Stochastische spelen

Hutter, P.C.

Stochastische spelen

Christian Hutter

Bachelorscriptie - Voorjaar 2005

Inhoudsopgave

1 Inleiding **3**

2 Markov Beslissings theorie **4**
2.1 Sommeerbare Markov beslissings processen 4
2.1.1 Verdiscontering 4
2.1.2 Stopkans 5
2.2 Eindige horizon 5
2.3 Lineaire Programmering bij verdiscontering 6
2.4 Gemiddelde opbrengst 8
2.4.1 Irreducibel 8
2.5 Toepassing: Het Hamiltonkringprobleem 9
2.6 Niet-stationaire strategieën 11
2.7 Strategieverbetering 12
2.8 Uitwerking opgaven 14

3 Stochastische Spelen **22**
3.1 Verdiscontering 22
3.1.1 Matrix Spelen 23
3.2 Lineaire Programmering 23
3.2.1 Overheersende beslisser 24
3.2.2 Toestand onafhankelijk en gescheiden opbrengst . . . 25
3.2.3 Wisselende overheersende beslisser 26
3.3 Aangepaste methode van Newton 26
3.4 The Big Match 30
3.5 Nulsom spel met overheersende beslisser en gemiddelde opbrengst 32
3.6 Uitwerking opgaven 36

1 Inleiding

Bij mijn bachelorproject heb ik het boek "Competitive Markov Decision Processes"[1] van Jerzy Filar en Koos Vrieze bestudeerd. Hierin wordt allereerst de basis van Markov beslissings processen uitgelegd. Daarna komt uitgebreid de theorie van Stochastische Spelen aan bod. Het Stochastische Spel kunnen we zien als een generalisatie van het Markov beslissings proces. In mijn scriptie heb ik geprobeerd de theorie uit het eerste deel van dit boek te behandelen. Dit eerste deel is een afgerond geheel en bevat naast de Markov beslissings theorie een basis van de theorie over Stochastische Spelen. Ik heb de structuur van het boek aangehouden: in hoofdstuk 2 van mijn scriptie behandel ik de Markov Beslissings theorie en in hoofdstuk 3 Stochastische Spelen. Ook heb ik enkele opgaven uit dit boek uitgewerkt, deze zijn te vinden in sectie 2.8 en 3.6.
Met veel plezier heb ik aan deze scriptie gewerkt. Graag wil ik de heer L.C.M. Kallenberg hartelijk danken voor zijn begeleiding. Ik wens u veel leesplezier.

Christian Hutter
Studentnummer: 0214981
Adres: De Maroc 35
2291 JX Wateringen
E-mail: pchutter@gmail.com

[1] Jerzy Filar, Koos Vrieze *Competitive Markov Decision Processes.* Springer-Verlag, New York (1997)

2 Markov Beslissings theorie

2.1 Sommeerbare Markov beslissings processen

We bestuderen een proces Γ op discrete tijdstippen $t = 0, 1, 2, \ldots$. Dit proces bevindt zich op tijdstip t in een bepaalde toestand s. We hebben een eindige toestandsruimte $S = \{1, 2, \ldots, N\}$. Er is een beslisser die in toestand $s \in S$ een actie $a \in A(s) = \{1, 2, \ldots, m(s)\}$ kiest, waarbij $m(s)$ het aantal acties in toestand s is. Er is een directe opbrengst $r(s, a)$ en met kans $p(s'|s, a)$ bevindt het proces zich op het volgende tijdstip in toestand $s' \in S$. We zullen aannemen dat deze kansen onafhankelijk zijn van de tijd en van eerdere toestanden en acties. De beslisser kiest op ieder tijdstip in toestand $s \in S$ actie $a \in A(s)$ met een bepaalde kans die we aangeven met $f(s, a)$. Deze kansen mogen in principe van het tijdstip t afhangen en dat noteren we met $f_t(s, a)$. In de meeste gevallen zullen ze echter niet van t afhangen. Deze kansen moeten voldoen aan $\sum_{a=1}^{m(s)} f(s, a) = 1$ voor alle $s \in S$. We kunnen voor een toestand s deze kansen in een vector zetten, we krijgen dan $\mathbf{f}(s) = (f(s, 1), f(s, 2), \ldots, f(s, m(s)))$. Als we deze $\mathbf{f}(s)$ voor alle toestanden in een vector zetten, krijgen we $\mathbf{f} = (\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(s), \ldots, \mathbf{f}(N))$ en we noemen $\mathbf{f}$ een strategie. Een strategie heet deterministisch als $f(s, a) \in \{0, 1\} \, \forall a \in A(s), s \in S$. Als een strategie niet van de tijd afhangt, heet deze stationair. De verzameling van alle deterministische en stationaire strategieën geven we aan met respectievelijk $\mathbf{F}_D$ en $\mathbf{F}_S$. In deze scriptie zullen we voornamelijk naar stationaire strategieën kijken.
Met een strategie kunnen we een overgangsmatrix definiëren:

$$P(\mathbf{f}) = \big(p(s'|s, \mathbf{f})\big)_{s,s'=1}^{N}$$

waarbij geldt $p(s'|s, \mathbf{f}) = \sum_{a=1}^{m(s)} p(s'|s, a) f(s, a)$. Omdat het proces op elk tijdstip naar een bepaalde toestand gaat, geldt $\sum_{s'=1}^{N} p(s'|s, a) = 1$. Dus $P(\mathbf{f})$ is een stochastische matrix en hierbij hoort een unieke Markov keten.

2.1.1 Verdiscontering

De beslisser zal proberen een zo goed mogelijke strategie te vinden, dat wil zeggen een strategie die hem zo veel mogelijk opbrengt. Als we echter uit gaan van een oneindige tijdsperiode (oneindige horizon), krijgen we oneindig veel opbrengsten. Als we deze opbrengsten optellen, kan dit leiden tot een niet gedefinieerde uitkomst. Een manier om dit te verhelpen is verdiscontering.
Met R_t geven we de opbrengst op tijdstip t aan. De verwachte opbrengst op tijdstip t met begintoestand s en strategie $\mathbf{f}$ noteren we met $E_{s\mathbf{f}}[R_t]$. De verwachte waarde van het proces wordt nu als volgt gedefinieerd:

$$v_\beta(s, \mathbf{f}) = \sum_{t=0}^{\infty} \beta^t E_{s\mathbf{f}}[R_t] \tag{1}$$

Zoals gezegd gebruiken we in dit geval verdiscontering. Hierbij komt het erop neer dat we een bedrag dat we in de toekomst, zeg op tijdstip t, verkrijgen, vermenigvuldigen met β^t, waarbij de verdisconteringsfactor β een getal tussen 0 en 1 is. Met verdiscontering kunnen we opbrengsten over een oneindige periode bij elkaar optellen tot een eindig getal. Dit kunnen we als volgt inzien: laat $M = \max_{s\in S, a\in A(s)} |r(s,a)|$, dan geldt

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

Zoals gezegd geven we met $r(s,a)$ de opbrengst aan als het proces in toestand s is en actie a wordt gekozen. Voor elke toestand definiëren we $r(s,\mathbf{f}) = \sum_{a\in A(s)} r(s,a) f(s,a)$. Als we deze in een vector zetten, krijgen we de onmiddelijk verwachte opbrengst vector $\mathbf{r}(\mathbf{f}) = (r(1,\mathbf{f}), r(2,\mathbf{f}), \ldots, r(N,\mathbf{f}))^T$. Als we nu de verwachte waarde als vector $\mathbf{v}_\beta(\mathbf{f}) = (v_\beta(1,\mathbf{f}), \ldots, v_\beta(N,\mathbf{f}))^T$ schrijven, dan geldt

$$\mathbf{v}_\beta(\mathbf{f}) = \sum_{t=0}^{\infty} \beta^t P^t(\mathbf{f})\mathbf{r}(\mathbf{f}) = [I - \beta P(\mathbf{f})]^{-1}\, \mathbf{r}(\mathbf{f}) \tag{2}$$

De laatste gelijkheid volgt uit het feit dat $[I - \beta P(\mathbf{f})]$ een inverteerbare matrix is, waarvoor geldt $[I - \beta P(\mathbf{f})]^{-1} = I + \beta P(\mathbf{f}) + \beta^2 P^2(\mathbf{f}) + \ldots$ (zie ook opgave 2 bij de uitgewerkte opgaven).

2.1.2 Stopkans

Behalve verdiscontering zijn er ook andere manieren om de opbrengsten sommeerbaar te maken. In deze sectie doen we dat met de volgende aanname:

$$\sum_{s'=1}^{N} p(s'|s,a) < 1 \quad \forall a \in A(s), s \in S$$

Dit betekent dat er bij iedere $a \in A(s)$ een positieve stopkans is. De overgangsmatrix $P(\mathbf{f})$ heeft nu ook de eigenschap dat $\sum_{s'=1}^{N} p(s'|s,\mathbf{f}) < 1\ \forall s \in S$. Analoog aan het verdisconteerde geval kunnen we de waarde vector als volgt definiëren:

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

2.2 Eindige horizon

In het gedeelte tot nu toe zijn we uitgegaan van een oneindige horizon, dat wil zeggen dat het proces oneindig lang doorgaat. In bijna alle toepassingen zal echter een eindige tijd beschikbaar zijn. Daarom zullen we nu naar processen met eindige horizon kijken. Hierbij komt een tijdstip t dus uit de

eindige verzameling $\{0, 1, 2, \ldots, T\}$. Er zijn echter twee praktische redenen waarom het niet interessant is om het eindige horizon model te bekijken:

1. Als T klein is, bestaat er een mooie oplossing voor het proces, die we hierna zullen bespreken. Dit kan dus als opgelost beschouwd worden.

2. Als T groot is, is bovengenoemde oplossing niet meer berekenbaar.

Bij een eindige horizon is het kiezen van een bepaalde actie afhankelijk van de tijd die nog over is tot het einde van het proces. Op een gegeven moment kan je namelijk niet meer in een "slechte" toestand komen. Er is nu dus niet een stationaire strategie die we altijd toepassen, maar we hebben een rij strategieën $\pi = (f_0, f_1, \ldots, f_T)$ met f_t een stationaire strategie die we op tijdstip t toepassen. Wederom hebben we de volgende verwachte waarde met begintoestand s:

$$v_T(s, \pi) := \sum_{t=0}^{T} E_{s\pi}[R_t]$$

Als we de verwachte waardes voor alle toestanden in een vector $\mathbf{v}_T(\pi) = (v_T(1, \pi), \ldots, v_T(N, \pi))^T$ zetten, komt het er op neer dat we $\mathbf{v}_T$ willen maximaliseren over de π's. We zullen nu een algoritme geven dat gebruik maakt van dynamische programmering:

- **Stap 1** Laat $V_{-1}(s) = 0 \ \forall s \in S$ en definieer
  $f_T^*(s) := a_s^T = \text{argmax}_{A(s)} \left\{ r(s, a) + \sum_{s'=1}^{N} p(s'|s, a) V_{-1}(s') \right\}$
  en $V_0(s) := r(s, a_s^T) = \max_{A(s)} \{r(s, a) + 0\}$

- **Stap 2** Doe voor iedere $n = 1, 2, \ldots, T$ het volgende: bereken $\forall s \in S$
  $f_{T-n}^*(s) := a_s^{T-n} = \text{argmax}_{A(s)} \left\{ r(s, a) + \sum_{s'=1}^{N} p(s'|s, a) V_{n-1}(s') \right\}$
  en $V_n(s) := r(s, a_s^{T-n}) + \sum_{s'=1}^{N} p(s'|s, a_s^{T-n}) V_{n-1}(s')$

- **Stap 3** Construeer een strategie $\pi^* = (f_0^*, f_1^*, \ldots, f_T^*)$.

In dit algoritme is $V_n(s, \pi)$ de verwachte waarde van de laatste n tijdstippen, gegeven dat de toestand op tijdstip $(T - n)$ s is. Er geldt dus $V_T(s, \pi) = v_T(s, \pi)$. Een bewijs van de correctheid van dit algoritme is te vinden in het boek, op bladzijde 20 en 21.

2.3 Lineaire Programmering bij verdiscontering

We willen graag een strategie vinden die een zo groot mogelijke waarde oplevert. Dit komt neer op het maximaliseren van $\mathbf{v}_\beta(\mathbf{f})$ over de $\mathbf{f}$. We zullen hier alleen stationaire strategieën bekijken. Het blijkt dat voor het verdisconteerde geval met oneindige horizon zowel de optimale strategie, $\mathbf{f^0}$,

als de grootste waarde vector, $\mathbf{v}_\beta(\mathbf{f}^0)$ te vinden zijn met behulp van Lineaire Programmering. Uit (2) volgt:

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

Als we veronderstellen dat $\mathbf{v}_\beta := \mathbf{v}_\beta(\mathbf{f}^0) = \max_{\mathbf{f}} \mathbf{v}_\beta(\mathbf{f})$ bestaat en als we weten hoe we vanaf het volgende tijdstip tot het eind moeten spelen, moeten we op het huidige tijdstip het probleem lokaal optimaliseren, dus:

$$v_\beta(s) = \max_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,a) v_\beta(s') \right\} \tag{3}$$

Dit suggereert dat $v_\beta(s)$ voldoet aan de volgende lineaire vergelijkingen, die we uitdrukken in termen van de vector $\mathbf{v} = (v(1), \ldots, v(N))^T$:

$$v(s) \geq r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,a) v(s') \quad \forall a \in A(s)$$

Dan geldt voor een willekeurige strategie $\mathbf{f} \in \mathbf{F}_S$

$$\begin{aligned} v(s) &\geq r(s,\mathbf{f}) + \beta \sum_{s'=1}^{N} p(s'|s,\mathbf{f}) v(s') \\ \mathbf{v} &\geq \mathbf{r}(\mathbf{f}) + \beta P(\mathbf{f})\mathbf{v} \end{aligned}$$

Als we $\mathbf{v}$ aan de rechterkant van de laatste ongelijkheid k keer substitueren, krijgen we:

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

en wanneer we de limiet als $k \to \infty$ nemen:

$$\mathbf{v} \geq [I - \beta P(\mathbf{f})]^{-1} \mathbf{r}(\mathbf{f}) = \mathbf{v}_\beta(\mathbf{f}) \tag{4}$$

Dit is de achterliggende gedachte voor het Lineaire Programmeringsprobleem om $\mathbf{v}_\beta$ te vinden, dat er als volgt uitziet:

$$\min \left\{ \sum_{s=1}^{N} \frac{1}{N} v(s) \mid v(s) \geq r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,a) v(s'); a \in A(s), s \in S \right\}$$

De coefficienten $\frac{1}{N}$ kunnen worden gezien als de kans om in een bepaalde toestand te beginnen. Het lijkt in eerste instantie wellicht vreemd dat we een minimaliseringsprobleem hebben terwijl we iets willen maximaliseren. We willen datgene dat in de beperking staat maximaliseren, zie ook (3). We

eisen dat een toegelaten oplossing van het Lineaire Programmerings probleem hierboven ligt, zie ook (4). Dus we moeten de oplossing minimaliseren. Het duale probleem ziet er als volgt uit:

$$\max \left\{ \sum_{s=1}^{N} \sum_{a=1}^{m(s)} r(s,a) x_{sa} \mid \begin{array}{ll} \sum_{s=1}^{N} \sum_{a=1}^{m(s)} [\delta(s,s') - \beta p(s'|s,a)] x_{sa} = \frac{1}{N}; & s' \in S \\ x_{sa} \geq 0; & a \in A(s), s \in S \end{array} \right\}$$

Laat $\mathbf{x}^0 = \left\{ x^0_{sa} | a \in A(s), s \in S \right\}$ een optimale oplossing van het duale probleem zijn en definieer $x^0_s = \sum_{a=1}^{m(s)} x^0_{sa}$, dan is de optimale strategie gelijk aan $f^0(s,a) = \frac{x^0_{sa}}{x^0_s}$. Voor het bewijs hiervan en de juistheid van het lineaire en duale probleem verwijs ik naar blz. 25-29 van het boek.

2.4 Gemiddelde opbrengst

Naast verdiscontering en een positieve stopkans kunnen we ook op een andere manier proberen de opbrengsten over een oneindige tijdsperiode te sommeren, namelijk met de gemiddelde opbrengst. We definiëren de waarde van strategie $\mathbf{f}$ met begintoestand s als volgt:

$$v_\alpha(s, \mathbf{f}) = \lim_{T \to \infty} \left[\frac{1}{T+1} \sum_{t=0}^{T} E_{s\mathbf{f}}[R_t] \right]$$

De waarde vector $\mathbf{v}_\alpha(\mathbf{f}) = (v_\alpha(1, \mathbf{f}), \ldots, v_\alpha(N, \mathbf{f}))^T$ kunnen we in dit geval schrijven als

$$\mathbf{v}_\alpha(\mathbf{f}) = \lim_{T \to \infty} \left[\frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f}) \mathbf{r}(\mathbf{f}) \right] = \left[\lim_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f}) \right] \mathbf{r}(\mathbf{f})$$

De laatste gelijkheid volgt uit de eigenschap van Markovketens dat er een Markov matrix $Q(\mathbf{f})$ bestaat zodat $Q(\mathbf{f}) = \lim_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f})$. Deze matrix wordt ook wel de Cesaro-limiet matrix van $P(\mathbf{f})$ genoemd. Nu geldt $\mathbf{v}_\alpha(\mathbf{f}) = Q(\mathbf{f})\mathbf{r}(\mathbf{f})$.

2.4.1 Irreducibel

Een moeilijkheid bij modellen met gemiddelde opbrengst is dat er absorberende toestanden kunnen bestaan. Dat zijn toestanden waarbij we met kans 1 in die toestand blijven. We zullen daarom eerst irreducibele modellen bekijken, want daarbij doet dit probleem zich niet voor. Irreducibel betekent dat er voor iedere $s, s' \in S$ een positieve t bestaat waarvoor geldt $\left[P^t(\mathbf{f})\right]_{s,s'} > 0$ voor iedere $\mathbf{f}$. Dus elke toestand zal met kans 1 oneindig vaak bezocht worden. Nu geldt dat de matrix $Q(\mathbf{f})$ uit sectie 2.4 identieke rijen heeft. Zo'n rij noteren we met $\mathbf{q} = (q_1, \ldots, q_N)$ en dit is de stationaire verdeling van het Markov proces (zie opgave 6 bij de uitgewerkte opgaven voor het lemma en

een bewijs). We definiëren $x_{sa}(\mathbf{f}) = q_s(\mathbf{f})f(s,a)$ wat kan worden gezien als de gemiddelde frequentie van het paar (s,a) bij strategie $\mathbf{f}$. En we definiëren $x_s(\mathbf{f}) = \sum_{a\in A(S)} x_{sa}(\mathbf{f}) = q_s(\mathbf{f})$ wat kan worden gezien als de gemiddelde frequentie van bezoeken aan s. De "toestand-actie frequentie vector" $\mathbf{x}(\mathbf{f})$ is de vector met op de s^e plaats de vector $\mathbf{x}_s(\mathbf{f}) = \big(x_{s1}(\mathbf{f}), x_{s2}(\mathbf{f}), \ldots, x_{sm(s)}(\mathbf{f})\big)^T$. Analoog is de "toestand frequentie vector" $\overline{\mathbf{x}}(\mathbf{f}) = (x_1(\mathbf{f}), x_2(\mathbf{f}), \ldots, x_N(\mathbf{f}))$. We kunnen nu een afbeelding van de strategieruimte $M : \mathbf{F}_S \to \mathbf{R}^m$ met $m = \sum_{s=1}^N m(s)$ maken waarbij $M(\mathbf{f}) = \mathbf{x}(\mathbf{f})$.
Omdat $P(f)\mathbf{q} = \mathbf{q}$ (zie opgave 6) geldt $\mathbf{q}(\mathbf{f})[I - P(\mathbf{f})] = \mathbf{0}$. Dit kunnen we als volgt uitschrijven:

$$\begin{aligned}\sum_{s=1}^{N}(\delta(s,s') - p(s'|s,\mathbf{f}))q_s(\mathbf{f}) &= \sum_{s=1}^{N}\sum_{a\in A(S)}(\delta(s,s') - p(s'|s,a))q_s(\mathbf{f})f(s,a) \\ &= \sum_{s=1}^{N}\sum_{a\in A(S)}(\delta(s,s') - p(s'|s,a))x_{sa}(\mathbf{f}) = 0\end{aligned}$$

met $s' \in S$. Hiermee kunnen we een verzameling $\mathbf{X}$ definiëren met de volgende lineaire beperkingen:

1. $\sum_{s=1}^N \sum_{a\in A(S)}(\delta(s,s') - p(s'|s,a))x_{sa} = 0$
2. $\sum_{s=1}^N \sum_{a\in A(S)} x_{sa} = 1$
3. $x_{sa} \geq 0$

Er geldt nu dat M een inverteerbare afbeelding van $\mathbf{F}_S$ naar $\mathbf{X}$ is en dat $M^{-1}(\mathbf{x}) = \mathbf{f}_\mathbf{x}$, met $f_\mathbf{x}(s,a) = \frac{x_{sa}}{x_s}$ $(x_s > 0\ \forall s \in S)$. Laat $\mathbf{x}^0$ een optimale oplossing zijn van het LP-probleem $\max \sum_{s=1}^N \sum_{a\in A(s)} r(s,a)x_{sa}$ onder de voorwaarden die hierboven genoemd zijn, dan is $\mathbf{f}_{\mathbf{x}^0} = M^{-1}(\mathbf{x}^0)$ een optimale strategie. Zie voor het bewijs blz. 37-39 van het boek.

2.5 Toepassing: Het Hamiltonkringprobleem

Nu zullen we kennis over het Markov beslissingsproces toepassen op het Hamiltonkringprobleem. Dit probleem houdt in dat we in een graaf G met N knooppunten een enkelvoudige kring van N takken willen vinden, dat is een kring die elk knooppunt precies één keer aandoet. We wandelen door deze graaf waarbij we gestuurd worden door een functie f die de verzameling knooppunten $S = \{1, 2, \ldots, N\}$ afbeeldt op de verzameling takken A. We kunnen de verzameling knooppunten zien als de toestandsruimte van een Markov beslissingsproces met de actieruimte $A(s) = \{a = s'|(s,s') \in A\}$. Knooppunt 1 zullen we zien als het beginknooppunt.
Als we eisen dat $f(s) \in A(s)$, dan kunnen we f zien als een deterministische strategie. De overgangsmatrix $P(f)$ is irreducibel als f een Hamiltonkring

induceert en er geldt voor elke toestand $x_s(f) = \frac{1}{N}$. We zullen in deze sectie alleen *unichained* processen bekijken. Bij zo'n proces bevat de overgangsmatrix slechts één ergodic klasse, zie opgave 16 bij de uitgewerkte opgaven. Om een proces unichained te maken zullen we de overgangskansen zo aanpassen dat voor elk paar knooppunten s, s' (beide verschillend van 1) de "deterministische" tak (s, s') vervangen wordt door twee "stochastische" takken $(s, 1)$ en (s, s') met gewichten ϵ en $(1 - \epsilon)$ $(\epsilon \in (0, 1))$:

$$p_\epsilon(s'|s,a) = \begin{cases} 1 & \text{als } s = 1 \text{ en } a = s' \\ 0 & \text{als } s = 1 \text{ en } a \neq s' \\ 1 & \text{als } s > 1 \text{ en } a = s' = 1 \\ \epsilon & \text{als } s > 1,\ a \neq s' \text{ en } s' = 1 \\ 1 - \epsilon & \text{als } s > 1,\ a = s' \text{ en } s' > 1 \\ 0 & \text{als } s > 1,\ a \neq s' \text{ en } s' > 1 \end{cases}$$

Omdat voor iedere $\mathbf{f} \in \mathbf{F}_S$ geldt $\mathbf{x}(\mathbf{f}) \in \mathbf{X}$ kunnen we een afbeelding $M : \mathbf{F}_S \to \mathbf{X}$ zoals in sectie 2.4 definiëren. Hier zullen we ook een afbeelding $\hat{M} : \mathbf{X} \to \mathbf{F}_S$ definiëren met

$$f_{\mathbf{x}}(s,a) = \begin{cases} \frac{x_{sa}}{x_s} & \text{als } x_s = \sum_{a \in A(s)} x_{sa} > 0 \\ 1 & \text{als } x_s = 0 \text{ en } a = a_1 \\ 0 & \text{als } x_s = 0 \text{ en } a \neq a_1 \end{cases}$$

voor iedere $a \in A(s), s \in S$, waarbij a_1 de eerst mogelijke actie voor een gegeven toestand volgens een bepaalde ordening aangeeft. In opgave 16 zullen we het volgende lemma (lemma 2.5.1) nagaan:

1. Er geldt $\mathbf{X} = \{\mathbf{x}(\mathbf{f}) \mid \mathbf{f} \in \mathbf{F}_S\}$.
2. Voor iedere $\mathbf{x} \in \mathbf{X}$ geldt $M(\hat{M}(\mathbf{x})) = x$, maar de inverse van M hoeft niet te bestaan.
3. Als $\mathbf{x}$ een extreme van $\mathbf{X}$ is, dan $\mathbf{f}_{\mathbf{x}} = \hat{M}(\mathbf{x}) \in \mathbf{F}_D$.
4. Als $\mathbf{f} \in \mathbf{F}_D$ een Hamiltonkring is, dan is $\mathbf{x}(\mathbf{f})$ een extreme van $\mathbf{X}$.

Met iedere $\mathbf{f} \in \mathbf{F}_D$ kunnen we een subgraaf $G_{\mathbf{f}}$ van G associëren: pijl$(s, s') \in G_{\mathbf{f}} \iff f(s) = s'$. Een enkelvoudige kring van lengte m beginnend in 1 noteren we met $c_m^1 = \{(s_1 = 1, s_2), (s_2, s_3), \ldots, (s_m, s_{m+1} = 1)\}$ waarbij $m = 2, 3, \ldots, N$. Als $G_{\mathbf{f}}$ een kring c_m^1 bevat, schrijven we $G_{\mathbf{f}} \supset c_m^1$. Laat $C_m = \left\{\mathbf{f} \in \mathbf{F}_D | G_{\mathbf{f}} \supset c_m^1\right\}$, dan is C_N de verzameling van strategieën die met een Hamiltonkring corresponderen. De deterministische strategieën die nu belangrijk voor ons zijn kunnen we noteren als

$$\mathbf{F}_D = \left[\bigcup_{m=2}^{N} C_m\right] \bigcup B \tag{5}$$

waarbij B alle deterministische strategieën bevat die niet in één van de C_m's zitten. Zo'n strategie uit B ziet er uit als
$\{(s_1 = 1, s_2), \ldots, (s_{k-1}, s_k), (s_k, s_{k+1}), \ldots, (s_{l-1}, s_l), (s_l, s_k)\}$. Alle strategieën in een gegeven verzameling in (5) induceren dezelfde frequentie van bezoeken aan knooppunt 1, namelijk:

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

met $d_m(\epsilon) = 1 + \sum_{i=2}^{m} (1-\epsilon)^{i-2}$ voor $m = 2, 3, \ldots, N$ en $\mathbf{f} \in \mathbf{F}_D$. Nu geldt dat als $\mathbf{f} \in \mathbf{F}_D$ een Hamiltonkring is in de graaf G, dan $G_{\mathbf{f}} = c_N^1$, $\mathbf{x}(\mathbf{f})$ is een extreem punt van $\mathbf{X}$ (met $\mathbf{X}$ zoals in sectie 2.4.1) en $x_1(\mathbf{f}) = \frac{1}{d_N(\epsilon)}$. Omgekeerd, als $\mathbf{x}$ een extreem punt van $\mathbf{X}$ is en $\sum_{a \in A(1)} x_{1a} = \frac{1}{d_N(\epsilon)}$, dan is $f = \hat{M}(\mathbf{x})$ een Hamiltonkring in G. Voor de bewijzen verwijs ik naar blz. 46-49 van het boek.

2.6 Niet-stationaire strategieën

Om Markov beslissingsprocessesen op te lossen is het niet voldoende om alleen naar deterministische strategieën te kijken. In bepaalde gevallen hebben we stationaire strategieën nodig. We zullen in deze sectie bekijken of $\mathbf{F}_S$ wel voldoende is. Daartoe zullen we twee meer algemene klassen van "niet-stationaire" strategieën introduceren.
Met S_t, A_t geven we respectievelijk de toestand en de actie op tijdstip t aan. Laat $h_t = (s_0, a_0, s_1, a_1, \ldots, a_{t-1}, s_t)$ de *geschiedenis* tot tijdstip t zijn en laat H_t de verzameling van alle mogelijke geschiedenissen tot tijdstip t zijn. Met $A = \bigcup_{s \in S} A(s)$ geven we de totale actieruimte aan en $\mathcal{P}(A)$ is de verzameling van alle kansverdelingen van A. We definiëren een *beslisregel* op tijdstip t als een functie $f_t : H_t \to \mathcal{P}(A)$ zodanig dat
$f_t(h_t, a) = \begin{cases} P_{f_t}[A_t = a | h_t] & \text{als } a \in A(s_t) \\ 0 & \text{als } a \notin A(s_t) \end{cases}$
Een strategie π geven we nu aan met een rij beslisregels $\pi = (f_0, f_1, \ldots, f_t, \ldots)$ en $\mathbf{F}_B$ is de klasse van deze strategieën. Bij een Markov of geheugenloze strategie π hangt elke beslisregel f_t alleen af van de huidige toestand, dus $f_t(s_t, a) = P_{f_t}[A_t = a | S_t = s_t] = P_{f_t}[A_t = a | S_0 = s_0, A_0 = a_0, \ldots, A_{t-1} = a_{t-1}, S_t = s_t] = f_t(h_t, a)$. Met $\mathbf{F}_M$ geven we de klasse van Markov strategieën aan. Een stationaire strategie is een Markov strategie waarbij alle beslisregels onafhankelijk van de tijd zijn, dus $f_t = \mathbf{f}$. Een deterministische strategie is een stationaire strategie waarbij er voor iedere $s \in S$ een actie $a_s \in A(s)$ bestaat zodat $f(s, a) = 0$ voor $a \neq a_s$. Er geldt $\mathbf{F}_B \supseteq \mathbf{F}_M \supseteq \mathbf{F}_S \supseteq \mathbf{F}_D$.
De definitie van de verdisconteerde waarde in (1) gaat ook op voor strategieën $\pi \in \mathbf{F}_B$. De waarde van een strategie bij gemiddelde opbrengst moeten

we iets aanpassen: $v_\alpha(s,\pi) = \liminf_{T\to\infty}\left[\frac{1}{T+1}\sum_{t=0}^{T} E_{s\pi}[R_t]\right]$ met begintoestand s en strategie $\pi \in \mathbf{F}_B$.
Neem een willekeurige $\pi \in \mathbf{F}_B$, dan bestaat er voor iedere begintoestand $s_0 \in S$ een Markov strategie $\overline{\pi} \in \mathbf{F}_M$ zodat voor alle $a \in A, s \in S$ en $t = 0, 1, 2, \ldots$ geldt $P_\pi[S_t = s, A_t = a|S_0 = s_0] = P_{\overline{\pi}}[S_t = s, A_t = a|S_0 = s_0]$. En er geldt $v_\beta(s,\pi) = v_\beta(s,\overline{\pi})$, $v_\tau(s,\pi) = v_\tau(s,\overline{\pi})$ en $v_\alpha(s,\pi) = v_\alpha(s,\overline{\pi})$. Dus we kunnen ons beperken tot de klasse van Markov strategieën.

2.7 Strategieverbetering

In sectie 2.3 hebben we gezien dat voor het verdisconteerde model een optimale deterministische strategie $\mathbf{f}^0 \in \mathbf{F}_D$ kan worden verkregen met een Lineair Programmerings probleem. Nu zullen we laten zien dat deze ook optimaal is voor de klasse $\mathbf{F}_B$. Dat wil zeggen dat we laten zien, voor alle $s \in S$:

$$v_\beta(s,\mathbf{f}^0) = \sup_{\mathbf{F}_B} v_\beta(s,\pi) = \sup_{\mathbf{F}_M} v_\beta(s,\pi) = \max_{\mathbf{F}_S} v_\beta(s,\mathbf{f}) = \max_{\mathbf{F}_D} v_\beta(s,\mathbf{f}) \qquad (6)$$

De tweede en vierde gelijkheid in (6) hebben we reeds bewezen, we zullen nu dus de derde gelijkheid aantonen. Bij een strategie $\pi = (f_0, f_1, \ldots, f_t, \ldots) \in \mathbf{F}_M$ kan elke f_t gezien worden als een op één tijdstip toegepaste stationaire strategie $\mathbf{f}_t$. Dit geeft aanleiding tot de overgangsmatrix $P(f_t)$ en opbrengstvector $\mathbf{r}(f_t)$. Nu kunnen we de overgangsmatrix voor π als volgt definiëren: $P_t(\pi) = P(f_0)P(f_1)\ldots P(f_{t-1})$ voor $t = 1, 2, \ldots$ en $P_0(\pi) = I_N$. De verdisconteerde waarde vector kunnen we nu schrijven als

$$\mathbf{v}_\beta(\pi) = \sum_{t=o}^{\infty} \beta^t P_t(\pi)\mathbf{r}(f_t)$$

We zeggen dat π^1 π^2 domineert (strict domineert), genoteerd met $\pi^1 \geq \pi^2$ ($\pi^1 > \pi^2$), dan en slechts dan als $\mathbf{v}_\beta(\pi^1) \geq \mathbf{v}_\beta(\pi^2)$ ($\mathbf{v}_\beta(\pi^1) > \mathbf{v}_\beta(\pi^2)$).
Laat $\pi^0 = (f_0^0, f_1^0, \ldots, f_t^0, \ldots) \in \mathbf{F}_M$ zodat voor elke beslisregel f geldt $\pi^0 \geq (f, \pi^0)$, dan is π^0 een optimale strategie. Hierbij is (f, π^0) de strategie die f op tijdstip 0 gebruikt en π^0 vanaf tijdstip 1. De optimale deterministische strategie die we bij het LP-probleem in sectie 2.3 verkregen is optimaal voor de klasse $\mathbf{F}_B$. Voor een bewijs van deze twee beweringen verwijs ik naar bladzijde 58 en 59 van het boek.
We zullen nu een algoritme bespreken om een strategie lokaal te verbeteren. Hiervoor hebben we eerst de volgende twee resultaten nodig. Laat $\pi \in \mathbf{F}_M$ en f een beslisregel zodat $(f, \pi) > \pi$, dan $\mathbf{f} > \pi$ waarbij $\mathbf{f}$ de beslisregel f op ieder tijdstip gebruikt. Laat $\mathbf{f} \in \mathbf{F}_S$ zodat er voor tenminste één s een a_s bestaat zodat $r(s,a_s) + \beta\sum_{s'=1}^{N} p(s'|s,a_s)v_\beta(s',\mathbf{f}) > v_\beta(s,\mathbf{f})$. Dan

$\mathbf{g} \in \mathbf{F}_S > \mathbf{f}$ met

$$g(s,a) = \begin{cases} f(s,a) & \text{als het bovenstaande niet geldt} \\ 1 & \text{als het bovenstaande wel geldt en } a = a_s \\ 0 & \text{anders} \end{cases}$$

Ook voor een bewijs hiervan verwijs ik naar het boek, bladzijde 59 en 60. Het algoritme ziet er als volgt uit:

- **Stap 0** Laat $k = 0$ en neem een deterministische strategie $\mathbf{f}^0$. Bepaal $\mathbf{v}^0 = \mathbf{v}_\beta(\mathbf{f}^0) = \left[I - \beta P(\mathbf{f}^0)\right]^{-1} \mathbf{r}(\mathbf{f}^0)$.

- **Stap 1** In het algemeen hebben we nu $\mathbf{f}^k \in \mathbf{F}_D$ en $\mathbf{v}^k = \mathbf{v}_\beta(\mathbf{f}^k)$. Laat a_s^k de actie zijn die $\mathbf{f}^k$ selecteert voor toestand s. Als

  $$r(s,a_s^k) + \beta \sum_{s'=1}^{N} p(s'|s,a_s^k) v^k(s') = \max_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,a) v^k(s') \right\}$$

  geldt voor iedere $s \in S$, dan stoppen we. De strategie $\mathbf{f}^k$ is optimaal.

- **Stap 2** Laat $\overline{S}$ de niet-lege deelverzameling zijn van toestanden waarvoor de vergelijking in stap 1 niet geldt. Neem $\overline{a}_s^k = \arg\max_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,a) v^k(s') \right\}$ voor iedere $s \in \overline{S}$ en een nieuwe strategie $\mathbf{g} \in \mathbf{F}_D$ met

  $$g(s,a) = \begin{cases} f(s,a) & \text{als } s \notin \overline{S} \\ 1 & \text{als } s \in \overline{S} \text{ en } a = \overline{a}_s^k \\ 0 & \text{anders} \end{cases}$$

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

- **Stap 3** Neem $k = k + 1$ en ga terug naar Stap 1 met $\mathbf{f}^k = \mathbf{f}^{k+1}$ en $\mathbf{v}^k = \mathbf{v}^{k+1}$.

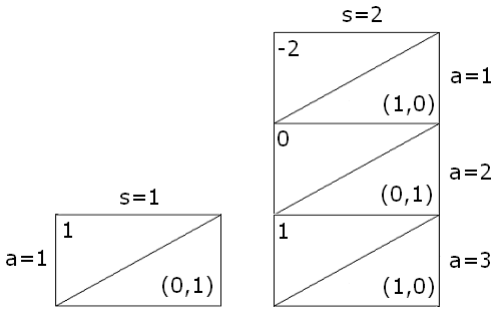
Dit algoritme eindigt in niet meer dan $\mu = \prod_{s=1}^{N} m(s)$ stappen en heeft dan een optimale deterministische strategie. Het is in zekere zin equivalent met de methode van Newton. Hiervoor verwijs ik naar blz.61-63 van het boek.

2.8 Uitwerking opgaven

Opgave 2 We beschouwen een Markov beslissings proces met oneindige horizon. De waarde van een stationaire strategie $\mathbf{f}$ met begintoestand s wordt gedefinieerd als:

$$v_\sigma(s, \mathbf{f}) = \sum_{t=0}^{\infty} E_{s\mathbf{f}}(R_t)$$

Allereerst zal ik een voorbeeld geven waaruit blijkt dat er drie stationaire strategieën $\mathbf{f}^1, \mathbf{f}^2, \mathbf{f}^3$ bestaan zodat geldt: $v(s, \mathbf{f}^1) = -\infty,\ -\infty < v(s, \mathbf{f}^2) < \infty,\ v(s, \mathbf{f}^3) = \infty$. Laat $S = \{1, 2\}, A(1) = \{1\}, A(2) = \{1, 2, 3\}$. De opbrengsten en overgangskansen zijn:



Elke rechthoek komt overeen met een bepaalde actie in een toestand. Linksboven staat de opbrengst van die bepaalde actie en rechtsonder staan de overgangskansen. Met $\mathbf{f}^1 = ((1), (1, 0, 0))\,,\ \mathbf{f}^2 = ((1), (0, 1, 0))\,,\ \mathbf{f}^3 = ((1), (0, 0, 1))$ krijgen we: $v(s, \mathbf{f}^1) = -\infty,\ v(s, \mathbf{f}^2) = 1,\ v(s, \mathbf{f}^3) = \infty$.

Stel dat voor elke strategie $\mathbf{f}$ en elk paar toestanden $(s, s') \in S \times S$ geldt:

$$\sum_{t=1}^{\infty} p^{(t)}(s'|s, \mathbf{f}) < \infty \tag{7}$$

Dan zijn alle toestanden transiënt. We zullen bewijzen dat nu geldt: $\mathbf{v}_\sigma(\mathbf{f}) = [I - P(\mathbf{f})]^{-1}\,\mathbf{r}(\mathbf{f})$.
Er geldt $\mathbf{v}_\sigma(\mathbf{f}) = \sum_{t=0}^{\infty} P^t(\mathbf{f})\mathbf{r}(\mathbf{f})$ en we zullen aantonen dat

$$\sum_{t=0}^{\infty} P^t(\mathbf{f}) = I + P(\mathbf{f}) + P^2(\mathbf{f}) + \ldots = [I - P(\mathbf{f})]^{-1}$$

Door het uitwerken van haakjes zien we dat $(I - P(\mathbf{f}))(I + P(\mathbf{f}) + P^2(\mathbf{f}) + \ldots + P^{t-1}(\mathbf{f})) = I - P^t(\mathbf{f})$. Uit (7) volgt $\left[P^t(\mathbf{f})\right]_{ss'} \to 0$ als $t \to \infty$, dus we krijgen $(I - P(\mathbf{f}))(I + P(\mathbf{f}) + P^2(\mathbf{f}) + \ldots) = I$. De determinant van I is gelijk aan 1, dus we zien dat $\det(I - P(\mathbf{f})) \times \det(I + P(\mathbf{f}) + P^2(\mathbf{f}) + \ldots) = \det(I) = 1$.

Dus $\det(I - P(\mathbf{f})) \neq 0$, dus de inverse van $(I - P(\mathbf{f}))$ bestaat en is gelijk aan $(I + P(\mathbf{f}) + P^2(\mathbf{f}) + \ldots)$.

Als we in plaats van (7) veronderstellen dat er scalairen $\mu_1, \mu_2, \ldots, \mu_N > 0$ en $\gamma \in [0, 1)$ bestaan zodat de overgangskansen voldoen aan

$$\sum_{s'=1}^{N} p(s'|s, a)\mu_{s'} \leq \gamma \mu_s \tag{8}$$

voor alle $a \in A(s)$ en $s, s' \in S$, dan wordt het proces contracterend genoemd. Er geldt dat een contracterend proces ook transiënt is. Het is namelijk na te gaan[2] dat in een contracterend model elke strategie transiënt is. Optimale transiënte strategieën zijn ook optimaal in de klasse van alle strategieën.

Tot slot zullen we bewijzen dat als we de overgangskansen als volgt herdefiniëren

$$\overline{p}(s'|s, a) := \beta p(s'|s, a)$$

voor alle $a \in A(s)$ en $s, s' \in S$, het verdisconteerde model gezien kan worden als een speciaal geval van het contracterende model.
We nemen $\mu_i = 1$ voor $i = 1, 2, \ldots, N$ en $\gamma = \beta$. Dan geldt:

$$\sum_{s'=1}^{N} \overline{p}(s'|s, a) = \sum_{s'=1}^{N} \beta p(s'|s, a) = \beta \sum_{s'=1}^{N} p(s'|s, a) \leq \beta = \gamma \mu_s$$

Opgave 6 We bewijzen het volgende lemma (lemma 2.4.1): Laat P de overgangsmatrix zijn van een irreducibele Markov keten en $Q = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t$ de bijbehorende Cesaro-limiet matrix. Dan geldt:

1. Q heeft identieke rijen
2. Laat $\mathbf{q} = (q_1, \ldots, q_N)$ een rij van Q zijn. Dan is elk element van $\mathbf{q}$ strict positief en $\mathbf{q}$ is de unieke oplossing van de lineaire vergelijkingen:

$$\mathbf{q}P = \mathbf{q} \tag{9}$$
$$\mathbf{q}\mathbf{1} = \mathbf{1} \tag{10}$$

   De vector $\mathbf{q}$ wordt de stationaire verdeling van een irreducibele Markov keten genoemd.

Laat $s \in S$. Onze Markov keten is irreducibel, dus de keten is één recurrente klasse. Stel dat deze klasse periode d heeft. Laat $l \in S$ en stel dat l in

[2] zie L.C.M. Kallenberg. *Linear Programming and Finite Markovian Control Problems.* Mathematical Center Tracts 148, Amsterdam (1983). Stelling 3.4.2, bladzijde 42.

de cyclische deelverzameling S_k zit. Zij $f_l(t) = \min_{s\in S} p(l|s)^{td}$ en $g_l(t) = \max_{s\in S} p(l|s)^{td}$, dan geldt:

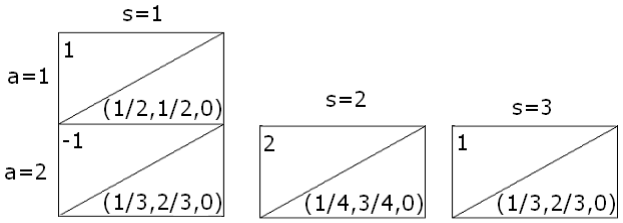
$$\pi_l := \lim_{t\to\infty} p(l|l)^{td} \geq \lim_{t\to\infty} f_l(td) = \lim_{t\to\infty} g_l(td) > 0$$

Omdat $p(l|s)^{(n)} = 0$ als $n \neq r(s,l)(mod\ d)$ kunnen we nu schrijven:

$$\begin{aligned} q_l &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p(l|s)^{(k)} = \lim_{m\to\infty} \frac{1}{md} \sum_{k=0}^{m-1} p(l|s)^{(kd+r(s,l))} \\ &= \frac{1}{d} \lim_{k\to\infty} p(l|s)^{(kd+r(s,l))} = \frac{1}{d}\pi_l > 0 \end{aligned}$$

En deze is dus onafhankelijk van s. Dus Q heeft identieke rijen[3].
We hebben al gezien dat ieder element van $\mathbf{q}$ strict positief is. Er geldt $\sum_{s\in S} q_s = 1$ en $p(s'|s)^{n+1} = \sum_{t\in S} p(t|s)^n p(s'|t)$. Als $n \to \infty$, krijgen we dan $q_s = \sum_{s'\in S} q_{s'} p(s|s')$ voor $s \in S$. In vectornotatie krijgen we $\mathbf{q} = \mathbf{q}P$. Dus $\mathbf{q}$ is een stationaire verdeling van de Markov keten. Stel dat $\mathbf{p}$ ook een stationaire verdeling is, dan geldt $p_s = \sum_{s'\in S} p_{s'} p(s|s')^n$ en als $n \to \infty$ in de Cesaro-limiet krijgen we $p_s = \sum_{s'\in S} p_{s'} q_s = q_s$. Dus de stationaire verdeling is uniek.

Opgave 16 Een Markov beslissings proces heet *unichained* als voor elke deterministische stationaire strategie $\mathbf{f}$ de Markov keten voortgebracht door $\mathbf{f}$ één ergodic klasse bevat. Allereerst zal ik een voorbeeld geven waaruit blijkt dat een unichained proces niet irreducibel hoeft te zijn. Laat $S = \{1,2,3\}$, $A(1) = \{1,2\}$, $A(2) = A(3) = \{1\}$ en:



Dit model is duidelijk niet irreducibel, we kunnen nooit vanuit toestand 1 en 2 in toestand 3 terecht komen. Toestanden 1 en 2 vormen een ergodic klasse en het model is dus unichained.
We kunnen lemma 2.4.1 (zie opgave 6) toepassen op het unichained geval. We hebben dan niet te maken met een irreducibele, maar met een unichained Markov keten. Het bewijs gaat analoog aan het bewijs in opgave 6 met

[3]Zie ook L.C.M. Kallenberg, dictaat Besliskunde 1, voorjaar 2004. Stelling 6.12, bladzijde 165.

$q_s = 0$ voor transiënte toestanden s.
We beschouwen een unichained model met gemiddelde opbrengst en de verzameling $\mathbf{X}$ (als in sectie 2.4.1 en opgave 21). Neem $\mathbf{x} \in \mathbf{X}$ en laat $\mathbf{f_x}$ uit $\mathbf{x}$ geconstrueerd zijn volgens

$$f_{\mathbf{x}}(s,a) = \begin{cases} x_{sa}/x_s & \text{als } x_s = \sum_{a \in A(s)} x_{sa} > 0 \\ \text{willekeurig} & \text{als } x_s = 0 \end{cases}$$

We zullen bewijzen dat als $\mathbf{x}^0$ een optimale oplossing is van het LP-probleem

$$\max \left\{ \sum_{s=1}^{N} \sum_{a \in A(s)} r(s,a) x_{sa} \mid \mathbf{x} \in \mathbf{X} \right\}$$

dan is $\mathbf{f}_{\mathbf{x}^0}$ optimaal.
Volgens het zojuist genoemde lemma 2.4.1 heeft $Q(\mathbf{f})$ identieke rijen. Dus $v_\alpha(s, \mathbf{f})$ is onafhankelijk van de begintoestand s. Er geldt nu voor iedere $s' \in S$ $v_\alpha(s', \mathbf{f}) = [Q(\mathbf{f})\mathbf{r}(\mathbf{f})]_{s'} = \sum_{s=1}^{N} q_s(\mathbf{f}) r(s, \mathbf{f})$. Dus er geldt voor iedere $s' \in S, \mathbf{f} \in \mathbf{F}_S$

$$v_\alpha(s', \mathbf{f}) = \sum_{s=1}^{N} \sum_{a \in A(s)} r(s,a) q_s(\mathbf{f}) f(s,a) = \sum_{s=1}^{N} \sum_{a \in A(s)} r(s,a) x_{sa}(\mathbf{f})$$

Stel dat $\mathbf{f}_{\mathbf{x}^0}$ niet optimaal is. Dan bestaat er een $\hat{\mathbf{f}} \in \mathbf{F}_S$ zodanig dat $v_\alpha(s', \hat{\mathbf{f}}) > v_\alpha(s', \mathbf{f}_{\mathbf{x}^0})$. Maar omdat $\mathbf{x}^0 = M(\hat{M}(\mathbf{x}^0))$ optimaal is en

$$\sum_{s=1}^{N} \sum_{a \in A(s)} r(s,a) x_{sa}(\hat{\mathbf{f}}) > \sum_{s=1}^{N} \sum_{a \in A(s)} r(s,a) x_{sa}(\mathbf{f}_{\mathbf{x}^0}) = \sum_{s=1}^{N} \sum_{a \in A(s)} r(s,a) x^0_{sa}$$

hebben we een tegenspraak. Dus $\mathbf{f}_{\mathbf{x}^0}$ is optimaal. Als $x_s = 0$ is s transiënt en maakt het niet uit wat we doen.

We zullen nu lemma 2.5.1 (zie sectie 2.5) nagaan voor een unichained Markov Beslissingsproces. Punt 1 en 2 zijn per definitie duidelijk. Volgens opgave 21 (zie hierna) geldt dat voor een extreem punt $\mathbf{x} \in \mathbf{X}$ in een unichained proces de cardinaliteiten van $\mathbf{S_x}$ en $\overline{\mathbf{S}}_{\mathbf{x}}$ gelijk zijn. Dus per toestand is er maar één $a \in A(s)$ waarvoor geldt $x_{sa} > 0$. Hieruit volgt dat $\mathbf{f_x} = \hat{M}(\mathbf{x})$ deterministisch is. Tenslotte kunnen we punt 4 als volgt inzien. Als $\mathbf{f} \in \mathbf{f}_D$ een Hamiltonkring beschrijft, dan is de hele Markov keten één ergodic klasse. Deze klasse wordt volgens opgave 21 uniek geïdentificeerd door een extreem punt $\mathbf{x} \in \mathbf{X}$.

Opgave 21 We beschouwen de verzameling $\mathbf{X} = \{\mathbf{x}|W\mathbf{x} = \mathbf{0}, \mathbf{1}^T\mathbf{x} = 1, \mathbf{x} \geq 0\}$, waarbij W een $N \times m$-matrix is met op plaats $(s', (s, a))$ het element $w_{s'(s,a)} = \delta(s, s') - p(s'|s, a)$. Voor $\mathbf{x} \in \mathbf{X}$ definiëren we $\mathbf{S_x} = \left\{s \in S | \sum_{a \in A(s)} x_{sa} > 0\right\}$ en $\overline{\mathbf{S}}_\mathbf{x} = \{(s, a)|x_{sa} > 0, a \in A(s), s \in S\}$. We zeggen dat $\mathbf{x}$ een unieke ergodic klasse identificeert als

1. De cardinaliteiten van $\mathbf{S_x}$ en $\overline{\mathbf{S}}_\mathbf{x}$ zijn gelijk
2. Alle toestanden van $\mathbf{S_x}$ vormen een ergodic klasse onder een stationaire strategie $\mathbf{f_x}$ gedefinieerd door:

$$f_\mathbf{x}(s, a) = \begin{cases} 1 & \text{als } (s, a) \in \overline{\mathbf{S}}_\mathbf{x} \\ 0 & \text{als } (s, a) \notin \overline{\mathbf{S}}_\mathbf{x}, s \in \mathbf{S_x} \\ \text{willekeurig} & \text{als } s \notin \mathbf{S_x} \end{cases}$$

We zullen bewijzen dat elk extreem punt $\mathbf{x} \in \mathbf{X}$ een unieke ergodic klasse identificeert.

$$A_{sa} = \begin{pmatrix} -p(1|s, a) \\ \vdots \\ 1 - p(s|s, a) \\ \vdots \\ -p(N|s, a) \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

Voor een punt $\mathbf{x} \in \mathbf{X}$ geldt

$$\sum_{(s,a) \in \overline{\mathbf{S}}_\mathbf{x}} A_{sa} x_{sa} = b \tag{11}$$

Uit $W\mathbf{x} = \mathbf{0}$ volgt dat $p(s'|s, a) = 0$ voor $(s, a) \in \overline{\mathbf{S}}_\mathbf{x}$ en $s' \notin \mathbf{S_x}$. Laat $\overline{\mathbf{S}}^1_\mathbf{x}$ een deelverzameling van $\overline{\mathbf{S}}_\mathbf{x}$ zijn zodanig dat deze voor iedere $s \in \mathbf{S_x}$ precies één element (s, a_s) bevat zodat de verzameling $\left\{p(s'|s, a)|(s, a) \in \overline{\mathbf{S}}^1_\mathbf{x}, s' \in \mathbf{S_x}\right\}$ de overgangskansen van een Markov keten bevat. Deze Markov keten heeft tenminste één ergodic klasse met toestanden $\mathbf{S}^2_\mathbf{x} \subset \mathbf{S_x}$ en overgangskansen die gespecificeerd worden door $\overline{\mathbf{S}}^2_\mathbf{x} = \left\{(s, a) \in \overline{\mathbf{S}}^1_\mathbf{x} | s \in \mathbf{S}^2_\mathbf{x}\right\}$.
Laat y_s de kans zijn dat iemand het systeem in toestand s aantreft met $s \in \mathbf{S}^2_\mathbf{x}$. Dan is $z_{sa} = y_s$ voor $(s, a) \in \overline{\mathbf{S}}^2_\mathbf{x}$ de unieke oplossing van

$$\sum_{(s,a) \in \overline{\mathbf{S}}^2_\mathbf{x}} A_{sa} z_{sa} = b \tag{12}$$

Als we (12) van (11) aftrekken krijgen we (omdat $\overline{\mathbf{S}}^2_\mathbf{x} \subset \overline{\mathbf{S}}_\mathbf{x}$)

$$\sum_{(s,a) \in \overline{\mathbf{S}}_\mathbf{x} - \overline{\mathbf{S}}^2_\mathbf{x}} A_{sa} x_{sa} + \sum_{(s,a) \in \overline{\mathbf{S}}^2_\mathbf{x}} A_{sa}(x_{sa} - z_{sa}) = 0 \tag{13}$$

Omdat $\mathbf{x}$ een extreem punt van $\mathbf{X}$ is, is de verzameling $\left\{A_{sa}|(s,a) \in \overline{\mathbf{S}}_{\mathbf{x}}\right\}$ lineair onafhankelijk. Dus elke coefficiënt in (13) moet nul zijn. Dus $\overline{\mathbf{S}}_{\mathbf{x}} = \overline{\mathbf{S}}_{\mathbf{x}}^2$ en $x_{sa} = z_{sa}$ voor $(s,a) \in \overline{\mathbf{S}}_{\mathbf{x}}$. Hiermee is bewezen dat $\mathbf{x}$ een unieke ergodic klasse identificeert.

Opgave 25 Een Markov beslissings proces wordt communicerend genoemd als er voor elk paar $(s,s') \in S \times S$ een strategie $\mathbf{f} \in \mathbf{F}_D$ en een geheel getal $\tau \geq 1$ bestaan (beide mogen afhangen van (s,s')) zodanig dat $p^{(\tau)}(s'|s,\mathbf{f}) > 0$. We beschouwen een communicerend Markov beslissings model met gemiddelde opbrengst. Allereerst zullen we bewijzen dat de overgangsmatrix $P(\mathbf{f})$ irreducibel is voor alle stationaire strategieën waarvoor geldt $f(s,a) > 0$ voor alle $a \in A(s), s \in S$.
Het is gemakkelijk na te gaan dat er een stationaire strategie $\mathbf{f}^0$ bestaat zodat $P(\mathbf{f}^0)$ irreducibel is: volgens bovenstaande definitie bestaat er voor elk paar (s,s') een deterministische strategie $\mathbf{f}$ met $p^{(\tau)}(s'|s,\mathbf{f}) > 0$ voor een geheel getal $\tau \geq 1$. Als we deze strategieën combineren, krijgen we een stationaire strategie $\mathbf{f}^0$ waarvoor geldt $p^{(\tau)}(s'|s,\mathbf{f}^0) > 0$ voor alle $s, s' \in S$ voor een geheel getal $\tau \geq 1$ en dus is $P(\mathbf{f}^0)$ irreducibel. Neem nu een strategie $\mathbf{f}^*$ met $f^*(s,a) > 0$ voor alle $s \in S, a \in A(s)$. Er geldt $p^{(\tau)}(s'|s,\mathbf{f}^0) = \sum_a p^{(\tau)}(s'|s,a) f^0(s,a) > 0$ voor een geheel getal $\tau \geq 1$ en voor iedere $s, s' \in S$. Omdat $f^*(s,a) > 0$ geldt nu $\sum_a p^{(\tau)}(s'|s,a) f^*(s,a) = p^{(\tau)}(s'|s,\mathbf{f}^*) > 0$. Dus $P(\mathbf{f}^*)$ is irreducibel.

We zullen vervolgens bewijzen dat een optimale oplossing voor een communicerend model met gemiddelde opbrengst gevonden kan worden uit een optimale oplossing van hetzelfde primaal-duale paar van LP dat wordt gebruikt om een irreducibel model op te lossen.
Bij het algemene LP-probleem voor Markov beslissings processen hebben we variabelen $g, w \in \mathbf{R}^S$. Dit probleem ziet er als volgt uit:

$$\min \sum_{s \in S} \beta_s g_s \tag{14}$$

onder de voorwaarden:

(a) $g_s - \sum_{s'} g_{s'} p(s'|s,a) \geq 0; \quad a \in A(s), s, s' \in S$

(b) $g_s + w_s - \sum_{s'} w_{s'} p(s'|s,a) \geq r(s,a); \quad a \in A(s), s, s' \in S$

met constanten $\beta_s > 0$ voor $s \in S$ en $\sum_{s \in S} \beta_s = 1$. We hebben duale variabelen $x = \{x_{sa}|a \in A(s), s \in S\}$ en $y = \{y_{sa}|a \in A(s), s \in S\}$. Het duale probleem ziet er als volgt uit:

$$\max \sum_{s \in S} \sum_{a \in A(s)} r(s,a) x_{sa} \tag{15}$$

onder de voorwaarden:

(c) $\sum_s \sum_a (\delta(s,s') - p(s'|s,a))x_{sa} = 0; \quad s,s' \in S, a \in A(s)$

(d) $\sum_a x_{s'a} + \sum_s \sum_a (\delta(s,s') - p(s'|s,a))y_{sa} = \beta_{s'}; \quad s,s' \in S, a \in A(s)$

(e) $x_{sa}, y_{sa} \geq 0; \quad s \in S, a \in A(s)$

Bij het irreducibele geval hebben we primale variabelen g (een scalair) en $w \in \mathbf{R}^S$. Dit probleem ziet er als volgt uit:

$$\min g \tag{16}$$

onder de voorwaarden:

(f) $g + w_s - \sum_{s'} w_{s'} p(s'|s,a) \geq r(s,a); \quad a \in A(s), s, s' \in S$

We hebben duale variabelen $x = \{x_{sa} | a \in A(s), s \in S\}$ en het duale probleem ziet er als volgt uit:

$$\max \sum_{s \in S} \sum_{a \in A(s)} r(s,a) x_{sa} \tag{17}$$

onder de voorwaarden:

(g) $\sum_s \sum_a (\delta(s,s') - p(s'|s,a))x_{sa} = 0; \quad s,s' \in S, a \in A(s)$

(h) $\sum_s \sum_a x_{sa} = 1; \quad s \in S, a \in A(s)$

(i) $x_{sa} \geq 0; \quad s \in S, a \in A(s)$

Laat (g^0, w^0) en x^0 optimale oplossingen van (16) en (17) zijn voor een communicerend model. Definieer $g^* \in \mathbf{S}$ door $g^*_s = g^0$ voor $s \in S$. Omdat optimale oplossingen voor communicerende modellen onafhankelijk zijn van de begintoestand is (g^*, w^0) optimaal voor (14). Omdat x dezelfde voorwaarden in (15) en (17) heeft, is x^0, die optimaal is voor (17), ook optimaal voor (15). We moeten nu alleen nog laten zien dat we in (15) een oplossing van de vorm (x^0, y^0) krijgen. Dus we moeten laten zien dat er een niet-negatieve y^0 bestaat die aan beperkingen (d) en (e) voldoet. Voor deze y^0 moet gelden:

$$\sum_{a \in A(s')} \left(x^0_{s'a} + y^0_{s'a}\right) - \sum_{s \in S} \sum_{a \in A(s)} y^0_{sa} p(s'|s,a) = \beta_{s'}$$

met $s' \in S$. We definiëren $b \in \mathbf{R}^S$ door $b_s = \beta_s - \sum_{a \in A(s)} x^0_{sa}$ met $s \in S$ waarbij $\sum_{s \in S} b_s = 0$. We zullen aantonen dat er nu een y^0 bestaat zodanig dat $y^0_{as} \geq 0$ voor $a \in A(s), s \in S$ en

$$\sum_{a \in A(s')} y^0_{as'} - \sum_{s \in S} \sum_{a \in A(s)} y^0_{as} p(s'|s,a) = b_{s'}$$

met $s' \in S$.
Zoals we gezien hebben bestaat er een stationaire strategie $\mathbf{f}^0$ zodat $P(\mathbf{f}^0)$ irreducibel is. Laat $\pi^0 > 0$ de evenwichtsverdeling zijn voor $P(\mathbf{f}^0)$ en $Z(\mathbf{f}^0) = \left[I - P(\mathbf{f}^0) + Q(\mathbf{f}^0)\right]^{-1}$ de fundamentele matrix, met Q de Cesaro-limiet matrix. We definiëren $d \in \mathbf{R}^S$ door $d = bZ(\mathbf{f}^0) + c\pi^0$, met $c \geq 0$ groot genoeg zodat $d \geq 0$. Neem $y^0_{as} = d_s f^0(s,a)$. Omdat d en $\mathbf{f}^0$ niet-negatief zijn, geldt $y^0 \geq 0$. Nu hebben we

$$d_{s'} - \sum_{s \in S} d_s p(s'|s, \mathbf{f}^0) = b_{s'}$$

met $s' \in S$. Deze y^0 voldoet omdat $\pi^0[I - P(\mathbf{f}^0)] = 0$ en omdat $Z(\mathbf{f}^0)[I - P(\mathbf{f}^0)] = I - Q(\mathbf{f}^0)$ en $bQ(\mathbf{f}^0) = 0$.
Omgekeerd, als (g^*, w^0) en (x^0, y^0) optimaal zijn voor (14) en (15), dan heeft g^* identieke componenten. In (16) is de variabele g een scalair. Omdat we in (14) en (16) dezelfde beperkingen hebben is (g^*_1, w^0) optimaal voor (16). Omdat x in (15) en (17) aan dezelfde voorwaarden moet voldoen is het duidelijk dat x^0 optimaal is voor (17).

3 Stochastische Spelen

3.1 Verdiscontering

In dit hoofdstuk zal ik de theorie van Stochastische Spelen behandelen. Deze spelen kunnen gezien worden als competitieve Markov beslissings processen waarbij er twee beslissers (meestal spelers genoemd) zijn. Zij kiezen onafhankelijk van elkaar hun eigen acties, maar de overgangskansen en opbrengsten hangen in het algemeen van beide acties af. We nemen aan dat de spelers elkaars opbrengstfuncties kennen, maar dat ze alleen hun eigen opbrengst willen maximaliseren. In deze sectie zullen we het verdisconteerde model bekijken.
Wederom hebben we een proces dat zich op tijdstip t in toestand $s \in S = \{1, 2, \ldots, N\}$ bevindt. In s kiezen spelers 1 en 2 onafhankelijk acties $a^1 \in A^1(s)$ en $a^2 \in A^2(s)$ en ze ontvangen $r^1(s, a^1, a^2)$ respectievelijk $r^2(s, a^1, a^2)$. De overgangskansen hebben nu de vorm $p(s'|s, a^1, a^2)$. We hebben de verzameling van stationaire strategieën $\mathbf{F}_S$ voor speler 1 zoals die in sectie 2.1 is gedefinieerd. Analoog hebben we de verzameling stationaire strategieën $\mathbf{G}_S$ voor speler 2. Met R_t^1 en R_t^2 geven we de opbrengst op tijdstip t aan voor speler 1 en 2. We noteren de directe opbrengst vector voor speler $k = 1, 2$ met $\mathbf{r}^k(\mathbf{f}, \mathbf{g})$ die hoort bij het paar strategieën $(\mathbf{f}, \mathbf{g}) \in \mathbf{F}_S \times \mathbf{G}_S$. De verwachte opbrengst op tijdstip t voor speler $k = 1, 2$ met strategieën $(\mathbf{f}, \mathbf{g})$ en begintoestand s schrijven we als $E_{s\mathbf{fg}}(R_t^k)$. We definiëren de verdisconteerde waarde van $(\mathbf{f}, \mathbf{g})$ voor speler k als volgt:

$$v_\beta^k(s, \mathbf{f}, \mathbf{g}) = \sum_{t=0}^{\infty} \beta^t E_{s\mathbf{fg}}(R_t^k) \tag{1}$$

waarbij $\beta \in [0, 1)$. We zullen hier alleen stationaire strategieën bekijken. We kunnen nu niet zoals in hoofdstuk 2 de waarde van een strategie voor bijvoorbeeld speler 1 maximaliseren, omdat deze in het algemeen zal afhangen van de strategie van speler 2.
Het paar $(\mathbf{f}^0, \mathbf{g}^0) \in \mathbf{F}_S \times \mathbf{G}_S$ heet een *Nash evenwichtspaar* als:

$$\mathbf{v}_\beta^1(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta^1(\mathbf{f}^0, \mathbf{g}^0) \ \forall \mathbf{f} \in \mathbf{F}_S$$
$$\mathbf{v}_\beta^2(\mathbf{f}^0, \mathbf{g}) \leq \mathbf{v}_\beta^2(\mathbf{f}^0, \mathbf{g}^0) \ \forall \mathbf{g} \in \mathbf{G}_S$$

Het is dus nooit optimaal voor speler 1 (2) om als de ander strategie $\mathbf{g}^0$ ($\mathbf{f}^0$) gebruikt af te wijken van $\mathbf{f}^0$ ($\mathbf{g}^0$). Helaas kunnen er meerdere evenwichtsparen bestaan die verschillende waarden hebben. Deze moeilijkheid verdwijnt als we naar "nulsomspelen" kijken. In dat geval geldt $r^1(s, a^1, a^2) + r^2(s, a^1, a^2) = 0$. Dan hebben we één opbrengstfunctie
$r(s, a^1, a^2) := r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)$ en één waardefunctie $\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) := \mathbf{v}_\beta^1(\mathbf{f}, \mathbf{g}) = -\mathbf{v}_\beta^2(\mathbf{f}, \mathbf{g})$. Als we nu een evenwichtspaar $(\mathbf{f}^0, \mathbf{g}^0)$ hebben geldt

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}) \tag{2}$$

voor alle $\mathbf{f} \in \mathbf{F}_S$ en $\mathbf{g} \in \mathbf{G}_S$. We noemen $\mathbf{f}^0$ ($\mathbf{g}^0$) een optimale strategie voor speler 1 (2). Als $(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \in \mathbf{F}_S \times \mathbf{G}_S$ een ander paar optimale strategieën is, geldt $\mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}}) = \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0)$, zoals in opgave 3 van dit hoofdstuk bij de uitgewerkte opgaven te zien is. Net zoals bij de Markov beslissings theorie geldt voor de verwachte waarde:

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = [I - \beta P(\mathbf{f}, \mathbf{g})]^{-1}\mathbf{r}(\mathbf{f}, \mathbf{g}) \tag{3}$$

3.1.1 Matrix Spelen

Tot zover hebben we stochastische spelen gezien als een generalisatie van Markov beslissings processen. We kunnen stochastische nulsomspelen echter ook beschouwen als een generalisatie van matrix spelen. We hebben dan voor elke toestand het volgende matrix spel

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

waarbij $m^k(s)$ het aantal acties voor speler k in toestand s is. In s spelen spelers 1 en 2 het matrix spel $R(s)$ waarbij ze niet alleen letten op $r(s, a^1, a^2)$ maar ook op $p(s'|s, a^1, a^2)$ en $R(s')$. We veronderstellen dat $\mathbf{v}_\beta$ bestaat. Als we kunnen aannemen dat we weten hoe we optimaal moeten spelen vanaf het volgende tijdstip, dan spelen we op het huidige tijdstip, in toestand s, het matrixspel

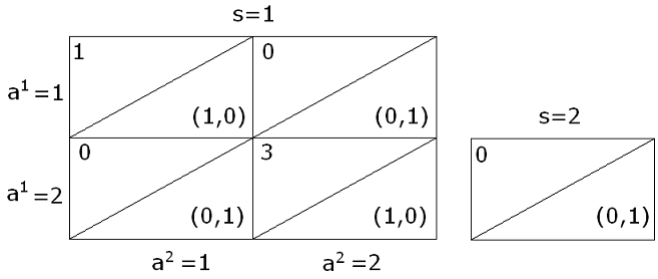
$$R(s, \mathbf{v}_\beta) = \left[r(s, a^1, a^2) + \beta \sum_{s' \in S} p(s'|s, a^1, a^2)\mathbf{v}_\beta(s')\right]_{a^1=1, a^2=2}^{m^1(s), m^2(s)} \tag{4}$$

De theorie van Shapley zegt dat $\mathbf{v}_\beta = \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0)$ met $(\mathbf{f}^0, \mathbf{g}^0)$ een evenwichtspaar bestaat en dat het de unieke oplossing is van $v(s) = \text{val}\,[R(s, \mathbf{v})]$ voor alle $s \in S$ met $\mathbf{v}^T = (v(1), \ldots, v(N))^T$. Met val$[A]$ geven we de waarde van het matrix-spel A aan. Ook zijn er volgens deze theorie optimale strategieën $\mathbf{f}^0$ en $\mathbf{g}^0$ voor respectievelijk speler 1 en 2. Voor een bewijs van de theorie van Shapley verwijs ik naar bladzijde 175-179 van het boek.

3.2 Lineaire Programmering

We weten dus dat de waardevector en optimale strategieën bestaan, maar de theorie van Shapley vertelt ons niet hoe we deze kunnen vinden. Bij het oplossen in de Markov beslissings theorie hebben we gebruik gemaakt van Lineaire Programmering (LP). Je zou daarom kunnen verwachten dat dit ook kan bij het oplossen van Stochastische Spelen. Helaas is dit niet het geval, dit zal ik laten zien aan de hand van een voorbeeld.

Laat $S = \{1, 2\}, A^1(1) = A^2(1) = \{1, 2\}, A^1(2) = A^2(2) = \{1\}, \beta = \frac{1}{2}$. De opbrengsten en overgangskansen zijn:



Dit plaatje moet net zo gelezen worden als de plaatjes in het vorige hoofdstuk, daarbij geldt dat de rijen overeenkomen met de acties van speler 1 en de kolommen met de acties van speler 2. Als je in toestand 2 zit, blijf je daar met kans 1 inzitten. De opbrengst van toestand 2 is 0, dus de waardevector voor toestand 2 is $v_{\frac{1}{2}}(2) = 0$. Volgens Shapley kunnen we de waarde voor toestand 1 als volgt schrijven:

$$v_{\frac{1}{2}}(1) = v = \text{val} \begin{bmatrix} 1 + \frac{1}{2}v & 0 \\ 0 & 3 + \frac{1}{2}v \end{bmatrix}$$

Het is duidelijk dat v niet-negatief is. Uit de speltheorie[4] weten we dat we de waarde van dit matrixspel als volgt kunnen berekenen: $v = \frac{(1+\frac{1}{2}v)(3+\frac{1}{2}v)}{4+v}$. Dit kunnen we omschrijven naar: $3v^2 + 8v - 12 = 0$. En dit kunnen we oplossen: $v = \frac{1}{3}(-4 + 2\sqrt{13})$. Dus de waardevector is nu: $\mathbf{v}_{\frac{1}{2}} = \left(\frac{1}{3}(-4 + 2\sqrt{13}), 0\right)^T$. We zien dat in de oplossing een wortel voorkomt. Hieruit kunnen we concluderen dat dit probleem niet met LP opgelost had kunnen worden. In LP werk je namelijk met rationale functies en rationale getallen. We kunnen ons nu afvragen of er misschien bepaalde gevallen zijn waarvoor je wel LP kan gebruiken. Dit blijkt inderdaad zo te zijn.

3.2.1 Overheersende beslisser

Als in een spel geldt dat de overgangskansen niet van de acties van speler 2 afhangen, dus als $p(s'|s, a^1, a^2) \equiv p(s'|s, a^1)$, dan blijkt het zo te zijn dat we Lineaire Programmering kunnen gebruiken. We zeggen dat in dit geval het spel overheerst wordt door speler 1. De verwachte waarde uit (3) wordt nu:

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = [I - \beta P(\mathbf{f})]^{-1}\mathbf{r}(\mathbf{f}, \mathbf{g}) \tag{5}$$

Dit probleem lijkt sterk op de problemen uit de Markov beslissingstheorie. Het enige verschil is dat de opbrengsten nu afhangen van de (niet-bekende)

[4] zie Appendix G in het boek

strategie $\mathbf{g}$ van speler 2. Het LP-probleem ziet er als volgt uit:

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

onder de voorwaarden:

(a) $v(s) \geq [R(s)\mathbf{g}(s)]_{a^1} + \beta \sum_{s'=1}^{N} p(s'|s,a^1)v(s'); \quad s \in S, a^1 \in A^1(s)$

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

(c) $g(s,a^2) \geq 0; \quad s \in S, a^2 \in A^2(s)$

Een optimale oplossing van het probleem geeft ook meteen een optimale strategie voor speler 2. De voorwaarden (b) en (c) zorgen ervoor dat deze strategie goed gedefinieerd is. Zodra de strategie voor speler 2 bekend is, is bovenstaand probleem hetzelfde als het LP-probleem in sectie 2.3. Het duale probleem ziet er als volgt uit:

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

onder de voorwaarden:

(d) $\sum_{s=1}^{N} \sum_{a^1 \in A^1(s)} [\delta(s,s') - \beta p(s'|s,a)]\, x_{sa^1} = \frac{1}{N}; \quad s' \in S$

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

(f) $x_{sa^1} \geq 0; \quad s \in S, a^1 \in A^1(s)$

Met behulp van het duale probleem kunnen we een optimale strategie $\mathbf{f}^0$ voor speler 1 berekenen, namelijk: $f^0(s,a) = \frac{x^0_{sa^1}}{x^0_s}$ met $s \in S$ en $a^1 \in A^1(s)$. In opgave 5 bij de uitgewerkte opgaven zal ik een voorbeeld van het spel met overheersende beslisser geven.

3.2.2 Toestand onafhankelijk en gescheiden opbrengst

Ook kunnen we het stochastische spel met LP oplossen als we het volgende aannemen:

$$r(s,a^1,a^2) = c(s) + \rho(a^1,a^2) \quad \text{en} \quad p(s'|s,a^1,a^2) = p(s'|a^1,a^2)$$

met $a^1 \in A^1(s), a^2 \in A^2(s), s, s' \in S$. De overgangskansen zijn dus niet meer afhankelijk van de huidige toestand. Dit heeft alleen betekenis als $m^1(s) = \mu$ en $m^2(s) = \nu$ voor alle $s \in S$. Er geldt dat alle opbrengsten een som zijn van een bedrag voor de huidige toestand en een bedrag voor de twee gekozen

acties.
Analoog aan (4) hebben we het volgende matrixspel

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

waarbij $\mathbf{c}^T = (c(1), c(2), \ldots, c(N))$. Laat $\rho := \text{val}\,[R(\mathbf{c})]$, en $\mathbf{x}^0 = (x_1^0 \ldots x_\mu^0)$ en $\mathbf{y}^0 = (y_1^0 \ldots y_\nu^0)^T$ optimale strategieën voor het matrixspel $R(\mathbf{c})$. Volgens de theorie van Shapley geldt nu dat $\mathbf{f}^0(s) = \mathbf{x}^0$ en $\mathbf{g}^0(s) = \mathbf{y}^0$ optimale strategieën voor het stochastische spel zijn. In opgave 6 bij de uitgewerkte opgaven zal ik laten zien dat

$$\mathbf{v}_\beta = \mathbf{c} + \left(\frac{\rho}{1-\beta} \right) \mathbf{1} \tag{6}$$

3.2.3 Wisselende overheersende beslisser

We kunnen ook LP gebruiken voor het oplossen van stochastische spelen waarbij de overgangskansen in sommige toestanden alleen van de ene en in de overige toestanden alleen van de andere speler afhangen. Dus we hebben twee niet-lege disjuncte verzamelingen S^1 en S^2 waarvoor geldt $S = S^1 \cup S^2$ en

$$p(s'|s, a^1, a^2) = \begin{cases} p(s'|s, a^1) & \text{als } s \in S^1 \\ p(s'|s, a^2) & \text{als } s \in S^2 \end{cases}$$

voor alle $a^1 \in A^1(s), a^2 \in A^2(s)$. We kunnen dit zien als een generalisatie van de overheersende beslisser (sectie 3.2.1). Helaas moeten we nu een eindig aantal LP-problemen oplossen om dit spel op te lossen.

3.3 Aangepaste methode van Newton

We hebben gezien dat Lineaire Programmering alleen te gebruiken is bij speciale gevallen van stochastische spelen. In deze sectie zullen we een aangepaste versie van de methode van Newton bespreken waarmee we algemene verdisconteerde stochastische spelen kunnen oplossen.
We beschouwen de afbeelding $L : \mathbf{R}^N \to \mathbf{R}^N$ gedefinieerd door

$$L(\mathbf{v})(s) := \text{val}\,[R(s, \mathbf{v})]$$

voor iedere $\mathbf{v} \in \mathbf{R}^N, s \in S$. De theorie van Shapley zegt dat de waardevector $\mathbf{v}_\beta$ de unieke oplossing is van de dekpunt vergelijkingen $L(\mathbf{v}) = \mathbf{v}$. Het vinden van dit dekpunt is hetzelfde als het vinden van het nulpunt van

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

en dit komt neer op het vinden van het globale minimum van de norm van $\psi(\mathbf{v})$. Dus we hebben het volgende mathematische programmerings probleem:

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

In het algoritme dat straks zal worden gegeven wordt, in stap k, de huidige benadering aangegeven met $\mathbf{v}^k$ en de zoekrichting met $\mathbf{d}^k$. We zullen de "klassieke" zoekrichting gebruiken, zoals in de methode van Newton: $\mathbf{d}^k = -\left[\psi'(\mathbf{v}^k)\right]^{-1}\psi(\mathbf{v}^k)$, met $\psi'(\mathbf{v})$ de gradient van $\psi(\mathbf{v})$. Echter, de "stapgrootte" van de zoekrichting zullen we zorgvuldig kiezen, dat wil zeggen dat de iteratieve stap in het algoritme er als volgt uitziet:

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \omega^k\left[\psi'(\mathbf{v}^k)\right]^{-1}\psi(\mathbf{v}^k) \tag{7}$$

waarbij $\omega^k \in (0,1]$ er voor moet zorgen dat convergentie optreedt. Er geldt (zie blz. 102 van het boek):

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

met $\mathbf{f}(\mathbf{v})$ en $\mathbf{g}(\mathbf{v})$ strategieën voor speler 1 en 2 respectievelijk. De eigenschappen van een overgangsmatrix en het feit dat $\beta \in [0,1)$ impliceren dat deze matrix inverteerbaar is en dus

$$\mathbf{d}^k = -\left[\psi'(\mathbf{v}^k)\right]^{-1}\left[\psi(\mathbf{v}^k)\right] = \left[I - \beta P(\mathbf{f}(\mathbf{v}^k),\mathbf{g}(\mathbf{v}^k))\right]^{-1}\psi(\mathbf{v}^k) \tag{8}$$

is goed gedefinieerd als $\psi'(\mathbf{v}^k)$ bestaat. Laat $J(\mathbf{v}) := \frac{1}{2}\left[\psi(\mathbf{v})\right]^T\psi(\mathbf{v})$, dan geldt:

$$\nabla J(\mathbf{v}) = -\left[\psi(\mathbf{v})\right]^T\left[I - \beta P(\mathbf{f}(\mathbf{v}),\mathbf{g}(\mathbf{v}))\right] \tag{9}$$

en als $\nabla J(\mathbf{v}^*) = 0$, dan geldt $\psi(\mathbf{v}^*) = 0$. Oftewel, $\mathbf{v}^*$ is het unieke dekpunt van $L(\mathbf{v})$. Hieronder zal ik het algoritme geven van de aangepaste versie van de methode van Newton. Voor de convergentie verwijs ik naar het boek.

- **Stap 0** Laat $k := 0$ en selecteer $\alpha \in (0,1)$ en $\mu \in [0,5; 0,8]$. Bepaal $\mathbf{v}^0$, de beginbenadering van de waarde vector.
- **Stap 1** Bepaal voor iedere $s \in S$ het matrixspel $R(s,\mathbf{v}^k)$, optimale strategieën $\mathbf{x}(s,\mathbf{v}^k)$ en $\mathbf{y}(s,\mathbf{v}^k)$ voor spelers 1 en 2 in dit matrixspel en $L(\mathbf{v}^k)(s)$. Bereken $L(\mathbf{v}^k)$, $\psi(\mathbf{v}^k)$ en $J(\mathbf{v}^k)$.
- **Stap 2** Als $J(\mathbf{v}^k) = 0$, dan stoppen we; $\mathbf{v}^k = \mathbf{v}_\beta$ en $\mathbf{x}(s,\mathbf{v}^k)$ en $\mathbf{y}(s,\mathbf{v}^k)$ zijn optimale strategieën voor spelers 1 en 2 respectievelijk.
- **Stap 3** Bereken $\mathbf{d}^k$ zoals in (8) en laat $\omega^k = 1$.

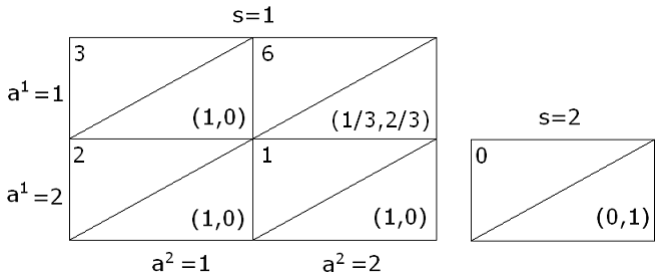
- **Stap 4** Test of aan de volgende ongelijkheid is voldaan:

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

  Als hieraan is voldaan, laat $\mathbf{v}^{k+1} = \mathbf{v}^k + \omega^k \mathbf{d}^k$ en $k := k+1$. Ga terug naar stap 2.

- **Stap 5** Laat $\omega^k := \mu\omega^k$ en ga terug naar stap 4.

Ter verduidelijking zal ik een voorbeeld geven. Laat $S = \{1,2\}, A^1(1) = A^2(1) = \{1,2\}, A^1(2) = A^2(2) = \{1\}, \beta = \frac{3}{4}$ en de opbrengsten en overgangskansen zijn:



Toestand 2 is absorberend met opbrengst 0, dus $v(2) = 0$. Volgens de theorie van Shapley geldt:

$$v(1) = \text{val}\begin{pmatrix} 3 + \frac{3}{4}v(1) & 6 + \frac{1}{4}v(1) + \frac{1}{2}v(2) \\ 2 + \frac{3}{4}v(1) & 1 + \frac{3}{4}v(1) \end{pmatrix}$$

We beginnen het algoritme met $k = 0$, $\alpha = 0,1$, $\mu = 0,5$ en $\mathbf{v}^0 = (0,0)^T$. Het eerste matrix spel in stap 1 is

$$R(1, \mathbf{v}^0) = \begin{pmatrix} 3 & 6 \\ 2 & 1 \end{pmatrix}$$

Deze matrix heeft een zadelpunt (een getal dat het grootst in zijn kolom en het kleinst in zijn rij is), dus er geldt:

$$L(\mathbf{v}^0)(1) = 3 \quad \text{en} \quad \mathbf{x}(1, \mathbf{v}^0) = (1,0) = \mathbf{y}(1, \mathbf{v}^0)^T$$

Omdat $L(\mathbf{v}^k)(2) = 0$ voor alle k geldt:

$$\psi(\mathbf{v}^0) = L(\mathbf{v}^0) - \mathbf{v}^0 = \begin{pmatrix} 3 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

Aangezien $J(\mathbf{v}^0) = \frac{1}{2}(9) \neq 0$ gaan we verder met stap 3. Uit

$$\psi'(\mathbf{v}^0) = -\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{3}{4}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] = \begin{pmatrix} -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} \end{pmatrix}$$

volgt

$$\mathbf{d}^0 = -\begin{pmatrix} -4 & 0 \\ 0 & -4 \end{pmatrix}\begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 12 \\ 0 \end{pmatrix}$$

Om de ongelijkheid in stap 4 te testen, berekenen we eerst

$$\psi\left((12,0)^T\right) = L\left((12,0)^T\right) - (12,0)^T = (-2,0)^T$$

en

$$J\left((12,0)^T\right) = \frac{1}{2}\left[\psi\left((12,0)^T\right)^T \psi\left((12,0)^T\right)\right] = 2$$

Nu vergelijken we

$$J\left(\mathbf{v}^0 + \begin{pmatrix} 12 \\ 0 \end{pmatrix}\right) - J(\mathbf{v}^0) = 2 - \frac{9}{2} = -\frac{5}{2}$$

met

$$(0,1)\left[\nabla J(\mathbf{v}^0)\begin{pmatrix} 12 \\ 0 \end{pmatrix}\right] = (0,1)\left[\left(-\frac{3}{4},0\right)\begin{pmatrix} 12 \\ 0 \end{pmatrix}\right] = -0,9$$

Omdat $-\frac{5}{2} < -0,9$ gaan we terug naar stap 2 met $\mathbf{v}^1 = (12,0)^T$. We hebben nu het matrix spel

$$R(1,\mathbf{v}^1) = \begin{pmatrix} 12 & 9 \\ 11 & 10 \end{pmatrix}$$

Deze matrix heeft een zadelpunt en dus $L(\mathbf{v}^1)(1) = 10$ en $\mathbf{x}(1,\mathbf{v}^1) = (0,1) = \mathbf{y}(1,\mathbf{v}^1)^T$. Omdat $J(\mathbf{v}^1) \neq 0$ berekenen we

$$\mathbf{d}^1 = -\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{3}{4}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]^{-1}\begin{pmatrix} -2 \\ 0 \end{pmatrix} = \begin{pmatrix} -8 \\ 0 \end{pmatrix}$$

Als we de ongelijkheid in stap 4 willen testen, krijgen we te maken met het matrixspel

$$R(1,(4,0)^T) = \begin{pmatrix} 6 & 7 \\ 5 & 4 \end{pmatrix}$$

Deze matrix heeft een zadelpunt en dus $L((4,0)^T)(1) = 6$ en $\mathbf{x}(1,(4,0)^T) = (1,0) = \mathbf{y}(1,(4,0)^T)^T$. Omdat $\psi\left((4,0)^T\right) = (6,0)^T - (4,0)^T$ geldt $J((4,0)^T) - J(\mathbf{v}^1) = 2 - 2 = 0$. Aan de andere kant geldt

$$\nabla J(\mathbf{v}^1) = -(-2,0)\begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} = \left(\frac{1}{2},0\right)$$

en

$$(0,1)\left[\nabla J(\mathbf{v}^1)\mathbf{d}^1\right] = (0,1)\left[\left(\frac{1}{2},0\right)\begin{pmatrix} -8 \\ 0 \end{pmatrix}\right] = -0,4$$

Omdat $0 > -0,4$ gaan we naar stap 5. We krijgen

$$\mathbf{v} := \mathbf{v}^1 + 0,5\mathbf{d}^1 = \begin{pmatrix} 12 \\ 0 \end{pmatrix} + \begin{pmatrix} -4 \\ 0 \end{pmatrix} = \begin{pmatrix} 8 \\ 0 \end{pmatrix}$$

We moeten nu opnieuw de ongelijkheid in stap 4 testen. Daarvoor berekenen we het volgende matrixspel

$$R(1, \mathbf{v}) = \begin{pmatrix} 9 & 8 \\ 8 & 7 \end{pmatrix}$$

Deze matrix heeft ook een zadelpunt en dus $\mathbf{x}(1, \mathbf{v}) = (1, 0), \mathbf{y}(1, \mathbf{v}) = (0, 1)^T$. Nu geldt $L(\mathbf{v}) = (8, 0)^T$ en $\psi(\mathbf{v}) = L(\mathbf{v}) - \mathbf{v} = (0, 0)^T$. Dit laatste impliceert onmiddellijk dat $\mathbf{v} = (8, 0)^T$ het dekpunt van $L(\mathbf{v})$ is en dus $\mathbf{v}_\beta = (8, 0)^T$.

3.4 The Big Match

In het gedeelte over Markov beslissings processen hebben we gezien dat de waarde met gemiddelde opbrengst moeilijker te analyseren is dan de verdisconteerde waarde. Echter, deze moeilijkheden konden we oplossen en we vonden bijvoorbeeld dat er in beide gevallen optimale stationaire strategieën bestaan en dat beide met lineaire programmering op te lossen zijn. Daarom zouden we kunnen verwachten dat ook stochastische spelen in beide gevallen gelijk zijn te benaderen. Helaas blijkt het zo te zijn dat we stochastische spelen met gemiddelde opbrengst niet hetzelfde kunnen behandelen als met verdiscontering. In deze sectie zal ik aan de hand van een voorbeeld (The Big Match) de moeilijkheden laten zien.
De waarde, ofwel de uitbetaling van speler 2 aan speler 1, wordt gegeven door:

$$v_\alpha(s, \mathbf{f}, \mathbf{g}) := \lim_{T\to\infty} \left[\frac{1}{T+1} \sum_{t=0}^{T} E_{s\mathbf{fg}}(R_t) \right] = [Q(\mathbf{f}, \mathbf{g})\mathbf{r}(\mathbf{f}, \mathbf{g})]_s$$

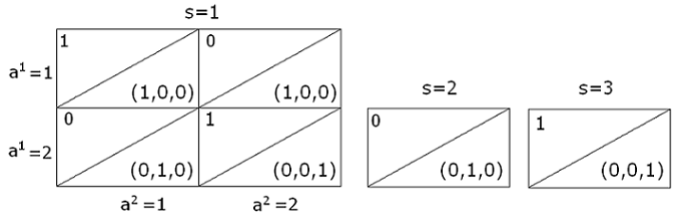
voor iedere $s \in S$ waarbij $Q(\mathbf{f}, \mathbf{g})$ de Cesaro-limiet matrix is van $P(\mathbf{f}, \mathbf{g})$. We noemen $\mathbf{f}^0 \in \mathbf{F}_S$ en $\mathbf{g}^0 \in \mathbf{G}_S$ optimale strategieën als voor alle $s \in S, \mathbf{f} \in \mathbf{F}_S, \mathbf{g} \in \mathbf{G}_S$ geldt:

$$v_\alpha(s, \mathbf{f}, \mathbf{g}^0) \leq v_\alpha(s, \mathbf{f}^0, \mathbf{g}^0) \leq v_\alpha(s, \mathbf{f}^0, \mathbf{g}) \tag{10}$$

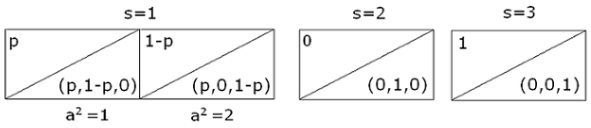
Dit komt overeen met (zie opgave 3)

$$v_\alpha(s, \mathbf{f}^0, \mathbf{g}^0) = \max_{\mathbf{f}\in\mathbf{F}_S} \min_{\mathbf{g}\in\mathbf{G}_S} v_\alpha(s, \mathbf{f}, \mathbf{g}) = \min_{\mathbf{g}\in\mathbf{G}_S} \max_{\mathbf{f}\in\mathbf{F}_S} v_\alpha(s, \mathbf{f}, \mathbf{g})$$

Laat $S = \{1, 2, 3\}, A^1(1) = A^2(1) = \{1, 2\}$ en $A^1(s) = A^2(s) = \{1\}$ voor $s = 2, 3$. De opbrengsten en overgangskansen zijn:



Toestanden 2 en 3 zijn, zoals te zien is, absorberende toestanden en dus $v_\alpha(2) = 0$ en $v_\alpha(3) = 1$. In toestand 1 staat speler 1 voor een moeilijke keuze: als hij actie 1 kiest, leidt dit tot een herhaling van hetzelfde spel en als hij actie 2 kiest absorbeert het spel in toestand 2 of 3, afhankelijk van wat speler 2 kiest. Toestand 2 en 3 hebben een zeer verschillend gevolg voor de uitbetaling aan speler 1.
Stel dat speler 1 gebuikt maakt van de stationaire strategie $\mathbf{f}_p = ((p, 1-p), (1), (1))$ met $p \in [0, 1]$. Dan heeft speler 2 te maken met het volgende Markov beslissings proces met gemiddelde opbrengst:

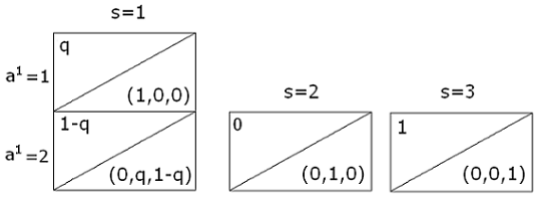


In toestand 1 is de opbrengst van de eerste actie bijvoorbeeld $r(1, 1) = p \times 1 + (1 - p) \times 0$. Nu zijn er twee gevallen:

- $p = 1$, dan neemt speler 1 nooit het risico om in een absorberende toestand te komen. Met $\mathbf{g}_0 = ((0, 1), (1), (1))^T$ zal speler 1 bijna overal 0 verdienen en er geldt $v_\alpha(1, \mathbf{f}_1, \mathbf{g}_0) = 0$.

- $0 < p < 1$, dan kiest speler 1 met kans $1 - p > 0$ actie 2, elke keer als toestand 1 zichzelf herhaalt. Met $\mathbf{g}_1 = ((1, 0), (1), (1))^T$ zal het spel uiteindelijk in toestand 2 komen met kans 1. Dus er geldt $v_\alpha(1, \mathbf{f}_p, \mathbf{g}_1) = 0$.

We zien dat er ongeacht de waarde van p geldt $\min_{\mathbf{g} \in \mathbf{G}_S} v_\alpha(1, \mathbf{f}_p, \mathbf{g}) = 0$. Als we veronderstellen dat iedere stationaire strategie voor speler 2 uit te drukken is in de vorm $\mathbf{g}_q = ((q, 1 - q), (1), (1))^T$ met $q \in [0, 1]$, dan heeft speler 1 te maken met het volgende Markov beslissings proces:



Als speler 1 nu strategie $\mathbf{f}_p$ gebruikt met $p < 1$, dan zal absorptie in toestanden 2 en 3 plaatsvinden met kansen q en $1-q$ respectievelijk. Als $p = 1$, dan zal toestand 1 zichzelf oneindig vaak herhalen. Dus we zien:

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

Als we $\max_{\mathbf{f}\in\mathbf{F}_S} v_\alpha(1, \mathbf{f}, \mathbf{g}_q)$ willen weten, zien we dat $p < 1$ voor $0 \leq q \leq \frac{1}{2}$ en dat $p = 1$ voor $\frac{1}{2} \leq q \leq 1$. Het minimum hiervan is $\frac{1}{2}$. We kunnen het volgende concluderen:

$$\max_{\mathbf{f}\in\mathbf{F}_S} \min_{\mathbf{g}\in\mathbf{G}_S} v_\alpha(1, \mathbf{f}, \mathbf{g}) = 0 < \frac{1}{2} = \min_{\mathbf{g}\in\mathbf{G}_S} \max_{\mathbf{f}\in\mathbf{F}_S} v_\alpha(1, \mathbf{f}, \mathbf{g})$$

We zien dat er geen optimale strategieën bestaan in dit voorbeeld. Overigens is het interessant om op te merken dat deze wel bestaan als we strategieën bekijken die afhangen van de geschiedenis van het spel, maar daar zullen we verder niet op ingaan.

3.5 Nulsom spel met overheersende beslisser en gemiddelde opbrengst

Omdat lineaire programmering een dominante rol speelt in de mathematische programmering is het fijn om bepaalde klassen van problemen met deze methode op te kunnen lossen. Zoals we in de vorige sectie gezien hebben is het stochastische spel met gemiddelde opbrengst in het algemeen een stuk lastiger te analyseren dan het verdisconteerde spel. Nu zullen we echter laten zien dat ook voor gemiddelde opbrengst het nulsom stochastische spel met overheersende beslisser met Lineaire Programmering is op te lossen. De overheersende beslisser is speler 1, dus er geldt:

$$p(s'|s, a^1, a^2) = p(s'|s, a^1)$$

voor alle $a^1 \in A^1(s), a^2 \in A^2(s), s, s' \in S$. De waarde van het strategiepaar $(\mathbf{f}, \mathbf{g})$ wordt gegeven door:

$$\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}) = Q(\mathbf{f}, \mathbf{g})\mathbf{r}(\mathbf{f}, \mathbf{g}) = Q(\mathbf{f})\mathbf{r}(\mathbf{f}, \mathbf{g})$$

met $Q(\mathbf{f}, \mathbf{g})$ de Cesaro-limiet matrix van $P(\mathbf{f}, \mathbf{g})$.
Voor elke toestand $s \in S$ definiëren we een $N \times m^1(s)$ matrix W_s waarbij het element op plaats $(s', (s, a^1))$ gegeven wordt door:

$$w_{s'(s,a^1)} := \delta(s, s') - p(s'|s, a^1)$$

voor iedere $s' \in S$ en $a^1 \in A^1(s)$. Ook definiëren we vijf $m^1(s) \times 1$ kolomvectoren:

$$\begin{aligned}
\mathbf{x}_s &= \left(x_{s1}, x_{s2}, \ldots, x_{sm^1(s)}\right)^T \\
\mathbf{y}_s &= \left(y_{s1}, y_{s2}, \ldots, y_{sm^1(s)}\right)^T \\
\mathbf{r}_s &= \left(r(s,1), r(s,2) \ldots, r(s,m^1(s))\right)^T \\
\mathbf{1}_s &= (1,1,\ldots,1)^T \\
\mathbf{0}_s &= (0,0,\ldots,0)^T
\end{aligned}$$

En we kunnen deze voor alle toestanden samenvoegen tot de volgende $1 \times m^1$ rijvectoren met $m^1 = \sum_{s=1}^{N} m^1(s)$:

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

We definiëren de volgende twee $N \times m^1$ matrices:

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

Tenslotte introduceren we de volgende $1 \times N$ rijvectoren:

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

waarbij iedere $\gamma(s) > 0$ en $\sum_{s=1}^{N} \gamma(s) = 1$. Omdat de opbrengsten wel nog afhangen van de acties van beide spelers zullen we hiervoor nog wat notatie invoeren. We introduceren de $m^1 \times m^2$ matrix

$$R = \text{diag}[R(1), R(2), \ldots, R(N)]$$

met

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

Voor een stationaire strategie $\mathbf{g}$ voor speler 2 geldt dat

$$R\mathbf{g} = \left[(R(1)\mathbf{g}(1))^T, (R(2)\mathbf{g}(2))^T, \ldots, (R(N)\mathbf{g}(N))^T\right]^T$$

een $m^1 \times 1$ blok-kolomvector is. Analoog is

$$\mathbf{f}R = [\mathbf{f}(1)R(1), \mathbf{f}(2)R(2), \ldots, \mathbf{f}(N)R(N)]$$

een $1 \times m^2$ blok-rijvector. We zullen nu het lineaire programmeringsprobleem geven:

$$\min \left[\gamma^T \mathbf{v}\right] \tag{11}$$

onder de voorwaarden:

(a)

$$\left(\mathbf{u}^T, \mathbf{v}^T, \mathbf{g}^T\right) \begin{pmatrix} W & \vdots & 0 \\ \ldots & \vdots & \ldots \\ J & \vdots & W \\ \ldots & \vdots & \ldots \\ -R^T & \vdots & 0 \end{pmatrix} \geq \left(\mathbf{0}^T, \mathbf{0}^T\right)$$

(b) $\mathbf{1}^T \mathbf{g}(s) = 1, \ s \in S$

(c) $\mathbf{g}(s) \geq \mathbf{0}, \ s \in S$

We zullen ook het duale probleem geven. De duale variabelen $\mathbf{x}, \mathbf{y}$ corresponderen met de twee "beperking blokken" uit (a) en de duale variabele $\mathbf{z}$ correspondeert met de beperkingen in (b).

$$\max \left[1^T \mathbf{z}\right] \tag{12}$$

onder de voorwaarden:

(d)

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

(e) $\left[-R^T \mathbf{x}\right]_s + z_s \mathbf{1}_{m^2(s)} \leq \mathbf{0}_{m^2(s)}, \ s \in S$

(f) $\mathbf{x}, \mathbf{y} \geq \mathbf{0}$

Waarbij $\mathbf{1}_a$ en $\mathbf{0}_a$ vectoren zijn met respectievelijk allemaal enen en allemaal nullen met dimensie a. We kunnen nu het stochastische spel als volgt oplossen:
Vind een optimale oplossing $\left(\hat{\mathbf{u}}^T, \hat{\mathbf{v}}^T, \hat{\mathbf{g}}^T\right)$ van (11) en een optimale oplossing $\left(\hat{\mathbf{x}}^T, \hat{\mathbf{y}}^T, \hat{\mathbf{z}}^T\right)$ van (12). Definieer de volgende verzameling van toestanden

$$S^* := \left\{ s \in S \mid \hat{x}_s := \sum_{a^1 \in A^1(s)} \hat{x}_{sa^1} > 0 \right\}$$

Construeer een stationaire strategie $\hat{\mathbf{f}} = \left(\hat{\mathbf{f}}(1), \hat{\mathbf{f}}(2), \ldots, \hat{\mathbf{f}}(N)\right)$ als volgt:

$$\hat{f}(s, a^1) = \begin{cases} \frac{\hat{x}_{sa^1}}{\hat{x}_s} & \text{als } s \in S^*, a^1 \in A^1(s) \\ \frac{\hat{y}_{sa^1}}{\hat{y}_s} & \text{als } s \in S \backslash S^*, a^1 \in A^1(s) \end{cases}$$

met $\hat{y}_s := \sum_{a^1 \in A^1(s)} \hat{y}_{sa^1}$. We hebben nu optimale stationaire strategieën $\hat{\mathbf{f}}$ en $\hat{\mathbf{g}}$ voor speler 1 en 2. Voor een volledig bewijs verwijs ik naar het boek, bladzijde 114-118. In opgave 11 bij de uitgewerkte opgaven zal ik een voorbeeld van dit type spel geven. Tot slot van deze sectie zal ik bewijzen dat als $\left(\overline{\mathbf{u}}^T, \overline{\mathbf{v}}^T, \overline{\mathbf{g}}^T\right)$ een toegelaten oplossing voor (11) is en $\mathbf{f}$ een stationaire strategie voor speler 1 is, dat dan geldt:

$$\overline{\mathbf{v}} \ \geq \ \mathbf{v}_\alpha(\mathbf{f}, \overline{\mathbf{g}})$$

Vanwege beperking (a) geldt $\overline{\mathbf{v}}^T W \geq \mathbf{0}_{m^1}^T$. Als we dit uitwerken, zien we dat geldt $\overline{v}(s) - [P(\mathbf{f})\overline{\mathbf{v}}]_s \geq 0$ en dus $\overline{v}(s) \geq [P(\mathbf{f})\overline{\mathbf{v}}]_s$ voor iedere $s \in S$. In vectornotatie ziet dat er als volgt uit:

$$\overline{\mathbf{v}} \ \geq \ P(\mathbf{f})\overline{\mathbf{v}} \tag{13}$$

In opgave 12 bij de uitgewerkte opgaven laat ik zien dat nu geldt:

$$\overline{\mathbf{v}} \ \geq \ Q(\mathbf{f})\overline{\mathbf{v}} \tag{14}$$

Vanwege het andere deel van beperking (a) geldt $\overline{\mathbf{u}}^T W + \overline{\mathbf{v}}^T J - \overline{\mathbf{g}}^T R^T \geq \mathbf{0}_{m^1}^T$. Door dit uit te werken zien we dat geldt $\overline{v}(s) + \overline{u}(s) \geq \mathbf{f}(s)R(s)\overline{\mathbf{g}}(s) + [P(\mathbf{f})\overline{\mathbf{u}}]_s$ voor iedere $s \in S$ en in vectornotatie kunnen we nu schrijven:

$$\overline{\mathbf{v}} + \overline{\mathbf{u}} \ \geq \ \mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) + P(\mathbf{f})\overline{\mathbf{u}} \tag{15}$$

In opgave 12 zal ik ook laten zien dat nu uit (14) en (15) volgt dat voor $\mathbf{f} \in \mathbf{F}_S$ geldt

$$\overline{\mathbf{v}} \ \geq \ Q(\mathbf{f})\mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) = \mathbf{v}_\alpha(\mathbf{f}, \overline{\mathbf{g}}) \tag{16}$$

3.6 Uitwerking opgaven

Opgave 3 Stel dat zowel $(\mathbf{f}^0, \mathbf{g}^0)$ als $(\hat{\mathbf{f}}, \hat{\mathbf{g}})$ in $\mathbf{F}_S \times \mathbf{G}_S$ voldoen aan (2). Dus beide paren van strategieën zijn optimaal. Er geldt $\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g})$ voor alle $\mathbf{f} \in \mathbf{F}_S$ en $\mathbf{g} \in \mathbf{G}_S$. Dus er geldt ook:

$$\mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \hat{\mathbf{g}}) \tag{17}$$

Analoog geldt ook $\mathbf{v}_\beta(\mathbf{f}, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g})$ voor alle $\mathbf{f} \in \mathbf{F}_S$ en $\mathbf{g} \in \mathbf{G}_S$. Dus:

$$\mathbf{v}_\beta(\mathbf{f}^0, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g}^0) \tag{18}$$

Uit (17) en (18) volgt nu:

$$\mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \hat{\mathbf{g}}) \leq \mathbf{v}_\beta(\hat{\mathbf{f}}, \mathbf{g}^0) \tag{19}$$

Dus in (19) geldt overal gelijkheid. We zullen nu aantonen dat

$$\mathbf{v}_\beta(s, \mathbf{f}^0, \mathbf{g}^0) = \max_{\mathbf{f} \in \mathbf{F}_S} \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) = \min_{\mathbf{g} \in \mathbf{G}_S} \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) \tag{20}$$

voor alle $s \in S$. Uit (2) weten we dat $\mathbf{v}_\beta(s, \mathbf{f}^0, \mathbf{g}^0) = \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}^0) = \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s, \mathbf{f}^0, \mathbf{g})$. Aangezien

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

voor iedere $\mathbf{f}^* \in \mathbf{F}_S$ geldt:

$$\max_{\mathbf{f} \in \mathbf{F}_S} \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) \geq \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s, \mathbf{f}^0, \mathbf{g}) = \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}^0) \geq \min_{\mathbf{g} \in \mathbf{G}_S} \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) \tag{21}$$

Omgekeerd geldt voor iedere $\mathbf{g}^* \in \mathbf{G}_S$:

$$\begin{aligned} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}^*) &\geq \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) \\ \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}^*) &\geq \max_{\mathbf{f} \in \mathbf{F}_S} \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) \\ \min_{\mathbf{g} \in \mathbf{G}_S} \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) &\geq \max_{\mathbf{f} \in \mathbf{F}_S} \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s, \mathbf{f}, \mathbf{g}) \end{aligned}$$

Dus in (21) geldt overal gelijkheid. Tot slot zullen we aantonen dat als

$$\max_{\mathbf{f} \in \mathbf{F}_S} \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s^0, \mathbf{f}, \mathbf{g}) \text{ en } \min_{\mathbf{g} \in \mathbf{G}_S} \max_{\mathbf{f} \in \mathbf{F}_S} \mathbf{v}_\beta(s^0, \mathbf{f}, \mathbf{g}) \tag{22}$$

bestaan en aan elkaar gelijk zijn voor sommige $s^0 \in S$, er stationaire strategieën $\mathbf{f}^0$ en $\mathbf{g}^0$ bestaan die aan (2) voldoen voor s^0.
Laat $\mathbf{v}_\beta := \mathbf{v}_\beta(s^0, \mathbf{f}^0, \mathbf{g}^0)$. Definieer $F_\beta(\mathbf{f}) := \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s^0, \mathbf{f}, \mathbf{g})$ met $\mathbf{f} \in \mathbf{F}_S$. Uit (22) volgt dat er een $\mathbf{f}^0 \in \mathbf{F}_S$ bestaat zodanig dat $F_\beta(\mathbf{f}^0) = \mathbf{v}_\beta$. Per definitie geldt nu

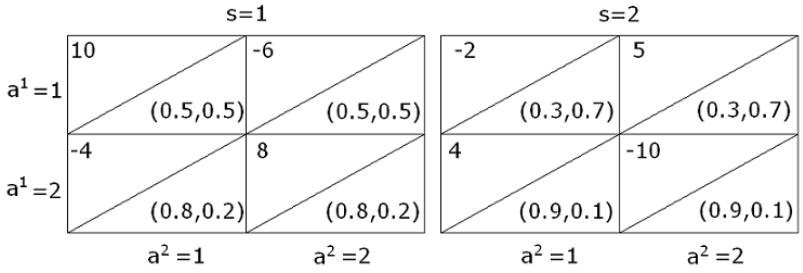
$$\mathbf{v}_\beta = \min_{\mathbf{g} \in \mathbf{G}_S} \mathbf{v}_\beta(s^0, \mathbf{f}^0, \mathbf{g}) \leq \mathbf{v}_\beta(s^0, \mathbf{f}^0, \mathbf{g}) \tag{23}$$

voor alle $\mathbf{g} \in \mathbf{G}_S$. Analoog geldt als we definiëren $G_\beta(\mathbf{g}) := \max_{\mathbf{f}\in\mathbf{F}_S} \mathbf{v}_\beta(s^0, \mathbf{f}, \mathbf{g})$ dat er een $\mathbf{g}^0 \in \mathbf{G}_S$ moet bestaan zodanig dat

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

voor alle $\mathbf{f} \in \mathbf{F}_S$. Dus er is aan (2) voldaan.

Opgave 5 We beschouwen het volgende verdisconteerde stochastische spel dat overheerst wordt door speler 1. Laat $S = \{1, 2\}, A^1(s) = A^2(s) = \{1, 2\}$ voor $s \in S$ en $\beta = 0,7$. De opbrengsten en de overgangskansen zijn:



We zien dat de overgangskansen in elke rij hetzelfde zijn. Dus het spel wordt inderdaad overheerst door speler 1, het maakt voor de overgangskansen namelijk niet uit welke actie speler 2 kiest. Het lineaire programmerings probleem ziet er als volgt uit:

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

onder de voorwaarden:

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

Ik heb dit probleem met het computerprogramma Orstat[5] opgelost: $\mathbf{v}^0_\beta = (4,9242; 2,6515)^T$ en $\mathbf{g}^0 = ((0,5170; 0,4830), (0,6688; 0,3312))^T$. Het duale lineaire programmerings probleem ziet er als volgt uit:

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

[5]ORSTAT 2000. Door Lucien M.J. Claassen, Erwin Kalvelagen, Peter Schram en Henk C. Tijms; Vrije Universiteit Amsterdam

onder de voorwaarden:

(d) $0,65x_{11} + 0,44x_{12} - 0,21x_{21} - 0,63x_{22} = \frac{1}{2}$
$-0,35x_{11} - 0,14x_{12} + 0,51x_{21} + 0,93x_{22} = \frac{1}{2}$

(e) $z(1) \geq 10x_{11} - 4x_{12}$
$z(1) \geq -6x_{11} + 8x_{12}$
$z(2) \geq -2x_{21} + 4x_{22}$
$z(2) \geq 5x_{21} - 10x_{22}$

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

Wederom heb ik dit met Orstat opgelost: $\mathbf{z}^0 = (3,7879; 0)$ en $x_{11}^0 = 0,8117, x_{12}^0 = 1,0823, x_{21}^0 = 0,9596, x_{22}^0 = 0,4798$. Met $x_1^0 = x_{11}^0 + x_{12}^0 = 1,894$ en $x_2^0 = x_{21}^0 + x_{22}^0 = 1,4394$ vinden we $f^0(1,1) = \frac{x_{11}^0}{x_1^0} = \frac{0,8117}{1,894} = 0,429$, $f^0(1,2) = 0,571$, $f^0(2,1) = 0,667$ en $f^0(2,2) = 0,333$. En dus $\mathbf{f}^0 = ((0,429; 0,571), (0,667; 0,333))$.
Tot slot zal ik nagaan dat $\mathbf{f}^0$ en $\mathbf{g}^0$ voldoen aan vergelijking (2). Eerst zal ik laten zien dat $\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0)$ voor iedere $\mathbf{f} \in \mathbf{F}_S$. Omdat alle overgangskansen alleen van speler 1 afhangen geldt:

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) = [I - 0,7P(\mathbf{f})]^{-1}\mathbf{r}(\mathbf{f}, \mathbf{g}^0) \tag{24}$$

voor iedere $\mathbf{f} \in \mathbf{F}_S$. Aangezien $P(\mathbf{f}) = (p(s'|s, \mathbf{f}))_{s,s'=1}^2$ en $p(s'|s, \mathbf{f}) = \sum_{a^1=1}^2 p(s'|s, a^1)f(s, a^1)$ geldt:

$$P(\mathbf{f}) = \begin{pmatrix} 0,5f(1,1) + 0,8f(1,2) & 0,5f(1,1) + 0,2f(1,2) \\ 0,3f(2,1) + 0,9f(2,2) & 0,7f(2,1) + 0,1f(2,2) \end{pmatrix}$$

We kunnen elke stationaire strategie van speler 1 schrijven als
$\mathbf{f} = ((p, 1-p), (q, 1-q))$ met $p, q \in [0,1]$, dus:

$$P(\mathbf{f}) = \begin{pmatrix} 0,8 - 0,3p & 0,2 + 0,3p \\ 0,9 - 0,6q & 0,1 + 0,6q \end{pmatrix}$$

En

$$I - 0,7P(\mathbf{f}) = \begin{pmatrix} 0,44 + 0,21p & -0,14 - 0,21p \\ -0,63 + 0,42q & 0,93 - 0,42q \end{pmatrix}$$

En

$$[I - 0,7P(\mathbf{f})]^{-1} = \begin{pmatrix} \frac{0,93-0,42q}{0,321-0,126q+0,063p} & \frac{0,14+0,21p}{0,321-0,126q+0,063p} \\ \frac{0,63-0,42q}{0,321-0,126q+0,063p} & \frac{0,44+0,21p}{0,321-0,126q+0,063p} \end{pmatrix} \tag{25}$$

We kunnen ook $\mathbf{r}(\mathbf{f}, \mathbf{g}^0)$ uitdrukken in p en q. Er geldt $\mathbf{r}(1, \mathbf{f}, \mathbf{g}^0) = \sum_{a^1} \sum_{a^2} r(1, a^1, a^2)f(1, a^1)g(1, a^2) = 10p0,5170 - 6p0,4830 - 4(1-p)0,5170 +$

$8(1-p)0,4830 = 1,796 + 0,476p$. Analoog geldt $\mathbf{r}(2, \mathbf{f}, \mathbf{g}^0) = -0,6368 + 0,9552q$. Dus

$$\mathbf{r}(\mathbf{f}, \mathbf{g}^0) = \begin{pmatrix} 1,796 + 0,476p \\ -0,6368 + 0,9552q \end{pmatrix} \tag{26}$$

Uit (24), (25) en (26) volgt nu:

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) = \begin{pmatrix} \frac{(0{,}93-0{,}42q)(1{,}796+0{,}476p)+(0{,}14+0{,}21p)(-0{,}6368+0{,}9552q)}{0{,}321-0{,}126q+0{,}063p} \\ \frac{(0{,}63-0{,}42q)(1{,}796+0{,}476p)+(0{,}44+0{,}21p)(-0{,}6368+0{,}9552q)}{0{,}321-0{,}126q+0{,}063p} \end{pmatrix} \tag{27}$$

Na enig rekenwerk zien we dat voor $p, q \in [0, 1]$ (27) nooit groter zal worden dan $\mathbf{v}^0_\beta = (4,9242; 2,6515)^T$. Tot slot zal ik laten zien dat $\mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g})$. Er geldt

$$\mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}) = [I - 0,7P(\mathbf{f}^0)]^{-1}\mathbf{r}(\mathbf{f}^0, \mathbf{g}) \tag{28}$$

Analoog aan het voorafgaande geldt

$$[I - 0,7P(\mathbf{f}^0)]^{-1} = \begin{pmatrix} 2,4617 & 0,8716 \\ 1,3253 & 2,008 \end{pmatrix} \tag{29}$$

en als $\mathbf{g} = ((s, 1-s), (t, 1-t))^T$

$$\mathbf{r}(\mathbf{f}^0, \mathbf{g}) = \begin{pmatrix} 0,12s + 1,994 \\ 0 \end{pmatrix} \tag{30}$$

Uit (28), (29) en (30) volgt:

$$\mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}) = \begin{pmatrix} 0,02954s + 4,9087 \\ 0,01590s + 2,6433 \end{pmatrix} \tag{31}$$

Wederom zien we na enig rekenwerk dat voor $s, t \in [0, 1]$ (31) nooit kleiner zal worden dan $\mathbf{v}^0_\beta = (4,9242; 2,6515)^T$. En dus is aan vergelijking (2) voldaan.

Opgave 6 In deze opgave beschouwen we het stochastische spel dat toestand onafhankelijk is en gescheiden opbrengst heeft (sectie 3.2.2). Ik zal laten zien dat vergelijking (6) klopt. De verdisconteerde waarde van een stochastisch spel is

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

en met de aannames uit sectie 3.2.2 kunnen we dit voor iedere $s \in S$ schrijven als

$$v_\beta(s, \mathbf{f}, \mathbf{g}) = c(s) + \mathbf{f}(s)R(\mathbf{c})\mathbf{g}(s) + \beta \sum_{s' \in S} p(s'|\mathbf{f}, \mathbf{g})(v_\beta(s', \mathbf{f}, \mathbf{g}) - c(s'))$$

met $R(\mathbf{c})$ het matrixspel dat in sectie 3.2.2 geïntroduceerd is. Als we $\mathbf{r}(\mathbf{c}, \mathbf{f}, \mathbf{g}) := \mathbf{f}(s)R(\mathbf{c})\mathbf{g}(s)$ definiëren voor iedere $s \in S$, dan kunnen we dit als volgt in vectornotatie schrijven

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

en als we hierin $\mathbf{v}_\beta(\mathbf{f}, \mathbf{g})$ zelf invullen, krijgen we:

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = \mathbf{c} + [I - \beta P(\mathbf{f}, \mathbf{g})]^{-1}[r(\mathbf{c}, \mathbf{f}, \mathbf{g})\mathbf{1}] \tag{32}$$

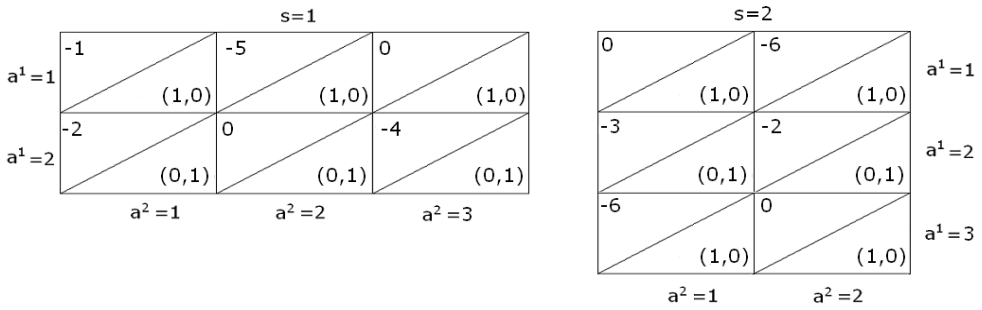
De tweede term van (32) hangt alleen van toestand s af in de keuze van de strategieën. We hebben in sectie 3.2.2 gezien dat $\mathbf{f}(s) = \mathbf{x}$ en $\mathbf{g}(s) = \mathbf{y}$, dus de strategie in toestand s hangt niet van s af.
Omdat de overgangskansen niet van de huidige toestand afhangen heeft $P(\mathbf{f}, \mathbf{g})$ identieke rijen. Het is gemakkelijk in te zien dat nu geldt $P(\mathbf{f}, \mathbf{g}) = P^t(\mathbf{f}, \mathbf{g})$ voor iedere $t \in \mathbf{N}$. Als we namelijk een matrix met zichzelf vermenigvuldigen, moeten we de rijen met de kolommen elementsgewijs vermenigvuldigen. In een kolom staan allemaal dezelfde elementen en een rij sommeert tot 1. De kolommen in de nieuwe matrix zijn dus precies gelijk aan de kolommen in de oude matrix. We weten (zie onder andere opgave 2 uit hoofdstuk 2) dat $[I - \beta P(\mathbf{f}, \mathbf{g})]^{-1} = \sum_{t=0}^{\infty} \beta^t P^t(\mathbf{f}, \mathbf{g})$. Dus er geldt:

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

Vanwege onze aannames en het feit dat $\mathbf{f}(s) = \mathbf{x}$ en $\mathbf{g}(s) = \mathbf{y}$ heeft $\mathbf{r}(\mathbf{c}, \mathbf{f}, \mathbf{g})$ identieke elementen. Dus er geldt $P(\mathbf{f}, \mathbf{g})\mathbf{r}(\mathbf{c}, \mathbf{f}, \mathbf{g}) = \mathbf{r}(\mathbf{c}, \mathbf{f}, \mathbf{g})$, aangezien de rijen van $P(\mathbf{f}, \mathbf{g})$ tot 1 sommeren. Uit (32) volgt nu:

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

Opgave 11 In deze opgave zal ik een voorbeeld uitwerken van het stochastische spel met gemiddelde opbrengst en speler 1 als overheersende beslisser. Laat $S = \{1, 2\}, A^1(1) = A^2(2) = \{1, 2\}, A^1(2) = A^2(1) = \{1, 2, 3\}$ en $\gamma^T = (\frac{1}{2}, \frac{1}{2})$. De opbrengsten en overgangskansen zijn:



De overgangskansen zijn in iedere rij hetzelfde, dus deze hangen niet van de keuze van speler 2 af. Het lineaire programmerings probleem ziet er als volgt uit:

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

onder de voorwaarden:

(a)

$$\left(\begin{array}{ccccc:ccccc}
0 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 \\
0 & 0 & 1 & 1 & 1 & 0 & -1 & 1 & 0 & 1 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 3 & 6 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 6 & 2 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right)^T \begin{pmatrix} u(1) \\ u(2) \\ v(1) \\ v(2) \\ g(1,1) \\ g(1,2) \\ g(1,3) \\ g(2,1) \\ g(2,2) \end{pmatrix} \geq \mathbf{0}$$

(b) $g(1,1) + g(1,2) + g(1,3) = 1$
$g(2,1) + g(2,2) = 1$

(c) $g(1,1), g(1,2), g(1,3), g(2,1), g(2,2) \geq 0$

Ik heb dit probleem met het programma Orstat opgelost en ik kreeg als antwoord $v(1) = v(2) = -2.500$ en $g(1,1) = 0, g(1,2) = g(1,3) = g(2,1) = g(2,2) = 0.500$. Dus $\hat{\mathbf{v}} = (\frac{-5}{2}, \frac{-5}{2})^T$ is de waarde en $\hat{\mathbf{g}} = \left((0, \frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2})\right)^T$ is een optimale strategie voor speler 2. Het duale probleem is:

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

onder de voorwaarden:

(d)

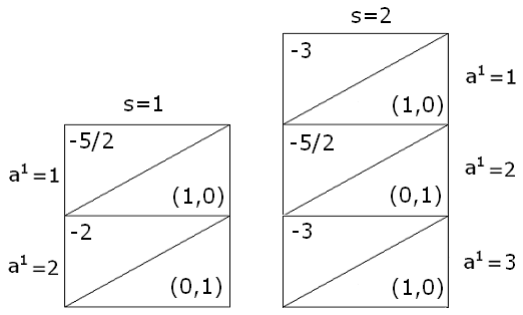
$$\left(\begin{array}{ccccc:ccccc} 0 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 & 0 & -1 & 1 & 0 & 1 \end{array}\right) \left(\begin{array}{c} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \\ x_{23} \\ y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{array}\right) = \left(\begin{array}{c} 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{array}\right)$$

(e) $x_{11} + 2x_{12} + z(1) \leq 0$
$5x_{11} + z(1) \leq 0$
$4x_{12} + z(1) \leq 0$
$3x_{22} + 6x_{23} + z(2) \leq 0$
$6x_{21} + 2x_{22} + z(2) \leq 0$

(f) $x_{11}, x_{12}, x_{21}, x_{22}, x_{23}, y_{11}, y_{12}, y_{21}, y_{22}, y_{23} \geq 0$

Orstat gaf bij dit probleem de volgende oplossing: $x_{11} = 0,2857, x_{12} = 0,3571$ en $x_{22} = 0, x_{21} = x_{23} = 0,1786$. Dus $x_1 = 0,2857 + 0,3571 = 0,6428$ en $x_2 = 2 \times 0,1786 = 0,3572$. Er volgt $f(1,1) = \frac{0,2857}{0,6428} = 0,444, f(1,2) = 0,556$ en $f(2,2) = 0, f(2,1) = f(2,3) = \frac{1}{2}$. Dus $\hat{\mathbf{f}} = \big((\frac{4}{9}, \frac{5}{9}), (\frac{1}{2}, 0, \frac{1}{2})\big)$.

Als we bovenstaande $\hat{\mathbf{g}}$ fixeren en speler 1 weet dat speler 2 deze strategie zal gebruiken, dan is het stochastische spel terug gebracht tot een Markov beslissings proces voor speler 1. Dat proces ziet er als volgt uit:



Voor de eerste actie in toestand 1 geldt bijvoorbeeld $r(1,1) = \frac{1}{2} \times -5 + \frac{1}{2} \times 0$. We kunnen dit Markov beslissings proces met gemiddelde opbrengst oplossen

met Lineaire Programmering zoals we in hoofdstuk 2 hebben gezien. Dit probleem luidt:

$$\max \frac{-5}{2}x_{11} - 2x_{12} - 3x_{21} - \frac{5}{2}x_{22} - 3x_{23}$$

onder de voorwaarden:

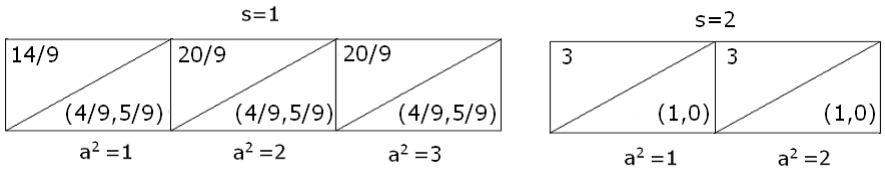
(a) $0x_{11} + 1x_{12} - 1x_{21} + 0x_{22} - 1x_{23} = 0$
$0x_{11} - 1x_{12} + 1x_{21} + 0x_{22} + 1x_{23} = 0$

(b) $x_{11} + x_{12} + x_{21} + x_{22} + x_{23} = 1$

(c) $x_{11}, x_{12}, x_{21}, x_{22}, x_{23} \geq 0$

Met het programma Orstat vond ik de volgende oplossing: $x_{12} = x_{21} = \frac{1}{2}$ en $x_{11} = x_{22} = x_{23} = 0$. Dit komt overeen met een optimale strategie $\mathbf{f}^0 = ((0,1)(1,0,0))$. We zien gemakkelijk in dat $\mathbf{f}^0 = ((0,1)(\frac{1}{2},0,\frac{1}{2}))$, aangezien de eerste en de derde actie in toestand 2 dezelfde opbrengsten en overgangskansen hebben. Na enig rekenwerk kunnen we ook inzien dat $\mathbf{f}^0 = ((\frac{4}{9},\frac{5}{9})(\frac{1}{2},0,\frac{1}{2}))$ en dus $\mathbf{f}^0 = \hat{\mathbf{f}}$. Dit rekenwerk houdt in dat we voor beide strategieën de waarde in toestand 1 berekenen. Omdat de tijdsperiode oneindig is, wordt dit redelijk complex. Wel is makkelijk te zien dat na twee tijdseenheden geldt $v^0(1) = \frac{-2-3}{2} = \frac{-5}{2}$ en $\hat{v}(1) = \frac{\frac{5}{9}(-2-3)+\frac{4}{9}(\frac{-5}{2}+\frac{4}{9}\times\frac{-5}{2}+\frac{5}{9}\times-2)}{2} = -2,44$.

Omgekeerd kunnen we ook $\hat{\mathbf{f}}$ fixeren. Speler 2 heeft dan te maken met het volgende Markov beslissings proces met gemiddelde opbrengst:



De opbrengsten zijn hier positief. Dit heeft te maken met het feit dat de opbrengsten in het oorspronkelijke probleem de betekenis hebben dat speler 2 deze aan speler 1 betaalt. Aangezien deze in het oorspronkelijke probleem negatief waren, werden deze bedragen dus aan speler 2 door speler 1 betaald. Het zijn dus positieve opbrengsten voor speler 2. We kunnen dit op dezelfde manier met Lineaire Programmering oplossen:

$$\max \frac{14}{9}y_{11} + \frac{20}{9}y_{12} + \frac{20}{9}y_{13} + 3y_{21} + 3y_{22}$$

onder de voorwaarden:

(a) $\frac{5}{9}y_{11} + \frac{5}{9}y_{12} + \frac{5}{9}y_{13} - y_{21} - y_{22} = 0$
$\frac{-5}{9}y_{11} - \frac{5}{9}y_{12} - \frac{5}{9}y_{13} + y_{21} + y_{22} = 0$

(b) $y_{11} + y_{12} + y_{13} + y_{21} + y_{22} = 1$

(c) $y_{11}, y_{12}, y_{13}, y_{21}, y_{22} \geq 0$

Orstat gaf als oplossing $y_{12} = 0,6428, y_{21} = 0,3572$ en $y_{11} = y_{13} = y_{22} = 0$. Dit geeft een optimale strategie $\mathbf{g}^0 = ((0,1,0),(1,0))^T$. Het is eenvoudig in te zien dat $\mathbf{g}^0 = ((0,\frac{1}{2},\frac{1}{2}),(\frac{1}{2},\frac{1}{2}))^T$, aangezien de twee acties in toestand 2 dezelfde opbrengsten en overgangskansen hebben en dit ook geldt voor de tweede en de derde actie van toestand 1. We kunnen dus concluderen $\mathbf{g}^0 = \hat{\mathbf{g}}$.

Opgave 12 We beschouwen het stochastische nulsomspel met overheersende beslisser en gemiddelde opbrengst (sectie 3.5). We zullen eerst bewijzen dat voor $\mathbf{f} \in \mathbf{F}_S$ geldt

$$\overline{\mathbf{v}} \geq P(\mathbf{f})\overline{\mathbf{v}} \implies \overline{\mathbf{v}} \geq Q(\mathbf{f})\overline{\mathbf{v}}$$

Omdat $\overline{\mathbf{v}} \geq P(\mathbf{f})\overline{\mathbf{v}}$, geldt er

$$\overline{\mathbf{v}} \geq P(\mathbf{f})\overline{\mathbf{v}} \geq P^2(\mathbf{f})\overline{\mathbf{v}} \geq P^3(\mathbf{f})\overline{\mathbf{v}} \geq \ldots \geq P^T(\mathbf{f})\overline{\mathbf{v}}$$

met $T \in \mathbf{N}$. Dus

$$\begin{aligned}
\overline{\mathbf{v}} &\geq P(\mathbf{f})\overline{\mathbf{v}} \\
\overline{\mathbf{v}} + \overline{\mathbf{v}} &\geq P(\mathbf{f})\overline{\mathbf{v}} + P^2(\mathbf{f})\overline{\mathbf{v}} \\
3\overline{\mathbf{v}} &\geq \sum_{t=1}^{3} P^t(\mathbf{f})\overline{\mathbf{v}} \\
&\vdots \\
T\overline{\mathbf{v}} &\geq \sum_{t=1}^{T} P^t(\mathbf{f})\overline{\mathbf{v}}
\end{aligned}$$

voor iedere $T \in \mathbf{N}$. Aangezien $\overline{\mathbf{v}} = I \times \overline{\mathbf{v}} = P^0(\mathbf{f})\overline{\mathbf{v}}$ kunnen we nu schrijven:

$$\begin{aligned}
(T+1)\overline{\mathbf{v}} &\geq \sum_{t=0}^{T} P^t(\mathbf{f})\overline{\mathbf{v}} \\
\overline{\mathbf{v}} &\geq \frac{1}{T+1}\sum_{t=0}^{T} P^t(\mathbf{f})\overline{\mathbf{v}}
\end{aligned}$$

voor iedere $T \in \mathbf{N}$. Omdat de laatste ongelijkheid geldt voor iedere $T \in \mathbf{N}$ kunnen we de limiet voor $T \to \infty$ nemen en we concluderen:

$$\overline{\mathbf{v}} \geq \lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} P^t(\mathbf{f})\overline{\mathbf{v}} = Q(\mathbf{f})\overline{\mathbf{v}} \tag{33}$$

We gebruiken dit resultaat en (15) om (16) te bewijzen. Uit (15) volgt $\overline{\mathbf{v}} \geq \mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) + P(\mathbf{f})\overline{\mathbf{u}} - \overline{\mathbf{u}}$. Met behulp van (33) kunnen we nu schrijven:

$$\begin{aligned} \overline{\mathbf{v}} \;\geq\; Q(\mathbf{f})\overline{\mathbf{v}} \;&\geq\; Q(\mathbf{f})(\mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) + P(\mathbf{f})\overline{\mathbf{u}} - \overline{\mathbf{u}}) \\ &=\; Q(\mathbf{f})\mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) + Q(\mathbf{f})P(\mathbf{f})\overline{\mathbf{u}} - Q(\mathbf{f})\overline{\mathbf{u}} \end{aligned}$$

Per definitie geldt:

$$Q(\mathbf{f}) \;=\; \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P(\mathbf{f})^t \tag{34}$$

$$Q(\mathbf{f})P(\mathbf{f}) \;=\; \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=1}^{T+1} P(\mathbf{f})^t \tag{35}$$

Deze twee gemiddelden verschillen van elkaar door de termen

$$\frac{P(\mathbf{f})^{T+1}}{T+1} \quad \text{en} \quad \frac{P(\mathbf{f})^0}{T+1}$$

Maar als $T \to \infty$, dan gaan deze allebei naar nul. Dus (34) en (35) hebben dezelfde limiet en er geldt $Q(\mathbf{f})P(\mathbf{f}) = Q(\mathbf{f})$. We kunnen nu concluderen dat:

$$\begin{aligned} \overline{\mathbf{v}} \;&\geq\; Q(\mathbf{f})\mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) + Q(\mathbf{f})\overline{\mathbf{u}} - Q(\mathbf{f})\overline{\mathbf{u}} \\ &=\; Q(\mathbf{f})\mathbf{r}(\mathbf{f}, \overline{\mathbf{g}}) \end{aligned}$$