, we can rewrite this as follows:

$$\begin{aligned}
\sum_{i=1}^{N} (\delta(i,j) - p_{ij}(\mathbf{f})) q_i(\mathbf{f}) &= 0,\ j \in \mathbf{S}, \iff \\
\sum_{i=1}^{N} \sum_{a \in A(i)} (\delta(i,j) - p_{ij}(a)) q_i(\mathbf{f}) f(i,a) &= 0,\ j \in \mathbf{S}, \iff \\
\sum_{i=1}^{N} \sum_{a \in A(i)} (\delta(i,j) - p_{ij}(a)) x_{ia}(\mathbf{f}) &= 0,\ j \in \mathbf{S},
\end{aligned}$$

where $\delta(ij)$ is the Kronecker delta. Furthermore, since

$$\sum_{i=1}^{N} \sum_{a \in A(i)} x_{ia}(\mathbf{f}) = \sum_{i=1}^{N} \sum_{a \in A(i)} q_i(\mathbf{f}) f(i,a) = \sum_{i=1}^{N} q_i(\mathbf{f}) = 1$$

we consider the polyhedral set $\mathbf{X}$ defined by the linear constraints

(i) $\sum_{i=1}^{N} \sum_{a \in A(i)} (\delta(i,j) - p_{ij}(a)) x_{ia} = 0,\ j \in \mathbf{S}$

(ii) $\sum_{i=1}^{N} \sum_{a \in A(i)} x_{ia} = 1$

(iii) $x_{ia} \geq 0,\ a \in A(i),\ i \in \mathbf{S}.$

Defining $m := \sum_{i=1}^{N} m(i)$, we can write this in matrix notation as

$$\mathbf{X} = \left\{\mathbf{x} | W\mathbf{x} = 0,\ \mathbf{1}^T\mathbf{x} = 1,\ \mathbf{x} \geq 0\right\},$$

where $\mathbf{x}$, $\mathbf{1}$ are both $m$-vectors, and $W$ is an $N \times m$ matrix whose $(j,(i,a))$-th entry is

$$w_{j(i,a)} := \delta(i,j) - p_{ij}(a).$$

It is important to note that (1.40) defines a map of the strategy space $M : \mathbf{F}_S \to \mathbb{R}^m$ with

$$M(\mathbf{f}) := \mathbf{x}(\mathbf{f}). \tag{1.42}$$

We will now show that the set $\mathbf{X}$ is the "frequency space" of $\Gamma_\alpha$. If we can also prove that $\mathbf{X}$ is in 1:1 correspondence with the space of stationary strategies $\mathbf{F}_S$, we may as well consider the transformed problem

$$\max\ \mathbf{v}_\alpha(M^{-1}(\mathbf{x}))$$

subject to:

$$\mathbf{x} \in \mathbf{X}$$

over the *long-run frequency space* $\mathbf{X}$. The next couple of theorems will state these results. In the next section we will take a look at them in the broader setting of unichained Markov chains.

**Lemma 1.5.2.** *Let $\Gamma_\alpha$ be an irreducible AMD model and $\mathbf{X}$ be the corresponding polyhedral set defined by (i)-(iii). Let $\mathbf{x}$ be any vector in $\mathbf{X}$ and consider the row vector $\overline{\mathbf{x}} = (x_1, x_2, \ldots, x_N)$ constructed from $\mathbf{x}$ by: $x_i := \sum_{a \in A(i)} x_{ia}$, $i \in \mathbf{S}$. Then we may conclude that $\overline{\mathbf{x}} > 0$ (i.e., $x_i > 0$ for all $i \in \mathbf{S}$).*

**Theorem 1.5.2.** *Let $\Gamma_\alpha$ be an irreducible AMD model, $\mathbf{X}$ be the polyhedron defined by (i)-(iii), and $M : \mathbf{F}_S \to \mathbb{R}^m$ be defined by (1.40) and (1.42). Then $M$ is an invertible map of $\mathbf{F}_S$ onto $\mathbf{X}$ with the inverse map defined by $M^{-1}(\mathbf{x}) = \mathbf{f}_\mathbf{x}$, where $f_\mathbf{x}(i,a) := \frac{x_{ia}}{x_i}$ for all $a \in A(i)$, $i \in \mathbf{S}$.*

**Theorem 1.5.3.** *Let $\Gamma_\alpha$ be an irreducible AMD model, $\mathbf{X}$ be its long-run frequency space defined by (i)-(iii), and $M^{-1} : \boldsymbol{X} \to \mathbf{F}_S$ be as in (1.5.3). Furthermore, let $\mathbf{x}^0$ be an optimal solution of the linear program*

$$\max \sum_{i=1}^{N} \sum_{a \in A(i)} r(i,a) x_{ia}$$

*subject to:*

$$W\mathbf{x} = \mathbf{0} \tag{1.43}$$

$$\mathbf{1}\mathbf{x} = 1 \tag{1.44}$$

$$\mathbf{x} \geq \mathbf{0}. \tag{1.45}$$

*Then $\mathbf{f}^0 := \mathbf{f}_{\mathbf{x}^0} = M^{-1}(\mathbf{x}^0)$ is an optimal strategy for the original (limiting average) optimal control problem.*

**Corollary 1.5.1.**

1. *Let* $\mathbf{x}$ *be any extreme point of* $\boldsymbol{X}$*. Then each block* $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im(i)})^T$ *of* $\mathbf{x}$ *contains exactly one positive element.*
2. *Let* $\mathbf{x}^0$ *be any basic optimal solution of the linear program (1.45). Then* $\mathbf{f}^0 = M^{-1}(\mathbf{x}^0)$ *is a pure optimal strategy.*

We can prove that we can use the same primal-dual pair of linear programs for communicating Markov decision processes. A Markov decision process is called *communicating* if for every pair of states $(i, j) \in \mathbf{S} \times \mathbf{S}$ there exists a control $\mathbf{f} \in \mathbf{F}_D$ and an integer $\tau \geq 1$ (both of which may depend on $i$ and $j$) such that $p_{ij}^{(\tau)}(\mathbf{f}) > 0$.

**Theorem 1.5.4.** *Consider a communicating MDP with the limiting average criterion. Then:*

1. *The transition matrix* $P(\mathbf{f})$ *is irreducible for all* $\mathbf{f} \in \mathbf{F}_S$ *such that* $f(i, a) > 0$ *for all* $a \in \mathbf{A}(i),\ i \in \mathbf{S}$.
2. *An optimal solution to a communicating limiting average MDP can be found from an optimal solution to the same primal-dual pair of linear programs that are used to solve an irreducible MDP ([7])*

*Proof.*

(i) Let $\mathbf{F}_C$ denote the set of all *completely mixed (stationary) policies*, where $f(i, a) > 0$ for all $a \in A(i),\ i \in \mathbf{S}$. Clearly a $\mathbf{f}^0 \in \mathbf{F}_S$ exists such that $P(\mathbf{f}^0)$ is irreducible. According to the definition of a communicating MDP, for every $(i, j) \in \mathbf{S} \times \mathbf{S}$ a control $\mathbf{f} \in \mathbf{F}_S$ and an integer $\tau \geq 1$ exist such that $p_{ij}^{\tau}(\mathbf{f}) > 0$. By combining these strategies we can find a policy $\mathbf{f}^0$ and an integer $\tau \geq 1$ such that $p_{ij}^{\tau}(\mathbf{f}^0) > 0$ for every $(i, j) \in \mathbf{S} \times \mathbf{S}$.
Now let $\mathbf{f}^* \in \mathbf{F}_C$ be arbitrary. The irreducibility of $P(\mathbf{f}^*)$ follows from that of $P(\mathbf{f}^0)$ because $p_{ij}(\mathbf{f}^0) > 0$ implies $p_{ij}(\mathbf{f}^*) > 0$ for any $i, j \in \mathbf{S}$.

(ii) We will first prove the following theorem:
**Theorem 1.5.5.** *If a policy* $\mathbf{f}^0 \in \mathbf{F}_S$ *exists such that* $P(\mathbf{f}^0)$ *is irreducible, then the following condition is satisfied:*

**Condition***: For every* $b = (b_1, \ldots, b_N) \in \mathbb{R}^N$ *such that*

$$\sum_{i=1}^{N} b_i = 0 \tag{1.46}$$

*there exists* $y = (y_{ia} \mid a \in A(i),\ i \in \mathbf{S})$ *(y may depend on b) such that:*

$$y_{ia} \geq 0, \qquad \text{for } a \in A(i),\ i \in \mathbf{S} \tag{1.47}$$

*and*

$$\sum_{a \in A(j)} y_{ja} - \sum_{i=1}^{N} \sum_{a \in A(i)} y_{ia} p_{ij}(a) = b_j,\ j \in \mathbf{S}. \tag{1.48}$$

*Proof.* Let $\mathbf{f}^0$ induce irreducible $P(\mathbf{f}^0)$, let $\pi > 0$ be the equilibrium distribution for $P(\mathbf{f}^0)$, and let $Z(\mathbf{f}^0) = \left[I - P(\mathbf{f}^0) + P^*(\mathbf{f}^0)\right]^{-1}$ be the fundamental matrix for $P(\mathbf{f}^0)$. Let $b \in \mathbb{R}^N$ satisfy (1.46) of the condition. Define $d \in \mathbb{R}^N$ by

$$d = bZ(\mathbf{f}^0) + c\pi^0$$

with $c \geq 0$ sufficiently large to assure $d \geq 0$. Take $y^0_{ia} = d_i f^0(i,a)$ for $a \in \mathbf{A}(i),\ i \in S$. Since both $d$ and $\mathbf{f}^0$ are non-negative, $y^0 \geq 0$ (i.e. (1.47) of the condition is satisfied). Finally, (1.48) of the condition takes the form

$$d_j - \sum_{i=1}^{N} d_i p_{ij}(\mathbf{f}^0) = b_j,\ j \in S.$$

Since $\pi^0 \left[I - P(\mathbf{f}^0)\right] = 0$, satisfaction of (1.48) follows from $Z(\mathbf{f}^0)\left[I - P(\mathbf{f}^0)\right] = I - P^*(\mathbf{f}^0)$ and from $bP^*(\mathbf{f}^0) = 0$ (using (1.46) and $p_{ij}(\mathbf{f}^0) = \pi^0_j$ since $P(\mathbf{f}^0)$ is irreducible) □

The general MDP (LP1) and the simpler irreducible MDP (LP2) are defined as follows:

**LP1:** Let $g, w \in \mathbb{R}^S$ be the primal variables, and the dual variables $x = \{x_{ia} \mid a \in \mathbf{A}(i),\ i \in \mathbf{S}\}$, $y = \{y_{ia} \mid a \in \mathbf{A}(i),\ i \in \mathbf{S}\}$.
Minimize $\sum_{i=1}^{N} \beta_i g_i$
subject to:

$$\begin{aligned} g_i - \sum_{j=1}^{N} g_j p_{ij}(a) &\geq 0 \\ g_i + w_i - \sum_{j=1}^{N} w_j p_{ij}(a) &\geq r(i,a) \end{aligned}$$

for $a \in \mathbf{A}(i),\ i \in S$, with constants $\beta_i > 0$ for $i \in \mathbf{S}$ and $\sum_{i \in S} \beta_i = 1$

**DLP1:**
Maximize $\sum_{i \in \mathbf{S}} \sum_{a \in A(i)} r(i,a) x_{ia}$
subject to:

$$\begin{aligned} \sum_{i \in \mathbf{S}} \sum_{a \in A(i)} \left(\delta(i,j) - p_{ij}(a)\right) x_{ia} &= 0;\ j \in \mathbf{S} \\ \sum_{a \in A(i)} x_{ja} + \sum_{i \in \mathbf{S}} \sum_{a \in A(i)} \left(\delta(i,j) - p_{ij}(a)\right) y_{ia} &= \beta_j;\ j \in \mathbf{S} \\ x_{ia}, y_{ia} &\geq 0;\ i \in \mathbf{S}, a \in A(i). \end{aligned}$$

**LP2:** Let $g$ (a scalar) and $w \in \mathbb{R}^S$ the primal variables, and the dual variables $x = \{x_{ia} \mid a \in \mathbf{A}(i), i \in \mathbf{S}\}$.
Minimize $g$
subject to:

$$g + w_i - \sum_{j=1}^{N} w_j p_{ij}(a) \geq r(i,a),\ a \in \mathbf{A}(i), i \in \mathbf{S}.$$

**DLP2:**
Maximize $\sum_{i \in \mathbf{S}} \sum_{a \in A(i)} r(i,a) x_{ia}$
subject to:

$$\begin{aligned} \sum_{i \in \mathbf{S}} \sum_{a \in A(i)} (\delta(i,j) - p_{ij}(a))\, x_{ia} &= 0;\ j \in \mathbf{S} \\ \sum_{i \in \mathbf{S}} \sum_{a \in A(i)} x_{ia} &= 1 \\ x_{ia} &\geq 0;\ i \in \mathbf{S}, a \in A(i). \end{aligned}$$

Let $(g^0, w^0)$ and $x^0$ optimally solve LP2 and its dual, respectively, for a communicating MDP. Define $g^* \in \mathbb{R}^S$ by $g^*_i = g^0$ for $i \in \mathbf{S}$. Since optimal gain values in communicating MDPs are independent of the starting state, $(g^*, w^0)$ are optimal in LP1. The objectives and dual constraints corresponding to $w$ of LP1 and LP2 are identical. It only remains to show that the LP1 dual has a feasible solution of the form $(x^0, y^0)$, i.e., that a non-negative $y^0 = \left\{y^0_{ia} \mid a \in A(i), i \in \mathbf{S}\right\}$ exists such that:

$$\sum_{a \in \mathbf{A}(j)} \left[x^0_{ja} + y^0_{ja}\right] - \sum_{i=1}^{N} \sum_{a \in A(i)} y^0_{ia} p_{ij}(a) = \beta_j,\ j \in \mathbf{S}.$$

Such an $y^0$ can be created by defining $b \in \mathbb{R}^S$ by

$$b_i = \beta_i - \sum_{a \in A(i)} x^0_{ia},\ i \in \mathbf{S}.$$

then applying the previous theorem. Note that the LP2 dual constraint corresponding to $g$ ensures that (1) of the condition mentioned in the previous theorem holds for the constructed $b$. Conversely, if LP1 and its dual are solved for a communicating MDP with $(g^*, w^0)$ and $(x^0, y^0)$ respectively, then $g^*$ has identical components and $(g^*_1, w^0)$ and $x^0$ solve LP2 and its dual.

□

## 1.6 Application: The Hamiltonian Cycle Problem

The problem here is to find a Hamilton cycle in a directed graph, or establish that it does not exist. In other words: find a simple cycle with $N$ arcs in a directed graph with $N$ nodes.

In this section we will try to reformulate this problem as a Markov Decision Problem. To do this we consider the directed path on the graph $G$ defined by a function $f$ mapping the set of nodes $S = \{1, 2, \ldots, N\}$ into the set of arcs $A$. The set of nodes can be regarded as the state space of a Markov decision process $\Gamma_\alpha$ where, for each state/node $i$, the action space

$$A(i) = \{a = j | (i, j) \in A\}$$

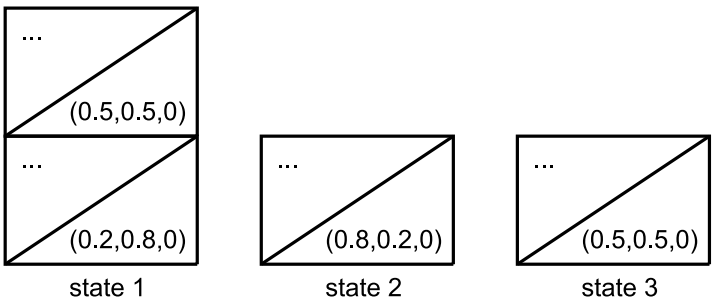
is in 1-1 correspondence with the set of arcs starting from that node. We will take node 1 as starting node in $G$.

If we restrict $f$ such that $f(i) \in A(i)$, for each $i \in \mathbf{S}$, then $f$ can be thought of as a deterministic strategy $\mathbf{f}$ in $\Gamma$. We can now say that $\mathbf{f}$ is a Hamiltonian cycle in $G$ if the set of arcs $\{(1, f(1)), (2, f(2)), \ldots, (N, f(N))\}$ is a Hamiltonian cycle in $G$. If the set of arcs contains cycles of length less than $N$, $\mathbf{f}$ has subcycles in G.

If $\mathbf{f}$ is a Hamiltonian cycle, then $P(\mathbf{f})$ is irreducible and the long-run frequency of visits to any state $x_i(\mathbf{f}) = \frac{1}{N}$. A problem arises however if $\mathbf{f}$ has subcycles in $G$. This means that $P(\mathbf{f})$ contains multiple ergodic classes, which complicates the analysis of the Markov decision process $\Gamma_\alpha$.

We can circumvent this problem by using unichained Markov decision processes. A Markov decision process is *unichained* if for every deterministic stationary control $\mathbf{f}$, $P(\mathbf{f})$ contains only a single ergodic class and possibly a nonempty set of transient states. That a unichained Markov decision process does not necessarily mean that the process is irreducible as shown by the following example.

**Example 1.6.1.** Let $S = \{1, 2, 3\}$, $A(1) = 1, 2$, $A(2) = (1)$, $A(3) = 1$. The transitional probabilities are given by




It is clear that this process is not irreducible, since we cannot reach state 3 via state 1 or 2. The only ergodic class consists of state 1 and 2. This means the process is unichained.

A way to destroy multiple ergodic classes and induce a unichained Markov decision process, is to perturb the transition probabilities of $\Gamma_\alpha$ slightly to create an $\varepsilon$-perturbed process $\Gamma_\alpha(\varepsilon)$ (for $0 < \varepsilon < 1$) defined by:

$$p_{ij}^{(\varepsilon)} = \begin{cases} 1 & \text{if } i = 1 \text{ and } a = j \\ 0 & \text{if } i = 1 \text{ and } a \neq j \\ 1 & \text{if } i > 1 \text{ and } a = j = 1 \\ \varepsilon & \text{if } i > 1,\ a \neq j, \text{ and } j = 1 \\ 1 - \varepsilon & \text{if } i > 1,\ a = j, \text{ and } j > 1 \\ 0 & \text{if } i > 1,\ a \neq j, \text{ and } j > 1. \end{cases}$$

With the above perturbation, for each pair of nodes $(i, j)$ (neither equal to 1) corresponding to the original arc $(i, j)$ the perturbation replaces that arc by a pair of stochastic arcs $(i, 1)$ and $(i, j)$ with weights $\varepsilon$ and $(1 - \varepsilon)$, with $\varepsilon \in (0, 1)$. We can interprete the perturbation that the decision to move along arc $(i, j)$ results in movement along $(i, j)$ only with probability $(1-\varepsilon)$ and with probability $\varepsilon$ that the process will return to the home node 1.

We will now analyse the Hamiltonian cycle problem in the "frequency" space of the perturbed process $\Gamma_\alpha(\varepsilon)$. Via (1.40)-(1.42) we know that with every $\mathbf{f} \in \mathbf{F}_S$ we can associate the long-run frequency vector $\mathbf{x}(\mathbf{f})$. We will now show that, if we set $q_i = 0$ for those states $i$ that are transient, Lemma (1.5.1) can be extended to matrices containing a single ergodic class.

**Lemma 1.6.1.** *Let $P$ be the probability transition matrix of an unichained Markov chain and $Q$ be the corresponding Cesaro-limit matrix. Then*

1. *$Q$ has identical rows.*
2. *Let $\mathbf{q} = (q_1, \ldots, q_N)$ be a row of $Q$. Then every entry of $\mathbf{q}$ is strictly positive for the recurrent states and zero for the transient states, and $\mathbf{q}$ is the unique solution of the linear system of equations:*

$$\mathbf{q}P = \mathbf{q}$$
$$\mathbf{q}\mathbf{1} = 1,$$

*where $\mathbf{1}$ is an $N$-dimensional column vector with unity in every entry. The vector $\mathbf{q}$ is called the "stationary distribution" of the unichained Markov chain.*

For a proof of this lemma we refer to [11]. Now, as in the previous section, consider the polyhedral set $\mathbf{X}(\varepsilon)$ defined by the constraints

(i) $\sum_{i=1}^{N} \sum_{a \in A(i)} (\delta(i, j) - p_{ij}^{(\varepsilon)}(a)) x_{ia} = 0,\ j \in \mathbf{S}$

(ii) $\sum_{i=1}^{N} \sum_{a \in A(i)} x_{ia} = 1$

(iii) $x_{ia} \geq 0,\ a \in A(i),\ i \in \mathbf{S}.$

Since for every $\mathbf{f} \in \mathbf{F}_S$, $\mathbf{x}(\mathbf{f}) \in \mathbf{X}(\varepsilon)$ we see that (1.40)-(1.42) define a map $M : \mathbf{F}_S \to \mathbf{X}(\varepsilon)$. This leads us to the following theorem:

**Theorem 1.6.1.** *Consider an unichained limiting average model $\Gamma_\alpha$ and the polyhedral set $\mathbf{X}$ as above. Let $\mathbf{x} \in \mathbf{X}$ and $\mathbf{f}_\mathbf{X}$ be constructed from $\mathbf{x}$ according to*

$$f_\mathbf{X}(i,a) = \begin{cases} x_{ia}/x_i, & \text{if } x_i = \sum_{a \in A(i)} x_{ia} > 0 \\ \text{arbitrary}, & \text{if } x_i = 0. \end{cases}$$

*If $\mathbf{x}_0$ is an optimal solution of the linear program*

$$\max \sum_{i=1}^{N} \sum_{a \in A(i)} r(i,a) x_{ia}$$

*subject to:*

$$\mathbf{x} \in \mathbf{X},$$

*then $\mathbf{f}_{\mathbf{X}^0}$ is optimal in $\Gamma_\alpha$*

*Proof.* Due to the adapted version of Lemma 2.4.1, $v_\alpha(i, \mathbf{f})$ is independent of the starting state $i$, because $Q(\mathbf{f})$ has identical rows $\mathbf{q}(\mathbf{f})$, and hence $\mathbf{v}_\alpha(\mathbf{f}) = Q(\mathbf{f})\mathbf{r}(\mathbf{f})$ implies that for every $\mathbf{f} \in \mathbf{F}_S$ and any $j \in S$

$$v_\alpha(j, \mathbf{f}) = [Q(\mathbf{f})\mathbf{r}(\mathbf{f})]_j = \sum_{i=1}^{N} q_i(\mathbf{f}) r(i, \mathbf{f}) = \sum_{i=1}^{N} \sum_{a \in A(i)} q_i(\mathbf{f}) r(i,a) f(i,a).$$

Hence by $\mathbf{x}_{ia(\mathbf{f})} := q_i(\mathbf{f}) f(i,a)$

$$v_\alpha(j, \mathbf{f})) = \sum_{i=1}^{N} \sum_{a \in A(i)} r(i,a) x_{ia}(\mathbf{f}) \tag{1.49}$$

for every $j \in S$ and $\mathbf{f} \in \mathbf{F}_S$.

Suppose now that there exists a control $\hat{\mathbf{f}} \in \mathbf{F}_S$ that is superior to $\mathbf{f}^0$. That is,

$$v_\alpha(j, \hat{\mathbf{f}}) > v_\alpha(j, \mathbf{f}^0).$$

Now define the map $\hat{M} : \mathbf{X}(\varepsilon) \to \mathbf{F}_S$ by

$$f_\mathbf{X}(i,a) = \begin{cases} \frac{x_{ia}}{x_i}, & \text{if } \sum_{a \in A(i)} x_{ia} > 0 \\ 1, & \text{if } x_i = 0 \text{ and } a = a_1 \\ 0, & \text{if } x_i = 0 \text{ and } a \neq a_1 \end{cases} \tag{1.50}$$

for every $a \in A(i),\ i \in S$, where $a_1$ denotes the first available action in a given state according to some ordering. This means that $\mathbf{x}^0 = M(\hat{M}(\mathbf{x}^0))$ is optimal. But now

$$\begin{aligned}\sum_{i=1}^{N}\sum_{a\in A(i)} r(i,a)x_{ia}(\hat{\mathbf{f}}) &> \sum_{i=1}^{N}\sum_{a\in A(i)} r(i,a)x_{ia}(\mathbf{f}^0)\\ &= \sum_{i=1}^{N}\sum_{a\in A(i)} r(i,a)x^0_{ia},\end{aligned}$$

which contradicts the optimality of $\mathbf{x}^0$. So $\mathbf{f}_{\mathbf{X}^0}$ is optimal. □

In order to be able to prove that an optimal control can derived from any optimal solution of a linear program, we will first need to prove the following theorem.

**Theorem 1.6.2.** *Consider the polyhedral set*

$$\mathbf{X} = \{\mathbf{x} \mid W\mathbf{x} = \mathbf{0},\ \mathbf{1}^{\mathbf{T}}\mathbf{x} = \mathbf{1},\ \mathbf{x} \geq \mathbf{0}\}$$

*introduced in Section 1.5, but without the irreducibility assumption on the transition probabilities. Let* $\mathbf{x} \in \mathbf{X}$, *and define* $\mathbf{S}_{\mathbf{X}} = \left\{i \in \mathbf{S} \mid \sum_{a\in A(i)} x_{ia} > 0\right\}$ *and* $\overline{\mathbf{S}}_{\mathbf{X}} = \{(i,a) \mid x_{ia} > 0,\ a \in A(i),\ i \in \mathbf{S}\}$. *We shall say that* $\mathbf{x}$ *identifies a unique ergodic class if*

1. *The cardinalities of* $\mathbf{S}_{\mathbf{X}}$ *and* $\overline{\mathbf{S}}_{\mathbf{X}}$ *are equal, and*
2. *All of the states of* $\mathbf{S}_{\mathbf{X}}$ *form an ergodic class under a stationary control* $\mathbf{f}_{\mathbf{X}}$ *defined by*

$$\mathbf{f}_{\mathbf{X}}(i,a) = \begin{cases} 1, & \text{if } (i,a) \in \overline{\mathbf{S}}_{\mathbf{X}},\ i \in \mathbf{S}_{\mathbf{X}} \\ 0, & \text{if } (i,a) \notin \overline{\mathbf{S}}_{\mathbf{X}},\ i \in \mathbf{S}_{\mathbf{X}} \\ \text{arbitrary}, & \text{if } i \notin \mathbf{S}_{\mathbf{X}}. \end{cases}$$

*Now every extreme point of* $\mathbf{X}$ *identifies a unique ergodic class ([3]).*

*Proof.* The column $\mathbf{w}_{(i,a)}$ of DLP2 on page 20 corresponding to variable $x_{ia}$ and the activity vector $b$ are given by

$$\mathbf{w}_{(i,a)} = \begin{vmatrix} -p_{i1}(a) \\ \vdots \\ 1 - p_{ii}(a) \\ \vdots \\ -p_{iN}(a) \\ 1 \end{vmatrix},\quad b = \begin{vmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 1 \end{vmatrix}.$$

A basic feasible solution to Program I has

$$\sum_{(i,a)\in\overline{\mathbf{S}}_{\mathbf{x}}} \mathbf{w}_{(i,a)}\mathbf{x}_{ia} = b. \tag{1.51}$$

Note from $W\mathbf{x} = 0$ that $p_{ij}(a) = 0$ when $(i,a) \in \overline{\mathbf{S}}_{\mathbf{x}}$ and $j \notin \mathbf{S}_{\mathbf{x}}$, in which sense "escape" from $\mathbf{S}_{\mathbf{x}}$ is impossible. Let $\overline{\mathbf{S}}_{\mathbf{x}}^{(1)}$ be a subset of $\overline{\mathbf{S}}_{\mathbf{x}}$ containing exactly one element $(i,a_i)$ for each $i$ in $\mathbf{S}_{\mathbf{x}}$, so that the set $\left\{p_{ij}(a) \mid (i,a) \in \overline{\mathbf{S}}_{\mathbf{x}}^{(1)},\ j \in \mathbf{S}_{\mathbf{x}}\right\}$ contains the transition probabilities of a Markov chain. This Markov chain has at least one ergodic chain consisting of states $\mathbf{S}_{\mathbf{x}}^{(2)} \subset \mathbf{S}_{\mathbf{x}}$ and transition probabilities specified by

$$\overline{\mathbf{S}}_{\mathbf{x}}^{(2)} = \left\{(i,a) \in \overline{\mathbf{S}}_{\mathbf{x}}^{(1)} : i \in \mathbf{S}_{\mathbf{x}}^{(2)}\right\}.$$

Confining attention to this ergodic chain, let $y_i$ be the probability that a random observer finds the last observed state to be state $i$. With $z_{ia} = y_i$ for $(i,a) \in \overline{\mathbf{S}}_{\mathbf{x}}^{(2)}$, $\{z_{ia}\}$ is the unique solution of an equation like

$$zP_\delta = z, \qquad z\mathbf{e} = 1, \tag{1.52}$$

which can be written as

$$\sum_{i=1}^{N} \sum_{a\in A(i)} \mathbf{w}_{ia} z_{ia} = b. \tag{1.53}$$

Subtract (1.53) from (1.51). Since $\overline{\mathbf{S}}_{\mathbf{x}}^{(2)} \subset \overline{\mathbf{S}}_{\mathbf{x}}$,

$$\sum_{(i,k)\in\overline{\mathbf{S}}_{\mathbf{x}}-\overline{\mathbf{S}}_{\mathbf{x}}^{(2)}} \mathbf{w}_{ia}x_{ia} + \sum_{(i,a)\in\overline{\mathbf{S}}_{\mathbf{x}}^{(2)}} \mathbf{w}_{ia}(x_{ia} - z_{ia}) = 0. \tag{1.54}$$

Since $\{x_{ia}\}$ is a basic feasible solution, the set $\left\{\mathbf{w}_{ia} : (i,a) \in \overline{\mathbf{S}}_{\mathbf{x}}\right\}$ is linearly independent, and every coefficient in (1.54) must therefore be zero. Hence $\overline{\mathbf{S}}_{\mathbf{x}} = \overline{\mathbf{S}}_{\mathbf{x}}^{(2)}$ and $x_{ia} = z_{ia}$ for $(i,a) \in \overline{\mathbf{S}}_{\mathbf{x}}$, the latter implying that $\sum_{i=1}^{N}\sum_{a\in A(i)} x_{ia}r_{ia}$ is the gain rate for the identified chain. □

**Lemma 1.6.2.**

*(i) The set* $\mathbf{X}(\varepsilon) = \{\mathbf{x}(\mathbf{f}) | \mathbf{f} \in \mathbf{F}_S\}$ *and will from now on be called the (long-run) "frequency space" of* $\Gamma_\alpha(\varepsilon)$.

*(ii) For every* $\mathbf{x} \in \mathbf{X}(\varepsilon)$,

$$M\left(\hat{M}(\mathbf{x})\right) = \mathbf{x},$$

*but the inverse of* $M$ *need not exist, in general.*

*(iii) If* $\mathbf{x}$ *is an extreme point of* $\mathbf{X}(\varepsilon)$*, then*

$$\mathbf{f}_{\mathbf{x}} = \hat{M}(\mathbf{x}) \in \mathbf{F}_D.$$

*(iv) If* $\mathbf{f} \in \mathbf{F}_D$ *is a Hamiltonian cycle, then* $\mathbf{x}(\mathbf{f})$ *is an extreme point of* $\mathbf{X}(\varepsilon)$*.*

*Proof.* (i) follows from the definition of $M$. (ii) follows from the fact that, since $M(x) = 0$ for any transient state $x$, there is not always an inverse of $M$. According to theorem 1.6.2, for every extreme point in $X$, the cardinalities of $\mathbf{S}_{\mathbf{x}}$ and $\overline{\mathbf{S}}_{\mathbf{x}}$ are equal, which means that there is only one $a \in A(i)$ for which $x_{ia} > 0$. This means $f_x = \hat{M}(x)$ is deterministic. (iv) follows from the fact that if $\mathbf{f} \in \mathbf{F}_D$ is a Hamilton cycle, the Markov chain consists of one ergodic class. According to theorem 1.6.2, $\mathbf{x}$ identifies a unique ergodic class. □

We shall now derive a useful partition of the class $\mathbf{F}_D$ of deterministic strategies that is based on the graphs they "trace out" in $G$. With each $\mathbf{f} \in \mathbf{F}_D$ we can associate a subgraph $G_{\mathbf{f}}$ of $G$ defined by

$$\text{arc}(i,j) \in G_{\mathbf{f}} \iff f(i) = j.$$

We shall denote a simple cycle of length $m$ and beginning at 1 by a set of arcs

$$c_m^1 = \{(i_1 = 1, i_2), (i_2, i_3), \ldots, (i_m, i_{m+1} = 1)\}\,;\ m = 2, 3, \ldots, N.$$

Now $c_1^N$ is a Hamiltonian cycle. If $G_{\mathbf{f}}$ contains a cycle $c_m^1$, we write $G_{\mathbf{f}} \supset c_m^1$. Let $C_m := \{\mathbf{f} \in \mathbf{F}_D | G_{\mathbf{f}} \supset c_m^1\}$ the set of deterministic strategies that trace out a simple cycle of length $m$, beginning at 1, for each $m = 2, 3, \ldots, N$. $C_N$ is the set of strategies that correspond to Hamiltonian cycles and any single $C_m$ can be empty, depending on the structure of the original graph $G$.

The partition of the deterministic strategies that seems to be most relevant for our purposes is

$$\mathbf{F}_D = \left[\bigcup_{m=2}^{N} C_m\right] \bigcup B, \tag{1.55}$$

where $B$ contains all of the deterministic strategies that are not in any of the $C_m$'s. All strategies in a given set in the partition (1.55) induce the same long-run frequency $x_1(\mathbf{f})$ of visits to the home node 1. This observation is captured in the following proposition.

**Proposition 1.6.1.** *Let* $\varepsilon \in (0, 1]$, $\mathbf{f} \in \mathbf{F}_D$*, and* $\mathbf{x}(\mathbf{f})$ *be its long-run frequency vector (that is,* $\mathbf{x}(\mathbf{f}) = M(\mathbf{f})$*). The long-run frequency of visits to the home state 1 is given by*

$$x_1(\mathbf{f}) = \sum_{a \in A(1)} x_{1a}(\mathbf{f}) = \begin{cases} \frac{1}{d_m(\varepsilon)}, & \text{if } \mathbf{f} \in C_m,\ m = 2, 3, \ldots, N \\ \frac{\varepsilon}{1+\varepsilon}, & \text{if } \mathbf{f} \in B, \end{cases}$$

*where* $d_m(\varepsilon) = 1 + \sum_{i=2}^{m}(1-\varepsilon)^{i-2}$ *for* $m = 2, 3, \ldots, N$.

This proposition leads to the following characterization of the Hamiltonian cycles of a directed graph.

**Theorem 1.6.3.**

*(i) Let* $\mathbf{f} \in \mathbf{F}_D$ *be a Hamiltonian cycle in the graph* $G$. *Then* $G_{\mathbf{f}} = c_N^1$, $\mathbf{x}(\mathbf{f})$ *is an extreme point of* $\mathbf{X}(\varepsilon)$ *and* $\mathbf{x}_1(\mathbf{f}) = \frac{1}{d_N(\varepsilon)}$.

*(ii) Conversely, suppose that* $\mathbf{x}$ *is an extreme point of* $\mathbf{X}(\varepsilon)$ *and that* $x_1 = \sum_{a\in A(1)} x_{1a} = \frac{1}{d_N(\varepsilon)}$. *Then* $\mathbf{f} = \hat{M}(\mathbf{x})$ *is an Hamiltonian cycle in* $G$.

**Corollary 1.6.1.** *Hamiltonian cycles of the graph* $G$ *are in 1:1 correspondence with those points of* $\mathbf{X}(\varepsilon)$ *that satisfy*

*(i)* $x_1 = \sum_{a\in A(1)} x_{1a} = \frac{1}{d_N(\varepsilon)}$

*(ii) For every* $i \in S$, $x_i = \sum_{a\in A(1)} x_{ia} > 0$ *and* $\frac{x_{ia}}{x_a} \in \{0, 1\}$ *for each* $a \in A(i)$, $i \in S$.

Now, let $D = \mathrm{diag}(D_1, D_2, \ldots, D_N)$ be a block-diagonal matrix with its $i$th block equal to $D_i$ for $i = 1, 2, \ldots, N$. Suppose that $D_i$ is an $m(i) \times m(i)$ matrix with all the diagonal elements equal to 0 and off-diagonal elements equal to 1 (where $m(i)$ is the cardinality of $A(i)$), for each $i \in S$. Consider the following (indefinite) quadratic program:

$$\min\ \mathbf{x}^T D\mathbf{x}$$

subject to: $(QP)$

$$\mathbf{x} \in \mathbf{X}(\varepsilon)$$

$$x_1 = \sum_{a\in A(1)} x_{1a} = \frac{1}{d_N(\varepsilon)}.$$

**Theorem 1.6.4.**

*(i) Let* $\mathbf{f}$ *be a Hamiltonian cycle in* $G$. *Then* $\mathbf{x}(\mathbf{f})$ *is a global minimum of* $(QP)$ *such that* $(\mathbf{x}^*)^T D\mathbf{x}^* = 0$.

*(ii) Conversely, let* $\mathbf{x}^*$ *be a global minimum of* $(QP)$ *such that* $(\mathbf{x}^*)^T D\mathbf{x}^* = 0$. *Then* $\mathbf{f}_{\mathbf{x}^*} = \hat{M}(\mathbf{x}^*)$ *is a deterministic strategy that traces out a Hamiltonian cycle in* $G$.

## 1.7 Behaviour and Markov Strategies

When we are trying to solve AMD models, it is not sufficient to use pure strategies only. Sometimes we have to use randomized controls in $\mathbf{F}_S$. A logical question at this point is whether the class $\mathbf{F}_S$ is sufficient, or whether the performance of the system could be improved by using a possibly more complex control that does not belong to $\mathbf{F}_S$? In order to provide an answer to this question we will take a look at two more general classes of nonstationary strategies and set up the result needed to compare the performance of the system when controlled by strategies selected from these classes.

Let $S_t$, $A_t$ denote the random variables representing, respectively, the state and action at time $t$. Let $h_t = (i_0, a_0, i_1, a_1, \ldots, a_{t-1}, i_t)$ be the *history* up to time $t$, and let $\mathcal{H}_t$ be the set of all possible histories up to time $t$. Let $\mathbf{A} = \left(\bigcup_{i \in S} A(i)\right)$ be the *total action space of the process*, and let $\mathcal{P}(\mathbf{A})$ be the set of all probability distributions on the finite set $\mathbf{A}$.

Define a *decision rule at time t* to be a function

$$f_t : \mathcal{H}_t \to \mathcal{P}(\mathbf{A})$$

such that

$$f_t(h_t, a) := \begin{cases} \mathbb{P}[A_t = a | h_t] & \text{if } a \in A(i_t) \\ 0 & \text{if } a \notin A(i_t). \end{cases} \tag{1.56}$$

We now have a history dependent class of strategies $\mathbf{F}_B$, called behaviour strategies. A *Markov (or memoryless) strategy* $\pi$ is a behaviour strategy in which every decision rule $f_t$ depends on only the current state, that is, for every $t = 0, 1, 2, \ldots$, and for all histories $h_t = (i_0, a_0, \ldots, a_{t-1}, i_t) \in \mathbf{H}_t$

$$\begin{aligned} f_t(i_t, a) &:= \mathbb{P}_{f_t}[A_t = a | S_t = i_t] \\ &= \mathbb{P}_{f_t}[A_t = a | S_0 = i_0, A_0 = a_0, \ldots, A_{t-1} = a_{t-1}, S_t = i_t] \\ &= f_t(h_t, a). \end{aligned}$$

Denote the class of all Markov strategies by $\mathbf{F}_M$. A *stationary strategy* $\pi$ is a Markov strategy in which all decision rules are independent of time, that is, $f_t = \mathbf{f}$ for every $t$. A stationary strategy $\pi = (\mathbf{f}, \mathbf{f}, \mathbf{f}, \ldots)$ can be represented by the already introduced block-row vector $\mathbf{f}$. The class of all stationary strategies is the same as the class $\mathbf{F}_S$ discussed in the previous sections. A *pure stationary* (or *deterministic*) strategy $\mathbf{f}$ is a stationary strategy such that for each $i \in S$ an action $a_i \in A(i)$ exists such that $\mathbf{f}(i, a) = 0$ whenever $a \neq a_i$. The class of all pure stationary strategies will be denoted by $\mathbf{F}_D$. It follows from the construction of these various classes that

$$\mathbf{F}_B \supseteq \mathbf{F}_M \supseteq \mathbf{F}_S \supseteq \mathbf{F}_D. \tag{1.57}$$

In all these classes the finiteness of states and actions ensures that the expected rewards/outputs at each time $t$ remain well defined and satisfy

$$\mathbb{E}_{i_0\pi}\{R_t\} = \sum_{i=1}^{N} \sum_{a\in A(i)} \mathbb{P}_\pi\left[S_t = i, A_t = a | S_0 = i_0\right] r(i,a)$$

for each $\pi \in \mathbf{F}_B$ and $i_0 \in S$. The definition (1.13) extends naturally to the strategies $\pi \in \mathbf{F}_B$. The same goes for the terminating performance criterion. The limiting average value of a control $\pi$ (see (1.35)) however needs to be modified. For every $\pi \in \mathbf{F}_B$ and initial state $i$ we now define

$$v_\alpha(i,\pi) := \liminf_{T\to\infty} \left[\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{i\pi}[R_t]\right] \tag{1.58}$$

which coincides with (1.35) whenever $\pi \in \mathbf{F}_S$.

The following important theorem shows that the performance of an arbitrary behaviour control may be "simulated" by a Markov control.

**Theorem 1.7.1.** *Let $\pi \in \mathbf{F}_B$ be an arbitrary behaviour strategy. Then for every initial state $i_0 \in S$ a Markov strategy $\overline{\pi} \in \mathbf{F}_M$ exists such that for all $a \in A, i \in S$, and $t = 0, 1, 2, \ldots,$*

$$\mathbb{P}\left[S_t = i, A_t = a | S_0 = i_0\right] = \mathbb{P}_{\overline{\pi}}\left[S_t = i, A_t = a | S_0 = i_0\right].$$

*In general $\overline{\pi}$ depends on the initial state.*

**Corollary 1.7.1.** *Fix an arbitrary $i \in S$. Let $\pi \in \mathbf{F}_B$ be an arbitrary behaviour strategy and $\overline{\pi} \in \mathbf{F}_M$ be constructed from $\pi$ by*

$$\overline{f}(i_t, a) = \mathbb{P}_\pi\{A_t = a | S_0 = i_0, S_t = i_t\}$$

*for all $a \in A(i_t)$, $i_t, i_0 \in S$, and $t = 1, 2, \ldots$. Let $\Gamma_\beta, \Gamma_\tau$, and $\Gamma_\alpha$ denote the discounted, terminating, and limiting average models considered so far. Then there is no loss of generality in restricting analysis to $\mathbf{F}_M$ since*

*(i)* $v_\beta(i,\pi) = v_\beta(i,\overline{\pi})$, $v_\tau(i,\pi) = v_\tau(i,\overline{\pi})$, *and* $v_\alpha(i,\pi) = v_\alpha(i,\overline{\pi})$

*(ii)* $\sup_{\pi\in\mathbb{F}_B} v_\beta(i,\pi) = \sup_{\pi\in\mathbb{F}_M} v_v(i,\pi) = \sup_{\pi\in\mathbb{F}_B} v_\tau(i,\pi) = \sup_{\pi\in\mathbb{F}_M} v_\tau(i,\pi)$ *and* $\sup_{\pi\in\mathbb{F}_B} v_\alpha(i,\pi) = \sup_{\pi\in\mathbb{F}_M} v_\alpha(i,\pi)$.

*Furthermore, (i) and (ii) hold for any other performance criterion that aggregates the sequence $\{\mathbb{E}_{i,\pi}[R_t]\}_{t=0}^{\infty}$; $\pi \in \mathbf{F}_B$, $i \in S$ .*

## 1.8 Policy Improvement and Newton's Method in Summable MDPs

We will now investigate whether we can improve on the optimal deterministic control $\mathbf{f}_0 \in \mathbf{F}_D$ of the discounted Markov decision model found in section 1.3 by using a behaviour strategy. In this section we will show that $\mathbf{f}_0 \in \mathbf{F}_D$ is also optimal in the class of all behaviour strategies. This means that for all $i \in S$

$$v_\beta(i, \mathbf{f}^0) = \sup_{\pi \in \mathbf{F}_B} v_\beta(i, \pi) = \sup_{\pi \in \mathbf{F}_M} v_\beta(i, \pi) = \sup_{\mathbf{f} \in \mathbf{F}_S} v_\beta(i, \mathbf{f}) = \sup_{\mathbf{f} \in \mathbf{F}_D} v_\beta(i, \mathbf{f}). \tag{1.59}$$

The second and fourth equalities in (1.59) were already established in the Corollaries (1.7.1) and (1.4.1), so will only need to prove the third equality.

Let $\pi = (f_0, f_1, \ldots, f_t, \ldots) \in \mathbf{F}_M$. Every $\mathbf{f}_t$ now defines a transition matrix $P(f_t)$ and an immediate expected reward vector $\mathbf{r}(f_t)$. This results in the following $t$-stage transition matrices:

$$\begin{aligned} P_t(\pi) &:= P(f_0)P(f_1)\ldots P(f_{t-1}) \text{ for } t = 1, 2, \ldots \\ P_0(\pi) &:= I_N. \end{aligned}$$

The (discounted) value vector of $\pi = (f_0, f_1, \ldots, f_t, \ldots)$ can now be written as

$$\mathbf{v}_\beta(\pi) = \sum_{t=0}^{\infty} \beta^t P_t(\pi)\mathbf{r}(f_t). \tag{1.60}$$

Further, we shall associate with $\pi$ a Markov control $\pi^+ := (f_1, f_2, \ldots, f_{t+1}, \ldots)$ that uses the decision rule $f_{t+1}$ at time $t$ for each $t = 0, 1, 2, \ldots$. It follows from (1.60) that

$$\begin{aligned} \mathbf{v}_\beta(\pi) &= \mathbf{r}(\mathbf{f}_0) + \beta P(f_0) \sum_{t=1}^{\infty} \beta^{t-1} P_{t-1}(\pi^+)\mathbf{r}(f_t) && (1.61) \\ &= \mathbf{r}(f_0) + \beta P(f_0)\mathbf{v}_\beta(\pi^+). && (1.62) \end{aligned}$$

We will write $\pi^1 \geq \pi^2$ ($\pi^1 > \pi^2$), respectively, if and only if

$$\mathbf{v}_\beta(\pi^1) \geq \mathbf{v}_\beta(\pi^2) \quad (\mathbf{v}_\beta(\pi^1) > \mathbf{v}_\beta(\pi^2)).$$

If $\pi = (f_0, f_1), \ldots, f_t, \ldots)$ is some Markov control, then $(g_0, g_1, \ldots, g_{t-1}, \pi)$ is the Markov control which uses decision rules $g_0, g_1, \ldots, g_{t-1}$ during the first $t$ stages, and then switches to $\pi$ thereafter, i.e. $g_t = f_0, g_{t+1} = f_1, \ldots$.

The following proposition states that if a control cannot be improved by a deviation at the initial stage, then it must be an optimal control.

**Proposition 1.8.1.** *Consider the discounted process $\Gamma_\beta$, and let $\pi^0 = (f_0^0, f_1^0, \ldots, f_t^0, \ldots) \in \mathbf{F}_M$ be such that for any one-stage decision rule $f$, $\pi^0 \geq (f, \pi^0)$, then $\pi^0$ is an optimal strategy.*

As a result of this theorem and the linear programming formulation of Section 1.3 we can show that a pure stationary optimal control exists.

**Theorem 1.8.1.** *Let $\Gamma_\beta$ be the discounted Markov decision model and $\mathbf{f}^* \in \mathbf{F}_D$ be a pure stationary strategy constructed in Corollary (1.4.1), then $\mathbf{f}^*$ is optimal in the entire class of behaviour strategies. That is,*

$$\mathbf{v}^\beta(\mathbf{f}^*) = \max_{\pi \in \mathbf{F}_B} \mathbf{v}_\beta(\pi).$$

**Corollary 1.8.1.** *(Local improvement Step)*

*(i) Let $\pi \in \mathbf{F}_M$ and $f$ be a decision rule such that $(f, \pi) > \pi$. then $\mathbf{f} > \pi$, where $\mathbf{f}$ uses the decision rule $f$ at every stage.*

*(ii) Let $\mathbf{f} \in \mathbf{F}_S$ be such that for at least one $i$ an action $a_i$ exists such that*

$$r(i, a_i) + \beta \sum_{j=1}^{N} p_{ij}(a_i) v_\beta(j, \mathbf{f}) > v_\beta(i, \mathbf{f}). \tag{1.63}$$

*Define $\boldsymbol{g} \in \mathbf{F}_S$ by*

$$g(i,a) = \begin{cases} f(i,a), & \text{if (1.63) does not hold at } i \\ 1, & \text{if } a = a_i \text{ and (1.63) holds at } i \\ 0, & \text{otherwise} \end{cases}$$

*Then $\boldsymbol{g} > \mathbf{f}$.*

We can now formulate the following policy improvement algorithm:

**Algorithm 1.8.1. Policy Improvement Algorithm**

**Step 1.** (Initialization) Set $k := 0$, select any pure stationary strategy $\mathbf{f}$, and set $\mathbf{f}^0 := \mathbf{f}$ and $\mathbf{v}^0 := \mathbf{v}_\beta(\mathbf{f}^0 = [I - \beta P(\mathbf{f}^0)]^{-1} \mathbf{r}(\mathbf{f}^0)$.

**Step 2.** (Check of Optimality) With general $k$ we have available $\mathbf{f}^k \in \mathbf{F}_D$ and $\mathbf{v}^k = \mathbf{v}_\beta(\mathbf{f}^k)$. Let $a_i^k$ be the action selected by $\mathbf{f}^k$ in state $i$ for each $i \in S$. If the optimality equation

$$r(i, a_i^k) + \beta \sum_{j=1}^{N} p_{ij}(a_i^k) v^k(j) = \max_{a \in A(i)} \left\{ r(i,a) + \beta \sum_{j=1}^{N} p_{ij}(a) v^k(j) \right\} \tag{1.64}$$

holds for each $i \in S$, STOP. The control $\mathbf{f}^k$ is a pure optimal strategy and $\mathbf{v}^k$ is the discounted value vector $\mathbf{v}_\beta$.

**Step 3.** (Policy Improvement) Let $\overline{\mathbf{S}}$ be the nonempty subset of states for which equality is violated in (1.64), that is, the left side is strictly smaller than the right side. Define

$$\overline{a}_i^k := \arg\max_{a\in A(i)} \left\{ r(i,a) + \beta \sum_{j=1}^{N} p_{ij}(a) v^k(j) \right\}$$

for each $i \in \overline{\mathbf{S}}$, and a new strategy $\mathbf{g} \in \mathbf{F}_D$ by

$$g(i,a) = \begin{cases} f(i,a), & \text{if } i \neq \overline{\mathbf{S}} \\ 1, & \text{if } i \in \overline{\mathbf{S}} \text{ and } a = \overline{a}_i^k \\ 0, & \text{otherwise.} \end{cases}$$

Set $\mathbf{f}^{k+1} := \mathbf{g}$, $\mathbf{v}^{k+1} := \mathbf{v}_\beta(\mathbf{g})$.

**Step 4.** (Iteration) Set $k := k+1$ and return to Step 1 with $\mathbf{f}^k := \mathbf{f}^{k+1}$, $\mathbf{v}^k := \mathbf{v}^{k+1}$.

**Theorem 1.8.2.** *The policy improvement algorithm terminates in no more than*

$$\mu = \prod_{i=1}^{N} m(i)$$

*steps, with an optimal deterministic policy* $\mathbf{f}^0$.

The question may arise if this algorithm is practical, because of its exponentially fast growth. We will now show that this is not an issue because the policy improvement algorithm is in more or less equivalent to the Newton's method for unconstrained optimization. Since this method converges to the global minimum with quadratic rate of convergence, so does the policy improvement algorithm.

Consider a variable vector $\mathbf{v} = (v(0), \ldots, v(N))$ and an operator $L : \mathbb{R}^N \to \mathbb{R}^N$ defined by

$$[L(\mathbf{v})]_i := \max_{a\in A(i)} \left\{ r(i,a) + \beta \sum_{j=1}^{N} p_{ij}(a) v(j) \right\} \tag{1.65}$$

for each $i \in S$. The right-hand side of (1.65) can be interpreted as defining some deterministic decision rule $g_{\mathbf{v}}$ for selecting the maximizing action $a_i$ in each state $i$. We now have

$$L(\mathbf{v}) = L(g_{\mathbf{v}})(\mathbf{v}),$$

where $L(g_{\mathbf{v}})(\mathbf{u}) = \mathbf{r}(g_{\mathbf{v}}) + \beta P(g_{\mathbf{v}})\mathbf{u}$, as before. In addition, define a vector-valued function on $\mathbb{R}^N$ by

$$\Psi(\mathbf{v}) := L(\mathbf{v}) - \mathbf{v}, \quad \mathbf{v} \in \mathbb{R}^N.$$

Since the optimality equation is valid, we have that $L(\mathbf{v}_\beta) = \mathbf{v}_\beta$ at the value vector $\mathbf{v}_\beta$. This means the search for this unique solution is the same as the search for the unique zero of $\Psi(\mathbf{v})$. This brings us to the following unconstrained minimization problem:

$$\min \ \frac{1}{2}||\Psi(\mathbf{v})||^2$$

subject to: $(M)$

$$\mathbf{v} \in \mathbb{R}^N.$$

**Proposition 1.8.2.**

*(i) Let* $\Psi'(\mathbf{v}) := \left[\frac{\partial[\Psi(\mathbf{v})]_{\bar{i}}}{\partial v(i)}{}^{N}_{\bar{i},i=1}\right]$ *wherever these partial derivatives are defined. Then,*

$$\Psi'(\mathbf{v}) = -\left[I - \beta P(g_{\mathbf{v}})\right].$$

*(ii) If* $J(\mathbf{v}) := \frac{1}{2}\left[\Psi(\mathbf{v})\right]^T\left[\Psi(\mathbf{v})\right]$*, then (wherever it is defined) the gradient of* $J(\mathbf{v})$ *is given by the row vector*

$$\nabla J(\mathbf{v}) = -\left[\Psi(\mathbf{v})\right]^T\left[I - \beta P(g_v)\right].$$

*(iii) with* $J(\mathbf{v})$ *as in (ii),* $J(\mathbf{v}) = 0$ *if and only if* $\Psi(\mathbf{v}) = 0$*.*

**Corollary 1.8.2.** *Consider the policy improvement algorithm and its typical update of the current estimate* $\mathbf{v}^k$ *of the value vector, that is*

$$\mathbf{v}^{k+1} = \mathbf{v}_\beta(\mathbf{f}^{k+1})$$

*then* $\mathbf{v}^{k+1}$ *also can be obtained by one step of the Newton's method applied to the unconstrained minimization problem* $(M)$*. That is,*

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \left[\Psi'(\mathbf{v}^k)\right]^{-1}\Psi(\mathbf{v}^k).$$

## 1.9 Connection Between the Discounted and the Limiting Average Models

The difference between the discounted Markov decision model $\Gamma_\beta$ and the limiting average model $\Gamma_\alpha$ (see (1.13) and (1.35)) is the difference between Abel summability and Cesaro summability. From the theory of summability it can be shown that these two are closely interlinked. We will briefly touch upon this connection here, and investigate it further in chapter 3. The

reason for investigating this is that it provides us with tools for analyzing the more difficult limiting average Markov decision process. In particular, we shall make use of the inequalities

$$\begin{aligned}\liminf_{T\to\infty}\left(\frac{1}{T+1}\right)\sum_{t=0}^{T}d_t &\leq \liminf_{\beta\to 1^-}(1-\beta)\sum_{t=0}^{\infty}\beta^t d_t \\ &\leq \limsup_{\beta\to 1^-}(1-\beta)\sum_{t=0}^{\infty}\beta^t d_t \\ &\leq \limsup_{T\to\infty}\left(\frac{1}{T+1}\right)\sum_{t=0}^{T}d_t \qquad (1.66)\end{aligned}$$

where $\{d_t\}_{t=0}^{\infty}$ is an arbitrary bounded sequence of real numbers.

Now if we take $d_t := \mathbb{E}_{i\pi}(R_t)$ for each $t = 0, 1, 2, \ldots$ for each fixed $\pi \in \mathbf{F}_B$ and $i \in S$ we immediately see that

$$v_\alpha(i,\pi) \leq \liminf_{\beta\to 1^-}(1-\beta)v_\beta(i,\pi). \qquad (1.67)$$

We have already seen that each $\mathbf{f}$ defines a Markov chain with the transition matrix $P(\mathbf{f})$ and the stationary distribution matrix $Q(\mathbf{f})$, the Cesaro-limit matrix of $P(\mathbf{f})$. Let us define the *deviation matrix* $D(\mathbf{f})$ of $P(\mathbf{f})$ by

$$D(\mathbf{f}) := \lim_{\beta\to 1^-}\sum_{t=0}^{\infty}\beta^t\left[P^t(\mathbf{f}) - Q(\mathbf{f})\right]$$

where the existence of the above limit follows from the next result.

**Proposition 1.9.1.** *Given a Markov matrix $P(\mathbf{f})$ defined by a stationary strategy $\mathbf{f}$, we have that:*

*(i) $Q(\mathbf{f})$ is well defined and satisfies*

$$Q(\mathbf{f})P(\mathbf{f}) = P(\mathbf{f})Q(\mathbf{f}) = Q(\mathbf{f})Q(\mathbf{f}) = Q(\mathbf{f}).$$

*(ii)* $\lim_{\beta\to 1^-}\left\{(1-\beta)\sum_{t=0}^{\infty}\left[P^t(\mathbf{f}) - Q(\mathbf{f})\right]\right\} = 0$

*(iii) The inverse* $[I - P(\mathbf{f}) + Q(\mathbf{f})]^{-1}$ *exists and*

$$[I - P(\mathbf{f}) + Q(\mathbf{f})]^{-1} = \lim_{\beta\to 1^-}\sum_{t=0}^{\infty}\left[\beta^t\left[P(\mathbf{f}) - Q(\mathbf{f})\right]\right]^t.$$

*(iv) The deviation matrix $D(\mathbf{f})$ is well defined and, in addition satisfies*

$$\begin{aligned}D(\mathbf{f}) &= [I - P(\mathbf{f}) + Q(\mathbf{f})]^{-1} - Q(\mathbf{f}) \\ &= lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{t}[P^{k-1}(\mathbf{f}) - Q(\mathbf{f})] \\ &= lim_{\beta\to 1^-}\sum_{t=0}^{\infty}[P^t(\mathbf{f}) - Q(\mathbf{f})].\end{aligned}$$

*(v)* $Q(\mathbf{f})D(\mathbf{f}) = D(\mathbf{f})Q(\mathbf{f}) = 0$ *and* $D(\mathbf{f})\mathbf{1} = \mathbf{0}$*, where* $\mathbf{1}$ *is a* $N$*-vector with 1 in every entry*

The above properties of $P(\mathbf{f})$, $Q(\mathbf{f})$ and $D(\mathbf{f})$ immediately lead to the following important connection between the discounted and the limiting average criteria.

**Proposition 1.9.2.** *Given any stationary strategy* $\mathbf{f} \in \mathbf{F}_S$ *and the associated Markov matrix* $P(\mathbf{f})$*, we have that:*

*(i) For any* $\beta \in (0, 1)$*,*

$$[I - \beta P(\mathbf{f})]^{-1} = \frac{1}{1-\beta} Q(\mathbf{f}) + D(\mathbf{f}) + E(\beta, \mathbf{f})$$

*where*

$$E(\beta, \mathbf{f}) := \sum_{t=0}^{\infty} \beta^t [P^t(\mathbf{f}) - Q(\mathbf{f})] - D(\mathbf{f}).$$

*(ii)* $\mathbf{v}_\beta(\mathbf{f}) = \frac{1}{1-\beta}\mathbf{v}_\alpha(\mathbf{f}) + \boldsymbol{u}(\mathbf{f}) + \boldsymbol{e}(\beta, \mathbf{f})$*, where*

$$\lim_{\beta \to 1^-} \boldsymbol{e}(\beta, \mathbf{f}) = \mathbf{0}$$

*and* $\boldsymbol{u}(\mathbf{f})$ *is an appropriate vector, called the bias vector of* $\mathbf{f}$*.*

*(iii)* $\lim_{\beta \to 1^-}(1-\beta)\mathbf{v}_\beta(\mathbf{f}) = \mathbf{v}_\alpha(\mathbf{f})$

The next theorem now states that a deterministic control $\mathbf{f}^0$ exists that is simultaneously optimal in the limiting average model $\Gamma_\alpha$ and in a whole family of discounted models $\Gamma_\beta$ for $\beta$ sufficiently near 1.

**Theorem 1.9.1.**

*(i)* $\beta^0 \in [0, 1)$ *and a deterministic control* $\mathbf{f}^0 \in \mathbf{F}_D$ *exist such that for all* $\beta \in [\beta^0, 1)$

$$\mathbf{v}_\beta(\mathbf{f}^0) = \max_{\pi \in \mathbf{F}_B} \mathbf{v}_\beta(\pi).$$

*(ii) With* $\mathbf{f}^0$ *as in part (i) above, we have that*

$$\mathbf{v}_\alpha(\mathbf{f}^0) = \max_{\pi \in \mathbf{F}_B} \mathbf{v}_\alpha(\pi).$$

From now on, a stationary control that is optimal in the discounted model for all values of the discount factor sufficiently near one will be called a *uniformly discount optimal control*. With the help of these uniformly discount optimal controls we now state two important properties of the limiting average value vector $\mathbf{v}_\alpha$. These properties will play an important role in the next section.

**Proposition 1.9.3.** *Let* $\mathbf{v}_\alpha$ *be the limiting average value vector. Then for all* $i \in S$ *and* $a \in A(i)$

$$v_\alpha(i) \geq \sum_{j=1}^{N} p_{ij}(a)\mathbf{v}_\alpha(j).$$

**Proposition 1.9.4.** *Let* $\mathbf{v}_\alpha$ *be the limiting average value vector. Then an* $N$*-vector* $\mathbf{u}$ *exists such that for all* $a \in A(i),\ i \in S$,

$$v_\alpha(i) + u(i) \geq r(i,a) + \sum_{j=1}^{N} p_{ij}(a)u(j).$$

## 1.10 Linear Programming and the Multichain Limiting Average Process

The general "multichain" limiting average Markov decision process can also be solved completely with the help of linear programs. These linear programs are structurally related to those that were previously developed for the discounted process $\Gamma_\beta$, and the irreducible limiting average process.
For a state $i \in S$, we define an $N \times m(i)$ matrix (block) $W_i$ whose $(j,(i,a))$th element is given by

$$w_{j(i,a)} := \delta(i,j) - p_{ij}(a) \tag{1.68}$$

for each $j \in S$ and $a = 1, 2, \ldots, m(i)$ in $A(i)$. Corresponding to this $i$th block we define five $m(i) \times 1$ column vectors:

$$\begin{aligned}
\mathbf{x}_i &= (x_{i1}, x_{i2}, \ldots, x_{im(i)})^T \\
\mathbf{y}_i &= (y_{i1}, y_{i2}, \ldots, y_{iN})^T \\
\mathbf{r}_i &= (r(i,1), r(i,2), \ldots, r(i,m(i)))^T \\
\mathbf{1}_i &= (1, 1, \ldots, 1)^T \\
\mathbf{0}_i &= (0, 0, \ldots, 0)^T.
\end{aligned}$$

If we put these blocks together we have

$$\begin{aligned}
\mathbf{x}^T &= (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T) \\
\mathbf{y}^T &= (\mathbf{y}_1^T, \ldots, \mathbf{x}_N^T) \\
\mathbf{r}^T &= (\mathbf{r}_1^T, \ldots, \mathbf{r}_N^T) \\
\mathbf{J}_1^T &= (\mathbf{1}_1^T, \mathbf{0}_2^T, \ldots, \mathbf{0}_N^T) \\
\mathbf{J}_2^T &= (\mathbf{0}_1^T, \mathbf{1}_2^T, \ldots, \mathbf{0}_N^T) \\
&\vdots \\
\mathbf{J}_N^T &= (\mathbf{0}_1^T, \mathbf{0}_2^T, \ldots, \mathbf{1}_N^T)
\end{aligned}$$

each of which is a $1 \times m$ row vector where $m = \sum_{i=1}^{N} m(i)$, as before. In addition, we define two $N \times m$ matrices:

$$\begin{aligned} W &:= \left( W_1 \vdots W_2 \vdots \ldots \vdots W_N \right) \\ J &:= \left( \mathbf{J}_1 \vdots \mathbf{J}_2 \vdots \ldots \vdots \mathbf{J}_N \right)^T . \end{aligned}$$

Finally, we introduce three $1 \times N$ row vectors:

$$\begin{aligned} \mathbf{v}^T &= (v(1), \ldots, v(N)) \\ \mathbf{u}^T &= (u(1), \ldots, u(N)) \\ \gamma^T &= (\gamma(1), \ldots, \gamma(N)) \end{aligned}$$

where each $\gamma(i) > 0$ and $\sum_{i=1}^{N} \gamma(i) = 1$.

Now we can use the following primal-dual pair of linear programs for an arbitrary limiting average Markov decision process $\Gamma_\alpha$:

$$\min \left[ \gamma^T \mathbf{v} \right]$$

subject to: $(P_\alpha)$

$$\left( \mathbf{u}^T, \mathbf{v}^T \right) \begin{pmatrix} W & \vdots & 0 \\ \cdots & \cdots & \cdots \\ J & \vdots & W \end{pmatrix} \geq \left( \mathbf{r}^T, \mathbf{0}^T \right)$$

and

$$\max \left[ \mathbf{r}^T \mathbf{x} \right]$$

subject to:

$$\begin{pmatrix} W & \vdots & 0 \\ \cdots & \cdots & \cdots \\ J & \vdots & W \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \gamma \end{pmatrix}$$

$$\mathbf{x},\ \mathbf{y} \geq \mathbf{0}.$$

The feasible region of the primal problem $(P_\alpha)$ consists precisely of the type of inequalities that appeared in Proposition (1.9.4) and (1.9.3), respectively. The following algorithm constructs an optimal strategy for a possibly multichain process $\Gamma_\alpha$.

**Algorithm 1.10.1.** Construction of an Optimal Strategy in $\Gamma_\alpha$

**Step 1.** Find any extreme optimal solution $\begin{pmatrix} \mathbf{x}^* \\ \mathbf{y}^* \end{pmatrix}$ of the dual linear program $(D_\alpha)$.

**Step 2.** Define the set of states

$$\mathbf{S}^* := \left\{ i \in \mathbf{S} | x_i^* = \sum_{a=1}^{m(i)} x_{ia}^* > 0 \right\}.$$

**Step 3.** If $i \in \mathbf{S}^*$, select any action $a_i$ in $A(i)$ such that $x_{ia}^* > 0$. If $i \notin \mathbf{S}^*$, select any action $a_i$ in $A(i)$ such that $y_{ia}^* > 0$.

**Step 4.** Construct $\mathbf{f}^* \in \mathbf{F}_D$ according to

$$f^*(i,a) = \begin{cases} 1, & \text{if } a = a_i \\ 0, & \text{otherwise.} \end{cases}$$

(Whenever it is convenient, we also shall write $f^*(i) = a_i$ rather than $f^*(i,a)$.)

To prove that $\mathbf{f}^*$ constructed above is indeed an optimal control in $\Gamma_\alpha$, we need to show that that $\mathbf{f}^*$ is well defined first, and that the value vector $\mathbf{v}_\alpha$ can be obtained from any optimal solution of the primal linear program $(P_\alpha)$.

**Proposition 1.10.1.**

*(i) Let $(\mathbf{u}^T, \mathbf{v}^T)$ be any feasible solution of $(P_\alpha)$. Then componentwise $\mathbf{v} \geq \mathbf{v}_\alpha$, where $\mathbf{v}_\alpha$ is the value vector of $\Gamma_\alpha$.*

*(ii) If $\left((\mathbf{u}^T)^T, (\mathbf{v}^*)^T\right)$ is any optimal solution of $(P_\alpha)$, then $\mathbf{v}^* = \mathbf{v}_\alpha$.*

*(iii) The dual problem $(D_\alpha)$ possesses a finite optimal solution, and the deterministic control $\mathbf{f}^*$ defined in Step 4 above is well defined.*

**Proposition 1.10.2.** *Let $\left((\mathbf{u}^*)^T, (\mathbf{v}^*)^T\right)^T$ and $\left((\mathbf{x}^*)^T, (\mathbf{y}^*)^T\right)^T$ be a pair of optimal solutions of $(P_\alpha)$ and $(D_\alpha)$, respectively, and $\mathbf{f}^* \in \mathbf{F}_D$ be the control constructed by Algorithm (1.10.1). Then*

*(i)* $[I - P(\mathbf{f}^*)]\mathbf{v}^* = (0)$
*and*

*(ii)* $[\mathbf{v}^*]_i + \{[I - P(\mathbf{f}^*)]\mathbf{u}^*\}_i = [\mathbf{r}(\mathbf{f}^*)]_i$ *for all* $i \in \mathbf{S}^*$.

**Proposition 1.10.3.** *Let $\left((\mathbf{x}^*)^T, (\mathbf{y}^*)^T\right)$ and $\mathbf{f}^* \in \mathbf{F}_D$ be as in the Proposition (1.10.1). Then:*

*(i) The set $\mathbf{S}^*$ is a closed set in the Markov chain $P(\mathbf{f}^*)$, that is,*

$$p_{ij}(a_i) = 0 \tag{1.69}$$

*whenever $i \in \mathbf{S}^*$ and $j \neq \mathbf{S}^*$.*

*(ii) The set* $\mathbf{S}_c^* = \mathbf{S} \backslash \mathbf{S}^*$ *consists of transient states of the Markov chain* $P(\mathbf{f}^*)$.

The main result now follows from the preceding propositions.

**Theorem 1.10.1.** *Let* $\left((\mathbf{u}^*)^T, (\mathbf{v}^*)^T\right)^T$ *be an optimal solution of the linear program* $(P_\alpha)$, $\left((\mathbf{x}^*)^T, (\mathbf{y}^*)^T\right)$ *be an extreme optimal solution of the dual* $(D_\alpha)$, *and* $\mathbf{f}^* \in \mathbf{F}_D$ *be constructed by the Algorithm 1.10.1. Then*

$$\mathbf{v}_\alpha(\mathbf{f}^*) = \mathbf{v}^* = \mathbf{v}_\alpha. \tag{1.70}$$

*That is,* $\mathbf{f}^*$ *is an optimal deterministic strategy in* $\Gamma_\alpha$.

# Chapter 2

# Stochastic Games

## 2.1 The Discounted Stochastic Games

### 2.1.1 Markov Decision Process Perspective

We will now take a look at stochastic games from the perspective of $\beta$-Discounted Markov Decision Models. We will restrict ourselves here to games with two controllers, referred to as player 1 and player 2, respectively. Furthermore, we will take the number of states and actions to be finite. The notation of Markov decision processes can easily be adapted to stochastic games. The actions of the players in state $i \in S = \{1, 2, \ldots, N\}$ at time $t$ will be denoted with $a^1 \in A^1(i)$ for player 1 and $a^2 \in A^2(i)$ for player 2, with the rewards being $r^1(i, a)$ and $r^2(i, a)$. The stationary transition probabilities generalize to:

$$p_{ij}(a^1, a^2) := \mathbb{P}\left\{S_{t+1} = j \mid S_t = i, A_t^1 = a^1, A_t^2 = a^2\right\} \tag{2.1}$$

for all $t = 0, 1, 2, \ldots$. The state at time $t$ is $S_t$, and $A_t^1, A_t^2$ denote the actions chosen by players 1 and 2 at time 2, respectively.

The set of stationary strategies $\mathbf{F}_S$ of player 1 is defined as in Section 2.2, and the set of stationary strategies $\mathbf{G}_S$ of player 2 is defined in the same way. If $\mathbf{g} = (\mathbf{g}(1), \mathbf{g}(2), \ldots, \mathbf{g}(N)) \in \mathbf{G}_S$, then each $\mathbf{g}(i)$ is an $m^2(i)$-dimensional probability vector, where $m^2(i) = |A^2(i)|$, the cardinality of $A^2(i)$. In this section we only consider stationary strategies.

The expected reward at stage $t$ to player $k$ (with $k \in \{1, 2\}$) resulting from $(\mathbf{f}, \mathbf{g})$ and an initial state $i$ now will be denoted by $\mathbb{E}_{i\mathbf{fg}}(R_t^k)$. Consequently, the overall *discounted value of a strategy pair* $(\mathbf{f}, \mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$ to player $k$ will be given by

$$v_\beta^k(i, \mathbf{f}, \mathbf{g}) := \sum_{t=0}^{\infty} \beta^t \, \mathbb{E}_{i\mathbf{fg}}(R_t) \tag{2.2}$$

where $\beta \in [0, 1)$ and $k \in \{1, 2\}$.

Within the space of strategies $\mathbf{F}_S \times \mathbf{G}_S$, we need to find a pair $(\mathbf{f}, \mathbf{g})$ of strategies that constitutes a solution to the game. The Markov control problem introduced in section 1.2 is no longer adequate because its solution will usually depend on player 2's strategy $\mathbf{g} \in \mathbf{G}_S$. This interdependence requires us to impose a "behavioural assumption" on the way that the controllers play this game.

We will assume that we are dealing with noncooperative games with complete information. That is, the players do not work together to maximize their individual overall reward function, and the players have precise knowledge about each other's presence in the game and reward functions. We can now use the Nash equilibrium stated below to formulate a solution to the game.

We shall say that $(\mathbf{f}^0, \mathbf{g}^0) \in \mathbf{F}_S \times \mathbf{G}_S$ is a *Nash equilibrium point* (or EP, for short) of the discounted stochastic game $\Gamma_\beta$ if

$$\mathbf{v}_\beta^1(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta^1(\mathbf{f}^0, \mathbf{g}^0) \quad \text{for all } \mathbf{f} \in \mathbf{F}_S \tag{2.3}$$

and

$$\mathbf{v}_\beta^2(\mathbf{f}^0, \mathbf{g}) \leq \mathbf{v}_\beta^2(\mathbf{f}^0, \mathbf{g}^0) \quad \text{for all } \mathbf{g} \in \mathbf{G}_S. \tag{2.4}$$

This seems to imply that there is no reason for either of them to deviate from $(\mathbf{f}^0, \mathbf{g}^0)$. A complication however is that, in general, there can be many Nash equilibria with very different payoffs to the players. To avoid this problem we will restrict ourselves to zero-sum games. A discounted stochastic game will be called *zero-sum* if

$$r^1(i, a^1, a^2) + r^2(i, a^1, a^2) = 0 \tag{2.5}$$

for all $i \in S, a^1 \in A^1(i), a^2 \in A^2(i)$. This means we can simply write:

$$r(i, a^1, a^2) := r^1(i, a^1, a^2) = -r^2(i, a^1, a^2)$$

for all $i \in S, a^1 \in A^1(i), a^2 \in A^2(i)$. A consistent extension of this definition leads to

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) := \mathbf{v}_\beta^1(\mathbf{f}, \mathbf{g}) = -\mathbf{v}_\beta^2(\mathbf{f}, \mathbf{g})$$

for all $\mathbf{f}, \mathbf{g} \in \mathbf{F}_S \times \mathbf{G}_S$, where the last equality follows immediately from the zero-sum property and (2.2).

In view of the above, and if $(\mathbf{f}^0, \mathbf{g}^0) \in \mathbf{F}_S \times \mathbf{G}_S$ is an EP, the two sets of inequalities defining an equilibrium point reduce to the single set of *saddle-point* inequalities

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, g^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}) \tag{2.6}$$

for all $\mathbf{f} \in \mathbf{F}_S$ and $\mathbf{g} \in \mathbf{G}_S$. In such a case we call $\mathbf{f}^0$ ($\mathbf{g}^0$) an *optimal stationary strategy* for player 1 (2). We now have the following theorem:

**Theorem 2.1.1.** *Consider the saddle-point optimality condition (2.6). Suppose this condition is satisfied by both* $(\mathbf{f}^0, \mathbf{g}^0)$ *and* $(\hat{f}, \hat{g})$ *in* $\mathbf{F}_S \times \mathbf{G}_S$. *We now have:*

*(i)* $\mathbf{v}_\beta = \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) = \mathbf{v}_\beta(\hat{f}, \hat{g}) = \mathbf{v}_\beta(\mathbf{f}^0, \hat{g}) = \mathbf{v}_\beta(\hat{f}, \mathbf{g}^0)$

*(ii)* $\mathbf{v}_\beta(i, \mathbf{f}^0, \mathbf{g}^0) = \max_{\mathbf{F}_S} \min_{\mathbf{G}_S} \mathbf{v}_\beta(i, \mathbf{f}, \mathbf{g}) = \min_{\mathbf{G}_S} \max_{\mathbf{F}_S} \mathbf{v}_\beta(i, \mathbf{f}, \mathbf{g})$ *for every* $i \in \mathbf{S}$

*(iii) Conversely, if the "minimaxes"*

$$\max_{\mathbf{F}_S} \min_{\mathbf{G}_S} \mathbf{v}_\beta(i^0, \mathbf{f}, \mathbf{g}) \text{ and } \min_{\mathbf{G}_S} \max_{\mathbf{F}_S} \mathbf{v}_\beta(i^0, \mathbf{f}, \mathbf{g}).$$

*exist and are equal for some fixed* $i^0 \in \mathbf{S}$, *then stationary strategies* $\mathbf{f}^0$ *and* $\mathbf{g}^0$ *exist for players 1 and 2, respectively, satisfying (2.6) for that same state* $i^0$. *This property (and (ii)) motivate the name minimax optimality that often is used to describe condition (2.6)*

*Proof.*

(i) Suppose that both $(\mathbf{f}^0, \mathbf{g}^0)$ and $(\hat{\mathbf{f}}, \hat{\mathbf{g}})$ satisfy condition (3.4). This means

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(f^0, \mathbf{g})$$

for all $\mathbf{f} \in \mathbf{F}_S$ and $\mathbf{g} \in \mathbf{G}_S$. This implies that

$$\mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(f^0, \hat{\mathbf{g}}). \tag{2.7}$$

On the other hand, we also have

$$\mathbf{v}_\beta(\mathbf{f}, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g}).$$

for all $\mathbf{f} \in \mathbf{F}_S$ and $\mathbf{g} \in \mathbf{G}_S$, and so

$$\mathbf{v}_\beta(\mathbf{f}^0, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g}^0). \tag{2.8}$$

Putting (2.7) and (2.8) together gives us the equality.

(ii) We already know from (2.6)

$$\max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(i, \mathbf{f}, \mathbf{g}^0) = \mathbf{v}_\beta(i, \mathbf{f}^0, \mathbf{g}^0) = \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(i, \mathbf{f}^0, \mathbf{g}).$$

Furthermore,

$$\max_{\mathbf{f} \in \mathbf{F}_S} \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(i, \mathbf{f}, \mathbf{g}) \geq \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(i, \mathbf{f}^*, \mathbf{g})$$

and

$$\min_{\mathbf{g} \in \mathbf{G}_S} \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(i, \mathbf{f}, \mathbf{g}) \leq \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(i, \mathbf{f}, \mathbf{g}^*)$$

for all $\mathbf{f}^* \in \mathbf{F}_S$ and $\mathbf{g}^* \in \mathbf{G}_S$. This leads to

$$\begin{aligned} \max_{\mathbf{f}\in\mathbf{F}_S} \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\beta(i,\mathbf{f},\mathbf{g}) &\geq \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\beta(i,\mathbf{f}^0,\mathbf{g}) \\ &= \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\beta(i,\mathbf{f},\mathbf{g}^0) \\ &\geq \min_{\mathbf{g}\in\mathbf{G}_S} \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\beta(i,\mathbf{f},\mathbf{g}) \end{aligned} \tag{2.9}$$

and

$$\begin{aligned} \min_{\mathbf{g}\in\mathbf{G}_S} \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\beta(i,\mathbf{f},\mathbf{g}) &\geq \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\beta(i,\mathbf{f}^0,\mathbf{g}) \\ &= \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\beta(i,\mathbf{f},\mathbf{g}^0) \\ &\geq \max_{\mathbf{f}\in\mathbf{F}_S} \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\beta(i,\mathbf{f},\mathbf{g}) \end{aligned} \tag{2.10}$$

which proves the equality.

(iii) Define $\mathbf{v}_\beta := \mathbf{v}_\beta(i^0,\mathbf{f}^0,\mathbf{g}^0)$, $F_\beta(\mathbf{f}) := \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\beta(i^0,\mathbf{f},\mathbf{g})$ and $G_\beta(\mathbf{g}) := \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\beta(i^0,\mathbf{f},\mathbf{g})$, with $\mathbf{f} \in \mathbf{F}_S, \mathbf{g} \in \mathbf{G}_S$. From our assumption it follows that there is a $\mathbf{f}^0 \in \mathbf{F}_S$ such that $F_\beta(\mathbf{f}^0) = \mathbf{v}_\beta$. This gives us

$$\mathbf{v}_\beta = \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\beta(i^0,\mathbf{f}^0,\mathbf{g}) \leq \mathbf{v}_\beta(i^0,\mathbf{f}^0,\mathbf{g}) \tag{2.11}$$

for all $\mathbf{g} \in \mathbf{G}_S$. Similarly there must be a $\mathbf{g}^0 \in \mathbf{G}_S$ for which

$$\mathbf{v}_\beta = G_\beta(\mathbf{g}^0) = \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\beta(i^0,\mathbf{f},\mathbf{g}^0) \geq \mathbf{v}_\beta(i^0,\mathbf{f},\mathbf{g}^0) \tag{2.12}$$

for all $\mathbf{f} \in \mathbf{F}_S$. This gives us the two stationary optimal strategies $\mathbf{f}^0$ and $\mathbf{g}^0$.

□

This means the discounted value vectors of all optimal strategy pairs coincide and will be called the *value vector* of the zero-sum game $\Gamma_\beta$ and denoted by

$$\mathbf{v}_\beta = (v_\beta(1), v_\beta(2), \ldots, v_\beta(N))^T.$$

### 2.1.2 Matrix Game Perspective

Up till now we have viewed a discounted stochastic game $\Gamma_\beta$ as a multi-controller generalization of the discounted Markov decision process, but we can also approach it from the perspective of static matrix games. If $m^1(i)$ (or $m^2(i)$) is the cardinality of $A^1(i)$ (or $A^2(i)$) for each $i \in S$, then we can define $N$ matrix games

$$R(i) = \left[r(i,a^1,a^2)\right]_{a^1=1,a^2=1}^{m^1(i),m^2(i)}$$

corresponding with the states of $\Gamma_\beta$. Now each game not only has a payoff $r(i, a^1, a^2)$ but also a probability transition $p_{ij}(a^1, a^2)$ leading to the next game. In the same way as in chapter 2 we can hypothesize that if we assume that $\mathbf{v}_\beta$ exists, and we know how to play optimally from the next stage onward, then we have the following matrix game in the current stage:

$$R(i, \mathbf{v}_\beta) = [r(i, a^1, a^2) + \beta \sum_{j \in S} p_{ij}(a^1, a^2)\mathbf{v}_\beta(j)]_{a^1=1, a^2=1}^{m^1(i), m^2(i)}. \tag{2.13}$$

This is also stated by the following theorem:

**Theorem 2.1.2.** (Shapley's Theorem)
*The discounted, zero-sum, stochastic game $\Gamma_\beta$ possesses the value vector $\mathbf{v}_\beta$ that is the unique solution of the equations*

$$v(i) = \text{val}[R(i, \mathbf{v})] \tag{2.14}$$

*for all $i \in S$, where $\mathbf{v}^T = (v_\beta(1), \ldots, v_\beta(N))^T$. Furthermore, if $(\mathbf{f}^0(i), \mathbf{g}^0(i))$ is an optimal (possibly mixed) strategy pair in the matrix game $R(i, \mathbf{v}_\beta)$ for each $i \in S$, then $\mathbf{f}^0 = (\mathbf{f}^0(1), \mathbf{f}^0(2), \ldots, \mathbf{f}^0(N))$ is an optimal strategy for player 1 in $\Gamma_\beta$, and $\mathbf{g}^0 = (\mathbf{g}^0(1), \mathbf{g}^0(2), \ldots, \mathbf{g}^0(N))$ is an optimal stationary strategy for player 2 in $\Gamma_\beta$.*

In the remainder of this section we will extend the notation of Chapter 2 to the discounted stochastic games.

For a fixed pair of stationary strategies $\mathbf{f} = (\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(N))$ and $\mathbf{g} = (\mathbf{g}(1), \mathbf{g}(2), \ldots, \mathbf{g}(N))$, for players 1 and 2, respectively, we shall adopt the convention that $\mathbf{f}$ is a block-row vector (as in Chapter 2) while $\mathbf{g}$ is a block-column vector. That is, if we define $m^2 := \sum_{i=1}^N m^2(i)$, then $\mathbf{g}$ is an $m^2$-dimensional column whose $i$th block $\mathbf{g}(i)$ is $m^2(i)$-dimensional. The following quantities will be used:

(i) $p_{ij}(\mathbf{f}, a^2) := \sum_{a^1=1}^{m^1(i)} p_{ij}(a^1, a^2) f(i, a^1);\ i, j \in S,\ a^2 \in A^2(i)$

(ii) $p_{ij}(a^1, \mathbf{g}) := \sum_{a^2=1}^{m^2(i)} p_{ij}(a^1, a^2) g(i, a^2);\ i, j \in S,\ a^1 \in A^1(i)$

(iii) $p_{ij}(\mathbf{f}, \mathbf{g}) := \sum_{a^1=1}^{m^1(i)} \sum_{a^2=1}^{m^2(i)} p_{ij}(a^1, a^2) f(i, a^1) g(i, a^2);\ i, j \in S$

(iv) The Markov probability transition matrix induced by $(\mathbf{f}, \mathbf{g})$:

$$P(\mathbf{f}, \mathbf{g}) := (p_{ij}(\mathbf{f}, \mathbf{g}))_{i,j=1}^N$$

(v) $$\begin{aligned} r(i, \mathbf{f}, a^2) &:= \sum_{a^1=1}^{m^1(i)} r(i, a^1, a^2)\mathbf{f}(i, a^1) \\ &= [\mathbf{f}(i)R(i)]_{a^2};\ i \in S,\ a^2 \in A^2(i) \end{aligned}$$

(vi) $$\begin{aligned} r(i, a^1, \mathbf{g}) &:= \sum_{a^2=1}^{m^2(i)} r(i, a^1, a^2)\mathbf{g}(i, a^2) \\ &= [R(i)\mathbf{g}(i)]_{a^1};\ i \in S,\ a^1 \in A^1(i) \end{aligned}$$

(vii) $r(i,\mathbf{f},\mathbf{g}) := \sum_{a^1=1}^{m^1(i)} \sum_{a^2=1}^{m^2(i)} r(i,a^1,a^2)\mathbf{f}(i,a^1)\mathbf{g}(i,a^2)$
$= \mathbf{f}(i)R(i)\mathbf{g}(i);\ i \in S$

(viii) $N$-dimensional column vector:

$$\mathbf{r}(\mathbf{f},\mathbf{g}) := (r(1,\mathbf{f},\mathbf{g}), r(2,\mathbf{f},\mathbf{g}), \ldots, r(N,\mathbf{f},\mathbf{g}))^T$$

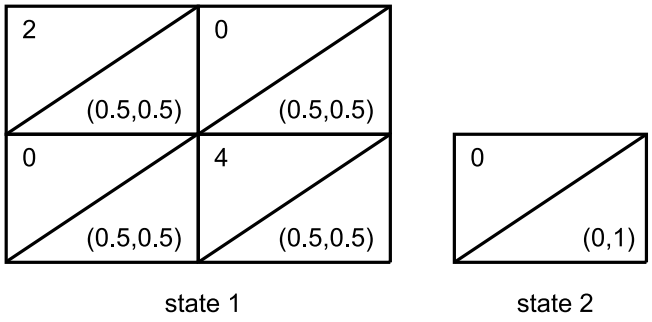
(ix) $N$-dimensional discounted value vector of the pair $(\mathbf{f},\mathbf{g})$:

$$\mathbf{v}_\beta(\mathbf{f},\mathbf{g}) := [I - \beta P(\mathbf{f},\mathbf{g})]^{-1}\,\mathbf{r}(\mathbf{f},\mathbf{g})$$

## 2.2 Linear Programming and the Discounted Stochastic Games

Now that we know that the value exists for $\Gamma_\beta$, a logical question is whether we can apply the linear programming formulation used in chapter 2 to $\Gamma_\beta$ as well. The following example however demonstrates that we run into problems when we try to do this.

**Example 2.2.1.** Let $\mathbf{S} = \{1,2\}, A^1(1) = A^2(1) = \{1,2\}, A^1(2) = a^2(2) = \{1\}, \beta = \frac{1}{4}$, and the reward and transition data be




The second state is absorbing, so $v_{\frac{1}{4}}(2) = 0$. The optimality equation (2.14) now reduces to finding $v := v_{\frac{1}{4}}(1)$ that satisfies

$$v = \text{val}\begin{bmatrix} 2+\frac{1}{4}v & 0 \\ 0 & 4+\frac{1}{4}v \end{bmatrix}.$$

Since $v$ is evidently nonnegative, we can use the following property of matrix games:
A game $\begin{pmatrix} r_{12} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}$ is completely mixed, if $r_{12} > r_{11}, r_{22}$ and $r_{21} > r_{11}, r_{22}$, or $r_{11} > r_{12}, r_{21}$ and $r_{22} > r_{12}, r_{21}$. It follows from the Shapley-Snow theorem (see [6], appendix G), that the formula for the value becomes

$$\text{val}\begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \frac{r_{11}r_{22} - r_{12}r_{21}}{r_{11} + r_{22} - r_{12} - r_{21}}.$$

Applying this to the example we get $v = \frac{(2+\frac{1}{4}v)(4+\frac{1}{4}v)}{6+\frac{1}{2}v}$, or equivalently, $7v^2 + 72v - 128 = 0$. It now follows that $v = \frac{1}{14}(72 + \sqrt{1829})$ or that the value vector is

$$\mathbf{v}_{\frac{1}{4}} = \left(\frac{1}{14}(72 + \sqrt{1829}), 0\right)^T.$$

Since linear programming can only produce rational values, the fact that the (unique) value vector above contains irrational entries implies that, in general, we cannot expect discounted stochastic games to be solved by linear programming.

There are, however, some interesting subclasses of stochastic games that can be solved by linear programming. We will take a look at three of them.

### 2.2.1 Single-Controller Discounted Games

These are the games where the transition probabilities depend on the actions of one player only. Thus $\Gamma_\beta(1)$ will be the player 1-controller game defined by the property

$$p_{ij}(a^1, a^2) \equiv p_{ij}(a^1) \tag{2.15}$$

for all $i, j \in S, a^1 \in A^1(i), a^2 \in A^2(i)$. The player 2-controlled game can be defined similarly.

Given a player 1-controlled game $\Gamma_\beta(1)$ some of the formulas of the previous section acquire a special form. In particular, $P(\mathbf{f}, \mathbf{g}) = P(\mathbf{f})$ and

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = [I - \beta P(\mathbf{f})]^{-1} \mathbf{r}(\mathbf{f}, \mathbf{g}) \tag{2.16}$$

for all $(\mathbf{f}, \mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$.

Due to the single-controller hypothesis (2.15) the game $\Gamma_\beta(1)$ can be expected to behave more like a Markov decision process with respect to player 1 than the general discounted game. Suppose now that player 2 follows some stationary strategy $\mathbf{g} \in \mathbf{G}_S$. Now a similar line of reasoning which led to the linear programs $(P_\beta)$ and $(D_\beta)$ in Section 2.3, leads us to the primal-dual pair of linear programs

$$\min \sum_{j=1}^{N} \frac{1}{N} v(j)$$

subject to: $(P_\beta(1))$

(a) $v(i) \geq [R(i)g(i)]_{a^1} + \beta \sum_{j=1}^{N} p_{ij}(a^1) v(j),\ i \in S, a^1 \in A^1(i)$

(b) $\sum_{a^2 \in A^2(i)} g(i, a^2) = 1,\ i \in S$

(c) $g(i, a^2) \geq 0,\ i \in S, a^2 \in A^2(i)$

and

$$\max \sum_{j=1}^{N} z(i)$$

subject to: $(D_\beta(1))$

(d) $\sum_{i=1}^{N} \sum_{a^1 \in A^1(i)} \left[\delta(i,j) - \beta p_{ij}(a^1)\right] x_{ia^1} = \frac{1}{N},\ j \in S$

(e) $z(i) \leq [\mathbf{x}(i)R(i)]_{a^2}\,,\ i \in S, a^2 \in A^2(i)$

(f) $x(i,a^1) \geq 0;\ i \in S, a^1 \in A^1(i)$

where $\mathbf{x}(i) = \big(x(i,1), x(i,2), \ldots, x(i,m^1(i))\big)$ for each $i \in S$. Just as in Section 2.3, the arguments given below also would be valid if the coefficients $\frac{1}{N}$ were replaced by some positive starting probabilities $\gamma(j)$ summing to 1.

We will verify first that the linear program $(D_\beta(1))$ is indeed the dual of $(P_\beta(1))$. We know that a linear program gives rise to two related optimization problems:

| Primal | | Dual | |
|---|---|---|---|
| maximize | $\mathbf{c}^T\mathbf{y}$ | minimize | $\mathbf{x}^T\mathbf{b}$ |
| subject to | $M\mathbf{y} \leq \mathbf{b}$ | subject to | $\mathbf{x}^T M \geq \mathbf{c}^T$ |
| | $\mathbf{y} \geq 0$ | | $\mathbf{x} \geq 0$ |

Furthermore, from [11] we know that an equality in the primal problem gives rise to a corresponding free variable in the dual problem, and vice versa. Now let us take $(P_\beta(1))$ as primal LP.

$$\min \sum_{j=1}^{N} \frac{1}{N} v(j)$$

subject to: $(P_\beta(1))$

(a) $v(i) \geq [R(i)g(i)]_{a^1} + \beta \sum_{j=1}^{N} p_{ij}(a^1) v(j),\ i \in S, a^1 \in A^1(i)$

(b) $\sum_{a^2 \in A^2(i)} g(i,a^2) = 1,\ i \in S$

(c) $g(i,a^2) \geq 0,\ i \in S, a^2 \in A^2(i).$

Now $W = \Big(W_1 \vdots \ldots \vdots W_N\Big)$, with $W_i$ an $N \times m^1(i)$ matrix for every $i = 1, \ldots, N$, whose $(j,(i,a^1))$th element is given by

$$w_{j(i,a^1)} = \delta(i,j) - \beta p_{ij}(a).$$

Using the notation used in Section 3.5 we can also write the primal problem as

$$\min(\frac{1}{N}\mathbf{1}, \mathbf{0})\,(\mathbf{v}, \mathbf{g})^T$$

subject to:

$$(\mathbf{v}^T, \mathbf{g}^T) \begin{pmatrix} W^{(\beta)} & \vdots & 0 \\ \cdots & \vdots & \cdots \\ -R^T & \vdots & 1 \end{pmatrix} \begin{pmatrix} \geq \\ = \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \end{pmatrix}$$

and $\mathbf{g}(i) \geq \mathbf{0}$.
The dual program can now be written as

$$\max \mathbf{z}$$

subject to:

$$\begin{pmatrix} W & \vdots & 0 \\ \cdots & \vdots & \cdots \\ -R^T & \vdots & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix} \begin{pmatrix} = \\ \leq \end{pmatrix} \begin{pmatrix} \frac{1}{N} \cdot \mathbf{1} \\ \mathbf{0} \end{pmatrix}$$

and $x(i, a^1) \geq 0$ for $i \in S, a^1 \in A^1(i)$. We see that the free variable $\mathbf{v}$ corresponds with the equality in the dual program, and the free variable $\mathbf{z}$ corresponds with the equality in the primal program.

We will now show that we can use the primal-dual pair of linear programs to solve the player 1-controlled discounted stochastic game $\Gamma_\beta(1)$.

**Theorem 2.2.1.** *Consider a player 1-controlled discounted stochastic game $\Gamma_\beta(1)$ and the primal-dual pair of linear programs $(P_\beta(1))$ and $(D_\beta(1))$. Further, let $(\mathbf{v}^0, \mathbf{g}^0)$ be an optimal solution of $(P_\beta(1))$ and $(\mathbf{z}^0, \mathbf{x}^0)$ be an optimal solution of $(D_\beta(1))$. Then:*

*(i) The value vector of $\Gamma_\beta(1)$ is $\mathbf{v}^0$, and $\mathbf{g}^0$ is an optimal stationary strategy for player 2; and*

*(ii) If $x_i^0 := \sum_{a^1 \in A^1(i)} x_{ia^1}^0$ and a stationary strategy $\mathbf{f}^0$ for player 1 is defined by*

$$f^0(i, a^1) = \frac{x_{ia^1}^0}{x_i^0} \qquad i \in S, a^1 \in A^1(i),$$

*then $\mathbf{f}^0$ is optimal for player 1.*

*Proof.* First we have to prove that finite optimal solutions to $(P_\beta(1))$ and $(D_\beta(1))$ exist. We know that $[R(i)\mathbf{g}(i)]_{a^1} = \sum_{a^2=1}^{m^2(i)} r(i, a^1, a^2) g(i, a^2)$. Now let

$$m := \min \left\{ [R(i)\mathbf{g}(i)]_{a^1} \mid a^1 \in A^1(i), i \in S \right\}$$
$$M := \max \left\{ [R(i)\mathbf{g}(i)]_{a^1} \mid a^1 \in A^1(i), i \in S \right\}$$

and $\mathbf{1} \in \mathbb{R}^N$ a vector whose entries are 1. Note that the vector $\mathbf{v} = \frac{M}{M-\beta}\mathbf{1}$ trivially satisfies the constraints of $P_\beta(1)$:

$$\left(\frac{M}{1-\beta}\right) - \beta \sum_{j=1}^{N} p_{ij}(a^1)\left(\frac{M}{1-\beta}\right) = M \geq [R(i)\mathbf{g}(i)]_{a^1}\,.$$

This means $(P_\beta(1))$ is feasible. Now let $\mathbf{v}$ be an arbitrary feasible solution and $\hat{i} \in S$ be such that $v(\hat{i}) \leq v(i)$ for all $i \in S$. We have from the constraints of $P_\beta(1)$

$$\begin{aligned} v(\hat{i}) &\geq \left[R(\hat{i})\mathbf{g}(\hat{i})\right]_{a^1} + \beta \sum_{j=1}^{N} p_{\hat{i}j}(a^1) v(j) \\ &\geq \left[R(\hat{i})\mathbf{g}(\hat{i})\right]_{a^1} + v(\hat{i})\beta \sum_{j=1}^{N} p_{\hat{i}j}(a^1) \\ &= \left[R(\hat{i})\mathbf{g}(\hat{i})\right]_{a^1} + \beta v(\hat{i}) \end{aligned}$$

for all $a^1 \in A^1(\hat{i})$. It now follows from the definitions of $m$ and $\hat{i}$ that for all $i \in S$ and $a^1 \in A^1(\hat{i})$

$$v(i) \geq v(\hat{i}) \geq \left(\frac{1}{1-\beta}\right)\left[R(\hat{i})\mathbf{g}(\hat{i})\right]_{a^1} \geq \frac{m}{(1-\beta)}.$$

This means every feasible solution of $(P_\beta(1))$ is bounded below by $\frac{m}{(1-\beta)}\mathbf{1}$. Since $(P_\beta(1))$ is feasible and bounded, it possesses a finite optimal solution.

We now mix the constraints (a) (with $\mathbf{v} = \mathbf{v}^0$ and $\mathbf{g} = \mathbf{g}^0$) with respect to an arbitrary $\mathbf{f} \in \mathbf{F}_S$. This means that every constraint in a group corresponding to a block $(i, a^1)$ (respectively $(i, a^2)$) is multiplied by $f(i, a^1)$ and all of the constraints in this group are then summed over $a^1 \in A^1(i)$. We now have

$$\sum_{a^1 \in A^1(i)} v(i) f(i, a^1) \geq \mathbf{f}(i) R(i) \mathbf{g}(i) + \beta \sum_{j=1}^{N} p_{ij}(\mathbf{f}) v(j)$$

or, equivalently (since $\sum_{a^1 \in A^1(i)} f(i, a^1) = 1$),

$$v(i) \geq r(i, \mathbf{f}, \mathbf{g}) + \beta \left[P(\mathbf{f})\mathbf{v}\right]_i \tag{2.17}$$

where $\mathbf{v} = (v(1), v(2), \ldots, v(N))^T$. This gives us

$$\mathbf{v}^0 \geq \mathbf{r}(\mathbf{f}, \mathbf{g}^0) + \beta P(\mathbf{f})\mathbf{v}^0.$$

Iterating this equation gives us

$$\mathbf{v}^0 \geq \left[I - \beta P(\mathbf{f})\right]^{-1} \mathbf{r}(\mathbf{f}, \mathbf{g}^0) = \mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \tag{2.18}$$

Now, note that constraints (d) imply that $x_i^0 \geq \frac{1}{N}$ for every $i \in S$ and so $\mathbf{f}^0$ is well defined. Note also that $f^0(i, a^1) > 0$ if and only if $x^0(i, a^1) > 0$ for all $i \in S, a^1 \in A^1(i)$.

When we mix the constraints (a) with respect to $\mathbf{f}^0$ now, the complementary slackness property of linear programs ensures that for every $i \in S$

$$v^0(i) = r(i, \mathbf{f}^0, \mathbf{g}^0) + \beta \left[P(\mathbf{f}^0)\mathbf{v}^0\right]_i$$

which leads to

$$\mathbf{v}^0 = \left[I - \beta P(\mathbf{f}^0)\right]^{-1} \mathbf{r}(\mathbf{f}^0, \mathbf{g}^0) = \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0).$$

This establishes the saddle point inequality (2.6), namely,

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0)$$

for all $\mathbf{f} \in \mathbf{F}_S$.

We now take the constraints (e) with $\mathbf{x} = \mathbf{x}^0$ and $\mathbf{z} = \mathbf{z}^0$, and mix them with respect to an arbitrary stationary strategy $\mathbf{g} \in \mathbf{G}_S$. This leads to

$$z^0(i) \leq \mathbf{x}^0(i) R(i) \mathbf{g}(i), \ i \in \mathbf{S}.$$

However, if $\mathbf{g} = \mathbf{g}^0$ were used above, then with the help of complementary slackeness we would have obtained

$$z^0(i) = \mathbf{x}^0(i) R(i) \mathbf{g}^0(i), \ i \in \mathbf{S}.$$

If the last two relations are divided by $x_i^0$ for each $i \in S$, we immediately observe that

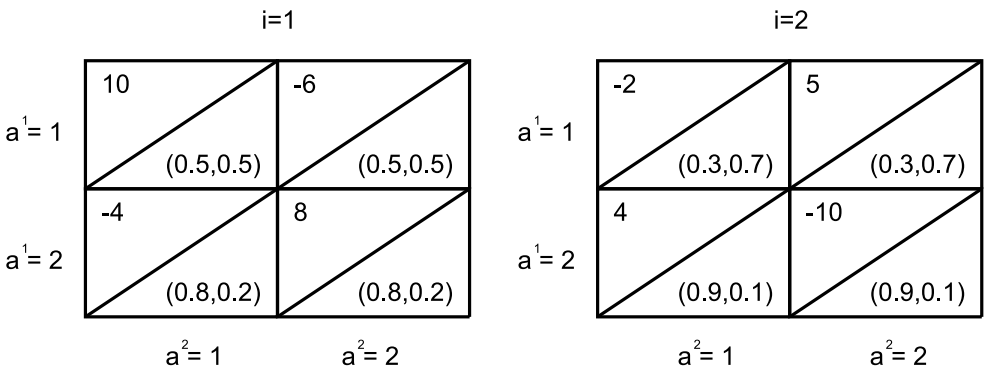
$$\mathbf{r}(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{r}(\mathbf{f}^0, \mathbf{g}), \ \mathbf{g} \in \mathbf{G}_S.$$

When we multiply this last equation with $\left[I - \beta P(\mathbf{f}^0)\right]^- 1$, we can (by (2.16)) conclude that

$$\mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}), \ \mathbf{g} \in \mathbf{G}_S$$

which completes the saddle point condition. □

**Example 2.2.2.** Let $S = \{1, 2\}, A^1(i) = A^2(i) = \{1, 2\}$ for $i \in S, \beta = 0.7$, and the reward and transition data be



Note that the player 1-controlled structure becomes apparent in the probability transition structure being the same in every cell in a given row in the data arrays above. The primal linear program($P_{.7}(1)$) for this problem now takes the form:

$$\min \left[\frac{1}{2}v(1) + \frac{1}{2}v(2)\right]$$

subject to:

(a) $v(1) \geq 10g(1,1) - 6g(1,2) + 0.35v(1) + 0.35v(2)$
$v(1) \geq -4g(1,1) + 8g(1,2) + 0.56v(1) + 0.14v(2)$
$v(2) \geq -2g(2,1) + 5g(2,2) + 0.21v(1) + 0.49v(2)$
$v(2) \geq 4g(2,1) - 10g(2,2) + 0.63v(1) + 0.07v(2)$

(b) $g(1,1) + g(1,2) = 1$
$g(2,1) + g(2,2) = 1$

(c) $g(1,1), g(1,2), g(2,1), g(2,2) \geq 0.$

We will solve this linear program and verify that it gives us the value vector $\mathbf{v}_{.7}$ and an optimal strategy $\mathbf{g}^0$ for player 2. Using the software package Maple to solve this program, we get the following results:

| Objective Function | : | $\frac{125}{33}$ |
|---|---|---|
| Variable | | Value |
| $v(1)$ | : | $\frac{325}{66}$ |
| $v(2)$ | : | $\frac{175}{66}$ |
| $g(1,1)$ | : | $\frac{91}{176}$ |
| $g(1,2)$ | : | $\frac{85}{176}$ |
| $g(2,1)$ | : | $\frac{103}{154}$ |
| $g(2,2)$ | : | $\frac{51}{154}$ |

So $\mathbf{v}^0_\beta = (\frac{325}{66}, \frac{175}{66})$ and $\mathbf{g}^0 = \big((\frac{91}{176}, \frac{85}{176}), (\frac{103}{154}, \frac{51}{154})\big)$.

The dual problem has the following form:

$$\max \left[z(1) + z(2)\right]$$

subject to:

(a)

$$\begin{array}{rcrcrcrcl} 0.65x_{11} & + & 0.44x_{12} & - & 0.21x_{21} & - & 0.63x_{22} & = & \frac{1}{2} \\ -0.35x_{11} & - & 0.14x_{12} & + & 0.51x_{21} & + & 0.93x_{22} & = & \frac{1}{2} \end{array}$$

(b)

$$\begin{array}{rcrcr} z(1) & \leq & 10x_{11} & - & 4x_{12} \\ z(1) & \leq & -6x_{11} & + & 8x_{12} \\ z(2) & \leq & -2x_{21} & + & 4x_{22} \\ z(2) & \leq & 5x_{21} & - & 10x_{22} \end{array}$$

(c) $x_{11}, x_{12}, x_{21}, x_{22} \geq 0$.

Solving this with Maple again, we get

| Objective Function | : | $\frac{125}{33}$ |
|---|---|---|
| Variable | | Value |
| $z(1)$ | : | $\frac{125}{33}$ |
| $z(2)$ | : | $0$ |
| $x_{11}$ | : | $\frac{125}{154}$ |
| $x_{12}$ | : | $\frac{250}{231}$ |
| $x_{21}$ | : | $\frac{95}{99}$ |
| $x_{22}$ | : | $\frac{95}{198}$ |

The optimal solution $\mathbf{f}^0$ now becomes

$$\begin{aligned} \mathbf{f}^0 &= \left(\left(\frac{x_{11}^0}{x_{11}^0 + x_{12}^0}, \frac{x_{12}^0}{x_{11}^0 + x_{12}^0}\right), \left(\frac{x_{21}^0}{x_{21}^0 + x_{22}^0}, \frac{x_{22}^0}{x_{21}^0 + x_{22}^0}\right)\right) \\ &= \left(\left(\frac{3}{7}, \frac{4}{7}\right), \left(\frac{2}{3}, \frac{1}{3}\right)\right). \end{aligned}$$

We will now verify that the $\mathbf{f}^0$ and $\mathbf{g}^0$ constructed above satisfy the saddle point condition (2.6). First we will show that $\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0)$ holds for all $\mathbf{f} \in \mathbf{F}_S$. We have already seen that in the case of a single-controller discounted stochastic game $\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0)$ can be written as

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) = \left[I - \frac{7}{10} P(\mathbf{f})\right]^{-1} \mathbf{r}(\mathbf{f}, \mathbf{g}^0) \tag{2.19}$$

for every $\mathbf{f} \in \mathbf{F}_S$. According to our definitions we have

$$\begin{aligned} P(\mathbf{f}) &= (p_{ij}(\mathbf{f}))_{i,j=1}^2 \\ p_{ij}(\mathbf{f}) &= \sum_{a^1=1}^{2} p_{ij}(a^1) f(i, a^1). \end{aligned}$$

Together with the observation that every strategy can also be written as $\mathbf{f} = ((p, 1-p), (q, 1-q))$, with $p, q \in [0, 1]$, this leads to the following transition probability matrix:

$$P(\mathbf{f}) = \begin{pmatrix} \frac{4}{5} - \frac{3}{10}p & \frac{1}{5} + \frac{3}{10}p \\ \frac{9}{10} - \frac{3}{5}q & \frac{1}{10} + \frac{3}{5}q \end{pmatrix}.$$

Using Maple again we can calculate

$$\left(I - \frac{7}{10} P(\mathbf{f})\right)^{-1} = \begin{pmatrix} -\frac{-310+140q}{107-42q+21p} & \frac{70}{3}\frac{2+3p}{107-42q+21p} \\ -\frac{-210+140q}{107-42q+21p} & \frac{10}{3}\frac{44+21p}{107-42q+21p} \end{pmatrix}. \tag{2.20}$$

This gives us

$$\begin{aligned}\mathbf{r}(1,\mathbf{f},\mathbf{g}^0) &= \mathbf{f}(1)R(1)\mathbf{g}^0(1) = \begin{pmatrix} p & 1-p \end{pmatrix}\begin{pmatrix} 10 & -6 \\ -4 & 8 \end{pmatrix}\begin{pmatrix} \frac{91}{176} \\ \frac{85}{176} \end{pmatrix} \\ &= \frac{21}{44}p + \frac{79}{44} \qquad (2.21)\\ \mathbf{r}(2,\mathbf{f},\mathbf{g}^0) &= \mathbf{f}(2)R(2)\mathbf{g}^0(2) = \begin{pmatrix} q & 1-q \end{pmatrix}\begin{pmatrix} -2 & 5 \\ 4 & -10 \end{pmatrix}\begin{pmatrix} \frac{103}{154} \\ \frac{51}{154} \end{pmatrix} \\ &= \frac{21}{22}q - \frac{7}{11}. \qquad (2.22)\end{aligned}$$

Putting everything together gives us

$$\mathbf{v}_\beta(\mathbf{f},\mathbf{g}^0) = \begin{pmatrix} \frac{(310-140q)\left(\frac{21}{44}p+\frac{79}{44}\right)}{107-42q+21p} + \frac{70}{3}\frac{(2+3p)\left(\frac{21}{22}q-\frac{7}{11}\right)}{107-42q+21p} \\ \frac{(210-140q)\left(\frac{21}{44}p+\frac{79}{44}\right)}{107-42q+21p} + \frac{10}{3}\frac{(44+21p)\left(\frac{21}{22}q-\frac{7}{11}\right)}{107-42q+21p} \end{pmatrix}. \qquad (2.23)$$

Using Maple again to calculate the gradient, we can verify that $\mathbf{f}^0 = \left(\left(\frac{3}{7},\frac{4}{7}\right),\left(\frac{2}{3},\frac{1}{3}\right)\right)$ is indeed the solution.

In the same way we can show $\mathbf{v}_\beta(\mathbf{f}^0,\mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0,\mathbf{g})$ holds for all $\mathbf{g} \in \mathbf{G}_S$. We now get

$$\mathbf{v}_\beta(\mathbf{f}^0,\mathbf{g}) = \left[I - \frac{7}{10}P(\mathbf{f}^0)\right]^{-1}\mathbf{r}(\mathbf{f}^0,\mathbf{g}) \qquad (2.24)$$

with

$$\left(I - \frac{7}{10}P(\mathbf{f}^0)\right)^{-1} = \begin{pmatrix} \frac{325}{132} & \frac{115}{132} \\ \frac{175}{132} & \frac{265}{132} \end{pmatrix}. \qquad (2.25)$$

Writing $\mathbf{g} = ((s,1-s),(t,1-t))$ we get

$$\mathbf{r}(\mathbf{f}^0,\mathbf{g}) = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \qquad (2.26)$$

$$\mathbf{v}_\beta(\mathbf{f}^0,\mathbf{g}) = \begin{pmatrix} \frac{325}{66} \\ \frac{175}{66} \end{pmatrix}. \qquad (2.27)$$

It is immediately clear now that $g^0$ is optimal.

### 2.2.2 Separable Reward State Independent Transition (SER-SIT) Discounted Stochastic Games

In this model we have $m^1(i) = \mu$ and $m^2(i) = \nu$ for all $i \in S$. The chosen actions $(a^1, a^2)$ are the same in every state and the rewards are a sum of state dependent part and a part depending on the action pair selected. In what follows, $\mathbf{c} = (c(1),\ldots,c(N))$. The assumptions are formulated in the following:

(SER) $r(i,a^1,a^2) = c(i) + \rho(a^1,a^2),\ a^1 \in A^1(i), a^2 \in A^2(i), i \in \mathbf{S}$

and

(SIT) $p_{ij}(a^1,a^2) = p_j(a^1,a^2),\ a^1 \in A^1(i), a^2 \in A^2(i),\ i,j \in \mathbf{S}.$

A solution of such SER-SIT games can be obtained via the following construction. With the vector $\mathbf{c}$ associate a single auxilliary matrix game similar in form to (2.13):

$$R(\mathbf{c}) := \left[\rho(a^1,a^2) + \beta \sum_{j\in S} p_j(a^1,a^2)c(j)\right]_{a^1=1,a^2=1}^{\mu,\nu}.$$

Here $R(\mathbf{c})$ does not depend on the state $i$. Let $\rho := val\,\{R(\mathbf{c})\}$, and $\mathbf{x}^0 = (x_1^0,\ldots,x_\mu^0)$ and $\mathbf{y}^0 = (y_1^0,\ldots,y_\nu^0)^T$ be a pair of optimal strategies in the matrix game $R(\mathbf{c})$.

$\mathbf{g}^0 \in \mathbf{G}_S$ (respectively, $\mathbf{f}^{,0} \in \mathbf{F}_S$) constructed by setting $\mathbf{g}^0(i) = \mathbf{y}^0$ (respectively, $\mathbf{f}^{,0}(i) = \mathbf{x}^0$) for every $i \in S$ are optimal stationary strategies in the SER-SIT discounted game. Together with the fact that the strategies do not depend on the state, we can see that $\mathbf{x}^0$ and $\mathbf{y}^0$ satisfy (2.6). Furthermore,

$$\mathbf{v}_\beta = \mathbf{c} + \left(\frac{\rho}{1-\beta}\right)\mathbf{1}. \tag{2.28}$$

We can rewrite this game in the same way we did with (1.22):

$$\mathbf{v}_\beta(\mathbf{f},\mathbf{g}) = \mathbf{r}(\mathbf{f},\mathbf{g}) + \beta P(\mathbf{f},\mathbf{g})\mathbf{v}_\beta(\mathbf{f},\mathbf{g})$$

with the help of the assumptions (SER) and (SIT). For every $i \in S$ we now have

$$\mathbf{v}_\beta(i,\mathbf{f},\mathbf{g}) = c(i) + \mathbf{f}(i)R(\mathbf{c})\mathbf{g}(i) + \beta \sum_{j\in S} p_j(\mathbf{f},\mathbf{g})(v_\beta(j,\mathbf{f},\mathbf{g}) - c(j)).$$

Setting $r(\mathbf{c},\mathbf{f},\mathbf{g}) := \mathbf{f}(i)R(\mathbf{c})\mathbf{g}(i)$ for every $i \in S$, we can write this in vector notation as

$$\mathbf{v}_\beta(\mathbf{f},\mathbf{g}) = \mathbf{c} + r(\mathbf{c},\mathbf{f},\mathbf{g})\mathbf{1} + \beta P(\mathbf{f},\mathbf{g})\left[\mathbf{v}_\beta(\mathbf{f},\mathbf{g}) - \mathbf{c}\right],$$

which, when solved for $\mathbf{v}_\beta(\mathbf{f},\mathbf{g})$, yields

$$\mathbf{v}_\beta(\mathbf{f},\mathbf{g}) = \mathbf{c} + \left[I - \beta P(\mathbf{f},\mathbf{g})\right]^{-1}\left[r(\mathbf{c},\mathbf{f},\mathbf{g})\mathbf{1}\right] \tag{2.29}$$

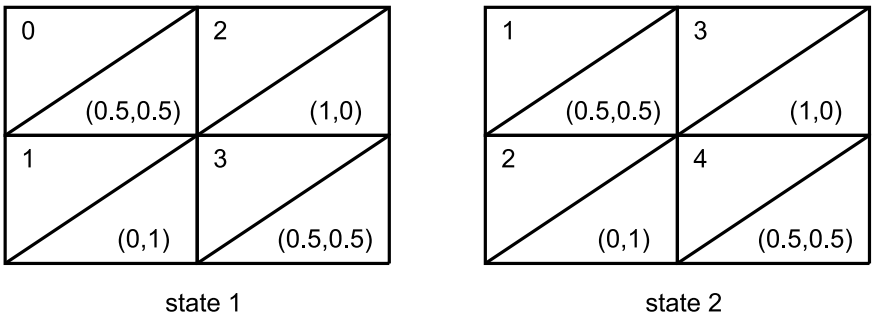
The second term of (2.29) only depends on the state $i$ through the choice of strategies. Since the rows of $P(\mathbf{f},\mathbf{g})$ are now all the same, and $P(\mathbf{f},\mathbf{g})$ is stochastic we have (see also section 1.2)

$$\left[I - \beta P(\mathbf{f},\mathbf{g})\right]^{-1} = \sum_{t=0}^{\infty} P^{(t)}(\mathbf{f},\mathbf{g})\beta^t = P(\mathbf{f},\mathbf{g})\sum_{t=0}^{\infty}\beta^t = \frac{1}{1-\beta}P(\mathbf{f},\mathbf{g})$$

and that $\mathbf{r}(\mathbf{f},\mathbf{g})$ has identical components. We now have (from (2.29)) that

$$\mathbf{v}_\beta(\mathbf{f},\mathbf{g}) = \mathbf{c} + \frac{1}{1-\beta}\left[\mathbf{x}^T R(\mathbf{c})\mathbf{y}\right]\mathbf{1}.$$

**Example 2.2.3.** Let $S = \{1,2\}, A^1(1) = A^2(1) = A^2(1) = A^2(2) = \{1,2\}$ and the transitions and rewards given by




In this example $\mathbf{c} = (0,1)$ and $\rho(a^1,a^2) = \begin{pmatrix} 0 & 2 \\ 1 & 3 \end{pmatrix}$.

This gives us

$$R(\mathbf{c}) = \left(\begin{pmatrix} 0 & 2 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} \frac{1}{2}\beta & 2 \\ 1+\beta & 3+\frac{1}{2}\beta \end{pmatrix}\right)$$

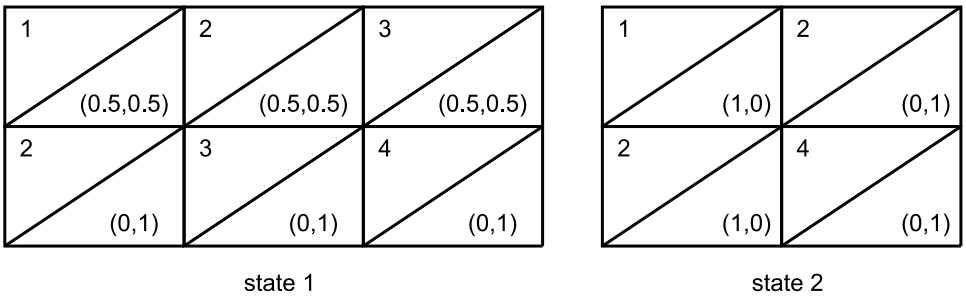
which, for $\beta = \frac{1}{2}$, gives us $\mathbf{x} = ((0,1),(0,1))$, $\mathbf{y} = ((1,0),(1,0))$, $\rho = \left(1, 1\frac{1}{2}\right)$ and $\mathbf{v}_{\frac{1}{2}} = (0,1) + \left(\frac{(1,1\frac{1}{2})}{1-\frac{1}{2}}\right)\mathbf{1} = (2,4)$.

### 2.2.3 Switching Controller Discounted Stochastic Games

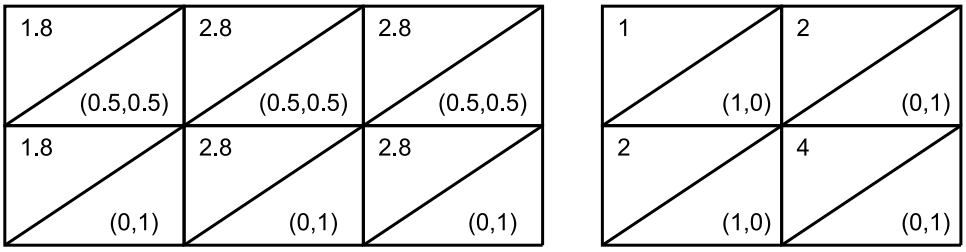
In these games the action space is divided in two partitions $\mathbf{S}_1$ and $\mathbf{S}_2$, with player 1 controlling the transition probabilities in $\mathbf{S}_1$ and player 2 controlling the transition probabilities in $\mathbf{S}_2$:

$$p_{ij}(a^1,a^2) = \begin{cases} p_{ij}(a^1) & \text{if } i \in \mathbf{S}^1 \\ p_{ij}(a^2) & \text{if } i \in \mathbf{S}^2. \end{cases} \tag{SW}$$

**Example 2.2.4.** Let $S = \{1,2\}, A^1(1) = A^1(2) = A^2(2) = \{1,2\}, A^2(1) = \{1,2,3\}$. Let the reward and transition data be



Player 1 controls the transitions in $S^1 = \{1\}$, and player 2 controls the transitions in $S^2 = \{2\}$. Furthermore, if player 1 were to fix her strategy in state 1 to, say $\hat{\mathbf{f}} = (0.2, 0.8)$, then the preceding construction would lead us to consider a player 2-controlled game $\Gamma_\beta(2, \mathbf{f}(1))$ given below.




The following algorithm solves the switching controller discounted stochastic game.

**Algorithm 2.2.1.**

**Step 1.** Set $k := 0$, choose an arbitrary $\mathbf{v}^0 = \big(v^0(1), \ldots, v^0(N)\big)^T$, and find an extreme optimal strategy $\mathbf{f}^0(i)$ for player 1 in the matrix game $R(i, \mathbf{v}^0)$ for each $i \in \mathbf{S}^1$.

**Step 2.** Set $k := k + 1$. Solve the player 2-controlled game $\Gamma_\beta(2, \mathbf{f}^{k-1})$, denote its value vector by $\mathbf{v}_\beta$, and set $\mathbf{v}^k := \mathbf{v}_\beta$.

**Step 3.** If $v^k(i) = \text{val}\left[R(i, \mathbf{v}^k)\right]$ for each $i \in S$, then stop. Otherwise, find an extreme optimal strategy $\mathbf{f}^k(i)$. for player 1, in the matrix game $R(i, \mathbf{v}^k)$ for ech $i \in \mathbf{S}^1$, and return to Step 2.

## 2.3 Modified Newton's Method and the Discounted Stochastic Games

When we tried to solve discounted stochastic games by means of Linear Programming, we had to restrict ourselves to special subclasses which ensured linearity. Another approach is to try extend the Newton-like method discussed in section 1.8 to stochastic games. The natural extension of the basic Newton's scheme such as the one presented in Corollary (1.8.2) fails to converge however in the case of $\Gamma_\beta$. We will deal with this by introducing a stepsize $\omega^k$ to ensure convergence.

To begin with, if we define an operator $L : \mathbb{R}^N \to \mathbb{R}^N$ by

$$L(\mathbf{v})(i) := \text{val}\left[R(i, \mathbf{v})\right] \tag{2.30}$$

for every $\mathbf{v} \in \mathbb{R}^N, i \in S$, Theorem 2.1.2 states that the value vector $\mathbf{v}_\beta$ is the unique solution of the fixed point equations

$$L(\mathbf{v}) = \mathbf{v}. \tag{2.31}$$

Finding a fixed point in (2.31) is equivalent to finding the zero of

$$\psi(\mathbf{v}) := L(\mathbf{v}) - \mathbf{v}$$

or to finding a global minimum of the norm of $\psi(\mathbf{v})$, which is equivalent to the following mathematical programming problem:

$$\min \ \frac{1}{2}\left[\psi(\mathbf{v})^T \psi(\mathbf{v})\right]$$

subject to: (M)

$$\mathbf{v} \in \mathbb{R}^N.$$

Now, if the gradient matrix of $\psi(\mathbf{v})$ is well-defined, it will be denoted by $\psi'(\mathbf{v})$. It is assumed that $\mathbf{v}^k$, the current estimate of the solution, is known, and a search direction $\mathbf{d}^k$ is selected. In the algorithm presented below, the search direction is selected by the classical Newton's scheme: $\mathbf{d}^k := -\left[\psi'(\mathbf{v}^k)\right]^{-1} \psi(\mathbf{v}^k)$, but the stepsize in that direction will be selected carefully to ensure descent. That is, the iterative step of the method will take the form

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \omega^k \left[\psi'(\mathbf{v}^k)\right]^{-1} \psi(\mathbf{v}^k) \tag{2.32}$$

where the stepsize $\omega^k \in (0, 1]$ is chosen according to a line search rule that ensures good convergence properties.[1]

Using a result of Shapley and Snow (see [6]) we can now formulate a result similar to Proposition 1.8.2. They proved that for each fixed $\mathbf{v} \in \mathbb{R}^N$ and $i \in \mathbf{S}$ a kernel (a square submatrix) $K(i, \mathbf{v})$ of $R(i, \mathbf{v})$ exists such that

$$\text{val}\left[R(i, \mathbf{v})\right] = \text{val}\left[K(i, \mathbf{v})\right] = \frac{|K(i, \mathbf{v})|}{\sum_s \sum_t \{K(i, \mathbf{v})_{st}\}} \tag{2.33}$$

where $\{K(i, \mathbf{v})\}_{st}$ is the $(s, t)$th cofactor of the kernel $K(i, \mathbf{v})$. This kernel uniquely determines a pair of extreme optimal strategies $\mathbf{x}(i, \mathbf{v})$ and $\mathbf{y}(i, \mathbf{v})$ for players 1 and 2, respectively, in the matrix game $R(i, v)$. Hence it is possible to define a pair of stationary strategies $\mathbf{f}(\mathbf{v})$ and $\mathbf{g}(\mathbf{v})$ for players 1 and 2 by setting the $i$th block of each of these strategies according to $\mathbf{f}(i, \mathbf{v}) = \mathbf{x}(i, \mathbf{v})$ and $\mathbf{g}(i, \mathbf{v}) = \mathbf{y}(i, \mathbf{v})$ for each $i \in S$.

In view of (2.30) and (2.33) we see that $\psi(\mathbf{v}) = L(\mathbf{v}) - \mathbf{v}$ can be differentiated at all points $\mathbf{v}$ except those where the kernels satisfying (2.33)

[1] We mention Armijo's rule as one popular choice (e.g., see McCormick, 1983)

change. Fortunately, the set of points where this occurs is a set of measure zero. Whenever the respective partial derivatives exist, they satisfy

$$\frac{\partial[\psi(\mathbf{v})]_i}{\partial v(i)} = \beta p_{ij}(\mathbf{f}(\mathbf{v}), \mathbf{g}(\mathbf{v})) - \delta(i,j)$$

for all $i, j \in S$. Hence, whenever it is well defined,

$$\psi'(\mathbf{v}) = -\left[I - \beta P(\mathbf{f}(\mathbf{v}), \mathbf{g}(\mathbf{v}))\right]. \tag{2.34}$$

The properties of a transition matrix and the fact that $\beta \in [0,1)$ imply that the above matrix is invertible, and hence that the Newton's search direction

$$\mathbf{d}^k = -\left[\psi'(\mathbf{v}^k)\right]^{-1} = \left[I - \beta P(\mathbf{f}(\mathbf{v}^k), \mathbf{g}(\mathbf{v}^k))\right]^{-1} \psi(\mathbf{v}^k) \tag{2.35}$$

is well defined whenever $\psi'(\mathbf{v}^k)$ exists. Now, if we let

$$J(\mathbf{v}) := \frac{1}{2}\left[\psi(\mathbf{v})\right]^T \psi(\mathbf{v})$$

then it follows that, just as in Proposition 1.8.2

$$\nabla J(\mathbf{v}) = -\left[\psi(\mathbf{v})\right]^T \left[I - \beta P(\mathbf{f}(\mathbf{v}), \mathbf{g}(\mathbf{v}))\right] \tag{2.36}$$

and hence that $\nabla J(\mathbf{v}^*) = 0$ implies that $\psi(\mathbf{v}^*) = 0$ or, equivalently, that $\mathbf{v}^* = \mathbf{v}_\beta$, the unique value vector of $\Gamma_\beta$.

The mathematical program (M) possesses the desirable property that there can only be one point where the gradient (2.36) is zero that is also the global minimum of $J(\mathbf{v})$ with objective funtion value equal to 0. This suggests that any good descent algorithm of unconstrained nonlinear programming might be expected to perform well in sovling (M) and thereby the discounted stochastic game $\Gamma_\beta$. Below we present one such algorithm.

**Algorithm 2.3.1. Modified Newton's Method**

**Step 1** Set $k := 0$ and select two parameter values: a "small" value of $\alpha \in (0,1)$ and $\mu \in [0.5, 0.8]$. Also select $\mathbf{v}^0$, the initial estimate of the value vector.

**Step 2** Calculate for each $i \in S$ the matrix game $R(i, \mathbf{v}^k)$, a pair of optimal extreme strategies $\mathbf{x}(i, \mathbf{v}^k)$ and $\mathbf{y}(i, \mathbf{v}^k)$ for player 1 and 2 in this matrix game, and its value $L(\mathbf{v}^k)(i)$. Hence calculate $L(\mathbf{v}^k), \psi(\mathbf{v}^k)$, and $J(\mathbf{v}^k)$.

**Step 3** If $J(\mathbf{v}^k) = 0$, stop; $\mathbf{v}^k = \mathbf{v}_\beta$.

**Step 4** Calculate $\mathbf{d}^k$ as in (2.35).

**Step 5** Set $\omega^k = 1$.

**Step 6** Test the inequality

$$J(\mathbf{v}^k + \omega^k \mathbf{d}^k) - J(\mathbf{v}^k) \leq \alpha\omega^k \left[\nabla J(\mathbf{v}^k)\mathbf{d}^k\right].$$

If the above inequality is satisfied, set $\mathbf{v}^{k+1} = \mathbf{v}^k + \omega^k \mathbf{d}^k, k := k+1$ and return to Step 2.

**Step 7** Set $\omega^k := \mu\omega^k$ and return to Step 5.

## 2.4 Limiting Average Stochastic Games: The Issues

In the same way we extended the discounted MDP to a discounted stochastic game, we can also extend the limiting average MDP to a limiting average stochastic game. We will define the *limiting average zero-sum stochastic game* $\Gamma_\alpha$ as having the same structure as the discounted game, except that the payoff by player 2 to player 1 corresponding to a strategy pair $(\mathbf{f}, \mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$ is given by

$$v_\alpha(i, \mathbf{f}, \mathbf{g}) := \lim_{T\to\infty} \left[\left(\frac{1}{T+1}\right) \sum_{t=0}^{T} \mathbb{E}_{i\mathbf{fg}}(R_t)\right] = [Q(\mathbf{f}, \mathbf{g})\mathbf{r}(\mathbf{f}, \mathbf{g})]_i \qquad (2.37)$$

for each $i \in S$, where $Q(\mathbf{f}, \mathbf{g})$ is the Cesaro-limit matrix of $P(\mathbf{f}, \mathbf{g})$. This limit exists:

$$\begin{aligned}
\mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}) &:= \lim_{T\to\infty} \left[\left(\frac{1}{T+1}\right) \sum_{t=0}^{T} \mathbb{E}_{i\mathbf{fg}}(R_t)\right]_i \\
&= \lim_{T\to\infty} \left[\left(\frac{1}{T+1}\right) \sum_{t=0}^{T} P^{(t)}(\mathbf{f}, \mathbf{g})\mathbf{r}(\mathbf{f}, \mathbf{g})\right] \\
&= \lim_{T\to\infty} \left[\left(\frac{1}{T+1}\right) \sum_{t=0}^{T} P^{(t)}(\mathbf{f}, \mathbf{g})\right] \mathbf{r}(i, \mathbf{f}, \mathbf{g}) \\
&= [Q(\mathbf{f}, \mathbf{g})\mathbf{r}(\mathbf{f}, \mathbf{g})]_i .
\end{aligned}$$

The validity of the last equality stems from a well-known property of Markov chains which ensures that a Markov matrix $Q(\mathbf{f})$ exists such that

$$Q(\mathbf{f}) := \lim_{T\to\infty} \sum_{t=0}^{T} P^t(\mathbf{f}).$$

We shall say that $\mathbf{f}^0 \in \mathbf{F}_S$ and $\mathbf{g}^0 \in \mathbf{G}_S$ are *optimal stationary strategies* if for all $i \in S, \mathbf{f} \in \mathbf{F}_S, \mathbf{g} \in \mathbf{G}_S$

$$v_\alpha(i, \mathbf{f}, \mathbf{g}^0) \leq v_\alpha(i, \mathbf{f}^0, \mathbf{g}^0) \leq v_\alpha(i, \mathbf{f}^0, \mathbf{g}^0). \qquad (2.38)$$

If a pair of optimal stationary strategies $(\mathbf{f}^0, \mathbf{g}^0)$ exists then so does the *undiscounted value vector* $\mathbf{v}_\alpha$ of the game $\Gamma_\alpha$ and $\mathbf{v}_\alpha := \mathbf{v}_\alpha(\mathbf{f}^0, \mathbf{g}^0)$. This can be deduced as follows:

$$\max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}^0) = \mathbf{v}_\alpha(i, \mathbf{f}^0, \mathbf{g}^0) = \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(i, \mathbf{f}^0, \mathbf{g}).$$

Furthermore,

$$\max_{\mathbf{f}\in\mathbf{F}_S} \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}) \geq \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(i, \mathbf{f}^*, \mathbf{g})$$

and

$$\min_{\mathbf{g}\in\mathbf{G}_S} \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}) \leq \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}^*)$$

for all $\mathbf{f}^* \in \mathbf{F}_S$ and $\mathbf{g}^* \in \mathbf{G}_S$. This leads to

$$\begin{aligned} \max_{\mathbf{f}\in\mathbf{F}_S} \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}) \geq \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(i, \mathbf{f}^0, \mathbf{g}) &= \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}^0) \\ &\geq \min_{\mathbf{g}\in\mathbf{G}_S} \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}) \end{aligned} \tag{2.39}$$

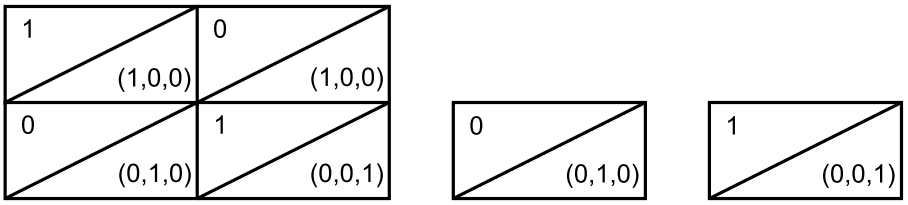
and

$$\begin{aligned} \min_{\mathbf{g}\in\mathbf{G}_S} \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}) \geq \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(i, \mathbf{f}^0, \mathbf{g}) &= \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}^0) \\ &\geq \max_{\mathbf{f}\in\mathbf{F}_S} \min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \mathbf{g}) \end{aligned} \tag{2.40}$$

which proves the equality.

In chapter 2 we mentioned that the existence of absorbing states complicates the analysis of an average limiting MDP. This observation is also valid for the average limiting stochastic games. The folowing example, first used by Gilette [8] and later elaborated on by Blackwell and Ferguson [2], illustrates this problem and has played an important role in the development of the theory of stochastic games.

**Example 2.4.1.** (The Big Match)
Let $S = \{1, 2, 3\}$, $A^1(1) = A^2(1) = \{1, 2\}$, $A^1(i) = A^2(i) = \{1\}$ for $i = 2, 3$ , and the reward and transition data be

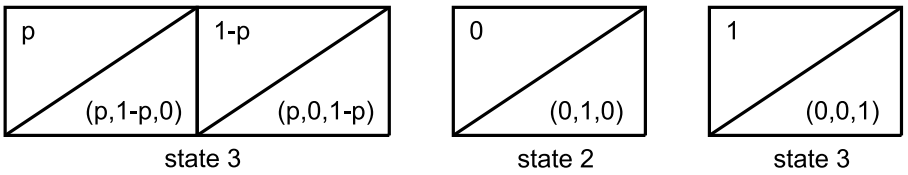


States 2 and 3 are absorbing, and so $v_\alpha(2) = 0$ and $v_\alpha(3) = 1$.

Suppose now that player 1 possesses a stationary optimal strategy

$$\mathbf{f}_p = ((p, 1-p), (1), (1))$$

for some fixed $p \in [0, 1]$. Against this strategy, player 2 is faced with minimizing limiting average Markov decision process




There are two cases now:

**C**ase **1.** $p = 1$, that is, player 1 does not take a chance and chooses to remain in state 1. However, in such a case, against $\mathbf{g}_0 = ((0, 1), (1), (1))^T$ player 1 will almost always earn 0 and hence $v_\alpha(1, \mathbf{f}_1, \mathbf{g}_0) = 0$.

**C**ase **2.** $0 < p < 1$, that is, player 1 risks choosing action 2 in state 1 with probability $1 - p > 0$ every time state 1 repeats itself. However, in such a case, against $\mathbf{g}_1 = ((1, 0), (1), (1))^T$ player 1 will ultimately be absorbed in state 2 with probability 1. In view of the nature of the limiting average payoff, this again results in $v_\alpha(1, \mathbf{f}_p, \mathbf{g}_1) = 0$.

This means that for all $p \in [0, 1]$ we can conclude that

$$\min_{\mathbf{g} \in \mathbf{G}_S} v_\alpha(1, \mathbf{f}_p, \mathbf{g}) = 0.$$

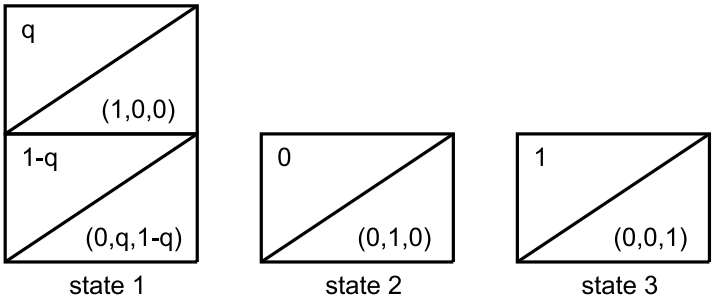
On the other hand, if player 2 uses a strategy $\mathbf{g}^* = \left((\frac{1}{2}, \frac{1}{2}), (1), (1)\right)^T$, we immediately see that, irrespective of what player 1 does in state 1,

$$v_\alpha(1, \mathbf{f}, \mathbf{g}^*) = \frac{1}{2}.$$

Furthermore, note that in this example every stationary strategy for player 2 can be expressed in the form

$$\mathbf{g}_p = ((q, 1-q), (1), (1))^T$$

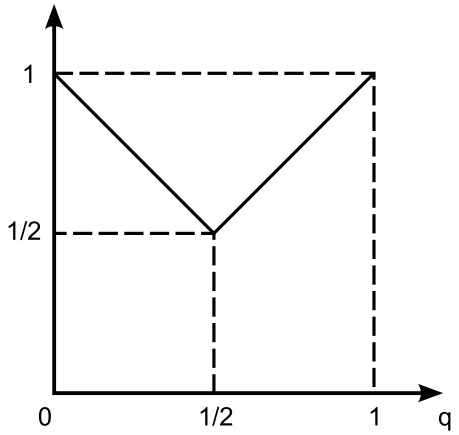
for some $q \in [0, 1]$. Of course, when player 2 fixes a strategy $\mathbf{g}_q$, then player 1 is facing the AMD process




Note that if player 1 uses strategy $\mathbf{f}_p$ with $p < 1$, then absorption in states 2 and 3 will occur with probabilities $q$ and $1-q$, respectively; but if $p = 1$, then state 1 will repeat itself infinitely often. It now follows that

$$v_\alpha(1, \mathbf{f}_p, \mathbf{g}_q) = \begin{cases} q & \text{if } p = 1 \\ 1 - q & \text{if } p < 1 \end{cases}$$

and hence that $\max_{\mathbf{f} \in \mathbf{F}_S} v_\alpha(1, \mathbf{f}, \mathbf{g}_q)$ is the function sketched below.




The minimum of the above function has a value of $\frac{1}{2}$. It should be clear from the discussion above that

$$0 = \max_{\mathbf{f} \in \mathbf{F}_S} \min_{\mathbf{g} \in \mathbf{G}_S} v_\alpha(1, \mathbf{f}, \mathbf{g}) < \frac{1}{2} = \min_{\mathbf{g} \in \mathbf{G}_S} \max_{\mathbf{f} \in \mathbf{F}_S} v_\alpha(1, \mathbf{f}, \mathbf{g}).$$

This strict inequality implies that optimal stationary strategies do not exist in the Big Match.

Broadening the class of strategies to include Markov strategies does not help either:

*Proof.* Let $\pi^1 = (\mathbf{f}_0, \mathbf{f}_1, \dots, \mathbf{f}_t, \dots) \in \mathbf{F}_M$ and $\pi_2 = (\mathbf{g}_0, \mathbf{g}_1, \dots \mathbf{g}_t, \dots) \in \mathbf{F}_M$ be a Markov strategy for player 1 resp. player 2 (see also section 1.8). If

the probability that player 1 will ever choose action 2 is 1, taking $\mathbf{g}_t = (1,0)$ for all $t$ gives us $v_\alpha(1,\pi^1,\pi^2) = 0$. If not, say $m \geq 0$ is the smallest initial number after which player 1 chooses action 2 with positive probability $\varepsilon$. Now define $\pi^2_\varepsilon$ as playing $\mathbf{g} = (1,0)$ in the first $m$ stages, then $\mathbf{g}_{m+1} = (0,1)$ and $\mathbf{g} = (\frac{1}{2},\frac{1}{2})$ thereafter. We now have $v_\alpha(1,\pi^1,\pi^2_\varepsilon) = (1-\varepsilon)\frac{1}{2}$.

As for the other side, taking $\mathbf{g}_t = (\frac{1}{2},\frac{1}{2})$ for all $t$ gives us $v_\alpha(1,\pi^1,\pi^2) = \frac{1}{2}$ for all $\pi^1 \in \mathbf{F}_M$. If we take $\mathbf{g}_t = (q, 1-q)$ with $q \leq \frac{1}{2}$ for all $t$, taking $\pi^1$ with $\mathbf{f}_t = (1,0)$ for all t gives us $v_\alpha(1,\pi^1,\pi^2) \geq \frac{1}{2}$. If at some stage $n$ we have that $q > \frac{1}{2}$ for the first time, taking a strategy $\pi^1$ with $\mathbf{f}_t = (1,0)$ for $t < n$ and $f_n = (0,1)$ gives us $v_\alpha(1,\pi^1,\pi^2) > \frac{1}{2}$. □

The question now arises whether the value and equilibrium solutions exist, and if so, how do we find optimal (or equilibrium) stationary strategies in those classes limiting average games that possess them? In the rest of this chapter we will focus our attention on the latter. In the next chapter we will take a look at the question of existence.

## 2.5 Zero-Sum Single-Controller Limiting Average Game

Just like in the case of $\beta$-discounted stochastic games a logical continuation is to look for subclasses of the limiting average stochastic games which are solvable by using linear programming. In this section we will analyze the Single-Controller case, and try the solve it along the lines of the approach used in section 1.10.

We have seen that for a general limiting average stochastic game $\Gamma_\alpha$ and a pair of stationary strategies $(\mathbf{f},\mathbf{g})$ the (limiting average) value vector of that strategy pair is given by:

$$\mathbf{v}_\alpha(\mathbf{f},\mathbf{g}) = Q(\mathbf{f},\mathbf{g})\mathbf{r}(\mathbf{f},\mathbf{g}) \tag{2.41}$$

where $Q(\mathbf{f},\mathbf{g})$ is the Cesaro-limit matrix of $P(\mathbf{f},\mathbf{g})$.

We get the player 1-controlled limiting average game $\Gamma_\alpha(1)$ by adding the restriction (same as (2.15)) that

$$p_{ij}(a^1,a^2) \equiv p_{ij}(a^1)$$

for all $a^1 \in A^1(i)$ and $a^2 \in A^2(i)$, $i,j \in S$. The player 2-controlled game $\Gamma_\alpha(2)$ is defined analogously

It follows immediately from the discussion above that in $\Gamma_\alpha(1)$ for every stationary strategy pair $(\mathbf{f},\mathbf{g})$

$$\mathbf{v}_\alpha(\mathbf{f},\mathbf{g}) = Q(\mathbf{f})\mathbf{r}(\mathbf{f},\mathbf{g}). \tag{2.42}$$

In order to introduce the linear programming formulation of the game $\Gamma_\alpha(1)$, we will need to exten slightly the notation used in Section 2.9.

**Notation:**

Just as in Section 2.9 we have $W = \Big( W_1 \vdots W_2 \vdots \ldots \vdots W_N \Big)$, where the $i$th block $W_i$ is an $N \times m^1(i)$ matrix whose $\big(j, (i, a^1)\big)$th element is given by

$$w_{j(i,a^1)} = \delta(i,j) - p_{ij}(a^1).$$

The $N \times m^1$ matrix $J = \Big( \mathbf{J}_1 \vdots \mathbf{J}_2 \vdots \ldots \vdots \mathbf{J}_N \Big)^T$ and the vectors $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}$, and $\gamma$ also will be exactly as in Section 2.9. However, since the rewards now depend on the actions of two players, we will introduce a block-diagonal matrix

$$R = \operatorname{diag}\left[R(1), R(2), \ldots, R(N)\right]$$

where

$$R(i) = \left[r(i, a^1, a^2)\right]_{a^1=1, a^2}^{m^1(i), m^2(i)}$$

as before. Since $m^k = \sum_{i \in \mathbf{S}} m^k(i)$ for $k \in \{1, 2\}$, we note that $R$ is an $m^1 \times m^2$ matrix. For a stationary strategy $\mathbf{g}$ for player 2,

$$R\mathbf{g} = \Big[ \left(R(1)\mathbf{g}(1)\right)^T, \left(R(2)\mathbf{g}(2)\right)^T, \ldots, \left(R(N)\mathbf{g}(N)\right)^T \Big]^T$$

is an $m^1 \times 1$ block-column vector. Similarly,

$$\mathbf{f}R = \left[\mathbf{f}(1)R(1), \mathbf{f}(2)R(2), \ldots, \mathbf{f}(N)R(N)\right]$$

is an $1 \times m^2$ block-row vector. Furthermore $\mathbf{0}_{m^k(i)}, \mathbf{1}_{m^k(i)}$ are $m^k(i)$-dimensional vectors with all entries 0 or 1 respectively, with $k = 1, 2$. When the dimension of the $\mathbf{0}$ or $\mathbf{1}$ vector is obvious from the context, it will not be specified.

We can now define a primal-dual pair of linear programs $(P_\alpha(1))$ and $(D_\alpha(1))$ corresponding to the limiting average player 1-controlled stochastic game $\Gamma_\alpha(1)$. They are the primal:

$$\min \left[\gamma^T \mathbf{v}\right]$$

subject to: $(P_\alpha(1))$

(a)

$$\left(\mathbf{u}^T, \mathbf{v}^T, \mathbf{g}^T\right) \begin{pmatrix} W & \vdots & 0 \\ \cdots & \vdots & \cdots \\ J & \vdots & W \\ \cdots & \vdots & \cdots \\ -R^T & \vdots & 0 \end{pmatrix} \geq \left(\mathbf{0}^T, \mathbf{0}^T\right)$$

(b) $\mathbf{1g}(i) = 1,\ i \in \mathbf{S}$

(c) $\mathbf{g}(i) \geq \mathbf{0},\ i \in \mathbf{S}$.

With dual variable vectors $\mathbf{x}, \mathbf{y}$ corresponding to the two constraint blocks in (a) and the dual variable vector $\mathbf{z}$ corresponding to the constraints in (b) we also have the dual:

$$\max \left[\mathbf{1}^T \mathbf{z}\right]$$

subject to: $(D_\alpha(1))$

(d)

$$\begin{pmatrix} W & \vdots & 0 \\ \cdots & \vdots & \cdots \\ J & \vdots & W \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \gamma \end{pmatrix}$$

(e) $\left[-R^T \mathbf{x}\right]_i + z_i \mathbf{1}_{m^2(i)} \leq \mathbf{0}_{m^2(i)},\ i \in S$

(f) $\mathbf{x}, \mathbf{y} \geq \mathbf{0}$.

Note that in (e) above $\left[-R^T \mathbf{x}\right]_i = -R(i)^T \mathbf{x}(i)$ is an $m^2(i) \times 1$ vector for each $i \in S$.

We can now formulate the following algorithm for solving $\Gamma_\alpha(1)$:

**Algorithm 2.5.1.**

**Step 1.** Find any optimal solution $\left(\hat{\mathbf{u}}^T, \hat{\mathbf{v}}^T, \hat{\mathbf{g}}^T\right)$ of $(P_\alpha(1))$ and any optimal solution $\left(\hat{\mathbf{x}}^T, \hat{\mathbf{y}}^T, \hat{\mathbf{z}}^T\right)$ of $(D_\alpha(1))$.

**Step 2.** Define the set of states

$$\mathbf{S}^* := \left\{ i \in S \mid \hat{x}_i := \sum_{a^1 \in A^1(i)} \hat{x}_{ia^1} > 0 \right\}.$$

**Step 3.** Construct a stationary strategy

$$\hat{\mathbf{f}} = \left(\hat{\mathbf{f}}(1), \hat{(\mathbf{f})}(2), \ldots, \hat{\mathbf{f}}(N)\right)$$

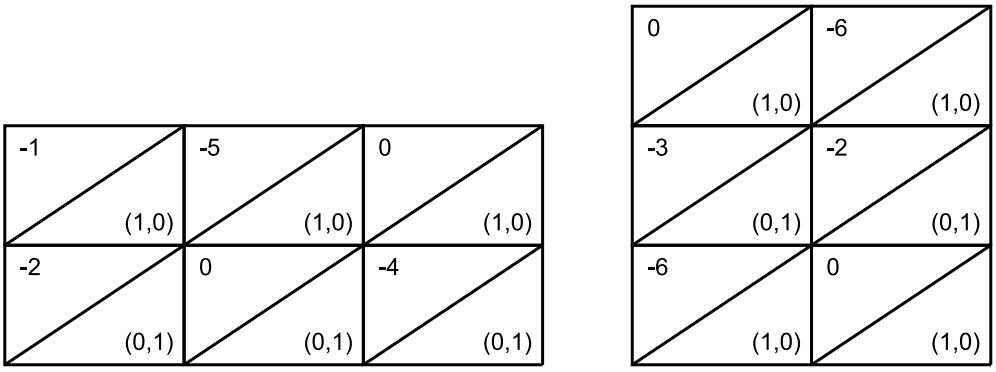
according to:

$$\hat{f}(i, a^1) = \begin{cases} \frac{\hat{x}_{ia^1}}{\hat{x}_i}, & i \in \mathbf{S}^*, a^1 \in A^1(i) \\ \frac{\hat{y}_{ia^1}}{\hat{y}_i}, & i \in \mathbf{S} \backslash \mathbf{S}^*, a^1 \in A^1(1) \end{cases}$$

where $\hat{y}_i := \sum_{a^1 \in A^1(i)} \hat{y}_{ia^1}$.

Unlike the case of limiting average Markov decision processes, the above algorithm will, in general, produce randomized optimal stationary strategies $\hat{\mathbf{f}}$ and $\hat{\mathbf{g}}$ for player 1 and 2 respectively.

**Example 2.5.1.** Let $\mathbf{S} = \{1, 2\}, A^1(1) = A^2(2) = \{1, 2\}, A^1(2) = A^2(1) = \{1, 2, 3\}$ and the reward and transition data be




We shall take the vector $\gamma^T$ to be $\left(\frac{1}{2}, \frac{1}{2}\right)$. The primal linear program reduces to:

$$\min \left[\frac{1}{2}v(1) + \frac{1}{2}v(2)\right]$$

subject to:

(a)

$$\begin{pmatrix} 0 & 1 & -1 & 0 & -1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 0 & 0 & 0 & \vdots & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 & \vdots & 0 & -1 & 1 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 2 & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 3 & 6 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 2 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T \begin{pmatrix} u(1) \\ u(2) \\ v(1) \\ v(2) \\ g(1,1) \\ g(1,2) \\ g(1,3) \\ g(2,1) \\ g(2,2) \end{pmatrix} \geq 0$$

(b) $g(1,1)+g(1,2)+g(1,3)=1$
$g(2,1)+g(2,2)=1$

(c) $g(i,a^2)\geq 0, i\in[1,2], a^2\in A^2(i)$.

Using Maple to solve this LP problem we get:

| Objective Function | : | $-2\frac{1}{2}$ |
|---|---|---|
| Variable | | Activity Level |
| $v(1)$ | : | $-2\frac{1}{2}$ |
| $v(2)$ | : | $-2\frac{1}{2}$ |
| $g_{11}$ | : | $0$ |
| $g_{12}$ | : | $\frac{1}{2}$ |
| $u(1)$ | : | $\frac{1}{2}$ |
| $u(2)$ | : | $0$ |
| $g_{13}$ | : | $\frac{1}{2}$ |
| $g_{22}$ | : | $\frac{1}{2}$ |
| $g_{21}$ | : | $\frac{1}{2}$ |

The dual program looks as follows:

$$\max z(1)+z(2)$$

subject to:

(a)

$$\begin{pmatrix} 0 & 1 & -1 & 0 & -1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 0 & 0 & 0 & \vdots & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 & \vdots & 0 & -1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \\ x_{23} \\ y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

(b) $x_{11}+2x_{12}+z(1)\leq 0$
$5x_{11}+z(1)\leq 0$
$4x_{12}+z(1)\leq 0$
$3x_{22}+6x_{23}+z(2)\leq 0$
$6x_{21}+2x_{22}+z(2)\leq 0$
$x_{ij},y_{ij}\geq 0,\ i=1,2;\ j=1,2,3.$

Using Maple again we get:

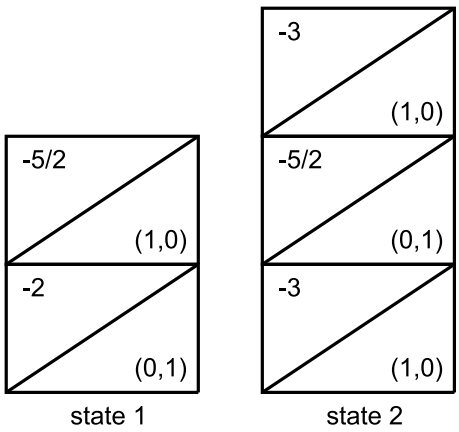
| Value Objective Function | : | $-2\frac{1}{2}$ |
|---|---|---|
| Variable | | Activity Level |
| $z(1)$ | : | $-\frac{10}{7}$ |
| $z(2)$ | : | $-\frac{15}{14}$ |
| $x_{11}$ | : | $\frac{2}{7}$ |
| $x_{12}$ | : | $\frac{5}{14}$ |
| $x_{21}$ | : | $\frac{5}{28}$ |
| $x_{22}$ | : | $0$ |
| $x_{23}$ | : | $\frac{5}{28}$ |
| $y_{12}$ | : | $0$ |
| $y_{21}$ | : | $0$ |
| $y_{23}$ | : | $\frac{1}{7}$ |

Using the results obtained for $(P_\alpha(1))$ and $(D_\alpha(1))$ we can now calculate the strategy $\hat{\mathbf{f}}$:

$$\begin{aligned}
\hat{f}(1,1) &= \frac{x_{11}}{x_{11}+x_{12}} = \frac{4}{9} \\
\hat{f}(1,2) &= \frac{x_{12}}{x_{11}+x_{12}} = \frac{5}{9} \\
\hat{f}(2,1) &= \frac{x_{21}}{x_{21}+x_{22}+x_{23}} = \frac{1}{2} \\
\hat{f}(2,2) &= \frac{x_{22}}{x_{21}+x_{22}+x_{23}} = 0 \\
\hat{f}(2,3) &= \frac{1}{2}.
\end{aligned}$$

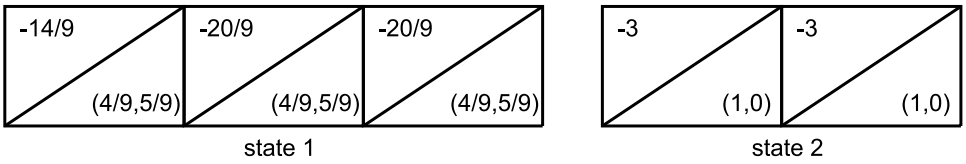
The optimal strategy $\hat{\mathbf{g}}$ is already mentioned in the solution of $(P_\alpha(1))$.

By fixing $\hat{\mathbf{g}}$ we can construct an AMD model that player 1 would be facing if she somehow knew that player 2 were going to use $\hat{\mathbf{g}}$. We can show that $\hat{\mathbf{f}}$ is optimal for player 1 in this model.



We have seen in chapter 2 that for a limiting average Markov Decision process the optimal strategy is pure. In this case all pure strategies will result in a value of $\mathbf{v} = -2\frac{1}{2}$. Since $\hat{\mathbf{f}}$ also results in $\mathbf{v} = -2\frac{1}{2}$ it must be optimal.

Of course we can also do this with the roles of player 1 and 2 reversed. We now get the AMD model




Now every pure strategy which does not involve choosing action 1 in state 1 is optimal, for example $\mathbf{f}^0 = ((0,1,0),(1,0))$. Since $\hat{\mathbf{g}} = \left((0,\frac{1}{2},\frac{1}{2}),(\frac{1}{2},\frac{1}{2})\right)$ results in the same value, it must be optimal.

In the remainder of this section we will validate Algorithm (2.5.1). We will begin by arguing that this algorithm is well defined. In particular, we note that taking an arbitrary $\mathbf{g} \in \mathbf{G}_S, \mathbf{u} = 0$ and $\mathbf{v} = M\mathbf{1}_N$ with $M := \max_{i,a^1,a^2}\left\{|r(i,a^1,a^2)|\right\}$, we obtain a feasible solution of $(P_\alpha(1))$. Clearly, the constraints (b) and (c) are satisfied trivially. The second block of constraints (a) reduces to

$$\mathbf{v}^T W = M\mathbf{1}^T \geq \mathbf{0}^T$$

while the first block of (a) becomes

$$\mathbf{u}^T W + \mathbf{v}^T J - \mathbf{g}^T R^T = \mathbf{0}^T + M\left[\mathbf{1}_N^T J\right] - [R\mathbf{g}]^T .$$

The last expression is greater than or equal to $\mathbf{0}^T$ because its $i$th block is

$$\left[M\mathbf{1}_m^1(i)\right] - [R(i)\mathbf{g}(i)]^T \geq \mathbf{0}_{m^1(i)}^T$$

for every $i \in S$

The existence of a finite optimal solution to $(P_\alpha(1))$ will be clear once we demonstrate that the $\mathbf{v}^T$-block of every feasible soution of $(P_\alpha(1))$ is bounded below. This will be a corollary of the next result.

**Proposition 2.5.1.** *Let $\overline{\mathbf{u}}^T, \overline{\mathbf{v}}^T, \overline{\mathbf{g}}^T$ be an arbitrary feasible solution of $(P_\alpha(1))$ and $\mathbf{f}$ be any stationary strategy of player 1. Then*

$$\overline{\mathbf{v}} \geq \mathbf{v}_\alpha(\mathbf{f}, \overline{\mathbf{g}}).$$

*Proof.* First consider the $\overline{\mathbf{v}}^T W \geq \mathbf{0}_{m^1}^T$ block of constraints (a). Since $W$ has the same block structure as $\mathbf{f}$, it is easy to verify that by mixing its $i$-th subblock with respect to $\mathbf{f}$ we obtain, for each $i \in \mathbf{S}$,

$$v(i) \geq [P(\mathbf{f})\overline{\mathbf{v}}]_i \,.$$

Equivalently,

$$\overline{\mathbf{v}} \geq P(\mathbf{f})\overline{\mathbf{v}};\ \mathbf{f} \in \mathbf{F}_S. \tag{2.43}$$

Furthermore, we can prove that

$$\overline{\mathbf{v}} \geq P(\mathbf{f})\overline{v} \ \Rightarrow\ \overline{\mathbf{v}} \geq Q(\mathbf{f})\overline{\mathbf{v}}. \tag{2.44}$$

Since $\overline{\mathbf{v}} \geq P(\mathbf{f})\overline{\mathbf{v}}$, we have

$$\overline{\mathbf{v}} \geq P(\mathbf{f})\overline{\mathbf{v}} \geq P^2(\mathbf{f})\overline{\mathbf{v}} \geq \ldots \geq P^k(\mathbf{f})\overline{\mathbf{v}}$$

This immediately leads to

$$\begin{aligned} T \cdot \overline{\mathbf{v}} &\geq \sum_{t=1}^{T} P^t(\mathbf{f})\overline{\mathbf{v}} \\ T \cdot \overline{\mathbf{v}} + \overline{\mathbf{v}} &\geq \sum_{t=1}^{T} P^t(\mathbf{f})\overline{\mathbf{v}} + \overline{\mathbf{v}} \\ (T+1)\overline{\mathbf{v}} &\geq \sum_{t=0}^{T} P^t(\mathbf{f})\overline{\mathbf{v}} \\ \overline{\mathbf{v}} &\geq \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f})\overline{\mathbf{v}} \end{aligned}$$

for every $T \in \mathbb{N}$. Taking the limit $T \to \infty$ now gives us the result needed:

$$\overline{\mathbf{v}} \geq \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f})\overline{\mathbf{v}} = Q(\mathbf{f})\overline{\mathbf{v}}. \tag{2.45}$$

Similarly, the $\overline{\mathbf{u}}^T W + \overline{\mathbf{v}}^T J - \overline{\mathbf{g}}^T R^T \geq \mathbf{0}_{m^1}^T$ block of constraints has the same structure as $\mathbf{f}$, and the mixing of its $i$-th subblock with respect to $\hat{\mathbf{f}}$ yields

$$\overline{v}(i) + \overline{u}(i) \geq \mathbf{f}(i)R(i)\overline{\mathbf{g}}(i) + [P(\mathbf{f})\overline{\mathbf{u}}]_i$$

for every $i \in \mathbf{S}$. Equivalently, in vector form we have

$$\overline{\mathbf{v}} + \overline{\mathbf{u}} \geq \mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) + P(\mathbf{f})\overline{\mathbf{u}};\ \mathbf{f} \in \mathbf{F}_S. \tag{2.46}$$

Multiplying both sides with $Q(\mathbf{f})$ gives us

$$Q(\mathbf{f})\overline{\mathbf{v}} + Q(\mathbf{f})\overline{\mathbf{u}} \geq Q(\mathbf{f})\mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) + Q(\mathbf{f})P(\mathbf{f})\overline{\mathbf{u}}.$$

Since $QP = Q$ we have

$$Q(\mathbf{f})\overline{\mathbf{v}} \geq Q(\mathbf{f})\mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) = \mathbf{v}_\alpha(\mathbf{f}, \overline{\mathbf{g}}).$$

Using (2.45), we now get

$$\overline{\mathbf{v}} \geq \mathbf{v}_\alpha(\mathbf{f}, \overline{\mathbf{g}}). \tag{2.47}$$

□

**Corollary 2.5.1.**

*(i) The objective function* $\left[\gamma^T \mathbf{v}\right]$ *of the linear program* $(P_\alpha(1))$ *is bounded below.*

*(ii) The programs* $(P_\alpha(1))$ *and* $(D_\alpha(1))$ *possess finite optimal solutions.*

Next we note that the second block of constraints (d) of $(D_\alpha(1))$, namely $J\mathbf{x} + W\mathbf{y} = \gamma$, has its $j$th entry (after a rearrangement of terms)

$$x_j + y_j = \left[\sum_{i \in S} \sum_{a^1 \in A^1(i)} p_{ij}(a^1) y_{ia^1}\right] + \gamma(j) \geq \gamma(j) > 0, \ i \in \mathbf{S}.$$

Thus, whenever $x_j = 0$ we have that $y_j > 0$, and hence $\hat{\mathbf{f}}$ in Step 3 of Algorithm (2.5.1) is well defined.

**Proposition 2.5.2.** *Let* $\left(\hat{\mathbf{u}}^T, \hat{\mathbf{v}}^T, \hat{\mathbf{g}}^T\right)$ *and* $\left(\hat{\mathbf{x}}^T, \hat{\mathbf{y}}^T, \hat{\mathbf{z}}^T\right)$ *be a dual pair of optimal solutions to* $(P_\alpha(1))$ *and* $(D_\alpha(1))$*, respectively, and let* $\hat{\mathbf{f}} \in \mathbf{F}_S$ *be a stationary strategy for player 1 as constructed by Step 3 of Algorithm (2.5.1). Also let* $\mathbf{S}^* \in S$ *be as in Step 2 of that algorithm. Then:*

*(i)* $\mathbf{S}^*$ *is the set of recurrent states in the Markov chain induced by* $P(\hat{\mathbf{f}})$

*(ii)* $\hat{\mathbf{v}} = P(\hat{\mathbf{f}})\hat{\mathbf{v}} = Q(\hat{\mathbf{f}})\hat{\mathbf{v}}$

*(iii)* $\hat{v}(i) + \left[\left(I - P(\hat{\mathbf{f}})\hat{\mathbf{u}}\right)\right]_i = \hat{\mathbf{f}}(i)R(i)\hat{\mathbf{g}}; \ i \in \mathbf{S}^*.$

*Proof.*

(i) First we show that $S^*$ is a closed set in the Markov chain $P(\mathbf{f}^*)$,that is

$$p_{ij}(a_i^1) > 0$$

for $i \in S^*$ and $j \notin S*$. Suppose that $S^*$ is not closed. This means that an $\bar{i} \in S^*, a_{\bar{i}}^1 \in A^1(\bar{i})$ and $j \notin S^*$ exist such that

$$p_{\bar{i}j}(a_{\bar{i}}^1) > 0.$$

The constraint $[W\mathbf{x}^*]_j = 0$ gives us

$$\sum_{a^1 \in A^1(j)} x^*_{ja^1} - \sum_{i=1}^{N} \sum_{a^1 \in A^1(i)} p_{ij}(a^1) x^*_{ia^1}.$$

Since $x^*_{\bar{i}a^1_{\bar{i}}} > 0$, the second summation above includes at least one strictly positive term corresponding to to the pair $(\bar{i}, a^1_{\bar{i}})$. Thus

$$\mathbf{x}^*_j = \sum_{a^1 \in A^1(j)} \mathbf{x}^*_{ja^1} > 0$$

which contradicts the assumption that $j \notin S^*$.
Now let $S^*_c = S \backslash S^*$. Suppose $S^*_c$ contains recurrent states of $P(\mathbf{f}^*)$. Then there is at least one ergodic class $\mathbf{E} = \{s_1, s_2, \ldots, s_p\}$ of $P(\mathbf{f}^*)$ that is completely contained in $S^*_c$ (otherwise, $\mathbf{E} \cap S^*$ is not empty, which contradicts the closedness of $S^*$). Since with each $i \in S^*_c$ the strategy $\mathbf{f}^*$ associates an action $a^1_i \in A^1(i)$ such that $y^*_{ia_i} > 0$, and since $\left((\mathbf{x}^*)^T, (\mathbf{y}^*)^T, (\mathbf{z}^*)^T\right)$ is an extreme optimal solution of $(D_\alpha(1))$, we must have that the columns of $\begin{pmatrix} 0 & W \end{pmatrix}^T$ corresponding to the pairs $(i_1, a_{i_1}), (i_2, a_{i_2}), \ldots, (i_p, a_{i_p})$ must be linearly independent. Let us denote these columns by $\begin{pmatrix} \mathbf{0} \\ \mathbf{w}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \mathbf{0} \\ \mathbf{w}_p \end{pmatrix}$. Note that the $N$ entries of each $\mathbf{w}_1, \ldots, \mathbf{w}_p$ can be partitioned into those corresponding to states $j \in \mathbf{E}$ and those corresponding to states $i \notin \mathbf{E}$. If $j \notin \mathbf{E}$, we have

$$[\mathbf{w}_t]_j = \delta(i_t, j) - p_{i_t j}(a^1_{i_t}) = 0 - 0 = 0$$

for each $t = 1, \ldots, p$, since $\mathbf{E}$ is assumed to be an ergodic class in $P(\mathbf{f}^*)$. Now let $\overline{\mathbf{w}}_1, \overline{\mathbf{w}}_2, \ldots, \overline{\mathbf{w}}_p$ be the truncations of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$ obtained by deleting the zero entries corresponding to the states not in $\mathbf{E}$. The truncated $p \times p$ matrix $\overline{W} = [\overline{\mathbf{w}}_1, \overline{\mathbf{w}}_2, \ldots, \overline{\mathbf{w}}_p]$ must be nonsingular. However, for each $t = 1, 2, \ldots, p$

$$\mathbf{1}^T \overline{\mathbf{w}}_t = \sum_{j \in \mathbf{E}} \left(\delta(i_t, j) - p_{i_t j}(a^1_{i_t})\right) = 1 - 1 = 0,$$

since $\mathbf{E}$ is closed under $P(\mathbf{f}^*)$. Thus $\mathbf{1}\overline{W} = \mathbf{0}^T$, contradicting the nonsingularity of $\overline{W}$.

(ii) From the second set of constraints of $(P_\alpha(1))$, namely $(\mathbf{v}^*)^T \geq \mathbf{0}^T$, it immediately follows that

$$\sum_{j=1}^{N} \left[\delta(j, i) - p_{ij}(a^1)\right] v^*(j) \geq 0$$

for all $a^1 \in A^1(i)$ and $i \in S$. Due to the complementary slackness we have again that for each $i \notin S^*$ and $a_i^1$ selected by $\mathbf{f}^*$ in that state $i$.

$$\sum_{j=1}^{N} \left(\delta(j,i) - p_{ij}(a_i^1)\right) \upsilon^*(j) = 0.$$

This gives us

$$[(I - P(\mathbf{f}^*))\mathbf{v}^*]_i = 0$$

whenever $i \notin S^*$ and is nonnegative otherwise.

Suppose that there exists some $\bar{i} \in S^*$ such that

$$0 < [(I - P(\mathbf{f}^*))\mathbf{v}^*]_{\bar{i}} = \sum_{j=1}^{N} \left(\delta(j,\bar{i}) - p_{\bar{i}j}(a_{\bar{i}}^1)\upsilon^*(j)\right).$$

Since $x^*_{\bar{i}a_{\bar{i}}} > 0$ for the pair $(\bar{i}, a_{\bar{i}})$, by the definition of $\mathbf{f}^*$ we have that

$$0 < \sum_{i=1}^{N} \sum_{a^1=1}^{m^1(i)} x^*_{ia} \sum_{j=1}^{N} (\delta(j,i) - p_{ij}(a^1))\upsilon^*(j)$$

since it contains as least one strictly positive term. If we change the order of summation however, the equation above yields

$$\sum_{j=1} \upsilon^*(j) \left[\sum_{j=1}^{N} \sum_{a^1=1}^{m^1(i)} (\delta(j,i) - p_{ij}(a^1))\, x^*_{ia^1}\right] = \sum_{j=1}^{N} \upsilon^*(j)\, [W\mathbf{x}^*]_j = 0$$

since $[W\mathbf{x}^*]_j = 0$ for all $j \in S$ according to the first set of constraints of $D_\alpha(1)$. That equality with zero contradicts the preceding strict inequality however. This completes the proof.

(iii) Since $x^*_{ia_i^1} > 0$ for each $i \in S^*$, the complementary slackness property of linear programming immediately yields (iii).

□

Let

$$Q(\hat{\mathbf{f}}) = \left(q_{ij}(\hat{\mathbf{f}})\right)_{i,j=1}^{N}$$

and note that by part (i) of the above proposition

$$q_{ij}(\hat{\mathbf{f}}) = 0 \text{ if } j \notin \mathbf{S}^*$$

where the latter is a consequence of properties of stationary distributions of Markov chains. Now consider

$$v_\alpha(i, \hat{\mathbf{f}}, \hat{\mathbf{g}}) = \left[Q(\hat{\mathbf{f}})\mathbf{r}(\hat{f}, \hat{\mathbf{g}})\right]_i = \sum_{j \in \mathbf{S}^*} q_{ij}(\hat{\mathbf{f}}) r(i, \hat{\mathbf{f}}, \hat{\mathbf{g}})$$

for any $i \in S$. We can apply Proposition 2.5.2 (iii) to the above equation to get that, for any $i \in S$,

$$\begin{aligned} v_\alpha(i, \hat{\mathbf{f}}, \hat{\mathbf{g}}) &= \sum_{j \in \mathbf{S}^*} q_{ij}(\hat{\mathbf{f}}) \hat{v}(j) + \sum_{j \in \mathbf{S}^*} q_{ij}(\hat{\mathbf{f}}) \left[\left(I - P(\hat{\mathbf{f}})\right) \hat{\mathbf{u}}\right]_i \\ &= \left[Q(\hat{\mathbf{f}})\hat{\mathbf{v}}\right]_i + \left[Q(\hat{\mathbf{f}}) \left(I - P(\hat{f})\right) \hat{\mathbf{u}}\right]_i \\ &= \hat{\mathbf{u}}_i, \end{aligned}$$

where the last equality follows from Proposition 2.5.2 (ii) and the fact that $Q(\hat{f}) = Q(\hat{f}P(\hat{f}))$. This means that an optimal solution $\hat{\mathbf{u}}^T, \hat{\mathbf{v}}^T, \hat{\mathbf{g}}^T$ of $(P_\alpha(1))$ satisfies

$$\hat{\mathbf{v}} = \mathbf{v}_\alpha(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \tag{2.48}$$

(2.48) and Proposition 2.5.1 together yield one half of the saddle point optimality condition, namely,

$$\mathbf{v}_\alpha(\mathbf{f}, \hat{\mathbf{g}}) \leq \mathbf{v}_\alpha(\hat{\mathbf{f}}, \hat{\mathbf{g}}), \ \mathbf{f} \in \mathbf{F}_S. \tag{2.49}$$

**Proposition 2.5.3.** *Let* $(\hat{\mathbf{u}}^T, \hat{\mathbf{v}}^T, \hat{\mathbf{g}}^T)$ *and* $(\hat{\mathbf{x}}^T, \hat{\mathbf{y}}^T, \hat{\mathbf{z}}^T)$, *as well as* $\hat{\mathbf{f}}$, *be as in Proposition 2.5.2. Then:*

*(i)* $\sum_{i \in S} \hat{z}_i = \gamma^T \mathbf{v}_\alpha(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \leq \gamma^T \mathbf{v}_\alpha(\hat{f}, \mathbf{g})$ *for all* $\mathbf{g} \in \mathbf{G}_S$, *and*

*(ii)* $\mathbf{v}_\alpha(\hat{f}, \hat{g}) \leq \mathbf{v}_\alpha(\hat{f}, \mathbf{g})$ *for all* $\mathbf{g} \in \mathbf{G}_S$.

*Proof.*

(i) (2.48) and the strong duality theorem of linear programming imply that

$$\sum_{i \in \mathbf{S}} \hat{z}_i = \gamma^T \mathbf{v}_\alpha(\hat{f}, \hat{\mathbf{g}}).$$

Now, from constraints (e) of $(D_\beta(1))$ divided by $\hat{x}_j$ we have that for each $j \in \mathbf{S}^*$

$$\hat{z}_j \mathbf{1}_{m^2(j)} \leq \hat{x}_j \left[\hat{\mathbf{f}}(j) R(j)\right]$$

so mixing the above constraints with respect to an arbitrary $\mathbf{g} \in \mathbf{G}_S$ yields for $j \in \mathbf{S}^*$

$$\hat{z}_j \leq \hat{x}_j r(j, \hat{\mathbf{f}}, \mathbf{g}). \tag{2.50}$$

We will now show that

$$\hat{x}_j = \left[\gamma^T Q(\hat{\mathbf{f}})\right]_j ;\ j \in \mathbf{S}. \tag{2.51}$$

Let $S_c^* := \mathbf{S}\backslash\mathbf{S}^*$ be the set of transient states induced by $P(\hat{\mathbf{f}})$ and $E_1, E_2, \ldots, E_m$ the ergodic sets. Let $n_k := |E_k|,\ k = 1, 2, \ldots, m$. We can now write $\gamma$ as

$$\gamma_l := \begin{cases} \frac{1}{n}, & \text{for } i \in \mathbf{S}_c^* \\ \frac{1}{n_k} \sum_{j \in \mathbb{E}_k} \left\{ \hat{x}_j - \frac{1}{n} \sum_{i \in \mathbf{S}_c^*} Q_{in}(\hat{\mathbf{f}}) \right\}, & \text{for } l \in E_k,\ k = 1, 2, \ldots, m. \end{cases} \tag{2.52}$$

This implies

$$\begin{aligned} \sum_{l \in \mathbf{S}} \sum_{j \in E_k} \gamma_l q_{lj}(\hat{\mathbf{f}}) &= \sum_{l \in \mathbf{S}_c^*} \sum_{j \in E_k} \gamma_l q_{lj}(\hat{\mathbf{f}}) + \sum_{i \in E_k} \gamma_i q_{ij}(\hat{\mathbf{f}}) \\ &= \frac{1}{n} \sum_{l \in \mathbf{S}} \sum_{j \in E_k} q_{lj}(\hat{\mathbf{f}}) + \sum_{i \in E_k} \gamma_i \\ &= \frac{1}{n} \sum_{l \in \mathbf{S}} \sum_{j \in E_k} q_{lj}(\hat{\mathbf{f}}) + \sum_{j \in \mathbb{E}_k} \left\{ \hat{x}_j - \frac{1}{n} \sum_{i \in \mathbf{S}_c^*} Q_{in}(\hat{\mathbf{f}}) \right\} \\ &= \sum_{j \in E_k} \hat{x}_j \end{aligned} \tag{2.53}$$

for $k = 1, 2, \ldots, m$. From the linear program $(D_\alpha(1))$ and the definition of $\hat{\mathbf{x}}$ it follows that $(\hat{\mathbf{x}})^T = (\hat{\mathbf{x}})^T P(\hat{\mathbf{f}})$ and, consequently, $\hat{\mathbf{x}} = (\hat{\mathbf{x}})^T Q(\hat{\mathbf{f}})$. Because $\hat{\mathbf{x}}_i = 0$ for all $i \in \mathbf{S}_c^*$, and $q_i(\hat{\mathbf{f}}) = 0$ for all $i \in \mathbf{S}_c^*$, we have

$$\hat{\mathbf{x}}_i = \left(\gamma^T Q(\hat{f})\right)_i = 0,\ i \in \mathbf{S}_c^*. \tag{2.54}$$

For any $i \in E_k$, we obtain using (2.53)

$$\begin{aligned} \hat{x}_i &= \sum_{j \in S} \hat{x}_j q_{ji}(\hat{f}) = \sum_{j \in E_k} \hat{x}_j q_{ji}(\hat{f}) = q_{ii}(\hat{\mathbf{f}}) \sum_{j \in E_k} \hat{x}_j \\ &= \sum_{l \in \mathbf{S}} q_{ii}(\hat{\mathbf{f}}) \sum_{l \in \mathbf{S}} \sum_{j \in E_k} \gamma_l q_{lj}(\hat{\mathbf{f}}) \\ &= \sum_{l \in \mathbf{S}} \gamma_l \sum_{j \in E_k} q_{lj}(\hat{\mathbf{f}}) q_{ji}(\hat{\mathbf{f}}) \\ &= \sum_{l \in \mathbf{S}} \gamma_l q_{li}(\hat{\mathbf{f}}) \\ &= \left(\gamma^T Q(\hat{f})\right)_i . \end{aligned} \tag{2.55}$$

Thus, (2.54) and (2.55) imply that $\hat{\mathbf{x}}^T = \gamma^T Q(\hat{f})$.

Noting that for $j \notin \mathbf{S}^*$ (again by constraints (e)) $\hat{z}_j \leq 0$, we can combine (2.50) and (2.51) to obtain

$$\begin{aligned}\sum_{j\in\mathbf{S}} \hat{z}_j &\leq \sum_{j\in\mathbf{S}} \hat{x}_j r(i,\hat{f},\mathbf{g}) \\ &= \sum_{j\in\mathbf{S}} \left[\gamma^T Q\hat{\mathbf{f}}\right]_j r(j,\hat{f},\mathbf{g}) \\ &= \gamma^T \left[Q(\hat{\mathbf{f}}\mathbf{r}(\hat{f},\mathbf{g}))\right] \\ &= \gamma^T v_\alpha(\hat{f},\mathbf{g}),\end{aligned}$$

which completes the proof of part (i).

(ii) Consider now the AMD model for player 2 induced by player 1 fixing $\hat{\mathbf{f}}$. Let $\mathbf{g}^0$ be optimal in that AMD, then componentwise

$$\min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(\hat{f},\mathbf{g}) = \mathbf{v}_\alpha(\hat{f},\mathbf{g}^0) \leq \mathbf{v}_\alpha(\hat{\mathbf{f}},\hat{\mathbf{g}}).$$

We claim that strict inequality is impossible in any component of the above vector inequality. If it were possible, then using the fact that $\gamma(i) > 0$ for every $i$ we could have

$$\gamma^T \mathbf{v}_\alpha(\hat{\mathbf{f}},\mathbf{g}) < \gamma^T \mathbf{v}_\alpha(\hat{\mathbf{f}},\hat{\mathbf{g}})$$

which would contradict (i). Thus

$$\min_{\mathbf{g}\in\mathbf{G}_S} \mathbf{v}_\alpha(\hat{\mathbf{f}},\mathbf{g}) = \mathbf{v}_\alpha(\hat{f},\hat{\mathbf{g}}),$$

which completes the proof of the proposition.

□

Finally, we have obtained, from (2.49) and Proposition 2.5.3 (ii), that

$$\mathbf{v}_\alpha(\mathbf{f},\hat{\mathbf{g}}) \leq \mathbf{v}_\alpha(\hat{\mathbf{f}},\hat{\mathbf{g}}) \leq \mathbf{v}_\alpha(\hat{\mathbf{f}},\mathbf{g})$$

for all $\mathbf{f} \in \mathbf{F}_S, \mathbf{g} \in \mathbf{G}_S$.

## 2.6 Nonlinear Programming and Zero-Sum Stochastic Games

In this section we will present a more general solution method of $\Gamma_\beta$ which is extendable to the general-sum and the limiting average case, and takes into account the inherent nonlinearity of stochastic games. We begin by briefly reviewing the elements of the optimality equations in Theorem 2.1.2.

For an arbitrary vector $\mathbf{v} \in \mathbb{R}^N$, Shapley's auxiliary matrix games $R(i, \mathbf{v})$ for $i \in S$ have the structure

$$R(i, \mathbf{v}) = R(i) + \beta T(i, \mathbf{v}) \tag{2.56}$$

where

$$T(i, \mathbf{v}) = \left[ \sum_{j \in S} p_{ij}(a^1, a^2) v(j) \right]_{a^1=1, a^2=1}^{m^1(i), m^2(i)} \tag{2.57}$$

for each $i \in S$. This is a convenient representation because $R(i)$ contains all of the reward information of state $i$, while $T(i, \mathbf{v})$ contains all of the transition information.

Whenever we search for a solution of $\Gamma_\beta$ in stationary strategies we are searching for vectors $\mathbf{f} \in \mathbb{R}^{m^1}, \mathbf{g} \in \mathbb{R}^{m^2}$ and $\mathbf{v} \in \mathbb{R}^N$ satisfying conditions of the type

$$R(i, \mathbf{v})\mathbf{g}(i) \leq v(i)\mathbf{1}_{m^1(i)}$$

or $\mathbf{f}(i)R(i, \mathbf{v}) \geq v(i)\mathbf{1}_{m^2(i)^T}$ for the other player, which capture the requirements of the optimality equation that the players must play optimally in the matrix games $R(i, \mathbf{v})$, provided that $\mathbf{v}$ were the value vector. The above inequalities can be viewed as

$$R(i)\mathbf{g}(i) + \beta T(i, \mathbf{v})\mathbf{g}(i) \leq v(i)\mathbf{1}_m^1(i), \ i \in \mathbf{S}. \tag{2.58}$$

The most natural extension of $(P_\beta(1))$ now becomes:

$$\min [\mathbf{1}\mathbf{v}]$$

subject to: $(NL_\beta(1))$

(a) $R(i)\mathbf{g}(i) + \beta T(i, \mathbf{v})\mathbf{g}(i) \leq v(i)\mathbf{1}_{m^1(i)}, \ i \in \mathbf{S}$

(b) $\mathbf{1}^T\mathbf{g}(i) = 1, \ i \in \mathbf{S}$

(c) $\mathbf{g}(i) \geq \mathbf{0}, \ i \in \mathbf{S}$.

If we assume the validity of Shapley's Theorem, then the following simple result immediately demonstrates that optimal solutions of $(NL_\beta(1))$ are closely related to the solution of $\Gamma_\beta$.

**Theorem 2.6.1.** *Let $\mathbf{v}_\beta$ be the value vector of the stochastic game and $\mathbf{g}^0$ be an optimal stationary strategy for player 2. Then $\left(\mathbf{v}_\beta^T, (\mathbf{g}^0)^T\right)$ is a global minimum of the nonlinear program $(NL_\beta(1))$*

Of course we can formulate $(NL_\beta(1))$ for player 2 as well. We now get:

$$\max [\mathbf{1}\mathbf{v}]$$

subject to: $(NL_\beta(2))$

(a) $\mathbf{f}(i)R(i) + \beta\mathbf{f}(i)T(i,\mathbf{v}) \geq v(i)\mathbf{1}_{m^2(i)},\ i \in \mathbf{S}$

(b) $\mathbf{f}(i)\mathbf{1} = 1,\ i \in \mathbf{S}$

(c) $\mathbf{f}(i) \geq \mathbf{0}^T,\ i \in \mathbf{S}$.

Clearly, the symmetric analogue of Theorem 2.6.1 holds for $(NL_\beta(2))$ with respect to player 1's optimal stationary strategy $\mathbf{f}^0$.

When we use the more general formulation of a Nash equilibrium (see (2.3) and (2.4)) instead of (2.6), we see that we can combine $(NL_\beta(1))$ and $(NL_\beta(2))$ into a single nonlinear program:

$$\min\left[\mathbf{1}^T(\mathbf{v}^1 + \mathbf{v}^2)\right]$$

subject to: $(NL_\beta)$

(a) $R(i)\mathbf{g}(i) + \beta T(i,\mathbf{v}^1)\mathbf{g}(i) \leq v^1(i)\mathbf{1}_{m^1(i)},\ i \in \mathbf{S}$

(b) $\mathbf{f}(i)(-R(i)) + \beta\mathbf{f}(i)T(i,\mathbf{v}^2) \leq v^2(i)\mathbf{1}_{m^2(i)},\ i \in \mathbf{S}$

(c) $(\mathbf{f},\mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$,

where (b) was obtained by multiplying (a) in $(NL_\beta(2))$ by -1 and replacing $\mathbf{v}$ with $-\mathbf{v}^2$. Constraints (b) and (c) from $(NL_\beta(1))$ and $(NL_\beta(2))$ are, together, equivalent to the constraint (c) in $(NL_\beta)$.

**Theorem 2.6.2.** *Let $(\hat{\mathbf{v}}^1, \hat{\mathbf{v}}^2, \hat{\mathbf{f}}, \hat{\mathbf{g}})$ be a global minimum of $(NL_\beta)$. Then:*

*(i)* $\mathbf{1}(\hat{\mathbf{v}}^1 + \hat{\mathbf{v}}^2) = 0$

*(ii)* $\hat{\mathbf{v}}^1 = \mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}})$ *and* $\hat{\mathbf{v}}^2 = -\mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}})$

*(iii)* $\hat{\mathbf{f}}$ *and* $\hat{\mathbf{g}}$ *are optimal stationary strategies for players 1 and 2, respectively, in* $\Gamma_\beta$.

### 2.6.1 Extensions to the Limiting Average Games

As we have seen in the Big Match example, optimal stationary strategies do not necessarily exist in limiting average stochastic games. This means we cannot simply translate Theorem 2.6.2 to the zero-sum limiting average stochastic game. To do this we need the concept of $\varepsilon$-optimal stationary strategies, developed by Blackwell and Ferguson in their paper on the Big Match [2]. Given an $\varepsilon > 0$, we shall say that $(\mathbf{f}_\varepsilon, \mathbf{g}_\varepsilon) \in \mathbf{F}_S \times \mathbf{G}_S$ are *$\varepsilon$-optimal stationary strategies* for players 1 and 2, respectively, in the game $\Gamma_\alpha$, if

$$\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_\varepsilon) - \varepsilon\mathbf{1} \leq \mathbf{v}_\alpha(\mathbf{f}_\varepsilon, \mathbf{g}_\varepsilon) \leq \mathbf{v}_\alpha(\mathbf{f}_\varepsilon, \mathbf{g}) + \varepsilon\mathbf{1} \tag{2.59}$$

for all $\mathbf{f} \in \mathbf{F}_S$ and $\mathbf{g} \in \mathbf{G}_S$. It is possible to show that if $\mathbf{v}_\alpha$ is the value vector of $\Gamma_\alpha$, then

$$||\mathbf{v}_\alpha(\mathbf{f}_\varepsilon, \mathbf{g}_\varepsilon) - \mathbf{v}_\alpha|| \leq \varepsilon.$$

In this sense, 0-optimality reduces to the usual mimimax optimality.

In order to evaluate the perfomance of an $\varepsilon$-optimal strategy we can define the following measure of *distance from optimality* for an arbitrary pair of stationary strategies $(\overline{f}, \overline{g})$:

$$\Delta(\overline{\mathbf{f}}, \overline{\mathbf{g}}) = \sum_{i \in S} \left[ \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\alpha(i, \mathbf{f}, \overline{\mathbf{g}}) - \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\alpha(i, \overline{\mathbf{f}}, \mathbf{g}) \right]. \tag{2.60}$$

Every term in the summation above is nonnegative and $\Delta(\overline{\mathbf{f}}, \overline{\mathbf{g}}) = 0$ if and only if $(\overline{\mathbf{f}}, \overline{\mathbf{g}})$ is a pair of optimal stationary strategies in $\Gamma_\alpha$. Furthermore, if $0 < \Delta(\overline{\mathbf{f}}, \overline{\mathbf{g}})$, then $(\overline{\mathbf{f}}, \overline{\mathbf{g}})$ is an $\varepsilon$-optimal pair of strategies for any $\varepsilon \leq \Delta(\overline{f}, \overline{\mathbf{g}})$.

We can now formulate an extension of the nonlinear program $(NL_\beta)$ to the case of a limiting average, zero-sum, stochastic game $\Gamma_\alpha$:

$$\min \left[ \mathbf{1}^T (\mathbf{v}^1 + \mathbf{v}^2) \right]$$

subject to: $(NL_\alpha)$

(a) $T(i, \mathbf{v}^1)\mathbf{g}(i) \leq v^1(i)\mathbf{1}_{m^1(i)},\ i \in \mathbf{S}$

(b) $R(i)\mathbf{g}(i) + T(i, \mathbf{u}^1)\mathbf{g}(i) \leq (v^1(i) + u^1(i))\mathbf{1}_{m^1(i)},\ i \in \mathbf{S}$

(c) $\mathbf{f}(i)T(i, \mathbf{v}^2) \leq v^2(i)\mathbf{1}^T_{m^2(i)},\ i \in \mathbf{S}$

(d) $\mathbf{f}(i)[-R(i)] + \mathbf{f}(i)T(i, \mathbf{u}^2) \leq (v^2(i) + u^2(i))\mathbf{1}_{m^2(i)},\ i \in \mathbf{S}$

(e) $(\mathbf{f}, \mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$.

The relationship between $\varepsilon$-optimality, in stationary strategies, and solutions of the nonlinear program $(NL_\alpha)$ is summarized in the following result.

**Theorem 2.6.3.** *Consider a zero-sum limiting average game $\Gamma_\alpha$ and a pair of $(\overline{\mathbf{f}}, \overline{\mathbf{g}})$ of stationary strategies for players 1 and 2, respectively.*

*(i) If there exists $\overline{\mathbf{u}}^1, \overline{\mathbf{u}}^2, \overline{\mathbf{v}}^1, \overline{\mathbf{v}}^2$ in $\mathbb{R}^N$ which together with $\overline{\mathbf{f}}$ and $\overline{\mathbf{g}}$ form a feasible point of $(NL_\alpha)$, and if $\mathbf{1}^T(\overline{\mathbf{v}}^1 + \overline{\mathbf{v}}^2) = \varepsilon$, then $\overline{\mathbf{f}}$ and $\overline{\mathbf{g}}$ are $\varepsilon$-optimal stationary strategies for players 1 and 2, respectively.*

*(ii) Conversely, if $(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \in \mathbf{F}_S \times \mathbf{G}_S$ is an $\varepsilon$-optimal strategy pair, then there exists $\hat{\mathbf{u}}^1, \hat{\mathbf{v}}^1, \hat{\mathbf{u}}^2, \hat{\mathbf{v}}^2 \in \mathbb{R}^N$ which together with $(\hat{f}, \hat{g})$ form a feasible point of $(NL_\alpha)$ that has an objective function value $\mathbf{1}^T(\hat{\mathbf{v}}^1 + \hat{\mathbf{v}}^2)$ that is $2N\varepsilon$, or less.*

**Corollary 2.6.1.** *If* $\hat{\mathbf{u}}^1, \hat{\mathbf{v}}^1, \hat{\mathbf{g}}, \hat{\mathbf{u}}^2, \hat{\mathbf{v}}^2, \hat{\mathbf{f}}$ *form a feasible solution of* $(NL_\alpha)$ *with* $\mathbf{1}(\hat{\mathbf{v}}^1 + \hat{\mathbf{v}}^2) = 0$*, then it is a global minimum and* $(\hat{\mathbf{f}}, \hat{\mathbf{g}})$ *is an optimal strategy pair in* $\Gamma_\alpha$*. Conversely, if optimal stationary strategies* $(\hat{\mathbf{f}}, \hat{\mathbf{g}})$ *exist in* $\Gamma_\alpha$*, then there is a global minimum with an objective function value equal to 0 in* $(NL_\alpha)$*.*

We now have a set of necessary and sufficient conditions for the existence of optimal stationary strategies, and the means to find them. We can begin with any reasonable pair of stationary strategies $(\mathbf{f}^0, \mathbf{g}^0)$, construct a feasible point of $(NL_\alpha)$, and apply any descent algorithm to that nonlinear program. The value of the objective function allows us to measure how much improvement has been made with the strategy pair $(\mathbf{f}^*, \mathbf{g}^*)$ when the algorithm terminates.

**Example 2.6.1.** We will now investigate how we can apply the results in this section to the Big Match example mentioned earlier. We have already seen that there is no optimal Markov strategy to be found for the Big Match. It is possible however to define an $\varepsilon$-optimal strategy, as was shown by [2].

**Theorem 2.6.4.** *The value of the Big Match is* $\frac{1}{2}$*. An optimal strategy for player 2 is to toss a fair coin every day. Player 1 has no optimal strategy, but for any nonnegative integer* $N$ *he can get*

$$v(\pi_N^1, \pi^2) = \frac{N}{2(N+1)}$$

*by using strategy* $\pi_N^1$*, defined as follows: let* $x_m \in \{1, 2\}$ *be the action chosen by player 2 at stage* $m$*. Then the history up to stage* $t+1$ *is defined by* $h_t = (x_1, x_2, \ldots, x_t)$*. We now calculate the excess* $k_t$ *of action 1 over action 2 in* $h_t$*, and choose action 2 with probability* $P(k_t + N)$*, where* $P(m) = \frac{1}{(m+1)^2}$*.*

*Proof.* Let $\pi^1 = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_t, \ldots) \in \mathbf{F}_M$ and $\pi_2 = (\mathbf{g}_1, \mathbf{g}_2, \ldots \mathbf{g}_t, \ldots) \in \mathbf{F}_M$ be a Markov strategy for player 1 resp. player 2. If player 2 plays $g_t = (\frac{1}{2}, \frac{1}{2})$ for all $t$, the expected payoff is $\frac{1}{2}$, no matter what he does.

Next, notice that strategy $N$ chooses action 2 at stage $t+1$ with certainty whenever $k_t = -N$.
Let $T$ denote the number of stages after which player 1 plays action 2 (at stage $T+1$ the game is essentially over). Let $T(m)$ denote the event $[T > m,$ or $T < m$ and $x_{T+1} = 2]$. We can now distinguish two types of strategies: one for which $k_t$ will eventually equal $-N$, and one for which it will not. (To distinguish these types it is assumed that player 1 will choose action 1 all the time.)

**Case A.** Let $\pi^2$ be a pure strategy for which we eventually have $k_t = -N$ for some $t$. By induction on $m$ we can show that $\mathbb{P}_{\pi_N^1, \pi^2}[T(m)] \geq$

$\frac{N}{2(N+1)}$ for all $m$.

$$\mathbb{P}_{\pi_N^1,\pi^2}[x_{T+1}=2]=\lim_{m\to\infty}\mathbb{P}_{\pi_N^1,\pi^2}[T(m)]\geq\frac{N}{2(N+1)} \tag{2.61}$$

(a) m=1. If $x_1=1$,

$$\mathbb{P}_{\pi_N^1,\pi^2}[T(1)]=1>v(\pi_N^1,\pi^2)=\frac{N}{2(N+1)}.$$

If $x_1=0$,

$$\begin{aligned}\mathbb{P}_{\pi_N^1,\pi^2}[T(1)] &= \mathbb{P}_{\pi_N^1,\pi^2}[t\geq 1]\\ &= 1-p(N)=\frac{N(N+2)}{(N+1)^2}\\ &\geq v(\pi_N^1,\pi^2).\end{aligned}$$

(b) Suppose $\mathbb{P}_{\pi_N^1,\pi^2}[T(m)]\geq v(\pi_N^1,\pi^2)$, for all $N$. If $x_1=1$,

$$\begin{aligned}\mathbb{P}_{\pi_N^1,\pi^2}[T(m+1)] &= p(N)+[1-p(N)]\,\mathbb{P}_{\pi_N^1,\pi^2}[T(m)]\\ &\geq p(N)+[1-p(N)]\,v(\pi_{N-1}^1,\pi^2)\\ &= v(\pi_N^1,\pi^2),\end{aligned}$$

where $\mathbb{P}_{\pi_{N-1}^1,\pi^2}[T(m+1)]\geq v(\pi_{N-1}^1,\pi^2)$ by induction since using strategy $N$ against $h=(1,x_2,x_3,\dots)$ is equivalent to choosing 1 initially with probability $p(N)$ and, with probability $1-p(N)$ predicting 0 initially and thereafter using strategy $N-1$ against $\omega'=(x_2,x_3,\dots)$. Similarly, if $x_1=0$,

$$\begin{aligned}\mathbb{P}_{\pi_N^1,\pi^2}[T(m+1)] &= [1-p(N)]\,\mathbb{P}_{\pi_{N+1}^1,\pi^2}[T(m)]\\ &\geq [1-p(N)]\,v(\pi_{N+1}^1,\pi^2)\\ &= v(\pi_N^1,\pi^2).\end{aligned}$$

So (2.61) is proved. Since for $\left(\pi_N^1,\pi^2\right)$ we have that $T<\infty$ with probability 1, we get

$$\mathbb{P}_{\pi_N^1,\pi^2}[x_{T+1}=2]=\lim_{m\to\infty}\mathbb{P}_{\pi_N^1,\pi^2}[T(m)]\geq\frac{N}{2(N+1)}.$$

**Case B.** Let $\pi^{2'}$ be a strategy for which $k_t>-N$ for all $t$. Define

$$\begin{aligned}\lambda(m) &= \mathbb{P}_{\pi_N^1,\pi^2}\{T<m \text{ and } x_t=1\}\\ \mu(m) &= \mathbb{P}_{\pi_N^1,\pi^2}\{T<m \text{ and } x_t=2\}\\ \lambda &= \lim_{m\to\infty}\lambda(m),\ \mu=\lim_{m\to\infty}\mu(m).\end{aligned}$$

Also define $\pi_m^{2'} = \big(g_1, g_2, \ldots, g_m, (\frac{1}{2}), (\frac{1}{2}, \frac{1}{2}), \ldots\big)$. Then, for each $m$ we have that $\pi_m^{2'}$ is a strategy of the type that is considered in Case A. Now observe that

$$\begin{aligned} v(\pi_N^1, \pi^{2'}) &\geq \mu + \frac{1}{2}(1 - \lambda - \mu) \\ &= \lim_{m\to\infty} \left[\mu(m) + \frac{1}{2}(1 - \lambda(m) - \mu(m))\right] \\ &= \lim_{m\to\infty} v(\pi_N^1, \pi^{2'}) \\ &\geq \frac{N}{2(N+1)}, \end{aligned}$$

where the first inequality follows from the fact that $k_t > -N$ for all $t$, which implies that player 1 will get at least $\frac{1}{2}$ if play does not absorb, and where we have used the result of Case A for the last inequality.

So, for every $\varepsilon > 0$ player 1 can ensure himself a limiting average reward of at least $\frac{1}{2} - \varepsilon$, by taking $N$ sufficiently large. As we have seen, player 2 can guarantee himself a limiting average reward of (at most) $\frac{1}{2}$. □

## 2.7 Nonlinear Programming and General-Sum Stochastic Games

In this section we shall discuss a generalization of the results of Section 2.6 to the general-sum, $K$-player, stochastic games, both discounted and limiting average. We will only derive the results for the case of two players, since the extension to K players only involves a more complex notation (see [5]).

We shall start with the 2-person, discounted general-sum stochastic games, denoted with $\Gamma_\beta$. Given $\varepsilon > 0$, we shall say that $(\mathbf{f}_\varepsilon, \mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$ forms an *$\varepsilon$-equilibrium* in stationary strategies if

$$\mathbf{v}_\beta^1(\mathbf{f}, \mathbf{g}_\varepsilon) - \varepsilon\mathbf{1} \leq \mathbf{v}_\beta^1(\mathbf{f}_\varepsilon, \mathbf{g}) \text{ and } \mathbf{v}_\beta^2(\mathbf{f}_\varepsilon, \mathbf{g}) - \varepsilon\mathbf{1} \leq \mathbf{v}_\beta^2(\mathbf{f}_\varepsilon, \mathbf{g}_\varepsilon) \tag{2.62}$$

for all $\mathbf{f} \in \mathbf{F}_S$ and $\mathbf{g} \in \mathbf{G}_S$. In this section we will assume the validity of the following existence theorem.

**Theorem 2.7.1.** *In a general-sum, discounted stochastic game $\Gamma_\beta$, a Nash equilibrium exists in stationary strategies.*

Let

$$\mathbf{z}^T = \big((\mathbf{v}^1)^T, (\mathbf{v}^2)^T, \mathbf{f}, \mathbf{g}^T\big)$$

be a $(2N + m^1 + m^2)$-dimensional vector of variables and consider the nonlinear program

$$\min \left\{ \sum_{k=1}^{2} \mathbf{1} \left[\mathbf{v}^k - \mathbf{r}^k(\mathbf{f}, \mathbf{g}) - \beta P(\mathbf{f}, \mathbf{g})\mathbf{v}^k\right] \right\}$$

subject to: $(GNL_\beta)$

(a) $R^1(i)\mathbf{g}(i) + \beta T(i, \mathbf{v}^1)\mathbf{g}(i) \leq v^1(i)\mathbf{1}_{m^1(i)},\ i \in \mathbf{S}$

(b) $\mathbf{f}(i)R^2(i) + \beta \mathbf{f}(i)T(i, \mathbf{v}^2) \leq v^2(i)\mathbf{1}^T_{m^2(i)},\ i \in \mathbf{S}$

(c) $(\mathbf{f}, \mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$,

where the superscript $k$ (1 or 2) refers to player 1 or 2,

$$R^k(i) = \left[r^k(i, a^1, a^2)\right]_{a^1=1, a^2=1}^{m^1(i), m^2(i)}$$

is the $k$th players's immediate reward matrix in state $i$, and the matrices $T(i, \mathbf{v}^k)$ are defined in (2.57). Let the objective function of $(GNL_\beta)$ be denoted by $\psi(\mathbf{z})$ for any $\mathbf{z}$ satisfying (a)-(c). We shall refer to the $\mathbf{f}$ and $\mathbf{g}$ in $\mathbf{z}$ as the "strategy part of $\mathbf{z}$".

**Theorem 2.7.2.** *Consider a point* $\hat{\mathbf{z}}^T = \left((\hat{\mathbf{v}}^1)^T, (\hat{\mathbf{v}}^2)^T, \hat{\mathbf{f}}, \hat{\mathbf{g}}^T\right)$. *Then the strategy part* $\left((\hat{\mathbf{f}}, \hat{\mathbf{g}})\right)$ *of* $\hat{\mathbf{z}}^T$ *forms a (Nash) equilibrium point of the general-sum discounted game* $\Gamma_\beta$ *if and only if* $\hat{\mathbf{z}}$ *is the global minimum of* $(GNL_\beta)$ *with* $\psi(\hat{\mathbf{z}}) = 0$.

**Corollary 2.7.1.** *Let* $\hat{\mathbf{z}}$ *be feasible for* $(GNL_\beta)$ *with an objective function value* $\psi(\hat{\mathbf{z}}) = \gamma > 0$. *Then* $(\hat{\mathbf{f}}, \hat{\mathbf{g}})$, *the strategy part of* $\hat{\mathbf{z}}$, *forms an* $\varepsilon$*-equilibrium with* $\varepsilon \leq \frac{\gamma}{1-\beta}$.

This corollary implies that descent algorithms applied to $(GNL_\beta)$ may lead to a good approximation of the equilibrium point, if $\gamma$ is sufficiently small relative to $(1 - \beta)$.

### 2.7.1 Extensions to the Limiting Average Noncooperative Stochastic Games

Similar to the derivation of $(NL_\alpha)$ from $\Gamma_\beta$ in Section 3.7, we will now generalize $(GNL_\beta)$ to the limiting average stochastic games $\Gamma_\alpha$.

The vector of variables will now be

$$\mathbf{z}^T = \left((\mathbf{u}^1)^T, (\mathbf{v}^1)^T, (\mathbf{w}^1)^T, \mathbf{g}^T, (\mathbf{u}^2)^T, (\mathbf{v}^2)^T, (\mathbf{w}^2)^T, \mathbf{f})\right)$$

which is of dimension $(6N + m^1 + m^2)$, and $(\mathbf{f}, \mathbf{g})$ again will be called the strategy part of $\mathbf{z}$. The nonlinear program that will characterize stationary equilibria (if any) of $\Gamma_\alpha$ is given below:

$$\min \sum_{k=1}^{2} \mathbf{1}^T \left[\mathbf{v}^k - P(\mathbf{f}, \mathbf{g})\mathbf{v}^k\right]$$

subject to: $(GNL_\alpha)$

(a) $T(i,\mathbf{v}^1)\mathbf{g}(i) \le v^1(i)\mathbf{1}_{m^1(i)},\ i \in S$

(b) $R^1(i)\mathbf{g}(i) + T(i,\mathbf{u}^1)\mathbf{g}(i) \le (v^1(i)+u^1(i))\mathbf{1}_{m^1(i)},\ i \in S$

(c) $\mathbf{f}(i)T(i,\mathbf{v}^2) \le v^2(i)\mathbf{1}^T_{m^2(i)},\ i \in \mathbf{S}$

(d) $\mathbf{f}(i)R^2(i) + \mathbf{f}(i)T(i,\mathbf{u}^2) \le (v^2(i)+u^2(i))\mathbf{1}^T_{m^2(i)},\ i \in \mathbf{S}$

(e) $\mathbf{r}^k(\mathbf{f},\mathbf{g}) + P(\mathbf{f},\mathbf{g})\mathbf{w}^k = \mathbf{v}^k + \mathbf{w}^k,\ k = 1,2$

(f) $(\mathbf{f},\mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$.

We shall denote the objective function of $(GNL_\alpha)$ by $\theta(\mathbf{z})$ for any $\mathbf{z}$ satisfying (a)-(e).

The following theorem characterizes the stationary equilibria (if any) of limiting average discounted games.

**Theorem 2.7.3.** *The stationary strategy pair $(\hat{\mathbf{f}},\hat{\mathbf{g}})$ forms a (Nash) equilibrium point of $\Gamma_\alpha$ if and only if there exists vectors $\hat{\mathbf{u}}^k,\hat{\mathbf{v}}^k,\hat{\mathbf{w}}^k \in R^N, k = 1,2$ which together with $(\hat{\mathbf{f}},\hat{\mathbf{g}})$ form a point $\hat{\mathbf{z}}$ that is a global minimum of $(GNL_\alpha)$ with $\theta(\hat{\mathbf{z}}) = 0$.*

## 2.8 Shapley's Theorem via Mathematical Programming

In the original proof of his theorem Shapley viewed the solution of the optimality equation as the unique fixed point of a suitably constructed contraction operator. The existence of such a fixed point is a consequence of Banach's fixed point theorem. Another approach involving mathematical programming was used by Vrieze [16]. In this section we will follow this proof based on properties of the nonlinear program

$$\min\left[\mathbf{1}^T\mathbf{v}\right]$$

subject to: $(NL_\beta(1))$

(a) $R(i)\mathbf{g}(i) + \beta T(i,\mathbf{v})g(i) \le v(i)\mathbf{1}_{m^1(i)},\ i \in \mathbf{S}$

(b) $\mathbf{1}^T\mathbf{g}(i) = 1,\ i \in \mathbf{S}$

(c) $\mathbf{g}(i) \ge \mathbf{0},\ i \in \mathbf{S}$

introduced in Section 3.7.

Using the same notation as in Section 3.5 we define

$$\begin{aligned} M_L &:= \min_{i,a^1,a^2}\left\{r(i,a^1,a^2)\right\} \\ M &:= \max_{i,a^1,a^2}\left\{r(i,a^1,a^2)\right\} \end{aligned}$$

and, without loss of generality, assume that for all $a^1 \in A^1(i), a^2 \in A^2(i),\ i \in \mathbf{S}$,

$$M \geq r(i, a^1, a^2) \geq M_L > 0. \tag{2.63}$$

Note that, irrespective of the player's choice of strategies, for all $i \in \mathbf{S}$

$$\frac{M_L}{1-\beta} \leq v_\beta(i, \mathbf{f}, \mathbf{g}) \leq \frac{M}{1-\beta}.$$

Throughout this section we shall regard $(NL_\beta(1))$ as an instance of this generic nonlinear program:

$$\min \theta(\mathbf{z})$$

subject to: $(NL)$

$$h_t(\mathbf{z}) \leq 0,\ t \in I,$$

where the index $t$ runs over all of the constraints (a), (b), and (c) of $(NL_\beta(1))$, suitably ordered.

The logical structure of this section is as follows:

1. We show that $(NL_\beta(1))$ possesses a global minimum $\hat{\mathbf{z}}^T = (\hat{\mathbf{v}}^T, \hat{\mathbf{g}}^T)$ (see Lemma 2.8.3).

2. We invoke a "constraint qualification" of nonlinear programming to claim that corresponding to $\hat{\mathbf{z}}$ there is a nonnegative vector of Lagrange multipliers $\hat{\mu}$ which, together with $\hat{\mathbf{z}}$, satisfy the Kuhn-Tucker conditions (see (2.65)-(2.69)).

3. We prove that $\hat{\mathbf{g}}$ is optimal for player 2, $\hat{\mathbf{v}}$ is the discounted value vector, and that an optimal stationary strategy for player 1 can be constructed from $\hat{\mu}$ (see Theorem 2.8.1).

**Lemma 2.8.1.** *The nonlinear program $(NL_\beta(1))$ is feasible. Let $\overline{\mathbf{z}}^T = (\overline{\mathbf{v}}^T, \overline{\mathbf{g}}^T)$ be any feasible point of $(NL_\beta(1))$. Then componentwise*

$$\overline{\mathbf{v}} \geq \mathbf{v}_\beta(\mathbf{f}, \overline{\mathbf{g}}) \tag{2.64}$$

*for all $\mathbf{f} \in \mathbf{F}_S$*

**Lemma 2.8.2.** *Let $\mathbf{z}^T = (\mathbf{v}^T.\mathbf{g}^T)$ be feasible for $(NL_\beta(1))$. Then*

*(i)* $v(i) \geq \frac{M_L}{1-\beta},\ i \in \mathbf{S}$

*(ii) If $(NL_\beta(1))$ possesses a local minimum $\overline{\mathbf{z}}^T = (\overline{\mathbf{v}}^T, \overline{\mathbf{g}}^T)$, then at $\mathbf{z}$ we have*

$$0 < \theta(\overline{\mathbf{z}}) := \mathbf{1}^T \overline{\mathbf{v}} \geq \left(\frac{M}{1-\beta}\right) N.$$

**Lemma 2.8.3.** *There exists a bounded global minimum* $\overline{\mathbf{z}}^T = (\overline{\mathbf{v}}^T, \overline{\mathbf{g}}^T)$ *of* $(NL_\beta(1))$.

Now that we have established that a global minimum $\hat{\mathbf{z}}$ of $\theta(\mathbf{z})$ over $\Omega$ exists, we will try to exploit the KKT conditions at $\hat{\mathbf{z}}$. Formally, for the generic nonlinear program $(NL)$, these conditions can be written as

(d) $\Delta\theta(\hat{\mathbf{z}}) + \sum_{t\in I} \hat{\mu}_t \nabla h_t(\hat{\mathbf{z}}) = \mathbf{0}$

(e) $\hat{\mu} h_t(\hat{\mathbf{z}}) = 0,\ t \in I$

(f) $\hat{\mu} \geq 0,\ t \in I.$

In order to be able to manipulate the above expressions in a meaningful way, it is convenient to reformulate and partition the set of constraints (a)-(c) of $(NL_\beta(1))$ as follows.

Let $I = I_1 \cup I_2 \cup I_3 \cup I_4$ and adopt the convention that with $\mathbf{z}^T = (\mathbf{v}^T, \mathbf{g}^T)$

(i) for $t = (i, a^1) \in I_1 := \big\{(i, a^1) | a^1 \in A^1(i,\ i \in \mathbf{S})\big\}$

$$\begin{aligned} h_t(\mathbf{z}) &= \sum_{a^2 \in A^2(i)} r(i, a^1, a^2) g(i, a^2) - v(i) \\ &\quad + \beta \sum_{j \in S} \sum_{a^2 \in A^2(i)} p_{ij}(a^1, a^2) g(i, a^2) v(j) \end{aligned}$$

(ii) for $t = i \in I_2 := S$

$$h_t(\mathbf{z}) = \sum_{a^2 \in A^2(i)} g(i, a^2) - 1$$

(iii) for $t = s \in I_3 := S$

$$h_t(\mathbf{z}) = 1 - \sum_{a^2 \in A^2(i)} g(i, a^2)$$

(iv) for $t = (i, a^2) \in I_4 := \big\{(i, a^2) \mid a^2 \in A^2(i,\ i \in \mathbf{S})\big\}$

$$h_t(\mathbf{z}) = -g(i, a^2).$$

With the above convention we can now also easily derive expressions for the partial derivatives making up the gradients $\nabla h_t(\mathbf{z}), t \in I$. In particular,

(v) If $t \in (i, a^1) \in I_1$, then

$$\begin{aligned} \frac{\partial h_t(\mathbf{z})}{\partial g(\bar{i}, \bar{a}^2)} &= \begin{cases} 0 & \text{if } i \neq \bar{i} \\ r(i, a^1, \bar{a}^2) + \beta \sum_{j \in S} p_{ij}(a^1, \bar{a}^2) v(j) & \text{if } i = \bar{i} \end{cases} \\ \frac{\partial h_t(\mathbf{z})}{\partial v(\bar{i})} &= \begin{cases} \beta \sum_{a^2 \in A^2(i)} p_{ij}(a^1, a^2) g(i, a^2), & \text{if } i \neq \bar{i} \\ -1 + \beta \sum_{a^2 \in A^2(i)} p_{ii}(a^1, a^2), & \text{if } i = \bar{i} \end{cases} \end{aligned}$$

(vi) If $t = i \in I_2$, then

$$\begin{aligned} \frac{\partial h_t(\mathbf{z})}{\partial v(\bar{i})} &= 0,\ \bar{i} \in \mathbf{S} \\ &\text{and} \\ \frac{\partial h_t(\overline{\mathbf{z}})}{\partial g(\bar{i}, \overline{a}^2)} &= -\delta(i, \bar{i}),\ \overline{a}^2 \in A^2(\bar{i}), \bar{i} \in \mathbf{S}, \end{aligned}$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta;

(vii) if $t = (i, a^2) \in I_4$, then

$$\begin{aligned} \frac{\partial h_t(\mathbf{z})}{\partial v(\bar{i})} &= 0,\ \bar{i} \in \mathbf{S} \\ \frac{\partial h_t(\mathbf{z})}{\partial g(\bar{i}, \overline{a}^2)} &= -\delta\left(\left(i, a^2\right), \left(\bar{i}, \overline{a}^2\right)\right),\ \overline{a}^2 \in A^2(\bar{i}), \bar{i} \in \mathbf{S}, \end{aligned}$$

where $\delta\left(\left(i, a^2\right), \left(\bar{i}, \overline{a}^2\right)\right) = 1$ only when $i = \bar{i}$ and $a^2 = \overline{a}^2$, and is 0 otherwise.

Returning to the first-order optimality conditions (d)-(f), we now assign convenient labels to the Lagrange multipliers $\hat{\mu}_t$ corresponding to the partition index set $I$. That is, we now define

$$\hat{\mu}_t := \begin{cases} \lambda(i, a^1), & \text{if } t = (i, a^1) \in I_1 \\ \omega(i), & \text{if } t = i \in I_2 \\ w(i), & \text{if } t = i \in I_3 \\ \gamma(i, a^2), & \text{if } t = (i, a^2) \in I_4. \end{cases}$$

With this new notation it is now possible to verify that the KKT condition (d) can be broken up into two parts:

$$\begin{aligned} 1 + \sum_{i \in S} \sum_{a^1 \in A^1(i)} \lambda(i, a^1) \quad & \left[\beta \textstyle\sum_{a^2 \in A^2(i)} p_{ij}(a^1, a^2) \hat{g}(i, a^2)\right] \\ & - \textstyle\sum_{a^1 \in A^1(\bar{i})} \lambda(\bar{i}, a^1) = 0,\ \bar{i} \in \mathbf{S} \end{aligned} \tag{2.65}$$

and

$$\begin{aligned} \sum_{a^1 \in A^1(\bar{i})} r(\bar{i}, a^1, \overline{a}^2) \quad & \lambda(\bar{i}, a^1) + \beta \textstyle\sum_{a^1 \in A^1(\bar{i})} \left[\sum_{j \in S} p_{\bar{i}j}(a^1, a^2) \hat{g}(i, a^2)\right] \\ & -\gamma(\bar{i}, \overline{a}^2) + \omega(\bar{i}) - w(\bar{i}) = 0,\ \overline{a}^2 \in A^2(\bar{i}), \bar{i} \in \mathbf{S}. \end{aligned} \tag{2.66}$$

Similarly, the second KKT condition (e) can be expressed in three parts as (recall the notation introduced at the end of Section 3.1 and the definitions

of $h_t(\overline{\mathbf{z}})$)

$$\lambda(i,a^1)\left[r(i,a^1,\hat{g})-\hat{v}(i)+\beta\sum_{j\in S}p_{ij}(a^1,\hat{\mathbf{g}})\hat{v}(j)\right]=0,$$
$$a^1\in A^1(i), i\in\mathbf{S} \tag{2.67}$$

$$\omega(i)\left[\sum_{a^2\in A^2(i)}\hat{g}(i,a^2)-1\right]=w(i)\left[-1+\sum_{a^2\in A^2(i)}\hat{g}(i,a^2)\right]=0,\quad i\in\mathbf{S} \tag{2.68}$$

$$-\gamma(i,a^2)\hat{g}(i,a^2)=0,\ a^2\in A^2(i),\ i\in\mathbf{S}. \tag{2.69}$$

We will now state the main result of the section. This result invokes a well-known constraint qualification condition which ensures that Lagrange multipliers exist that satisfy (d)-(f). This permits the use of equations (2.65)-(2.69) to help construct optimal strategies in $\Gamma_\beta$.

**Theorem 2.8.1.** *Consider a discounted stochastic game $\Gamma_\beta$ and the associated nonlinear program $(NL_\beta(1))$. Let $\hat{\mathbf{z}}^T=(\hat{\mathbf{v}},\hat{\mathbf{g}}^T)$ be a global minimum of $(NL_\beta(1))$. The following assertions hold:*

*(i) There is a nonnegative vector of Lagrange multipliers that satisfies (d)-(f), (or equivalently, (2.65)-(2.69)) and such that*

$$\hat{f}(i,a^1):=\frac{\lambda(i,a^1)}{\sum_{a^1\in A^1(i)}\lambda(i,a^1)},\ i\in\mathbf{S}$$

*defines a stationary strategy $\hat{\mathbf{f}}$ for player 1.*

*(ii) Stationary strategy pair $(\hat{\mathbf{f}},\hat{\mathbf{g}})$ is optimal in the discounted stochastic game $\Gamma_\beta$ and*

$$\hat{\mathbf{v}}=\mathbf{v}_\beta=\mathbf{v}_\beta(\hat{\mathbf{f}},\hat{\mathbf{g}}).$$

*(iii) Every local minimum of $(NL_\beta(1))$ is also a global minimum.*

**Corollary 2.8.1.** *Let $(\hat{f},\hat{g})\in\mathbf{F}_S\times\mathbf{G}_S$ be a pair of optimal strategies in $\Gamma_\beta$. Then $(\hat{\mathbf{f}}(i),\hat{\mathbf{g}}(i))$ is an optimal strategy pair in the auxiliary matrix game $R(i,\mathbf{v}_\beta)$ for every $i\in S$, and the optimality equations*

$$v_\beta(i)=val\left[R(i,\mathbf{v}_\beta)\right],\ i\in\mathbf{S}$$

*are valid.*

# Chapter 3

# Stochastic Games Have a Value

Mertens and Neyman were the first (in 1979) to prove that a stochastic game always has a value (published in 1981). With the value of a game we mean the following:

**Definition 3.0.1.** A two person zero-sum game is said to have a value $V$ if and only if

$$\sup_{\pi_1}\inf_{\pi_2} v_\alpha(i,\pi_1,\pi_2) = \inf_{\pi_2}\sup_{\pi_1} v_\alpha(i,\pi_1,\pi_2) = V_i,\ \forall i \in S.$$

They presented a simplification of their proof two years later, [12]. In this chapter we will give an overview of these results, but before we can do so, we need to briefly touch upon a couple of subjects they used in their article.

## 3.1 Preliminaries

### 3.1.1 A Few Notes on Summability

In Chapter 3 we defined the $\beta$-discounted Stochastic game as a game without nonzero stopping probabilities but where future payoffs are discounted by a discount factor $\beta$. The performance criterion we used to establish the value was:

$$v_\beta(i,\pi^1,\pi^2) := \sum_{t=0}^{\infty} \beta^t\, \mathbb{E}_{i\pi^1\pi^2}\left[r(S_t, A_t^1, A_t^2)\right]. \tag{3.1}$$

It turns out to be convenient in the definition of the discounted reward for a pair of strategies $(\pi^1,\pi^2)$ to introduce the normalization factor $(1-\beta)$:

$$v_\beta(i,\pi^1,\pi^2) := (1-\beta)\sum_{t=0}^{\infty} \beta^t\, \mathbb{E}_{i\pi^1\pi^2}\left[r(S_t, A_t^1, A_t^2)\right].$$

By this normalization factor the discounted reward can be seen to be a convex combination of the quantities $\mathbb{E}_{i\pi^1\pi^2}\left[r(S_t, A_t^1, A_t^2)\right]$ with $t \in \mathbb{N}_0$, and so

$$|v_\beta(i, \pi^1, \pi^2)| \leq |\max_{i,a^1,a^2} |r(i, a^1, a^2)|.$$

Now consider a sequence of real numbers $\{a_n\}_{n=0}^\infty$. Let $\sigma_N$ be defined by

$$\sigma_N := \frac{s_N}{N+1} = \frac{\sum_{n=0}^N a_n}{N+1}.$$

A sequence $\{a_n\}_{n=1}^\infty$ is said to be *C(esaro)-summable* to the limit $s$ if

$$\lim_{n\to\infty} \sigma_N = s.$$

A sequence $\{a_n\}_{n=1}^\infty$ is said to be *A(bel)-summable* to $a$ if

$$\lim_{\beta\to 1^-} (1-\beta) \sum_{n=0}^\infty \beta^n a_n = a.$$

Two classical results in the theory of summability are the following theorems.

**Theorem 3.1.1.**
*If the sequence of real numbers $\{a_n\}_{n=0}^\infty$ is C-summable to $s$, then it is also A-summable to $s$.*

*Proof.* First we show that $\sum_{n=0}^\infty (n+1)\sigma_n\beta^n$ exists for $0 \leq \beta < 1$. We know $\sigma_n$ converges, so there is a $K$ such that $|\sigma_k| < K$. We now have:

$$\begin{aligned}
\sum_{n=0}^\infty (n+1)\sigma_n\beta^n &\leq K \sum_{n=0}^\infty (n+1)\beta^n \\
&\leq K \frac{d}{d\beta} \sum_{n=0}^\infty \beta^{n+1} \\
&\leq K \frac{d}{d\beta} \left(\frac{\beta}{1-\beta}\right) \\
&\leq \frac{K}{(1-\beta)^2}
\end{aligned}$$

so $\sum_{n=0}^\infty (n+1)\sigma_n\beta^n$ exists.
Next we prove that $(1-\beta)\sum_{n=0}^\infty s_n\beta^n = \sum_{n=0}^\infty a_n\beta^n$.

$$\begin{aligned}
(1-\beta)\sum_{n=0}^\infty s_n\beta^n &= (1-\beta)\left[a_0 + (a_0+a_1)\beta + (a_0+a_1+a_2)\beta^2 + \dots\right] \\
&= \left[a_0 + (a_0+a_1)\beta + (a_0+a_1+a_2)\beta^2 + \dots\right] + \left[-a_0\beta - (a_0+a_1)\beta^2 - \dots\right] \\
&= a_0 + a_1\beta + a_2\beta^2 + \dots \qquad (3.2)
\end{aligned}$$

Since $\sigma_n := \frac{s_n}{n+1}$, we also have that $(1-\beta)\sum_{n=0}^{\infty}(n+1)\sigma_n\beta^n = \sum_{n=0}^{\infty} s_n\beta^n$. Combining these two equations we get:

$$\sum_{n=0}^{\infty} a_n\beta^n = (1-\beta)\sum_{n=0}^{\infty}\sigma_n(n+1)\beta^n. \tag{3.3}$$

Furthermore we have, for $0 < \beta < 1$ that

$$\cdot\frac{1}{(1-\beta)^2} = \sum_{n=0}^{\infty}(n+1)\beta^n$$

which can be written as

$$1 = (1-\beta)^2\sum_{n=0}^{\infty}(n+1)\beta^n. \tag{3.4}$$

Multiplying this with $s$ and subtracting the result from (3.3), we obtain

$$\begin{aligned}(1-\beta)\sum_{n=0}^{\infty} a_n\beta^n - s &= (1-\beta)^2\sum_{n=0}^{\infty}(n+1)(\sigma_n - s)\beta^n \\ &= (1-\beta)^2\sum_{n=0}^{N}(n+1)(\sigma_n - s)\beta^n \\ &\quad +(1-\beta)^2\sum_{n=N+1}^{\infty}(n+1)(\sigma_n - s)\beta^n.\end{aligned}$$

We can now, for any $\varepsilon > 0$, choose $N$ large enough such that

$$|\sigma_{n+1} - s| < \varepsilon, \text{ for } n > N.$$

This means the second term on the right-hand side becomes

$$(1-\beta)^2\sum_{n=N+1}^{\infty}(n+1)(\sigma_n - s)\beta^n \le \varepsilon(1-\beta)^2\sum_{n=0}^{\infty}(n+1)\beta^n = \varepsilon.$$

In the first term, take $M = \max_{0\le n\le N}|\sigma_n - s|$. We now get

$$(1-\beta)^2\sum_{n=0}^{N}(n+1)(\sigma_n - s)\beta^n \le M(1-\beta)^2\sum_{n=0}^{N}(n+1)\beta^n \le \varepsilon$$

for $\beta$ close enough to 1. This proves the theorem. □

The reverse is not always true. Take for example the series $\{(-1)^n n\}_{n=0}^{\infty}$. The sequence of partial sums is $s_n = 0, -1, 1, -2, 2, -3, 3, ...$, and $\sigma_n = \frac{1}{n+1}s_n = 0, -\frac{1}{2}, \frac{1}{3}, -\frac{2}{4}, \frac{2}{5}\ldots$. We can also write this sequence as $s_n = \frac{\frac{1}{2}n}{n+1}$ for

$n$ is even, and $s_n = -\frac{\frac{1}{2}(n+1)}{n+1} = -\frac{1}{2}$ for $n$ is odd. We now get $\liminf_{n\to\infty} \sigma_n = -\frac{1}{2}$ and $\limsup_{n\to\infty} \sigma_n = \frac{1}{2}$. In order to show that the series is $A$-summable, we need to show that $\lim_{\beta\uparrow 1}(1-\beta)\sum_{k=0}^{\infty}(-1)^k k\beta^k$ exists.

$$\begin{aligned}
\lim_{\beta\uparrow 1}(1-\beta)\sum_{k=0}^{\infty}(-1)^k k\beta^k &= \lim_{\beta\uparrow 1}(1-\beta)\beta\sum_{k=0}^{\infty}(-1)^k \frac{d}{d\beta}\beta^k \\
&= \lim_{\beta\uparrow 1}(1-\beta)\beta\frac{d}{d\beta}\sum_{k=0}^{\infty}(-1)^k\beta^k \\
&= \lim_{\beta\uparrow 1}(1-\beta)\beta\frac{d}{d\beta}\left[\frac{1}{1+\beta}\right] \\
&= \lim_{\beta\uparrow 1}(1-\beta)\frac{-\beta}{(1+\beta)^2} = 0.
\end{aligned}$$

This means the series is $A$-summable, but not $C$-summable.

**Theorem 3.1.2.**

$$s = \liminf_{N\to\infty}\sigma_N \le \liminf_{\beta\to 1^-} f(\beta) \le \limsup_{\beta\to 1^-} f(\beta) \le \limsup_{N\to\infty}\sigma_N = S,$$

*where* $f(\beta) := (1-\beta)\sum_{n=0}^{\infty}\beta^n a_n$.

*Proof.* We only need to prove the first and third inequality, since the second one is trivial. The proof resembles the one for theorem 3.1.1. First we prove the first inequality.

$$\begin{aligned}
(1-\beta)\sum_{n=0}^{\infty}a_n\beta^n - s &= (1-\beta)^2\sum_{n=0}^{\infty}\{s_n-(n+1)s\}\beta^n \\
&= (1-\beta)^2\sum_{n=0}^{N}\{s_n-(n+1)s\}\beta^n + (1-\beta)^2\sum_{n=N+1}^{\infty}\{s_n-(n+1)s\}\beta^n.
\end{aligned}$$

For the first term on the right hand side we have:

$$(1-\beta)^2\sum_{n=0}^{N}\{s_n-(n+1)s\}\beta^n \ge (1-\beta^2)\min_{0\le n\le N}\left\{\frac{1}{n+1}s_n - s\right\}\sum_{n=0}^{N}(n+1)\beta^n.$$

Taking $\min_{0\le n\le N}\left\{\frac{1}{n+1}s_n - s\right\} \ge -M$, with $M>0$, we have:

$$\begin{aligned}
&\ge (1-\beta)^2(-M)\sum_{n=0}^{N}(n+1)\beta^n \\
&\ge -\varepsilon \text{ for } \beta \text{ close enough to } 1.
\end{aligned}$$

Since $s = \liminf_{N\to\infty} \frac{1}{N+1} s_N \leq \frac{1}{N+1} s_N + \varepsilon$ for $N$ large enough, we have for the second term on the right hand side:

$$(1-\beta)^2 \sum_{n=N+1}^{\infty} \{s_n - (n+1)s\}\beta^n \geq (1-\beta)^2 \sum_{n=N+1}^{\infty} (-\varepsilon)(n+1)\beta^n$$
$$\geq (1-\beta)^2 \sum_{n=0}^{\infty} (-\varepsilon)(n+1)\beta^n = -\varepsilon.$$

Putting this together again we get

$$(1-\beta)\sum_{n=0}^{\infty} a_n\beta^n - s \geq -2\varepsilon.$$

The proof for the third inequality is almost the same.

$$\begin{aligned}(1-\beta)\sum_{n=0}^{\infty} a_n\beta^n - S &= (1-\beta)^2 \sum_{n=0}^{\infty} \{s_n - (n+1)S\}\beta^n \\ &= (1-\beta)^2 \sum_{n=0}^{N} \{s_n - (n+1)S\}\beta^n + (1-\beta)^2 \sum_{n=N+1}^{\infty} \{s_n - (n+1)S\}\beta^n.\end{aligned}$$

For the first term on the right hand side we have:

$$(1-\beta)^2 \sum_{n=0}^{N} \{s_n - (n+1)S\}\beta^n \leq (1-\beta^2) \max_{0\leq n\leq N} \left\{ \frac{1}{n+1} s_n - S \right\} \sum_{n=0}^{N} (n+1)\beta^n.$$

Taking $\max_{0\leq n\leq N} \left\{ \frac{1}{n+1} s_n - S \right\} \leq M$, with $M > 0$, we have:

$$\leq (1-\beta)^2 M \sum_{n=0}^{N} (n+1)\beta^n$$
$$\leq \varepsilon \text{ for } \beta \text{ close enough to 1.}$$

Since $S = \limsup_{N\to\infty} \frac{1}{N+1} s_N \geq \frac{1}{N+1} s_N - \varepsilon$ for $N$ large enough, we have for the second term on the right hand side:

$$(1-\beta)^2 \sum_{n=N+1}^{\infty} \{s_n - (n+1)s\}\beta^n \leq (1-\beta)^2 \sum_{n=N+1}^{\infty} (\varepsilon)(n+1)\beta^n$$
$$\leq (1-\beta)^2 \sum_{n=0}^{\infty} (-\varepsilon)(n+1)\beta^n = -\varepsilon.$$

Putting this together again we get

$$(1-\beta)\sum_{n=0}^{\infty} a_n\beta^n - S \leq 2\varepsilon.$$

□

The inequalities in this theorem can be strict. To see that $\liminf_{N\to\infty}\sigma_N < \liminf_{\beta\to 1^-} f(\beta)$ and $\limsup_{\beta\to 1^-} f(\beta) < \limsup_{N\to\infty}\sigma_N$ are possible, we can use the previous example. Since $\lim_{\beta\to 1^-} f(\beta) = 0$, both inequalities are satisfied.

For $\liminf_{\beta\to 1^-} f(\beta) < \limsup_{\beta\to 1^-} f(\beta)$, we will need theorem 3.1.3. According to this theorem, if we can find a bounded sequence for which $\liminf_{n\to\infty}\sigma_n < \limsup_{n\to\infty}\sigma_n$, $\liminf_{\beta\uparrow 1^-} f(\beta)$ cannot be the same as $\limsup_{\beta\uparrow 1^-} f(\beta)$. Consider the sequence $\{a_n\}_{n=0}^{\infty}$ as follows:

$$0,0,1,1,0,0,0,0,1,1,1,1,1,1,1,1,\ldots,$$

where each block of zeros or ones is of the same length as the sum of the lengths of the preceding blocks.

First we consider a block of ones and zeros (resp. a block of zeros and ones) as one block.

$$0,0,[1,1,0,0,0,0],[1,1,1,\ldots$$

Except for the first block, each of these blocks consists of 1s for $1/3$ (if block $k$ consists of $\frac{1}{2}N$ of ones and $N$ zeros, the following block will consist of $2N$ ones and $4N$ zeros). Furthermore, the length of a sequence of $k$ blocks is $2^{2k+1}-2$, with total length $2^{2k+1}-1$. This gives us

$$\lim_{k\to\infty}\sigma_{2^{2k+1}-1} = \lim_{k\to\infty}\frac{\frac{1}{3}\left(2^{2k+1}-2\right)}{2^{2k+1}-1} = \frac{1}{3}$$

In the same way we can take a block of zeros and ones:

$$0,[0,1,1],[0,0,0,0,1,1,1,1,1,1,1,1],\ldots$$

Each of these blocks consists of ones for $2/3$. This gives us:

$$\lim_{k\to\infty}\sigma_{2^{2k}} = \lim_{k\to\infty}\frac{\frac{2}{3}\left(2^{2k}-1\right)}{2^{2k}} = \frac{2}{3}$$

We now have $\liminf_{n\to\infty}\sigma_n < \limsup_{n\to\infty}\sigma_n$, and so $\liminf_{\beta\uparrow 1^-} f(\beta) < \limsup_{\beta\uparrow 1^-} f(\beta)$.

The relation between $C$-summability and $A$-summability suggests that it might be possible to study the limiting average stochastic games by studying the limiting behaviour of $\lim_{\beta\to 1^-}$, because $g(\pi^1,\pi^2) = \lim_{\beta\to 1^-}(1-\beta)v_\beta(\pi^1,\pi^2)$, with $\pi^1,\pi^2 \in F_S$. This is also suggested by the following result by Hardy and Littlewood [9] which is often used in the theories of MDPs and stochastic games.

**Theorem 3.1.3.** *(Hardy and Littlewood)*
*Let $\{a_n\}_{n=0}^{\infty}$ be a bounded sequence of real numbers and $\lim_{\beta\to 1^-} f(\beta) = a$. Then $\lim_{N\to\infty}\sigma_N = a$.*

*Proof.* $(1-\beta)\sum_{n=0}^{\infty}\beta^n = 1$ for all $|\beta|<1$ implies that $\lim_{\beta\to 1^-} f(\beta) = A$ if and only if

$$\lim_{\beta\to 1^-}(1-\beta)\sum_{n=0}^{\infty}\beta^n(a_n - A) = 0.$$

Furthermore, $\lim_{n\to\infty}\sigma_n = \sigma$ if and only if

$$\lim_{n\to\infty}\frac{[(a_0-\sigma)+\cdots+(a_n-\sigma)]}{n+1} = 0.$$

This means it is sufficient to show that

$$\lim_{\beta\to 1^-} f(\beta) = 0 \text{ implies that } \lim_{n\to\infty}\sigma_n = 0.$$

Consider

$$w_n = \begin{cases} \sigma_n - \sigma_{n-1} = \frac{na_n - s_{n-1}}{n(n+1)}, & \text{if } n \geq 1 \\ \sigma_0 = a_0, & \text{if } n = 0. \end{cases}$$

Since $\{a_n\}_{n=0}^{\infty}$ is bounded, $na_n = \mathcal{O}(n)$, and

$$s_{n-1} = \mathcal{O}(n), \text{ then } w_n = \mathcal{O}(1/n^2)\mathcal{O}(n) = \mathcal{O}(1/n).$$

Define an auxiliary function

$$g(\beta) := \sum_{n=0}^{\infty} w_n\beta^n = \sum_{n=0}^{\infty}\frac{a_n\beta^n}{n+1} - \sum_{n=1}^{\infty}\frac{s_{n-1}\beta^n}{n(n+1)}.$$

Thus

$$\begin{aligned} \beta g(\beta) &= \sum_{n=0}^{\infty}\frac{a_n\beta^{n+1}}{n+1} - \sum_{n=0}^{\infty}\frac{s_n\beta^{n+2}}{(n+1)(n+2)} \\ &= \int_0^{\beta}\left(\sum_{n=0}^{\infty}a_n y^n\right)dy - \sum_{n=0}^{\infty}s_n\int_0^{\beta}\left\{\int_0^t y^n dy\right\}dt \\ &= \int_0^{\beta}\left(\sum_{n=0}^{\infty}a_n y^n\right)dy - \int_0^{\beta}\left\{\int_0^t\left(\sum_{n=0}^{\infty}s_n y^n\right)dy\right\}dt. \end{aligned} \tag{ah1}$$

We will now use (3.2):

$$\sum_{n=0}^{\infty}a_n\beta^n = (1-\beta)\sum_{n=0}^{\infty}s_n\beta^n.$$

Together with $\frac{f(\beta)}{1-\beta} = \sum_{n=0}^{\infty}a_n\beta^n$, we get

$$\sum_{n=0}^{\infty}s_n\beta^n = \frac{f(\beta)}{(1-\beta)^2}.$$

$$(ah2)$$

Using the hypothesis that $\lim_{\beta\to 1^-} f(\beta) = 0$ we obtain from (ah1) and (ah2) that

$$\beta g(\beta) = \int_0^\beta \frac{f(y)}{(1-y)}dy - \int_0^\beta \left[\int_0^t \frac{f(y)}{(1-y)^2}dy\right] dt. \tag{3.5}$$

We now integrate the second term by parts ($\int_0^\beta F(t)dt = F(t)t|_0^\beta - \int_0^\beta tF'(t)dt$ with $F(t) = \int_0^t \frac{f(y)}{(1-y)^2}dy$):

$$\begin{aligned}\int_0^\beta \left[\int_0^t \frac{f(y)}{(1-y)^2}dy\right] dt &= t\int_0^t \frac{f(y)}{(1-y)^2}dy\Big|_0^\beta - \int_0^\beta y\frac{f(y)}{(1-y)^2}dy \\ &= \int_0^\beta \frac{(\beta-y)f(y)}{(1-y)^2}dy.\end{aligned}$$

This gives us

$$\begin{aligned}\beta g(\beta) &= \int_0^\beta \frac{f(y)}{(1-y)}dy - \int_0^\beta \frac{(\beta-y)f(y)}{(1-y)^2}dy \\ &= (1-\beta)\int_0^\beta \frac{f(y)}{(1-y)^2}dy \\ &= (1-\beta)\left[\int_0^{\beta-\varepsilon} \frac{f(y)}{(1-y)^2}dy + \int_{\beta-\varepsilon}^\beta \frac{f(y)}{(1-y)^2}dy\right] \\ &= (1-\beta)\left[C\cdot\frac{1}{1-y}\Big|_0^{\beta-\varepsilon} + o(1)\int_{\beta-\varepsilon}^\beta \frac{dy}{(1-y)^2}\right] \\ &= (1-\beta)\left[C' + o(1)\frac{1}{1-y}\Big|_{\beta-\varepsilon}^\beta\right] \\ &= (1-\beta)\left[C' + o(1)\frac{1}{1-\beta} + o(1)\frac{1}{1-\beta+\varepsilon}\right] \\ &= (1-\beta)\left[\frac{1}{1-\beta}o(1)\right] \to 0 \text{ as } \beta\to 1^-.\end{aligned}$$

We have now proved that $g(\beta) = \sum_{n=0}^\infty w_n\beta^n \to 0$ as $\beta \to 1^-$. Since $w_n = \mathcal{O}(1/n)$, we can now use Theorem 3.1.4 (see below) to obtain

$$\sum_{n=0}^\infty w_n = 0.$$

Since $\sigma_n = \sum_{k=0}^n w_k$ we have proved that $\lim_{n\to\infty}\sigma_n = 0$, which concludes the proof. □

**Theorem 3.1.4.** *(Littlewood)*
*Let* $\{u_n\}_{n=0}^{\infty}$ *be a sequence such that* $u_n = \mathcal{O}(1/n)$ *and* $\lim_{\beta\to 1^-} \sum_{n=0}^{\infty} \beta^n u_n$ *exists. Then the series* $\sum_{n=0}^{\infty} u_n$ *is convergent, and*

$$\sum_{n=0}^{\infty} u_n = \lim_{\beta\to 1^-} \beta^n u_n.$$

### 3.1.2 Puiseux Series

Before continuing to research the idea that we can study average limiting stochastic games by studying the limiting behaviour of $\beta$-discounted stochastic games, we will first take a brief look at Puiseux series. These were introduced to the study of stochastic games by Bewley and Kohlberg [1], whose results we will summarize in the next section.

Puiseux series in a variable $x$ are often mathematically defined to be Laurent series in another variable, say $y$, where $y = x^{1/d}$, for a fixed positive integer $d$; this $d$ is usually fixed for all the series under consideration. Formally, for a positive integer $M$, let

$$F_M := \left\{ \sum_{k=K}^{\infty} c_k x^{k/M} \right\}$$

with $K$ an integer, $c_k \in \mathbb{R}$, and such that the series $\sum_{k=K}^{\infty} c_k x^{k/M}$ converges for all sufficiently small but positive real numbers $x$.

Thus the members of $F_M$ are power series in $x^{1/M}$. Addition and multiplication in $F_M$ are defined in the same way as in the case of power series. The ordering on $F_M$ reflects the notion that $x$ represents an arbitrary small but positive real number. To be more specific:

$$\sum_{k=K_1}^{\infty} c_k x^{k/M} + \sum_{k=K_2}^{\infty} d_k x^{k/M} := \sum_{k=\min(K_1,K_2)}^{\infty} (c_k + d_k) x^{k/M}.$$

If $K_1 > K_2$, then we define $c_k = 0$ for $k = K_2, \ldots, K_1 - 1$ in the summation of the right-hand side expression:

$$\left( \sum_{k=K_1}^{\infty} c_k x^{k/M} \right) \left( \sum_{k=K_2}^{\infty} d_k x^{k/M} \right) := \sum_{k=K_1+K_2}^{\infty} \left( \sum_{i+j=k} c_i d_j \right) x^{k/M}.$$

Further, $\sum_{k=K} c_k x^{k/M} > 0$ if and only if $c_{k^*} > 0$, where $k^*$ is the smallest integer $k$ such that $c_k \neq 0$. One can verify that $F_M$ is an ordered field. Let $F := \cup_{M=1}^{\infty} F_M$. Then $F$ is also an ordered field, and $F$ is called the field of *real Puiseux series.*

If

$$v(x) = \sum_{k=K}^{\infty} c_k x^{k/M} \in F$$

then $\emptyset_\beta(v)$, called the *valuation* of $v$ at $\beta$, for $\beta \in (0,1)$ will denote the sum $\sum_{k=K}^{\infty} c_k(1-\beta)^{k/M}$.
Some properties of Puiseux series are:

> If $v \in F$, then $\emptyset_\beta(v)$ is well defined for $\beta$ sufficiently close to 1, and $v > 0$ if and only if $\emptyset_\beta(v) > 0$ for all $\beta$ sufficiently close to 1.

Furthermore, when $B$ is a matrix $[b_{ij}]$ with entries in $F$, then $\emptyset_\beta(B)$ denotes a matrix with entries $[\emptyset_\beta(b_{ij})]$ in $\mathbb{R}$. $F^N$ will denote the $N$-fold Cartesian product of $F$.
We are now able to define the limit discount equation.

**Definition 3.1.1.**
The set of $N$ equations (one for each state $i \in S$) in the variable $\mathbf{v} \in F^N$

$$v_\beta(i) = \text{val}\left[\beta r(i, a^1, a^2) + (1-\beta)\sum_{j=1}^{N} p_{ij}(a^1, a^2)v(j)\right]_{a^1=1,a^2=1}^{m^1(i),m^2(i)} \tag{3.6}$$

where $\beta \in (0,1)$, is called the limit discount equation.

### 3.1.3 Bewley and Kohlberg results

The main result of [1] is captured in the following theorem:

**Theorem 3.1.5.** *The limit discount equation has a unique solution*

$$\mathbf{v}^*_\beta = \sum_{k=0}^{\infty} \boldsymbol{c}_k(1-\beta)^{k/M} \in F^N \text{ with } \boldsymbol{c}_k \in \mathbb{R}^N \tag{3.7}$$

*for $\beta$ close enough to* 1.

This result can be interpreted as saying that in a left-sided neighbourhood of $\beta = 1$ the solutions to the system of equations are given by Puiseux series in the variable $(1-\beta)$ over the field of real numbers. It can be shown that a positive integer $M$ and a number $\beta_0 \in (0,1)$ exist such that for all $\beta \geq \beta_0$ and each $i = 1, \ldots, N$,

$$v_i(\beta) = \sum_{k=k_i}^{\infty} c_{ik}(1-\beta)^{k/M}$$

where $c_{ik}$ are real numbers and $k_i$ is an integer. The expression on the right-hand side is a Puiseux series. It is now clear that

$$\lim_{\beta \uparrow 1} v_i(\beta) = c_{i0}$$

which gives us a candidate for $(v_\alpha(\beta))_i = c_{i0}$.

The logical structure of the proof by Bewley and Kohlberg is outlined below:

1. Shapley's theorem shows that the value vector of the stochastic game $\Gamma_\beta$ is the solution of the fixed point equation $\mathbf{v}(\beta) = T_\beta(\mathbf{v}(\beta))$. This theorem can be viewed as a "valid elementary sentence" over the field of real numbers.

2. A theorem in formal logic know as "Tarski's principle" (Tarski 1951) says that, "An elementary sentence that is valid over one real closed field is valid over every real closed field." Note that an ordered field is by definition *real closed* if no proper algebraic extension is ordered.

3. The field of Puiseux series over the real numbers is real closed.

4. Therefore, by Tarski's principle, the fixed point equation also can be viewed as an elementary sentence over the real closed field of Puiseux series, thereby completing the proof.

## 3.2 Main Result

We can now begin to discuss the main result from the article by Neyman and Mertens. In addition to the notation used in the previous chapters we will use:

$$\lambda := 1 - \beta \text{ whenever } \beta \text{ denotes a discount factor.}$$

We will define a sequence $\{\lambda_t\}_{t=0}^{\infty}$ so that $\lambda_t$ is a function of the past history. The main result is captured in the following theorem:

**Theorem 3.2.1.** *For every stochastic game and for every $\varepsilon > 0$ there exists a strategy $\pi_\varepsilon^1$ of player 1 and a number $N > 0$ such that for every $n \in \{N, N+1, \ldots, \infty\}$, for every strategy $\pi^2$ of player 2, and for every initial state $i \in S$*

$$v_n(i) = \frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E}_{i\pi_\varepsilon^1\pi^2} \left[ r\left(S_t, A_t^1, A_t^2\right) \right] \geq \lim_{\beta \uparrow 1} \mathbf{v}_\beta(i) - \varepsilon.$$

If $\pi_\varepsilon^1$ is an optimal strategy for player 1 with respect to a $\beta_t$-discounted game, with $\beta_t$ still to be specified. Since $\pi_\varepsilon^1$ is optimal, we have $\mathbf{v}_\beta \leq \mathbf{v}_\beta(\pi_\varepsilon^1, \pi^2)\ \forall \pi^2$. Thus we have, with state $i_t$ and history $h_t$,

$$\begin{aligned} v_{\beta_t}(i_t) \quad \leq \quad & (1-\beta_t)\, r(i_t, \pi_\varepsilon^1(h_t), \pi^2(h_t)) \\ & + \beta_t \sum_{j=1}^{N} p_{i_t j}(\pi_\varepsilon^1(h_t), \pi^2(h_t)) v_{\beta_t}(j). \end{aligned} \tag{3.8}$$

Another way of putting this is

$$\mathbb{E}_{i,\pi^1_\varepsilon,\pi^2}\left[(1-\beta_t)\, r(S_t, A^1_t, A^2_t) + \beta_t v_{\beta_t}(S_{t+1}) - v_{\beta_t}(S_t) \mid h_t\right] \geq 0 \tag{3.9}$$

which can also be written as

$$\begin{aligned}\mathbb{E}_{i,\pi^1_\varepsilon,\pi^2}\big[(1-\beta_t)\,(r\left(S_t, A^1_t, A^2_t\right) - v_{\beta_t}(S_{t+1})) \\ +(v_{\beta_t}(S_{t+1}) - v_{\beta_t}(S_t)) \mid h_t\big] \geq 0.\end{aligned} \tag{3.10}$$

For the sake of notational convenience we will write $\mathbb{E} := \mathbb{E}_{i,\pi^1_\varepsilon,\pi^2}$ from now on. Let $K > M$, in order to satisfy

$$||\mathbf{v}_{\beta_t} - \mathbf{v}_{\beta_{t+1}}|| \leq |\lambda_t^{1/K} - \lambda_{t+1}^{1/K}|$$

for $\lambda_t$ and $\lambda_{t+1}$ small enough, which will be used in Lemma 3.2.2. To see that this inequality holds, write

$$f(\lambda_t) := \mathbf{v}_{\beta_t} = \sum_{k=0}^{\infty} c_k \lambda_t^{k/M}.$$

We now have, using the mean value theorem:

$$\begin{aligned}\left|\mathbf{v}_{\beta_t} - \mathbf{v}_{\beta_{t+1}}\right| &\leq |f'(\lambda_t)| \cdot \left|\lambda_t^{1/M} - \lambda_{t+1}^{1/M}\right| \\ &\leq C\left|\lambda_t^{1/M} - \lambda_{t+1}^{1/M}\right| \\ &= C\left|\left(\lambda_t^{1/NM} - \lambda_{t+1}^{1/NM}\right)\left(\lambda_t^{\frac{N-1}{NM}} + \lambda_t^{\frac{N-2}{NM}}\lambda_{t+1}^{\frac{1}{NM}} + \cdots + \lambda_{t+1}^{\frac{N-1}{NM}}\right)\right| \\ &\leq \left|\lambda_t^{1/K} - \lambda_{t+1}^{1/K}\right|.\end{aligned} \tag{3.11}$$

Here, the last inequality is due to the fact that we can ensure that

$$\left(\lambda_t^{\frac{N-1}{NM}} + \lambda_t^{\frac{N-2}{NM}}\lambda_{t+1}^{\frac{1}{NM}} + \cdots + \lambda_{t+1}^{\frac{N-1}{NM}}\right) \leq N \max\{\lambda_t, \lambda_{t+1}\}^{\frac{1}{M}} \leq \frac{1}{C}$$

by taking $\lambda_t$ small enough, and taking $K = MN$.
Define

$$k(\lambda) := \lambda^{-\frac{K-1}{K}} \quad \text{or} \quad \lambda(k) := k^{-\frac{K}{K-1}}. \tag{3.12}$$

Observe that $\lim_{\lambda\to 0} k(\lambda) = \infty$. Furthermore, define the 1-1 correspondence

$$y(\lambda) := (K-1)\lambda^{1/K} = (K-1)k(\lambda)^{-\frac{1}{K-1}}. \tag{3.13}$$

From (3.13) it follows that

$$\frac{dy}{dk} = (K-1)\frac{-1}{K-1}k^{-\frac{K}{K-1}} = -\lambda(k). \tag{3.14}$$

Since the process is history dependent, we have to specify a sequence of discount factors $\beta_0, \beta_1, \ldots$. We will do this recursively. Let $L, k_0 \in \mathbb{R}^+$ be arbitrary, but large enough to satisfy certain requirements which will be specified later (for $L$, see Lemma (3.2.1)(ii)), and assume that $k_0 \geq L$

Define for $t = 0, 1, 2, \ldots$:

$$\begin{aligned} k_{t+1} &:= \max\left\{L, k_t + r(i_t, a_t^1, a_t^2) - v_{\beta_t}(i_{t+1}) + 4\varepsilon\right\} && (3.15)\\ \lambda_{t+1} &:= \lambda(k_{t+1}) && (3.16)\\ \beta_{t+1} &= 1 - \lambda_{t+1}. && (3.17) \end{aligned}$$

Furthermore, in the following we will use

$$D = \max_{i,a^1,a^2} |r(i, a^1, a^2). \tag{3.18}$$

We will now prove some useful properties of $k, \lambda$ and $y$ which will be needed later on.

**Lemma 3.2.1.**
*For any realization it holds that:*

*(i)* $|k_{t+1} - k_t| \leq 6D$

*(ii)* $|\lambda_{t+1} - \lambda_t| \leq \frac{\varepsilon\lambda_t}{6D} \leq \varepsilon\lambda_t$

*(iii)* $y_t - y_{t+1} \geq \lambda_t(k_{t+1} - k_t) - \varepsilon\lambda_t$.

*Proof.*

(i) The definition in (3.15) implies that $k_t \geq L$ for all $t$. Taking $\varepsilon \leq D$ we get

$$|k_{t+1} - k_t| \leq |r(i_t, a_t^1, a_t^2) - v_{\beta_{t+1}}(i_{t+1}) + 4\varepsilon| \leq 6D$$

(ii)

$$\begin{aligned} |\lambda_{t+1} - \lambda_t| &= \lambda_t \left|\frac{\lambda_{t+1}}{\lambda_t} - 1\right| \\ &= \lambda_t \left|\left(\frac{k_t}{k_{t+1}}\right)^{\frac{K}{K-1}} - 1\right| \\ &= \lambda_t \left|\left(\frac{k_t}{k_t + D_t}\right)^{\frac{K}{K-1}} - 1\right| \end{aligned}$$

for some $D_t \leq 6D$ in view of (i). Since

$$\lim_{k_t \to \infty} \left(\frac{k_t}{k_t + D_t}\right)^{\frac{K}{K-1}} = 1$$

inequality (ii) follows from (3.15) when we choose $L$ large enough.

(iii) $y$ is a decreasing convex function of $k$. Using the mean value theorem and (3.14) we obtain:

$$y_t - y_{t+1} \geq (k_t - k_{t+1}) \frac{dy}{dk}\bigg|_{k=k_{t+1}} = (k_{t+1} - k_t)\lambda_{t+1}.$$

Using (i) and (ii) gives the desired expression.

□

Since $\lambda_t$ and $\beta_t$ change slowly, we suspect that $\mathbf{v}_{\beta_t}$ changes slowly as well. The following lemma confirms this idea.

**Lemma 3.2.2.** *For any realization it holds that*

$$\|\mathbf{v}_{\beta_t} - \mathbf{v}_{\beta_{t+1}}\| \leq \varepsilon\lambda_t.$$

*Proof.*
When we choose $K$ and $L$ large enough, we have for all $t$ (see (3.11)):

$$|\ \mathbf{v}_{\beta_t} - \mathbf{v}_{\beta_{t+1}}\| = \left\| \sum_{k=1}^{\infty} \mathbf{c}_k \lambda_t^{k/M} - \sum_{k=1}^{\infty} \mathbf{c}_k \lambda_{t+1}^{k/M} \right\| \leq \left| \lambda_t^{1/K} - \lambda_{t+1}^{1/K} \right|.$$

The proof will now be divided in two cases:

(i) $\lambda_t > \lambda_{t+1}$:
When we use the mean value theorem on the concave function $\lambda^{1/K}$, we get:

$$\begin{aligned} 0 < \lambda_t^{1/K} - \lambda_{t+1}^{1/K} \leq \frac{1}{K} \lambda_{t+1}^{-\frac{K-1}{K}} (\lambda_t - \lambda_{t+1}) &= \frac{k_{t+1}}{K} \left(1 - \frac{\lambda_{t+1}\lambda_t}{\lambda_t}\right) \lambda_t \\ &= \frac{1}{K} \frac{\left(k_{t+1}^{\frac{K}{K-1}} - k_t^{\frac{K}{K-1}}\right)}{k_{t+1}^{\frac{1}{K-1}}} \lambda_t. \end{aligned} \tag{3.19}$$

When we take $K$ to be even, we get:

$$\begin{aligned} &k_{t+1}^{\frac{K}{K-1}} - k_t^{\frac{K}{K-1}} \\ &= \left(k_{t+1}^{\frac{1}{K-1}} + k_t^{\frac{1}{K-1}}\right) \left(k_{t+1} - k_{t+1}^{\frac{K-2}{K-1}} \cdot k_t^{\frac{1}{K-1}} + k_{t+1}^{\frac{K-3}{K-1}} \cdot k_t^{\frac{2}{K-1}} - \cdots - k_t\right). \end{aligned}$$

Since, for $n = 1, 2, \ldots, K-1$, all of the numbers $k_{t+1}^{\frac{n}{K-1}} \cdot k_t^{\frac{K-1-n}{K-1}}$ lie between $k_{t+1}$ and $k_t$ in a monotonic order, and $\lambda_t > \lambda_{t+1}$ implies that $k(\lambda_t) < k(\lambda_{t+1})$, we get in view of (i) of Lemma (3.2.1):

$$k_{t+1}^{\frac{K}{K-1}} - k_t^{\frac{K}{K-1}} \leq \left(k_{t+1}^{\frac{1}{K-1}} + k_t^{\frac{1}{K-1}}\right)(k_{t+1} - k_t) \leq 3k_{t+1}^{\frac{1}{K-1}} 6D.$$

Substituting this in (3.19) gives us

$$0 < \lambda_t^{1/K} - \lambda_{t+1}^{1/K} \leq \frac{18D}{K}\lambda_t.$$

(ii) $\lambda_t < \lambda_{t+1}$:
This is similar to (i). Using the mean value theorem again on the concave function $\lambda^{1/K}$, we get:

$$\begin{aligned} 0 < \lambda_{t+1}^{1/K} - \lambda_t^{1/K} \leq \frac{1}{K}\lambda_t^{-\frac{K-1}{K}}(\lambda_{t+1} - \lambda_t) &= \frac{k_t}{K}\left(1 - \frac{\lambda_t}{\lambda_{t+1}}\right)\lambda_{t+1} \\ &= \frac{1}{K}\frac{\left(k_t^{\frac{K}{K-1}} - k_{t+1}^{\frac{K}{K-1}}\right)}{k_t^{\frac{1}{K-1}}}\lambda_{t+1}. \end{aligned} \tag{3.20}$$

When we take $K$ to be even, we get:

$$\begin{aligned} &k_t^{\frac{K}{K-1}} - k_{t+1}^{\frac{K}{K-1}} \\ &= \left(k_t^{\frac{1}{K-1}} + k_{t+1}^{\frac{1}{K-1}}\right)\left(k_t - k_t^{\frac{K-2}{K-1}} \cdot k_{t+1}^{\frac{1}{K-1}} + k_t^{\frac{K-3}{K-1}} \cdot k_{t+1}^{\frac{2}{K-1}} - \cdots - k_{t+1}\right). \end{aligned}$$

Since, for $n = 1, 2, \ldots, K-1$, all of the numbers $k_t^{\frac{n}{K-1}} \cdot k_{t+1}^{\frac{K-1-n}{K-1}}$ lie between $k_t$ and $k_{t+1}$ in a monotonic order, and $\lambda_{t+1} > \lambda_t$ implies that $k(\lambda_{t+1}) < k(\lambda_t)$, we get in view of (i) of Lemma (3.2.1):

$$k_t^{\frac{K}{K-1}} - k_{t+1}^{\frac{K}{K-1}} \leq \left(k_t^{\frac{1}{K-1}} + k_{t+1}^{\frac{1}{K-1}}\right)(k_t - k_{t+1}) \leq 3k_t^{\frac{1}{K-1}}6D.$$

Using $\lambda_{t+1} - \lambda_t \leq \varepsilon\lambda_t$ (see (ii) of Lemma (3.2.1)) we get:

$$0 < \lambda_{t+1}^{\frac{1}{K-1}} - \lambda_t^{\frac{1}{K-1}} \leq \frac{18D}{K}\lambda_{t+1} \leq \frac{18D}{K}(1+\varepsilon)\lambda_t.$$

So when we take $K$ large enough in order to ensure $\frac{18D}{K}(1+\varepsilon) \leq \varepsilon$, we get the lemma. □

For a fixed starting state and strategies $\pi_\varepsilon^1$ and $\pi^2$, the sequences $\beta_t, \lambda_t, k_t$ and $y_t, t = 0, 1, 2, \ldots$ are realizations of stochastic processes. Let $\overline{\beta}_t, \overline{\lambda}_t, \overline{k}_t$ and $\overline{y}_t$ denote the corresponding stochastic variables. Now let us assume that $L$ is large enough to ensure that $y_t \leq \varepsilon$ for all $t$ (since $\lambda_t \to 0 \iff k_t \to \infty \iff y_t \to 0$). We now define the random variables $Y_t := v_{\overline{\beta}_t}(S_t) - \overline{y}_t$, $t = 1, 2, \ldots$.

**Lemma 3.2.3.** *The sequence $Y_t,\ t = 1, 2, \ldots$ is a semi-martingale*[1] *and in particular*

$$\mathbb{E}\left[Y_{t+1} - Y_t \mid h_t\right] \geq 2\varepsilon\lambda_t.$$

*Proof.* From the definition of $k_{t+1}$ in (3.15) we can deduce for any realization that

$$r(i_t, a_t^1, a_t^2) - v_{\beta_t}(i_{t+1}) + 4\varepsilon \leq k_{t+1} - k_t$$

or

$$\mathbb{E}\left[r(S_t, A_t^1, A_t^2) - v_{\overline{\beta}_t}(S_{t+1}) + 4\varepsilon - (\overline{k}_{t+1} - \overline{k}_t) \mid h_t\right] \leq 0. \tag{3.21}$$

Now let us recall (3.10):

$$\mathbb{E}\left[\overline{\lambda}_t(r\left(S_t, A_t^1, A_t^2\right) - v_{\overline{\beta}_t}\left(S_{t+1}\right)) + (v_{\overline{\beta}_t}(S_{t+1}) - v_{\overline{\beta}_t}(S_t)) \mid h_t\right] \geq 0.$$

Multiplying (3.21) with $\overline{\lambda}_t$ we can see that replacing $\overline{\lambda}_t(r\left(S_t, A_t^1, A_t^2\right) - v_{\overline{\beta}_t}\left(S_{t+1}\right))$ by $\overline{\lambda}_t(\overline{k}_{t+1} - \overline{k}_t) - 4\varepsilon\overline{\lambda}_t$ in the equation above will lead to an increase on the left-hand side. Similarly, replacing $v_{\overline{\beta}_t}(S_{t+1})$ with $v_{\overline{\beta}_{t+1}}(S_{t+1})$ will decrease the left-hand side with at most $\varepsilon\overline{\lambda}_t$, since according to lemma (3.2.2) $\|\mathbf{v}_{\beta_t} - \mathbf{v}_{\beta_{t+1}}\| \leq \varepsilon\lambda_t$. These replacements lead to the following equation:

$$\mathbb{E}\left[\overline{\lambda}_t(\overline{k}_{t+1} - \overline{k}_t) - 4\varepsilon\overline{\lambda}_t + v_{\overline{\beta}_{t+1}}(S_{t+1}) - v_{\overline{\beta}_t}(S_t) + \overline{\lambda}_t\varepsilon \mid h_t\right] \geq 0.$$

We can now use part (iii) of Lemma 3.2.1 to get the following result:

$$\mathbb{E}\left[\overline{y}_t - \overline{y}_{t+1} + \varepsilon\overline{\lambda}_t - 4\varepsilon\overline{\lambda}_t + v_{\overline{\beta}_{t+1}}(S_{t+1}) - v_{\overline{\beta}_t}(S)_t) + \overline{\lambda}_t\varepsilon \mid h_t\right] \geq 0.$$

This is equivalent to $\mathbb{E}\left[Y_{t+1} - Y_t \mid h_t\right] \geq 2\varepsilon\lambda_t$. Since $\mathbb{E}\left[Y_t \mid h_t\right]$ is not stochastic, we also have $\mathbb{E}\left\{Y_{t+1} \mid h_t\right\} \geq Y_t + 2\varepsilon\lambda_t$. □

If we choose $L$ large enough (and so $k_t$ large enough), $|y_t| \leq D$. Together with $|v_\beta(i)| \leq D$ for all $\beta \in (0, 1)$ and $i \in S$ this implies that $Y_t$ is uniformly bounded. This leads to the following corollary.

**Corollary 3.2.1.** *The semi-martingale $Y_t,\ t = 0, 1, 2, \ldots$ converges with probability 1, say to $Y_\infty$.*

[1] **Definition 3.2.1.** $X$ is Martingale if $\mathbb{E}\left\{X_{t+1} \mid X_0, X_1, \ldots, X_t\right\} = X_t$, and semi-martingale if $\mathbb{E}\left\{X_{t+1} \mid X_0, X_1, \ldots, X_t\right\} \geq X_t$ or $\mathbb{E}\left\{X_{t+1} \mid X_0, X_1, \ldots, X_t\right\} \leq X_t$

*Proof.* The proof is a direct result of the following theorem (see [4]):

**Theorem 3.2.2.** *(Martingale Convergence Theorem)*
*Let* $\{X_n, \mathcal{H}_n,\ n = 0, 1, \dots\}$ *be a semi-martingale and suppose that* $\mathbb{E}\{|X_n|\}$ *is bounded. Then with probability one, there exists a finite integrable r.v.* $X_\infty$ *such that*

$$\lim_{n\to\infty} X_n = X_\infty \ \text{a.e.}$$

□

Since lemma 3.2.3 holds for every realization $h_t$, and all $t = 1, 2, \dots$, we also get for all $T$

$$\mathbb{E}\left[Y_T - Y_0 | h_t\right] \geq 2\varepsilon\, \mathbb{E}\left\{ \sum_{t=0}^{T-1} \lambda_t \;\middle|\; h_0 \right\}. \tag{3.22}$$

The convergence of $Y_t$ implies the convergence of other variables. We will show this in the following lemma.

**Lemma 3.2.4.**

*(i)* $\lim_{t\to\infty} \overline{\lambda}_t = 0$, *with probability 1.*
*(ii)* $\lim_{t\to\infty} \overline{y}_t = 0$, *with probability 1.*
*(iii)* $\lim_{t\to\infty} \overline{v}_{\beta_t}(S_t) = Y_\infty$
*(iv)* $\lim_{t\to\infty} \overline{v}_{\beta_t}(S_{t+1}) = Y_\infty$

*Proof.*

(i) Since $4D \geq \mathbb{E}\left[Y_T - Y_0 \mid h_0\right]$ for all $T$ and $L$ large enough, we get

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \overline{\lambda}_t \mid h_0\right] \leq \frac{2D}{\varepsilon}.$$

Because $\overline{\lambda}_t \in (0, 1)$, the monotone convergence theorem, which says that every monotone bounded sequence converges, can be applied (see also [4]). This gives us the required result.

(ii) Using (3.13) and (i) gives us $\lim_{t\to\infty} \overline{y}_t = 0$.

(iii) Since $Y_t := v_{\overline{\beta}_t}(S_t) - \overline{y}_t,\ t = 1, 2, \dots$, applying the result of (ii) gives us (iii)

(iv) We know from lemma 3.2.2 that

$$\left|v_{\beta_t}(S_{t+1}) - v_{\beta_{t+1}}(S_{t+1})\right| - \varepsilon\overline{\lambda}_t \leq 0$$

with probability 1; in view of (iii) and (i), (iv) follows.

□

We are now able to prove the $\varepsilon$-optimality of $\pi^1_\varepsilon$.

**Theorem 3.2.3.** *The strategy $\pi^1_\varepsilon$ is $8\varepsilon$-optimal for player 1, both with respect to the limiting average reward and with respect to the average of the limit distribution.*

*Proof.* Since $\lim_{t\to\infty} \overline{\lambda}_t = 0$, we have $\lim_{t\to\infty} \overline{\beta}_t = 1$. From the results of Bewley and Kohlberg, [1], we know that $\lim_{\beta\uparrow 1} \mathbf{v}_\beta = \mathbf{c}_0$. Using Lemma 3.2.2 and the semi-martingale property of $Y_t$:

$$\begin{aligned}
\mathbb{E}\left[v_{\overline{\beta}_t}(S_{t+1} \mid h_0)\right] &\geq \mathbb{E}\left[v_{\overline{\beta}_{t+1}}(S_{t+1}) - \varepsilon\overline{\lambda}_t \mid h_0\right] \\
&= \mathbb{E}\left[v_{\overline{\beta}_{t+1}}(S_{t+1}) - \overline{y}_{t+1} + \overline{y}_{t+1} - \varepsilon\overline{\lambda}_t \mid h_0\right] \\
&= \mathbb{E}\left[Y_{t+1} + \overline{y}_{t+1} - \varepsilon\overline{\lambda}_t \mid h_0\right].
\end{aligned}$$

Using lemma 3.2.3 we have

$$\begin{aligned}
&\geq \mathbb{E}\left[Y_0 + \overline{y}_{t+1} - \varepsilon\overline{\lambda}_t \mid h_0\right] \\
&= \mathbb{E}\left[v_{\beta_0}(S_0) - \overline{y}_0 + \overline{y}_{t+1} - \varepsilon\overline{\lambda}_t \mid h_0\right].
\end{aligned}$$

and now for $\beta_0$ sufficiently close to 1 we get

$$\geq \mathbb{E}\left[c_0(S_0) - \varepsilon - \overline{y}_0 + \overline{y}_{t+1} - \varepsilon\overline{\lambda}_t \mid h_0\right].$$

Since $\overline{y}_{t+1} \geq 0$ and $-\varepsilon\overline{\lambda}_t \geq -\varepsilon\lambda$, this implies for $L$ big enough

$$\begin{aligned}
&\geq c_0(i_0) - 2\varepsilon - y_0 \\
&\geq c_0(i_0) - 3\varepsilon. \qquad (3.23)
\end{aligned}$$

$L$ is chosen large enough (which means that the $k_t$'s are large enough, and the $\beta_t$'s are close enough to 1 (see (3.15))) in order to ensure that $|v_{\beta_0}(S_0) - c_0(S_0)| \leq \varepsilon$ and $y_0 \leq \varepsilon$. From the definition of $k_t$ and lemma 3.2.1,(i) it follows that

$$k_{t+1} - k_t \leq r(i_t, a^1_t, a^2_t) - v_\beta(i_{t+1}) + 4\varepsilon + 6DI(k_{t+1} = L) \qquad (3.24)$$

where $I$ denotes the indication function. Summing (3.24) over $0 \leq t \leq T$ gives us:

$$\begin{aligned}
&\sum_{t=0}^{T} r(i_t, a^1_t, a^2_t) \\
&\geq \sum_{t=0}^{T} v_{\beta_t}(s_{t+1}) + k_{T+1} - k_0 - 6D\sum_{t=0}^{T} I(k_{t+1} = L) - 4\varepsilon(T+1).
\end{aligned} \qquad (3.25)$$

Since $k_{T+1}$ is positive in the last inequality, we can leave it out. Furthermore, for $T$ large enough we also have:

$$-k_0 - 6D\sum_{t=0}^{T} I(k_{t+1} = L) \geq -\varepsilon(T+1). \tag{3.26}$$

This is because $k_0$ is constant, and since $\overline{\lambda}_t \to 0$ implies that $k_t \to \infty$, $\sum_{t=0}^{T} I(k_{t+1} = L)$ is bounded and monotone, and so exists for $T \to \infty$. Putting (3.25) and (3.26) together gives us

$$\sum_{t=0}^{T} r(i_t, a_t^1, a_t^2) \geq \sum_{t=0}^{T} v_{\beta_t}(s_{t+1}) - 5\varepsilon(T+1). \tag{3.27}$$

Combination of (3.27) and (3.23) yields for the limiting average expected rewards:

$$\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}\left[r(S_t, A_t^1, A_t^2) \mid h_0\right] \geq c_0(i_0) - 8\varepsilon. \tag{3.28}$$

From (3.27) we also get for any realization:

$$\liminf_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} r(i_t, a_t^1, a_t^2) \geq \liminf_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} v_{\beta_t}(i_{t+1}) - 5\varepsilon$$

which implies that

$$\begin{aligned} &\mathbb{E}\left[\liminf_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} r(S_t, A_t^1, A_t^2) \;\middle|\; h_0\right] \\ &\geq \mathbb{E}\left[\liminf_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} v_{\beta_t}(S_{t+1}) \;\middle|\; h_0\right] - 5\varepsilon. \end{aligned} \tag{3.29}$$

Lemma 3.2.4 states that $\lim_{t\to\infty} v_{\beta_t}(S_{t+1}) = Y_\infty$ with probability 1. This means that the limit of the right-hand side of (3.29) exists and equals $\mathbb{E}[Y_\infty|h_0] - 5\varepsilon$. Furthermore, using (3.22), and $\beta_0$ and $L$ large enough:

$$\mathbb{E}[Y_\infty|h_0] > \mathbb{E}[Y_0|h_0] = v_{\beta_0} - y_0 \geq c_0(i_0) - \varepsilon - y_0 \geq c_0(i_0) - 2\varepsilon.$$

We now have for the expected average limit distribibution:

$$\mathbb{E}\left[\liminf_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} r(S_t, A_t^1, A_t^2) \;\middle|\; h_0\right] \geq c_0(i_0) - 7\varepsilon \geq c_0(i_0) - 8\varepsilon. \tag{3.30}$$

The choices of $K$ and $L$ can be made quite independently of $\pi_2$. This means the analysis above holds for any $\pi_2$. (3.28) and (3.30) now show the theorem. □

**Corollary 3.2.2.**
*The stochastic game possesses a value vector with respect to the limiting average reward criterion as well as with respect to the average reward of the limiting distribution. In both cases this value equals* $\mathbf{c}_0 = \lim_{\beta \uparrow 1} \mathbf{v}_\beta$.

# Bibliography


[1] Truman Bewley and Elon Kohlberg. The asymptotic theory of stochastic games. *Math. Oper. Res.*, 1:197–208, 1976.

[2] D. Blackwell and T.S. Ferguson. The big match. *Ann. Math. Stat.*, 39:159–163, 1968.

[3] E.V. Denardo. On linear programming in a Markov decision problem. *Manage. Sci., Theory*, 16:281–288, 1970.

[4] Joseph L. Doob. *Stochastic processes.* New York: Wiley. 654 S. , 1953.

[5] J.A. Filar, T.A. Schultz, F. Thuijsman, and O.J. Vrieze. Nonlinear programming and stationary equilibria in stochastic games. *Math. Program., Ser. A*, 50(2):227–237, 1991.

[6] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes.* New York, NY: Springer. xii, 393 p. $ 69.00 , 1997.

[7] Jerzy A. Filar and Todd A. Schultz. Communicating MDPs: Equivalence and LP properties. *Oper. Res. Lett.*, 7(6):303–307, 1988.

[8] Dean Gillette. Stochastic games with zero stop probabilities. *Ann. Math. Stud.*, 39:179–187, 1957.

[9] G.H. Hardy and J.E. Littlewood. Notes on the theory of series. XVI.: Two Tauberian theorems. *J. Lond. Math. Soc.*, 6:281–286, 1931.

[10] R. Horn and C. Johnson. *Matrix Analysis.* Cambridge University Press, New York, 1985.

[11] L.Kallenberg. *Inleiding Besliskunde.* 2003.

[12] J.-F. Mertens and A. Neyman. Stochastic games. *Int. J. Game Theory*, 10:53–66, 1981.

[13] Abraham Neyman. Stochastic games: Existence of the minmax. In *Neyman, Abraham (ed.) et al., Stochastic games and applications. Lectures given at the NATO Advanced Study Institute on "Stochastic games*

*and applications", Stony Brook, NY, USA, July 1999. On the occasion of L. S. Shapley's eightieth birthday. Dordrecht: Kluwer Academic Publishers. NATO ASI Ser., Ser. C, Math. Phys. Sci. 570, 173-193* . 2003.

[14] Sylvain Sorin. Classification and basic tools. In *Neyman, Abraham (ed.) et al., Stochastic games and applications. Lectures given at the NATO Advanced Study Institute on "Stochastic games and applications", Stony Brook, NY, USA, July 1999. On the occasion of L. S. Shapley's eightieth birthday. Dordrecht: Kluwer Academic Publishers. NATO ASI Ser., Ser. C, Math. Phys. Sci. 570, 27-36* . 2003.

[15] W.W. Szczechla, S.A. Connell, J.A. Filar, and O.J. Vrieze. On the Puiseux series expansion of the limit discount equation of stochastic games. *SIAM J. Control Optimization*, 35(3):860–875, 1997.

[16] O.J. Vrieze. *Stochastic games with finite state and action spaces.* PhD thesis, CWI Tracts, 33. Centrum voor Wiskunde en Informatica. Amsterdam: Stichting Mathematisch Centrum. (Rev. version of thesis). VIII, 221 p.; Dfl. 34.20 , 1987.

[17] A. Zygmund. *Trigonometric Series.* Cambridge University Press, Cambridge, England, 1968.