TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

STOCHASTISCHE BESLISSINGSPROBLEMEN

door

Prof. Dr. J. Wessels

samengesteld door

Ir. J. van der Wal

Najaarssemester 1979

Technische Hogeschool Eindhoven

Onderafdeling der Wiskunde

Stochastische beslissingsproblemen

door prof.dr. J. Wessels

Wij verzoeken U, dit collegedictaat
niet mee te nemen buiten de leeszaal
en het na lezing terug te leggen op
de ladenkasten. Dank U!

samengesteld door ir. J. van der Wal

Dictaatnr. 2.217 Prijs f 4,50

Inhoudsbeschrijving

STOCHASTISCHE BESLISSINGSPROBLEMEN

Najaarssemester 1979

VERVOLG →

JdG, 22 Juli 2005

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

STOCHASTISCHE BESLISSINGSPROBLEMEN

door prof.dr. J. Wessels

samengesteld door ir. J. van der Wal

Najaarssemester 1979

1. Inleiding

Stochastische beslissingsproblemen zijn heel algemeen: problemen waarbij een systeem bestuurd moet worden en de ontwikkeling van het systeem mede door het toeval wordt bepaald, eventueel nog afhankelijk van de te nemen beslissingen.
Voorbeelden van stochastische beslissingsproblemen zijn schattings- en toetsingsproblemen, varianten op lineaire programmerings problemen met bijvoorbeeld i.p.v. beperking $Ax \leq b$ de stochastische restriatie $\mathbb{P}(Ax \leq b) \geq 1 - \alpha$, voorraad- en productie besturingsproblemen, vervangingsproblemen, besturing van een satelliet, enz.
Sommige beslissingsproblemen zijn statisch: er wordt eenmalig een beslissing genomen, vb. lp. probleem. Andere zijn dynamisch: het stochastisch proces ontwikkelt zich in de tijd en er moeten steeds opnieuw beslissingen genomen worden, vb. voorraadproblemen. Een dynamisch beslissingsprobleem kan verder zowel continu in de tijd, vb. satellietbesturing, als discreet in de tijd zijn, vb. wekelijks het produktie niveau vaststellen. We kunnen verder onderscheid maken tussen problemen met een sterke struktuur, vb. een proces waar de ontwikkeling door een (stochastische) differentiaalvergelijking wordt vastgelegd, en problemen vrijwel zonder enige structuur, vb. Markov ketens.
In dit college zullen we ons beperken tot een speciaal type dynamische stochastische beslissingsproblemen namelijk de Markov beslissingsproblemen die discreet zijn in de tijd en waarbij de voortgang van het proces wordt bepaald door een van de genomen beslissingen afhankelijke Markov matrix, vgl. Stochastische Processen I. Het gedeelte van dit college dat handelt over Markov ketens wordt bekend verondersteld.
Hoofdstuk 2 geeft een introductie tot het Markov beslissingsprobleem met eindige toestands- en beslissingsruimten. In hoofdstuk 3 wordt het Markov beslissingsprobleem met eindige tijdshorizon behandeld en er wordt aangetoond hoe met behulp van de dynamische programmeringstechniek optimale strategieën bepaald kunnen worden.
Hoofdstuk 4 beschouwt het probleem met oneindige tijdshorizon en verdiscontering. Eerst wordt aangetoond dat er optimale stationaire strategieën bestaan en daarna worden de methode van de successieve approximaties, de policy iteration methode en de lineaire programmeringsaanpak afgeleid.

Vervolgens wordt in hoofdstuk 5 het oneindig horizon probleem met als criterium de gemiddelde opbrengst per tijdseenheid beschouwd. Opnieuw worden de methode van de successieve approximaties, de policy iteration methode en de lineaire programmeringsaanpak ontwikkeld. In hoofdstuk 6 beschouwen we twee personen Markov spelen; de generalisatie van het Markov beslissingsprobleem met één beslisser tot een probleem met twee beslissers met eventueel tegengestelde belangen. Hoofdstuk 7 beschouwt het meer algemene totale kosten model met aftelbare toestandsruimte en willekeurige actieruimte. Onder de voorwaarde dat de som van alle positieve opbrengsten eindig is wordt o.a. aangetoond dat het beslissingsprobleem een waarde heeft en dat we ons voor elke begintoestand tot zuivere Markov strategieën kunnen beperken.

2. Markov beslissingsproblemen: introductie

2.1 Markov ketens met beslissingen

We hebben in Stochastische Processen I of Inleiding in de theorie van de Stochastische Processen gezien dat een Markov keten met toestandsruimte $I = \{1,2,\ldots,N\}$ volledig wordt vastgelegd door de beginverdeling $p = (p_1,\ldots,p_N)$ en de Markov matrix $P = (p_{ij})_{i,j=1}^{N}$.
Het gedrag van het stochastische proces (de Markov keten) konden we op geen enkele manier beïnvloeden.
Hier beschouwen we de situatie dat we het proces op elk tijdstip $t = 0,1,2,\ldots$ kunnen (bij)sturen. Het sturen bestaat daarbij uit het kiezen van een actie k uit een verzameling (die we hier voorlopig eindig en van de tijd t onafhankelijk zullen veronderstellen). Deze actie legt dan de overgangskansen p_{ij}^{k} vanuit toestand i vast.
Een Markov keten met beslissingen wordt dus gekarakteriseerd door:

I:= $\{1,2,\ldots,N\}$, de eindige toestandsruimte

K:= $\{1,2,\ldots,K_0\}$,de eindige beslissingsruimte

p: overgangsmechanisme, bij elke toestand $i \in I$ en beslissing $k \in K$ hoort een vector $(p_{i1}^{k},\ldots,p_{iN}^{k})$ met p_{ij}^{k} de kans dat het systeem, als in toestand i actie k gekozen is een tijdseenheid later in toestand j zit, dus $p_{ij}^{k} \geq 0$ en $\sum_j p_{ij}^{k} = 1$.

Merk op dat het feit dat het aantal beslissingen in elke toestand gelijk is geen wezenlijke beperking inhoudt. Men kan immers het aantal beslissingen in een toestand altijd uitbreiden door een beslissing (met ander etiket) nog eens in de beslissingsruimte van toestand i te stoppen. Bij veel praktische beslissingsproblemen zullen de 'natuurlijke' beslissingsverzamelingen per toestand variëren. Denk bijvoorbeeld aan een voorraadprobleem met eindige magazijn grootte G met in toestand i de beslissingen $0,1,\ldots,G-i$ bijbestellen.

2.2 Strategieën

De manier waarop het systeem wordt bestuurd, de wijze waarop de acties op elk moment is elke toestand gekozen worden, kunnen we aangeven met behulp van een strategie.

Definitie 2.1 Een *strategie* is een voorschrift dat voor elk beslissingstijdstip aangeeft met welke kans een bepaalde aktie, eventueel afhankelijk van het verleden van het proces, wordt gekozen.

We laten dus toe dat acties mede op grond van het verleden van het proces genomen worden. Ook laten we toe dat er geloot wordt tussen verschillende akties.

Laat s een strategie zijn dan is s feitelijk een rij functies $s_0, s_1, s_2, \ldots$ met

s_0: $I \to K$, waar K de verzameling kansverdelingen op K is. Dus $s_0(i)$ is een kansverdeling op K. En we noteren met $s_0(i,k)$ de kans dat volgens strategie s als het systeem op tijdstip 0 in toestand i zit beslissing k wordt gekozen.

s_1: $I \times K \times I \to K$. Dus als het proces in toestand i_0 is gestart, daar beslissing k_0 genomen is en het systeem nu (op tijdstip 1) in i_1 zit dan is $s_1(i_0,k_0,i_1)$ de door s voorgeschreven loting tussen de beslissingen uit K. $s_1(i_0,k_0,i_1,k)$ is dan weer de kans dat beslissing k wordt gekozen.

s_n: $(I \times K)^n \times I \to K$. s_n legt voor elke mogelijke historie $h_n = (i_0,k_0,i_1,k_1,\ldots,i_{n-1},k_{n-1})$ (de opeenvolging van vroegere toestanden en eerder genomen beslissingen) van het stochastisch proces tot tijdstip n de kansen $s_n(h_n,i,k)$ vast waarmee als het systeem nu in i zit en het verleden h_n is actie k gekozen wordt.

Sommige strategieën hebben een aanmerkelijk eenvoudiger structuur dan de strategie s.

Definitie 2.2 We noemen een strategie s *gemengd Markov* als de functies $s_n(h_n,i,k)$ niet van h_n afhangen.

Een gemengde Markov strategie legt dus voor elk tijdstip n de kansen $s_n(i,k)$ vast waarmee als het systeem op tijdstip n in toestand i zit actie k gekozen wordt zonder te letten op het verleden van proces.

Definitie 2.3 Een strategie heet *zuiver Markov* of kortweg Markov als de strategie gemengd Markov is en de $s_n(i,k)$ alleen de waarde 0 of 1 aannemen.

Feitelijk wordt er niet meer geloot tussen de verschillende beslissingen. Een Markov strategie is dus eigenlijk een rij afbeeldingen van S in K. Voor elk tijdstip eventueel een andere. Voor Markov strategieën gebruiken we daarom de notatie $s = (f_0,f_1,f_2,\ldots)$ met $f_n(i) = k$ als k juist de aktie is waarvoor $s_n(i,k) = 1$.

Zo een functie f_n noemen we een beslissingsregel.

Definitie 2.4 Een strategie heet *stationair* als s een Markov strategie is: $s = (f_0, f_1, \ldots)$ en de functies f_n allemaal gelijk zijn, $f_n \equiv f$. Notatie $f^{(\infty)}$

Een stationaire strategie schrijft dus in een toestand op elk tijdstip de zelfde aktie voor.

2.3 Het bij strategie s behorende stochastisch proces

We kunnen bij elke strategie s een stochastisch proces definiëren.
Laat $(p_1, \ldots, p_N)$ een beginverdeling op I zijn en laten de stochastische variabelen X_n en B_n de toestand respectievelijk de beslissing op tijdstip n zijn.
Dan definiëren we het bij strategie s behorende stochastische proces $(X_0, B_0, X_1, \ldots)$ door

$$\mathbb{P}_s(X_0 = i_0) = p_{i_0}$$

$$\mathbb{P}_s(X_0 = i_0,\ B_0 = k_0) = p_{i_0}\, s_0(i_0,k_0)$$

$$\mathbb{P}_s(X_0 = i_0,\ B_0 = k_0,\ X_1 = i_1) = p_{i_0}\, s_0(i_0,k_0) p_{i_0 i_1}^{k_0}$$

$$\mathbb{P}_s(X_0 = i_0,\ B_0 = k_0,\ X_1 = i_1,\ B_1 = k_1) = p_{i_0}\, s_0(i_0,k_0) p_{i_0 i_1}^{k_0}\, s_1(i_0,k_0,i_1,k_1)$$

etc.

We kunnen ook kijken naar alleen het stochastisch proces $(X_0, X_1, \ldots)$

$$\mathbb{P}_s(X_0 = i_0) = p_{i_0}$$

$$\mathbb{P}_s(X_0 = i_0,\ X_1 = i_1) = p_{i_0} \sum_{k_0} s_0(i_0,k_0) p_{i_0 i_1}^{k_0}$$

$$\mathbb{P}_s(X_0 = i_0,\ X_1 = i_1,\ X_2 = i_2) =$$

$$= p_{i_0} \sum_{k_0} s_0(i_0,k_0)\, p_{i_0 i_1}^{k_0} \sum_{k_1} s_1(i_0,k_0,i_1,k_1) p_{i_1 i_2}^{k_1}$$

We zien dat de overgangskansen $\mathbb{P}_s(X_{n+1} = j \mid X_0 = i_0, \ldots, X_n = i_n)$ van het proces $(X_0, X_1, \ldots)$ nu niet meer alleen van i_n afhangen maar, via de strategie s, ook van vroegere toestanden van het systeem. In het algemeen is het bij strategie s behorende stochastische proces dan ook geen Markov proces meer.
Ga na dat voor (gemengde) Markov strategieën het stochastische proces wel een Markov proces wordt.

2.4 Criteriumfunctie

Om een keus te kunnen maken tussen de verschillende manieren om het systeem te besturen moeten we de verschillende strategieën kunnen vergelijken. We hebben dus een nutsfunctie nodig die aan elke realisatie van het stochastisch proces een bepaalde waarde toekent.

Definitie 2.5 Een nutsfunktie u is een afbeelding van $(I \times K)^{\infty}$ in $\overline{\mathbb{R}}$ $(\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty,\infty\})$.

We kunnen nu het verwachte nut w(s) van een bepaalde strategie als criterium gebruiken om strategieën te vergelijken

$$w(s) = E_s\, u(X_0,B_0,X_1,\ldots)$$

Wij zullen hier verder niet naar zulke algemene nutsfuncties kijken maar het bijzondere geval bekijken dat bij elke overgang van een toestand i naar een toestand j een, van de beslissing k afhankelijke opbrengst, hoort.

Definitie 2.6. De opbrengststructuur, of functie van directe opbrengsten, is een functie

$$r\colon I \times K \times I \to \mathbb{R}$$

Interpretatie. Als de bestuurder van het systeem in toestand i beslissing k neemt en het systeem daarna een overgang maakt naar toestand j dat krijgt de bestuurder een directe opbrengst r(i,k,j).

In het vervolg zullen we met bovenstaande opbrengststructuur verschillende criteriumfuncties beschouwen:

(i) De totale verwachte opbrengst (speciaal voor problemen met eindige tijdshorizon)

(ii) De totale verdisconteerde opbrengst (de opbrengsten op de tijdstippen t worden met geometrisch afnemende factoren gewogen)

(iii) De gemiddelde opbrengst per periode.

3. Markov beslissingsproblemen met eindige tijdshorizon

In dit hoofdstuk zullen we het Markov beslissingsprobleem beschouwen met eindig veel beslissingstijdstippen, dat volledig gekarakteriseerd kon worden door

$I = \{1,2,\ldots N\}$, de toestandsruimte

$K = \{1,2,\ldots,K_0\}$, de beslissingsruimte

$m = 0,1,\ldots,M-1$, de beslissingstijdstippen

p: overgangsmechanisme ($p_{ij}^k = \mathbb{P}(X_{m+1} = j \mid X_m = i, B_m = k)$)

r: opbrengststructuur

Neem nog aan dat als het systeem na de laatste beslissing in toestand j terechtkomt we een eindopbrengst q(j) krijgen (q is bijvoorbeeld de restwaarde).

Het doel is nu een optimale strategie s te bepalen, dat wil zeggen een strategie die de criteriumfunctie

$$v_M(i,s) = E_{i,s}\left[\sum_{m=0}^{M-1} r(X_m,B_m,X_{m+1}) + q(X_M)\right]$$

maximaliseert.

In het college Inleiding in de Beslissingstheorie is de techniek van het dynamisch programmeren gebruikt bij het bestuderen van voorraadproblemen en in de speltheorie bij het bepalen van goede (optimale) strategieën in spelen met een boomstruktuur en volledige informatie.

Het is niet mogelijk een beslissingsboom voor het M-staps Markov beslissingsprobleem op te stellen omdat er door het toelaten van gemengde strategieën overaftelbaar veel takken zullen zijn.

We zullen hier de aanpak volgen die bij de voorraadproblemen is gehanteerd.

Om notatietechnische redenen nummeren we de beslissingstijdstippen in omgekeerde volgorde. Tijdstip M (feitelijk geen beslissingstijdstip) wordt nu tijdstip 0, M-1 wordt 1 enz., zodat er op tijdstip m nog m keer een beslissing genomen moet worden.

Definieer nu

$$v_0(i) := q(i),\ i \in I$$

$$v_1(i) := \max_{k \in K} \{ \sum_{j \in I} p_{ij}^k [r(i,k,j) + v_0(j)] \},\ i \in I$$

$$v_m(i) := \max_{k \in K} \{ \sum_{j \in I} p_{ij}^k [r(i,k,j) + v_{m-1}(j)] \},\ i \in I,\ m = 2,\ldots M$$

Het ligt voor de hand $v_m(i)$ te zien als de maximale verwachte opbrengst als er nog m stappen te gaan zijn.
In het resterende deel van dit hoofdstuk zullen we dat ook bewijzen.

Definieer de Markov strategie $s^* = (f_M, f_{M-1}, \ldots, f_1)$ met $f_m(i)$ voor alle i en m gelijk aan een maximaliserende aktie in de uitdrukking

$$\sum_{j \in I} p_{ij}^k [r(i,k,j) + v_{m-1}(j)]$$

Stelling 3.1 (i) $\max_s v_M(i,s) = v_M(i)$

(ii) $v_M(i,s^*) = v_M(i)$

Bewijs.
Definieer eerst:

$v_m(h_m,i,s) :=$ opbrengst vanaf tijdstip m (omgekeerde tijd) bij strategie s, als het systeem zich op tijdstip m in i bevindt en de historie $h_m = (i_M, k_M, \ldots, i_{m+1}, k_{m+1})$ heeft

$$v_m(h_m,i) := \max_s v_m(h_m,i,s)$$

Met inductie zullen we bewijzen dat

(3.1) $v_m(h_m,i) \leq v_m(i) = v_m(h_m,i,s^*)$ voor alle m, h_m en i.

Eerst m = 0. Op tijdstip 0 wordt geen beslissing meer genomen, maar alleen nog de eindopbrengst q(.) geïncasseerd. Daarmee is duidelijk dat geldt

$v_0(h_0,i,s) = q(i)$ voor alle h_0, i en s, dus ook

$v_0(h_0,i) \leq (=)\ v_0(i) = v_0(h_0,i,s^*)$.

Aannemend dat (3.1) voor m geldt bewijzen we nu dat (3.1) ook geldt voor m+1

$$
\begin{aligned}
v_{m+1}(h_{m+1},i,s) &= \sum_{k\in K} s_{m+1}(h_{m+1},i,k) \sum_{j\in I} p_{ij}^{k}[r(i,k,j) + \\
&\quad + v_m(h_{m+1},i,k,j,s)] \\
&\leq \sum_{k\in K} s_{m+1}(h_{m+1},i,k) \sum_{j\in I} p_{ij}^{k}[r(i,k,j) + v_m(j)] \\
&\leq \max_{k\in K} \sum_{j\in I} p_{ij}^{k}[r(i,k,j) + v_m(j)] = v_{m+1}(i)
\end{aligned}
\tag{3.2}
$$

Het eerste ongelijkteken volgt uit de inductieveronderstelling, het tweede uit het feit dat een convexe combinatie van getallen kleiner of gelijk is aan het maximale. Hiermee is dus bewezen

$$v_{m+1}(h_{m+1},i) = \max_{s} v_{m+1}(h_{m+1},i,s) \leq v_{m+1}(i)$$

We moeten nog na gaan of

$$v_{m+1}(h_{m+1},i,s^*) = v_{m+1}(i)$$

Dit volgt direct uit (3.2) voor $s = s^*$.

$$
\begin{aligned}
v_{m+1}(h_{m+1},i,s^*) &= \sum_{j\in I} p_{ij}^{f_{m+1}(i)} [r(i,f_{m+1}(i),j) + \\
&\quad + v_m(h_{m+1},i,f_{m+1}(i),j,s^*)] \\
&= \sum_{j\in I} p_{ij}^{f_{m+1}(i)} [r(i,f_{m+1}(i),j) + v_m(j)] \\
&= v_{m+1}(i)
\end{aligned}
$$

(Het tweede gelijkteken volgt uit de inductie veronderstelling en het derde uit de definitie van $f_{m+1}(i)$ als maximaliserende actie). Hiermee is het bewijs vrijwel voltooid. Neem in (3.1) m = M dan volgt met $(h_M,i) = (i)$ direct het resultaat. □

Deze stelling zegt dus dat in het M-staps Markov beslissingsprobleem met dynamische programmering een optimale Markov strategie kan worden bepaald. De dynamische programmeringsaanpak maakt gebruik van het feit dat de staart van een M-staps optimale strategie ook optimaal moet zijn in het 1-staps probleem, het 2-staps probleem enz. Dit principe, dat hier is bewezen, maar dat zeker niet algemeen geldt, staat in de literatuur bekend als 'het optimaliteits principe van Bellman'. Gevolg van de stelling is ook dat de ingevoerde gemengde beslissingen en de historie afhankelijke strategieën overbodig zijn. Voor willekeurige nutsfuncties zal dit in het algemeen niet het geval zijn.

In alle formules komen de grootheden $r(i,k,j)$ steeds voor in de vorm $\Sigma p_{ij}^{k}\ r(i,k,j)$. Om de notaties in het vervolg wat te vereenvoudigen definiëren we

$$r(i,k) := \sum_{j \in I} p_{ij}^{k}\ r(i,k,j)$$

$r(i,k)$ is dus de verwachte directe opbrengst als in toestand i actie k wordt genomen.

4. Markov beslissingsproblemen met oneindige tijdshorizon en verdiscontering

4.1 Inleiding

Als we overgaan van Markov beslissingsproblemen met eindige horizon naar problemen met oneindige horizon, dus met beslissingstijdstippen $0,1,2,\ldots$, dan zal de totale verwachte opbrengst, zo die al gedefinieerd is, meestal ∞ of $-\infty$ zijn. Een van de mogelijkheden is nu om de gemiddelde opbrengst per tijdseenheid als criterium te nemen. In dit hoofdstuk zullen we een andere criteriumfunctie nemen namelijk de totale verwachte verdisconteerde opbrengst. Dat wil zeggen dat we een opbrengst op tijdstip n niet volledig tellen maar die opbrengst vermenigvuldigen met een factor β^n, en voor zekere $0 \leq \beta < 1$.

Definitie 4.1. Laat s een willekeurige strategie zijn dan definiëren we $v_\beta(i,s)$ door

$$(4.1) \qquad v_\beta(i,s) := E_{i,s} \sum_{n=0}^{\infty} \beta^n r(X_n,B_n)$$

$v_\beta(i,s)$ is dus de *totale verwachte verdisconteerde opbrengst bij strategie* s als het proces in i start en β de verdisconteringsfactor is.

Definitie 4.2 . We definiëren de *waarde* v_β van het beslissingsproces door

$$v_\beta(i) = \sup_s v_\beta(i,s) \quad , \quad i \in I$$

We zullen in het vervolg zien dat er zelfs een stationaire strategie s^* bestaat met

$$v_\beta(i,s^*) = v_\beta(i), \quad i \in I$$

We kunnen in definitie 4.2 dus sup vervangen door max.

Definieer verder

$$A := \max_{i,k} |r(i,k)|$$

Dan zien we direct dat voor de n-de term van de som in (4.1) geldt

$$(4.2) \qquad |\beta^n r(X_n,B_n)| \leq A\beta^n$$

en voor de staart vanaf t

$$(4.3) \qquad \sum_{n=t}^{\infty} |\beta^n r(X_n,B_n)| \leq A\beta^t/(1-\beta)$$

Zij verder $v_{\beta,t}(i,s)$ de verwachte verdisconteerde opbrengst bij strategie s tot tijdstip t:

$$v_{\beta,t}(i,s) := E_{i,s} \sum_{n=0}^{t-1} \beta^n r(X_n,B_n)$$

Dan volgt met (4.3)

$$(4.4) \qquad |v_\beta(i,s) - v_{\beta,t}(i,s)| \leq A\, \beta^t/(1-\beta)$$

4.2 De benadering met eindig-staps problemen

We kunnen net als in het totale kostenmodel uit hoofdstuk 3 ook in het eindig-staps verdisconteerde Markov beslissingsprobleem met dynamische programmering de optimale opbrengst en een optimale strategie bepalen.

Definieer $v_{\beta,0}(i) = 0$, $i \in I$

$$(4.5) \text{ en } \quad v_{\beta,t}(i) = \max_k \{r(i,k) + \beta \sum p_{ij}^k\, v_{\beta,t-1}(j)\}, \quad i \in I, \quad t = 1,2,\ldots$$

en zij $\hat{s}_t = (\hat{f}_t,\ldots,\hat{f}_1)$ een Markov strategie met $\hat{f}_n(i)$ een maximaliseerde actie voor de uitdrukking

$$r(i,k) + \beta \sum p_{ij}^k\, v_{\beta,n-1}(j)$$

voor alle $i \in I$ en $n = 1,2,\ldots,t$

Dan geldt

Lemma 4.1

$$(i) \quad v_{\beta,t}(i) = \sup_s v_{\beta,t}(i,s)$$

$$(ii) \quad v_{\beta,t}(i,\hat{s}_t) = v_{\beta,t}(i)$$

Bewijs. Het bewijs verloopt geheel analoog aan het bewijs van stelling 3.1. □

We zijn echter niet zo zeer geïnteresseerd in het eindig staps probleem; we willen het ∞-staps probleem 'oplosssen'.

Het zou plezierig zijn als we het ∞-staps probleem konden benaderen door het eindig staps probleem. We zullen in het vervolg zien dat dit heel goed mogelijk is.

De volgende stelling is al een eerste belangrijk resultaat.

Stelling 4.1

$$\lim_{t\to\infty} v_{\beta,t}(i) = v_\beta(i) \quad , \quad i \in I$$

Bewijs. Laat s een willekeurige strategie zijn, dan volgt met behulp van (4.4)

$$v_{\beta,t}(i,s) - A\beta^t(1-\beta)^{-1} \le v_\beta(i,s) \le v_{\beta,t}(i,s) + A\beta^t(1-\beta)^{-1}$$

zodat ook

$$\sup_s v_{\beta,t}(i,s) - A\beta^t(1-\beta)^{-1} \le \sup_s v_\beta(i,s) \le \sup_s v_{\beta,t}(i,s) + A\beta^t(1-\beta)^{-1}$$

ofwel

$$v_{\beta,t}(i) - A\beta^t(1-\beta)^{-1} \le v_\beta(i) \le v_{\beta,t}(i) + A\beta^t(1-\beta)^{-1}$$

dus

$$|v_\beta(i) - v_{\beta,t}(i)| \le A\beta^t(1-\beta)^{-1}$$

waaruit we zien dat $v_{\beta,t}(i) \to v_\beta(i)$ als $t \to \infty$. □

Voordat we verder gaan voeren we eerste de vector-matrix notatie in.

Definieer

$r(f) :=$ kolomvector in $\mathbb{R}^N$ met i-de component $r(i,f(i))$, $i \in I$.
$P(f) :=$ $N \times N$ matrix met i,j-de element $p_{ij}^{f(i)}$.

Verder nemen we aan dat we, indien we in het vervolg de op de toestand betrekking hebbende variabele weglaten, overgaan op de vector notatie. Vergelijking (4.5) wordt met deze afspraken

$$v_{\beta,t} = \max_f \{r(f) + \beta P(f)\, v_{\beta,t-1}\} \tag{4.6}$$

Verder noteren we voor een vector $v \in \mathbb{R}^N$ met $\|v\|$ de maximumnorm van v:

$$\|v\| = \max_i |v(i)|$$

4.3 De functionaalvergelijking

We hebben in stelling 4.1 gezien dat $v_{\beta,t}$ naar v_β convergeert. Beschouwen we nu formule (4.6) dan leidt de limietovergang voor t naar ∞ tot het volgende resultaat.

Stelling 4.2

(i) $$v_\beta = \max_f \{r(f) + \beta P(f) v_\beta\}$$

(ii) Als

(4.7) $$w = \max_f \{r(f) + \beta P(f) w\}$$

dan $w = v_\beta$.

Vergelijking (4.7) heet de *functionaalvergelijking* van het beslissingsproces. Stelling 4.2 zegt dus dat v_β de unieke oplossing is van de functionaalvergelijking.

Bewijs. (i) Definieer $\varepsilon_t := v_{\beta,t} - v_\beta$. Daarmee kunnen we (4.6) als volgt herschrijven.

$$v_\beta + \varepsilon_t = \max_f \{r(f) + \beta P(f)(v_\beta + \varepsilon_{t-1})\}$$

Nu geldt voor alle f

$$P(f)\varepsilon_{t-1} \leq \|\varepsilon_{t-1}\| e$$

Dus

$$v_\beta + \varepsilon_t \leq \max_f \{r(f) + \beta P(f) v_\beta\} + \beta \|\varepsilon_{t-1}\| e$$

Aangezien voor $t \to \infty$ ook $\varepsilon_t \to 0$ (stelling 4.1(ii)) geldt ook

$$v_\beta \leq \max_f \{r(f) + \beta P(f) v_\beta\}$$

Analoog tonen we, gebruikmakend van $P(f)\varepsilon_{t-1} \geq -\|\varepsilon_{t-1}\| e$, aan dat

$$v_\beta \geq \max_f \{r(f) + \beta P(f) v_\beta\}$$

waarmee (i) is bewezen

(ii) Laat v en w twee oplossingen van (4.7) zijn.

En de beslissingsregels f en g voldoen aan respectievelijk

$$v = r(f) + \beta P(f) v$$

en $w = r(g) + \beta P(g) w$

Dan geldt dus

$$v - w = (r(f) + \beta P(f)v) - (r(g) + \beta P(g)w)$$

$$\geq (r(g) + \beta P(g)v) - (r(g) + \beta P(g)w)$$

$$= \beta P(g)(v - w) \geq \beta P(g) \min_i (v_i - w_i)e = \beta \min_i (v_i - w_i).e$$

Dus ook $\min_i (v_i - w_i) \geq \beta \min_i (v_i - w_i)$ waaruit volgt $v \geq w$.
Met verwisseling van v en w volgt natuurlijk ook $w \geq v$ dus $v = w$. □

Uit de stellingen 4.1 en 4.2 volgt dat de oplossing v_β van de functionaalvergelijking (4.7) benaderd kan worden met de procedure (4.5). Om deze reden noemen we de methode om v_β te benaderen met de waarden van eindig staps problemen *de methode van de successieve approximaties.*

4.4 Het bestaan van stationaire optimale strategieën

Een direct gevolg van het feit dat v_β aan de functionaalvergelijking (4.7) voldoet is dat er een stationaire optimale strategie bestaat.

Stelling 4.3. Als f^* een beslissingsregel is die voldoet aan

$$(4.8) \qquad r(f^*) + \beta P(f^*)v_\beta = v_\beta$$

(zo een f^* bestaat) dan geldt

$$v_\beta(f^{*(\infty)}) = v_\beta$$

Bewijs.

$$v_\beta(f^{*(\infty)}) = r(f^*) + \beta P(f^*)r(f^*) + \beta^2 P^2(f^*)r(f^*) + \ldots$$

$$= \sum_{n=0}^{\infty} \beta^n P^n(f^*)r(f^*)$$

$$\text{vgl.}(4.8) \quad = \sum_{n=0}^{\infty} \beta^n P^n(f^*)[v_\beta - \beta P(f^*)v_\beta]$$

$$= \sum_{n=0}^{\infty} \beta^n P^n(f^*)\, v_\beta - \sum_{n=1}^{\infty} \beta^n P^n(f^*)\, v_\beta = v_\beta.$$

(Merk op dat beide sommen absoluut convergent zijn, immers

$$\sum_{n=0}^{\infty} \beta^n \left| P^n(f^*)v_\beta \right| \leq \sum_{n=0}^{\infty} \beta^n P^n(f^*) \|v_\beta\| e = (1 - \beta)^{-1} \|v_\beta\| e).$$ □

4.5 Onder en bovengrenzen voor v_β en bijna optimale stationaire strategieën

We zullen nu laten zien hoe met de dynamische programmeringsaanpak bijna optimale stationaire strategieën en onder- en bovengrenzen voor v_β gevonden kunnen worden.
Maar eerst geven we nog twee vrij zwakke resultaten die we feitelijk al eerder hebben bewezen.

Stelling 4.4.

(i) $|v_\beta - v_{\beta,t}| \leq A\beta^t/(1-\beta).e$

(ii) Zij $\hat{s}_T$ een strategie die begint met $(\hat{f}_T,\ldots,\hat{f}_1)$ met daarna een willekeurig vervolg (met $\hat{f}_t$ een maximalisator in (4.6), $t = 1,\ldots,T$) dan geldt er

$$|v_\beta(\hat{s}_T) - v_\beta| \leq 2A\beta^T/(1-\beta).e$$

Bewijs. (i) zie bewijs stelling 4.1. (ii) direct uit $v_{\beta,T}(\hat{s}_T) = v_{\beta,T}$ en het bewijs van stelling 4.1. □

De volgende stelling is een belangrijk deelresultaat voor het vinden van bijna optimale stationaire strategieën en onder- en bovengrenzen voor v_β.

Stelling 4.5

(i) Als $u \geq r(f) + \beta P(f)w$ dan

$$v_\beta(f^{(\infty)}) \leq u + \beta(1-\beta)^{-1} \max(u_i - w_i).e$$

(ii) Als $u \leq r(f) + \beta P(f)w$ dan

$$v_\beta(f^{(\infty)}) \geq u + \beta(1-\beta)^{-1} \min(u_i - w_i).e$$

Bewijs. We zullen alleen (i) bewijzen. Het bewijs van (ii) gaat volkomen analoog. Met $r(f) \leq u - \beta P(f)w$ vinden we

$$\begin{aligned} v_\beta(f^{(\infty)}) &= \sum_{n=0}^{\infty} \beta^n P^n(f) r(f) \\ &\leq \sum_{n=0}^{\infty} \beta^n P^n(f)\ (u - \beta P(f)w) \\ &= u + \sum_{n=1}^{\infty} \beta^n P^n(f)(u - w) \\ &\leq u + \sum_{n=1}^{\infty} \beta^n P^n(f) \max_i (u_i - w_i).e \end{aligned}$$

$$= u + \beta(1 - \beta)^{-1} \max_i (u_i - w_i).e \qquad \square$$

Een bijna direct gevolg van deze stelling is dat we onder- en bovengrenzen voor v_β kunnen bepalen.

Stelling 4.6

Zij $u = \max_f\{r(f) + \beta p(f)w\}$, dan geldt er

$$u + \beta(1 - \beta)^{-1} \min(u_i - w_i).e \leq v_\beta \leq u + \beta(1 - \beta)^{-1} \max_i(u_i - w_i).e$$

Bewijs. Eerst de linker ongelijkheid. Zij f_0 zodanig dat $u = r(f_0) + \beta P(f_0)w$ dan geldt met stelling 4.5 (ii)

$$v_\beta \geq v_\beta(f_0^{(\infty)}) \geq u + \beta(1 -\beta)^{-1} \min_i(u_i - w_i).e$$

Uit $u = \max_f \{r(f) + \beta P(f)w\}$ volgt $u \geq r(f^*) + \beta P(f^*)w$, voor een beslissingsregel f^* die voldoet aan (4.8). Met stelling 4.5 (i) en stelling 4.3 volgt dan de rechter ongelijkheid. $\square$

Om een nauwkeurige schatting voor v_β te vinden is het noodzakelijk dat $\max_i (u_i - w_i) - \min_i(u_i - w_i)$ klein is. Beschouw nu eens het iteratieproces

$$v_{\beta,0} = 0, \quad v_{\beta,t} = \max_f\{r(f) + \beta P(f)v_{\beta,t-1}\}, \ t = 1,2,...$$

Gemakkelijk kan nu worden aangetoond (vgl. bewijs stelling 4.1) dat

$$\|v_{\beta,t} - v_{\beta,t-1}\| \leq \beta^{t-1} M$$

zodat ook $\max_i(v_{\beta,t} - v_{\beta,t-1})(i) - \min_i(v_{\beta,t} - v_{\beta,t-1})(i) \to 0$ als $t \to \infty$.

En voor een beslissingsregel $\hat{f}_t$ die voldoet aan

$$r(\hat{f}_t) + \beta P(\hat{f}_t)v_{\beta,t-1} = v_{\beta,t}$$

geldt nu (vgl. stelling 4.5 (ii) en 4.6)

$$v_{\beta,t} + \beta(1 - \beta)^{-1} \min_i(v_{\beta,t} - v_{\beta,t-1})(i)e \leq v_\beta(\hat{f}_t^{(\infty)}) \leq v_\beta$$

$$\leq v_{\beta,t} + \beta(1 -\beta)^{-1} \max_i(v_{\beta,t} - v_{\beta,t-1})(i).e$$

Dus als $t \to \infty$ dan gaat het verschil tussen de onder- en bovengrens voor v_β naar 0 zodat ook $\hat{f}_t^{(\infty)}$ een bijna optimale strategie wordt.

4.6 De $L_\beta(f)$- en de U_β-operator

Tot nu toe hebben we steeds gewerkt met de afbeeldingen $r(f) + \beta P(f)v$ en $\max_f r(f) + \beta P(f)v$.

We zullen in deze paragraaf notaties invoeren voor deze afbeeldingen en er enige eigenschappen voor bewijzen.

Definieer de operatoren $L(f)$ en U op $\mathbb{R}^N$ door

$$L_\beta(f)v = r(f) + \beta P(f)v, \quad v \in \mathbb{R}^N$$

$$U_\beta v = \max_f \{r(f) + \beta P(f)v\}, \quad v \in \mathbb{R}^N$$

Lemma 4.2. Zij $v,w \in \mathbb{R}^N$ dan geldt

(i) $L_\beta(f)v - L_\beta(f)w \leq \beta \max_i (v - w)\ (i).\ e$

(ii) $L_\beta(f)v - L_\beta(f)w \geq \beta \min_i (v - w)\ (i).\ e$

(iii) $\|L_\beta(f)v - L_\beta(f)w\| \leq \beta\|v - w\|$

(iv) $v_\beta(f^{(\infty)})$ is de unieke oplossing van $L_\beta(f)v = v$.

Bewijs. (i) - (iii) direct uit

$$L_\beta(f)v - L_\beta(f)w = \beta P(f)\ (v - w).$$

(iv) $$v_\beta(f^{(\infty)}) = \sum_{n=0}^{\infty} \beta^n P^n(f)\ r(f) = r(f) + \beta P(f) \sum_{n=0}^{\infty} \beta^n P^n(f) r(f)$$

$$= r(f) + \beta P(f) v_\beta(f^{(\infty)}).$$

Dus $v_\beta(f^{(\infty)})$ voldoet aan $L_\beta(f)v = v$.
De uniciteit volgt nu uit (iii) immers stel $L_\beta(f)v = v$ en $L_\beta(f)w = w$ dan geldt met (iii) $\|v - w\| \leq \beta \|v - w\|$ dus $v = w$. □

Lemma 4.3. Zij $v,w \in \mathbb{R}^N$ dan geldt

(i) $U_\beta v - U_\beta w \leq \beta \max_i (v - w)\ (i).e$

(ii) $U_\beta v - U_\beta w \geq \beta \min_i (v - w)\ (i).e$

(iii) $\|U_\beta v - U_\beta w\| \leq \beta\|v - w\|$

Bewijs. Zij f_v en f_w zodanig dat

$L_\beta(f_v)v = U_\beta v$ en $L_\beta(f_w)w = U_\beta w$. Dan geldt

$$U_\beta v - U_\beta w = r(f_v) + \beta P(f_v)v - [r(f_w) + \beta P(f_w)w]$$

$$\leq r(f_v) + \beta P(f_v)v - [r(f_v) + \beta P(f_v)w]$$

$$= \beta P(f_v)(v - w) \leq \beta \max_i (v - w)\ (i).e$$

Het eerste ongelijkteken volgt uit het feit dat f_w de uitdrukking $r(f) + \beta P(f)w$ maximaliseert zodat zeker $r(f_v) + \beta P(f_v)w \leq r(f_w) + \beta P(f_w)w$. Analoog geldt ook

$$U_\beta v - U_\beta w \geq r(f_w) + \beta P(f_w)v - [r(f_w) + \beta P(f_w)w]$$

$$= \beta P(f_w)\ (v - w) \geq \beta \min_i\ (v - w)\ (i).e.$$

Bewering (iii) volgt nu door combinatie van (i) en (ii). □

Het bewijs van lemma 4.3 is feitelijk al eerder geleverd nl. in het bewijs van stelling 4.2 (ii).
We zullen deze lemma's, vooral lemma 4.3, in het vervolg vaak gebruiken voor het afleiden van ongelijkheden.

4.7. Nog enkele resultaten

We hebben in stelling 4.1 gezien dat $v_{\beta,t}$ naar v_β convergeert voor $t \to \infty$. Maar als we v_β willen benaderen met optima van eindig staps problemen dan zou het wel eens veel verstandiger kunnen zijn om niet $v_{\beta,0} = 0$ te kiezen ($v_{\beta,0}$ heet restwaarde of eindopbrengst). Gezien het feit dat v_β aan de functionaalvergelijking voldoet is starten met een $v_{\beta,0}$ in de buurt van v_β wellicht beter. Dat ook in dit geval de methode van successieve approximaties convergeert is een gevolg van het volgende lemma.

Lemma 4.4. Zij $w \in \mathbb{R}^N$ dan geldt

$$\|U_\beta w - v_\beta\| \leq \beta \|w - v_\beta\|.$$

Bewijs. Direct met $U_\beta v_\beta = v_\beta$ (stelling 4.2 (i)) en lemma 4.3 (iii). □

Stelling 4.7. Zij $w \in \mathbb{R}^N$ en $w_{\beta,t}$ de rij successieve approximaties $w_{\beta,0} = w$, $w_{\beta,t} = U_\beta w_{\beta,t-1}$, $t = 1,2,\ldots$ ($w_{\beta,t}$ is dus de waarde van het t-staps Markov beslissingsprobleem met eindopbrengst w) dan geldt

$$w_{\beta,t} \to v_\beta \quad \text{als } t \to \infty$$

Bewijs. Direct uit

$$\|w_{\beta,t} - v_\beta\| = \|U_\beta w_{\beta,t-1} - U_\beta v_\beta\| \leq \beta \|w_{\beta,t-1} - v_\beta\|$$

$$\leq \ldots \leq \beta^t \|w_{\beta,0} - v_\beta\|. \qquad \square$$

In de volgende paragraaf zullen we na n successieve approximaties te hebben uitgevoerd vast een schatting willen hebben voor de (n + m)-de approximatie. Daartoe geven we het volgende lemma.

Lemma 4.5. Zij $v_{\beta,t}$, $t = 0,1,\ldots,n$ een rij successieve approximaties, met niet noodzakelijk $v_{\beta,0} = 0$, dan geldt

(i) $$v_{\beta,n+m} \leq v_{\beta,n-1} + (1 + \beta + \ldots + \beta^m) \max_i (v_{\beta,n} - v_{\beta,n-1})_{(i)}$$

(ii) $$v_{\beta,n+m+1} \geq v_{\beta,n} + (\beta + \ldots + \beta^{m+1}) \min_i (v_{\beta,n} - v_{\beta,n-1})_{(i)}$$

De formulering van dit lemma is alleen daarom asymmetrisch omdat we het lemma in de volgende paragraaf juist in deze vorm nodig hebben.

Bewijs. We bewijzen alleen (i), het bewijs van (ii) gaat analoog.

$$v_{\beta,n+m} = (v_{\beta,n+m} - v_{\beta,n+m-1}) + (v_{\beta,n+m-1} - v_{\beta,n+m-2}) + \ldots +$$

$$+ (v_{\beta,n} - v_{\beta,n-1}) + v_{\beta,n-1} \qquad (4.9)$$

Nu is

$$v_{\beta,n+k} - v_{\beta,n+k-1} = U_\beta v_{\beta,n+k-1} - U_\beta v_{\beta,n+k-2}$$

zodat ook met lemma 4.3 (i) geldt

$$\max_i (v_{\beta,n+k} - v_{\beta,n+k-1})_{(i)} \leq \beta \max_i (v_{\beta,n+k-1} - v_{\beta,n+k-2})_{(i)}$$

Herhaalde toepassing geeft dus

$$v_{\beta,n+k} - v_{\beta,n+k-1} \leq \beta^k \max_i (v_{\beta,n} - v_{\beta,n-1})_{(i).e}$$

Door elk van de termen in (4.9) zo af te schatten vinden we (i). $\square$

4.8. Suboptimale akties

Tijdens het iteratie proces moeten we in elke iteratieslag en in elke toestand een maximalisatie van de volgende vorm uitvoeren

$$\max_k \{r(i,k) + \beta \sum_j p_{ij}^k \, v(j)\}$$

Als het aantal toestanden wat groot wordt dan wordt ook het bepalen van $\sum_j p_{ij}^k \, v(j)$ kostbaar. Het zou dan heel plezierig zijn al van te voren te weten dat een aantal akties toch een slecht resultaat geven zodat die bij het bepalen van het maximum buiten beschouwing gelaten kunnen worden.
In deze paragraaf geven we twee methoden om deze onvoordelige acties, *suboptimale* acties te elimineren.

Opdat een actie k in de (n+m+1)- de iteratieslag optimaal is in toestand i moet gelden

$$r(i,k) + \beta \sum_j p_{ij}^k \, v_{\beta,n+m}(j) = v_{\beta,n+m+1}(i)$$

Dus als we al weten dat

$$r(i,k) + \beta \sum_j p_{ij}^k \, v_{\beta,n+m}(j) < v_{\beta,n+m+1}(i) \tag{4.10}$$

dan kunnen we actie k bij het bepalen van

$$\max_k \{r(i,k) + \beta \sum_j p_{ij}^k \, v_{\beta,n+m}(j)\}$$

buiten beschouwing laten.
Nu kennen we in ieder geval $v_{\beta,n+m+1}$ niet van te voren.
Maar als voor een bovengrens $\overline{v}_{\beta,n+m}$ voor $v_{\beta,n+m}$ en een ondergrens $\underline{v}_{\beta,n+m+1}$ voor $v_{\beta,n+m+1}$ geldt

$$r(i,k) + \beta \sum_j p_{ij}^k \, \overline{v}_{\beta,n+m}(j) < \underline{v}_{\beta,n+m+1}(i) \tag{4.11}$$

dan geldt zeker ook ongelijkheid (4.10) zodat k suboptimaal is in toestand i in iteratieslag n + m + 1.
Definiëren we voor een vector $v \in \mathbb{R}^N$ het span door

$$sp(v) = \max_i v(i) - \min_i v(i)$$

dan krijgen we het volgende resultaat

Stelling 4.8. (suboptimaliteitstest)

Als $v_{\beta,n}(i) - r(i,\hat{k}) - \beta \sum p_{ij}^{\hat{k}} v_{\beta,n-1}(j) > (\beta + \ldots + \beta^{m+1})\ sp(v_{\beta,n} - v_{\beta,n-1})$

dan is actie $\hat{k}$ suboptimaal in stap $n + m + 1$ in toestand i.

Bewijs. We hebben gezien dat $\hat{k}$ suboptimaal is als (4.11) geldt. Substitueren we nu in (4.11) de onder en bovengrenzen uit lemma 4.5 dan krijgen we

$$r(i,\hat{k}) + \beta \sum_j p_{ij}^{\hat{k}} v_{\beta,n-1}(j) + (\beta + \ldots + \beta^{m+1}) \max_i (v_{\beta,n} - v_{\beta,n-1})\ (i)$$

$$< v_{\beta,n}(i) + (\beta + \ldots + \beta^{m+1}) \min_i (v_{\beta,n} - v_{\beta,n-1})\ (i)$$

ofwel

$$v_{\beta,n}(i) - r(i,\hat{k}) - \beta \sum_j p_{ij}^{\hat{k}} v_{\beta,n-1}(j) > (\beta + \ldots + \beta^{m+1})\ sp(v_{\beta,n} - v_{\beta,n-1})$$

waarmee de stelling is bewezen. □

Dus als we na de berekening van $v_{\beta,n}(i)$ nog de verschillen $v_{\beta,n}(i) - r(i,k) - \beta \sum_j p_{ij}^{k} v_{\beta,n-1}(j)$ bepalen dan kunnen we door dit te vergelijken met $sp(v_{\beta,n} - v_{\beta,n-1})$ inzien of actie k in een of meer van de komende iteratiestappen suboptimaal is, en dus geëlimineerd kan worden.

Stelling 4.9. Als

$$(4.12) \qquad v_{\beta,n}(i) - r(i,k) - \beta \sum p_{ij}^{k} v_{\beta,n-1}(j) > \beta(1-\beta)^{-1}\ sp(v_{\beta,n} - v_{\beta,n-1})$$

dan is k niet alleen suboptimaal in toestand i in alle volgende iteratieslagen maar geldt ook als $f(i) = k$

$$V(f^{(\infty)})\ (i) < v_{\beta}(i).$$

Dus een stationaire strategie f met $f(i) = k$ is niet optimaal in het ∞-horizon probleem.

Bewijs. Opdat een stationaire strategie f optimaal is in het ∞-horizon probleem is het niet alleen voldoende dat f voldoet aan de functionaalvergelijking (stelling 4.3) het is ook nodig. (ga zelf na, vgl. bewijs van stelling 4.3, en stelling 4.5). Dus als voor onder-en bovengrenzen $\underline{v}_{\beta}$ en $\bar{v}_{\beta}$ voor v_{β} geldt

$$(4.13) \qquad r(i,k) + \beta \sum p_{ij}^{k} \bar{v}_{\beta}(i) < \underline{v}_{\beta}(i)$$

dan is k suboptimaal. Door nu in lemma 4.5 m naar ∞ te laten gaan vinden we de grenzen

$$v_{\beta,n} + \beta(1-\beta)^{-1} \min_i (v_{\beta,n} - v_{\beta,n-1})(i).e \leq v_\beta \leq v_{\beta,n-1} + (1-\beta)^{-1} \max_i (v_{\beta,n} - v_{\beta,n-1})(i).e$$

Invullen van deze grenzen in (4.13) geeft nu (4.12). □

Bij de suboptimaliteitstest uit stelling 4.8 beslissen we steeds van te voren voor hoeveel iteratieslagen de actie geëlimineerd wordt. Het gebruik van andere grenzen leidt tot een andere test waarbij per keer bekeken wordt of een actie geëlimineerd kan worden (blijven).

Stelling 4.10. Als

$$(4.14)\quad v_{\beta,n}(i) - r(i,k) - \beta \sum_j p_{ij}^k v_{\beta,n-1}(j) > \beta \sum_{l=0}^{m} sp(v_{\beta,n+l} - v_{\beta,n+l-1})$$

dan is actie k suboptimaal in stap n + m + 1 in toestand i.

Bewijs. Als we in de uitdrukking (4.9) voor $v_{\beta,n+m}$ de termen $v_{\beta,n+k} - v_{\beta,n+k-1}$ naar boven afschatten met hun maximale component dan krijgen we als bovengrens voor $v_{\beta,n+m}$:

$$\bar{v}_{\beta,n+m} = v_{\beta,n-1} + \sum_{\ell=0}^{m} \max_i (v_{\beta,n+\ell} - v_{\beta,n+\ell-1})\ (i).e$$

schatten we in

$$v_{\beta,n+m+1} = (v_{\beta,n+m+1} - v_{\beta,n+m}) + (v_{\beta,n+m} - v_{\beta,n+m-1})$$

$$+ \ldots + (v_{\beta,n+1} - v_{\beta,n}) + v_{\beta,n}$$

de termen $v_{\beta,n+\ell+1} - v_{\beta,n+\ell}$ naar beneden af met $\beta \min(v_{\beta,n+\ell} - v_{\beta,n+\ell-1})(i)$ dan vinden we als ondergrens voor $v_{\beta,n+m+1}$

$$\underline{v}_{\beta,n+m+1} = v_{\beta,n} + \beta \sum_{\ell=0}^{m} \min_i (v_{\beta,n+\ell} - v_{\beta,n+\ell-1})(i).e$$

Substitutie van deze onder- en bovengrenzen in (4.11) geeft dan (4.14). □

Hoe passen we nu deze suboptimaliteitstest toe?

Definieer $\varphi(0,i,k,n) := v_{\beta,n}(i) - r(i,k) - \beta \sum_j p_{ij}^k v_{\beta,n-1}(j)$

Om te zien of k suboptimaal is in stap n + 1 beschouwen we $\varphi(1,i,k,n) := \varphi(0,i,k,n) - \beta\, sp(v_{\beta,n} - v_{\beta,n-1})$. Is $\varphi(1,i,k,n) > 0$

dan is k suboptimaal in stap n + 1. Vervolgens bepalen we $v_{\beta,n+1}$. Beschouw nu $\varphi(2,i,k,n) := \varphi(1,i,k,n) - \beta\ sp(v_{\beta,n+1} - v_{\beta,n})$. Is $\varphi(2,i,k,n) > 0$ dan is k ook suboptimaal in stap n + 2. Etc. Als op zeker moment $\varphi(1,i,k,n) \leq 0$ wordt, dan nemen we actie k bij de bepaling van $v_{n+1}(i)$ weer in beschouwing. Daarna bepalen we weer $\varphi(0,i,k,n+1)$.

Deze test kost natuurlijk iets meer werk. Dat zal alleen dan lonend zijn als $sp(v_{\beta,n+1} - v_{\beta,n+1-1})$ echt sneller afneemt dan met de factor β.
Dat dit soms inderdaad het geval is wordt in het volgende voorbeeld aangetoond.

Voorbeeld 4.1.

Laat op zeker moment beslissingsregel f twee keer achtereen een maximalisator zijn. D.w.z.

$$L_\beta(f)v_{\beta,n-1} = U_\beta v_{\beta,n-1} = v_{\beta,n}\ ,\ \text{en}$$
$$L_\beta(f)v_{\beta,n} = U_\beta v_{\beta,n} = v_{\beta,n+1}.$$

En zij

$$P(f) = \begin{pmatrix} 1/3 & 2/3 \\ 3/4 & 1/4 \end{pmatrix}$$

Dan geldt

$$sp(v_{\beta,n+1} - v_{\beta,n}) = sp(r(f) + \beta P(f)v_{\beta,n} - r(f) - \beta\ P(f)v_{\beta,n-1})$$

$$= \beta sp(P(f)(v_{\beta,n} - v_{\beta,n-1}))$$

splitsen we nu P(f) in twee matrices Q(f) en R(f) alsvolgt

$$Q(f) = \begin{pmatrix} 1/3 & 1/4 \\ 1/3 & 1/4 \end{pmatrix}\ ,\quad R(f) = \begin{pmatrix} 0 & 5/12 \\ 5/12 & 0 \end{pmatrix}$$

Dan geldt dus

$$sp(v_{\beta,n+1} - v_{\beta,n}) = \beta sp(Q(f)(v_{\beta,n} - v_{\beta,n-1}) + R(f)\ (v_{\beta,n} - v_{\beta,n-1}))$$

Omdat $Q(f)(v_{\beta,n} - v_{\beta,n-1})$ een constante vector kan die bij de bepaling van het span worden weggelaten, zodat we vinden:

$$sp(v_{\beta,n+1} - v_{\beta,n}) = \beta sp(R(f)(v_{\beta,n} - v_{\beta,n-1})) = 5/12\,\beta\ sp(v_{\beta,n} - v_{\beta,n-1})$$

4.9. De policy iteration methode

In de voorgaande paragrafen hebben we gezien hoe we de waarde van het ∞-staps probleem konden benaderen met de waarde van een eindig staps probleem, met de verfijningen zoals betere grenzen, bijna optimale stationaire strategieën en suboptimaliteitstesten.
In deze paragraaf willen we een heel andere methode bekijken.
Een methode van het type strategieverbetering.

Laat $f_0^{(\infty)}$ een willekeurige stationaire strategie zijn.
Dan voldoet $v_\beta(f_0^{(\infty)})$ aan

$$v_\beta(f_0^{(\infty)}) = \sum_{n=0}^{\infty} \beta^n P^n(f_0)\, r(f_0) = r(f_0) + \beta P(f_0) \sum_{n=0}^{\infty} \beta^n P^n(f_0)\, r(f_0)$$

$$= r(f_0) + \beta P(f_0)\, v_\beta(f_0^{(\infty)}).$$

Dit is een eenduidig oplosbaar stelsel van N vergelijkingen met N onbekenden, 1 voor elke toestand. Bepaal nu $v_\beta(f_0^{(\infty)})$ door dit stelsel op te lossen. We kunnen nu proberen de strategie $f_0^{(\infty)}$ te verbeteren.
Zij f_1 zodanig dat

$$r(f_1) + \beta P(f_1)\, v_\beta(f_0^{(\infty)}) = \max_f \{r(f) + \beta P(f) v_\beta(f_0^{(\infty)})\}$$

Dan geldt voor zekere $\delta \in \mathbb{R}^N$, $\delta \geq 0$

$$r(f_1) + \beta P(f_1) v_\beta(f_0^{(\infty)}) = r(f_0) + \beta P(f_0) v_\beta(f_0^{(\infty)}) + \delta = v_\beta(f_0^{(\infty)}) + \delta$$

Vervolgens kunnen we weer $v_\beta(f_1^{(\infty)})$ bepalen.
Dat dit leidt tot een zinvol algorithme blijkt uit de volgende stelling

Stelling 4.11.
Als $L(f_1)v_\beta(f_0^{(\infty)}) = U_\beta v_\beta(f_0^{(\infty)}) = v_\beta(f_0^{(\infty)}) + \delta$ $(\delta \geq 0)$ dan

$$v_\beta(f_1^{(\infty)}) = v_\beta(f_0^{(\infty)}) + \sum_{n=0}^{\infty} \beta^n P^n(f_1)\delta \geq v_\beta(f_0^{(\infty)}).$$

En als $\delta = 0$ dan is $f_0^{(\infty)}$ een optimale stationaire strategie en geldt dus $v_\beta(f_0^{(\infty)}) = v_\beta$.

Bewijs.

$$v_\beta(f_1^{(\infty)}) = \sum_{n=0}^{\infty} \beta^n P^n(f_1)\, r(f_1) = \sum_{n=0}^{\infty} \beta^n P^n(f_1)[v_\beta(f_0^{(\infty)} + \delta - \beta P(f_1) v_\beta(f_0^{(\infty)}]$$

$$= v_\beta(f_0^{(\infty)}) + \sum_{n=0}^{\infty} \beta^n P^n(f_1)\delta \quad \text{(vgl. bewijs stelling 4.3)}$$

Dus $v_\beta(f_1^{(\infty)}) \geq v_\beta(f_0^{(\infty)})$ en $f_1^{(\infty)}$ is alleen dan niet echt beter dan $f_0^{(\infty)}$ als $\delta = 0$.
Maar $\delta = 0$ betekent $U_\beta v_\beta(f_0^{(\infty)}) = v_\beta(f_0^{(\infty)})$ dus $v_\beta(f_0^{(\infty)})$ voldoet aan de functionaalvergelijking (4.7) die de unieke oplossing v_β heeft.
Dus $v_\beta(f_0^{(\infty)}) = v_\beta$ en $f_0^{(\infty)}$ is een optimale stationaire strategie. □

Er zijn maar eindig veel toestanden (N) en maar eindig veel beslissingen per toestand (K_0) dus er zijn maar eindig veel stationaire strategieën (K_0^N). Zodat uit stelling 4.11 volgt dat we na eindig veel strategieverbeteringen (hoogstens K_0^N-1) de waarde en een optimale strategie zullen vinden.
In het algemeen zal het aantal benodigde iteratie veel lager zijn dat $K_0^N - 1$. Meestal zal dit aantal ook zelfs aanzienlijk lager zijn dat het aantal iteratieslagen dat bij successieve approximaties nodig is om een redelijke nauwkeurigheid te bereiken. Het nadeel van deze methode is dat er in iedere slag een stelsel van N vergelijkingen met N onbekenden moet worden opgelost om $v_\beta(f^{(\infty)})$ te bepalen. Als N groot wordt, wordt dit kostbaar of zelfs onuitvoerbaar. In praktijkproblemen waar N al gauw wat groter is heeft daarom meestal de methode van de successieve approximaties (met nog wat extra verfijningen die we hier niet hebben bekeken) de voorkeur.

4.10 Formulering van het verdisconteerde Markov beslissingsprobleem als linear programmeringsprobleem

We zullen in deze paragraaf laten zien dat we het verdisconteerde Markov beslissingsprobleem ook als lineair programmeringsprobleem kunnen formuleren.
Oplossen van het Markov beslissingsprobleem met verdiscontering houdt in het bepalen van een strategie s die de uitdrukking

$$(4.1) \qquad v_\beta(s) = \mathbf{E}_s \sum_{n=0}^{\infty} \beta^n r(X_n, B_n)$$

maximaliseert.

Met behulp van

$$(4.15) \qquad \mathbb{E}_s\, r(X_n,B_n) = \sum_{i,k} \mathbb{P}\,(X_n = i,\ B_n = k)\ r(i,k)$$

kunnen we (4.1) als volgt herschrijven

$$(4.16) \qquad v_\beta(s) = \sum_{n=0}^{\infty} \beta^n \sum_{i,k} \mathbb{P}\,(X_n = i,\ B_n = k)\ r(i,k)$$

Laat nu $\pi(0)$ een gegeven startvector zijn

$$\pi(0) = (\pi_1(0),\ldots,\pi_N(0))$$

met $\pi_i(0)$ de kans dat het beslissingsproces in toestand i start. En definieer $\pi_i^k(n)$ als de kans dat gegeven de beginverdeling $\pi(0)$ het proces bij het toepassen van strategie s op tijdstip n in i zit en daar beslissing k wordt genomen.
Notatie:

$$(4.17) \qquad \pi_i^k(n) = \mathbb{P}_{\pi(0),s}\,(X_n = i,\ B_n = k)$$

Gegeven de startverdeling $\pi(0)$ is het probleem dus de volgende uitdrukking te maximaliseren:

$$(4.18) \qquad \sum_{n=0}^{\infty} \beta^n \sum_{i,k} \pi_i^k(n)\ r(i,k)$$

Waarbij in ieder geval moet gelden

$$(4.19) \qquad \sum_k \pi_i^k(0) = \pi_i(0),\ \pi_i^k(0) \geq 0$$

Maar ook ten aanzien van de $\pi_i^k(n)$ zijn er een aantal voorwaarden. Deze volgen uit

$$(4.20) \qquad \mathbb{P}_{\pi(0),s}\,(X_{n+1} = j) = \sum_{i,\ell} \mathbb{P}_{\pi(0),s}\,(X_n = 1,\ B_n = \ell)\, p_{ij}^{\ell}$$

en

$$(4.21) \qquad \mathbb{P}_{\pi(0),s}\,(X_{n+1} = j) = \sum_k \mathbb{P}_{\pi(0),s}\,(X_{n+1} = j,\ B_{n+1} = k)$$

Bij gegeven beginverdeling $\pi(0)$ leiden (4.20) en (4.21) tot de volgende vergelijkingen

$$(4.22) \qquad \sum_k \pi_j^k(n+1) = \sum_{i,\ell} \pi_i^\ell(n) p_{ij}^\ell \quad , \; n = 0,1,\ldots.$$

We zien dus dat de bij een strategie s behorende collectie $\{\pi_i^k(n)\}$ een toegelaten oplossing is van het onderstaande lp. probleem.

$$I \begin{cases} \max\limits_{\{\pi_i^k(n)\}_{i,k,n}} \sum\limits_{n=0}^{\infty} \beta^n \sum\limits_{i,k} \pi_i^k(n)\, r(i,k) \\ \text{onder de voorwaarden} \\ \sum\limits_k \pi_i^k(0) = \pi_i(0), \; i \in I \\ \sum\limits_k \pi_j^k(n+1) = \sum\limits_{i,\ell} \pi_i^\ell(n)\, p_{ij}^\ell \quad , \; j \in I, \; k \in K, \; n \geq 0 \\ \pi_i^k(n) \geq 0 \quad , \qquad\qquad j \in I, \; k \in K, \; n \geq 0. \end{cases}$$

Opdat lp. probleem I ook equivalent is met het oorspronkelijke probleem, het maximaliseren van (4.18), is het voldoende dat er bij elke toegelaten oplossing $\{\pi_i^k(n)\}$ ook een strategie s bestaat die deze $\pi_i^k(n)$ oplevert. Er is zelfs al een gemengde Markov strategie die de gewenste $\pi_i^k(n)$ geeft, namelijk de strategie $s = (s_0, s_1, \ldots)$ met s_n gedefinieerd door

$$s_n(i,k) = \begin{cases} \pi_i^k(n) \;/\; \sum_k \pi_i^k(n) \text{ als } \sum_k \pi_i^k(n) > 0 \\ \text{willekeurig als } \sum_k \pi_i^k(n) = 0. \end{cases}$$

Nu heeft probleem I nog aftelbaar veel variabelen en aftelbaar veel beperkingen. Om probleem I in een meer geschikte vorm, met eindig veel variabelen en eindig veel beperkingen te brengen kunnen we de volgende variabelentransformatie uitvoeren

$$(4.23) \qquad x_i^k := \sum_{n=0}^{\infty} \beta^n \pi_i^k(n)$$

Ook voor de x_i^k kunnen we een vergelijking afleiden. Namelijk vermenigvuldigen we (4.22) met β^{n+1} en sommeren we over n dan vinden we

$$\sum_{n=0}^{\infty} \beta^{n+1} \sum_{k} \pi_j^k(n+1) = \sum_{n=0}^{\infty} \beta^{n+1} \sum_{i,\ell} \pi_i^\ell(n)\ p_{ij}^\ell$$

ofwel na verwisseling van de sommatie volgorde

$$\sum_{k} \sum_{n=0}^{\infty} \beta^{n+1}\ \pi_j^k(n+1) = \sum_{i,\ell} \beta p_{ij}^\ell \sum_{n=0}^{\infty} \beta^n\ \pi_i^\ell(n)$$

Hetgeen we met (4.23) kunnen herleiden tot

$$\sum_{k}(x_j^k - \pi_j^k(0)) = \sum_{i,\ell} \beta p_{ij}^\ell\ x_i^\ell$$

ofwel

$$\sum_{k} x_j^k - \beta \sum_{i,\ell} p_{ij}^\ell\ x_i^\ell = \pi_j(0) \tag{4.24}$$

Uit lp. probleem I vinden we zo dus met de transformatie (4.23) een tweede lp. probleem:

II

$$\max_{\{x_i^k\}_{i,k}} \sum_{i,k} x_i^k\ r(i,k)$$

onder de voorwaarden

$$\sum_{k} x_j^k - \beta \sum_{i,\ell} p_{ij}^\ell\ x_i^\ell = \pi_j(0), \quad j \in I$$

$$x_i^k \geq 0 \quad , i \in I,\ k \in K$$

De voorwaarden bij probleem II zijn zwakker dan de voorwaarden bij probleem I. Dat toch probleem II een geschikte vertaling van het oorspronkelijke probleem is blijkt uit het volgende:
Een basisoplossing van probleem II heeft precies N niet-nul variabelen. Beschouw nu eens een basisoplossing $\{x_i^k\}_{i,k}$. En stel eens dat voor zekere i

$$x_i^{k_1} > 0,\ x_i^{k_2} > 0 \text{ en } k_1 \neq k_2$$

Dan moet er een j zijn met $x_j^k = 0$ voor alle k.
Maar volgens (4.24) geldt $\Sigma_k x_j^k \geq \pi_j(0)$.
We hebben dus het volgende lemma

Lemma 4.6. Als $\{x_i^k\}_{i,k}$ een basisoplossing is van probleem II en als $\pi_i(0) > 0$ voor alle i dan is er bij elke i precies één k_i waarvoor $x_i^{k_i} > 0$.

Dus van de collecties $x_i^1,\ldots,x_i^{K_0}$ komt steeds precies één variabele in de basis voor. Gevolg hiervan is dat (als alle $\pi_i(0) > 0$) basisoplossingen alleen kunnen corresponderen met stationaire strategieën. Resteert nog de vraag of elke basisoplossing correspondeert met een (noodzakelijkerwijs unieke) stationaire strategie (het omgekeerde volgt direct uit de wijze waarop probleem II is afgeleid).

Lemma 4.7. Als $\pi_i(0) > 0$ voor alle i dan correspondeert met elke basisoplossing van probleem II precies één stationaire strategie en omgekeerd.

Bewijs. Laat $\underline{x} = (\underline{x}_1^{k(1)},\ldots,\underline{x}_N^{k(N)})$ een basisoplossing van II zijn dan voldoet $\underline{x}$ aan (4.24), dus

$$\underline{x}(I - \beta P(k)) = \pi(0)$$

En beschouwen we de stationaire strategie $k^{(\infty)}$ (actie k(i) in toestand i, $i \in I$). Dan volgen via (4.17) en (4.23) de bijbehorende x_i^k. Natuurlijk $x_i^k = 0$ als $k \neq k(i)$ en verder voldoet $x = (x_1^{k(1)},\ldots,x_N^{k(N)})$ weer aan (4.24)

$$x(I - \beta P(k)) = \pi(0)$$

Nu is $(I - \beta P)$ regulier, dus de oplossing van $y(I - \beta P) = \pi(0)$ is uniek, zodat $x = \underline{x}$.

Dus er is precies een basisoplossing $(y_1^{k(1)},\ldots,y_N^{k(N)})$ en die is juist gelijk aan de bij de stationaire strategie $k^{(\infty)}$ behorende $(x_1^{k(1)},\ldots,x_N^{k(N)})$. □

We weten nu dat extreme punten van het toegelaten gebied van II corresponderen met stationaire strategieën. Het is ook bekend dat het optimum van een lp probleem in een extreem punt wordt aangenomen. Dus, hoewel de beperkingen van probleem II zwakker zijn dan de beperkingen bij I, zijn de extrema gelijk.

We kunnen dus de waarde en een optimale stationaire strategie voor het Markov beslissingsprobleem bepalen door het lp. probleem II op te lossen. Hiermee is dus opnieuw aangetoond dat er een optimale stationaire strategie bestaat.

4.11 Relatie tussen lineaire programmering en policy iteration

We zullen in deze paragraaf laten zien dat er een grote overeenkomst is tussen het lp. probleem II en de policy iteration methode.
We kunnen voor lp. probleem II het volgende starttableau opstellen.

x_1^1	x_1^2	...	$x_1^{K_0}$	x_2^1	...	$x_2^{K_0}$	...	x_N^1	...	$x_N^{K_0}$	
$1-\beta p_{11}^1$	$1-\beta p_{11}^2$	...	$1-\beta p_{11}^{K_0}$	$-\beta p_{21}^1$	...	$-\beta p_{21}^{K_0}$	...	$-\beta p_{N1}^1$	...	$-\beta p_{N1}^{K_0}$	$\pi_1(0)$
$-\beta p_{12}^1$	$-\beta p_{12}^2$	...	$-\beta p_{12}^{K_0}$	$1-\beta p_{22}^1$	...	$1-\beta p_{22}^{K_0}$	...	$-\beta p_{N2}^1$	...	$-\beta p_{N2}^{K_0}$	$\pi_2(0)$
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
$-\beta p_{1N}^1$	$-\beta p_{1N}^2$	...	$-\beta p_{1N}^{K_0}$	$-\beta p_{2N}^1$	...	$-\beta p_{2N}^{K_0}$	...	$1-\beta p_{NN}^1$	...	$1-\beta p_{NN}^{K_0}$	$\pi_N(0)$
$r(1,1)$	$r(1,2)$	...	$r(1,K_0)$	$r(2,1)$	...	$r(2,K_0)$	...	$r(N,1)$	...	$r(N,K_0)$	

De grondgedachte van de simplexmethode is het vervangen van basisoplossingen door betere basisoplossingen tot dit niet langer kan. De laatst gevonden basisoplossing is dan optimaal.
Bij de simplexmethode vervangen we in iedere slag slechts één van de basisvariabelen. We weten hier dat steeds één van de variabelen $x_i^1 \ldots x_i^{K_0}$ in de basis zal zitten ($\pi(0) > 0$) en we kunnen dus proberen meerdere variabelen uit de basis tegelijk te vervangen.
Beschouw eens een willekeurige basisoplossing $(x_1^{k(1)},\ldots,x_N^{k(N)})$. Om te onderzoeken of deze basisoplossing optimaal is, bepalen we een vector van veegconstanten $v = (v_1,\ldots,v_N)$, zodat, als we de vergelijkingen van het tableau met deze veegconstanten vermenigvuldigen en vervolgens van de laatste rij aftrekken, we juist op de plaatsen $r(i,k(i))$, $i \in I$, nullen krijgen.

$$(4.25)\quad \begin{cases} v_1(1-\beta p_{11}^{k(1)}) + v_2(-\beta p_{12}^{k(1)}) + \ldots + v_N(-\beta p_{1N}^{k(1)}) - r(1,k(1)) = 0 \\ \vdots \\ v_1(-\beta p_{N1}^{k(N)}) + v_2(-\beta p_{N2}^{k(N)}) + \ldots + v_N(1-\beta p_{NN}^{k(N)}) - r(N,k(N)) = 0 \end{cases}$$

Merk op dat de vector van veegconstanten v die aan (4.25) voldoet uniek is. Immers v voldoet aan $(I - \beta P(k))v = r(k)$ welk stelsel de unieke oplossing $v_\beta(k^{(\infty)})$ heeft.
De vraag of de basisoplossing $(x_1^{k(1)},\ldots,x_N^{k(N)})$ optimaal is, is nu eenvoudig te beantwoorden. Immers de basisoplossing is optimaal als er na het vegen in de laatste rij geen negatieve elementen

meer voorkomen dus als

$$(4.26) \qquad r(i,k) + \beta \sum_j p_{ij}^k \, v_\beta(j,k^{(\infty)}) - v_\beta(i,k^{(\infty)}) =: \delta(i,k)$$

voor alle $k \in K$ en $i \in I$ kleiner of gelijk aan nul is.
Als voor sommige i en k nog geldt $\delta(i,k) > 0$ dan kunnen we een betere basisoplossing bepalen, b.v. de basis $(x_1^{\tilde{k}(1)},\ldots,x_N^{\tilde{k}(N)})$ met $\tilde{k}(i)$ zodanig dat

$$\begin{cases} \tilde{k}(i) = k(i) \text{ als } \max_k \delta(i,k) = 0 \quad (\delta(i,k(i)) = 0) \\ \\ \tilde{k}(i) \text{ is maximalisator van } \delta(i,k) \text{ als } \max_k \delta(i,k) > 0. \end{cases}$$

Definiëren we $\delta \in \mathbb{R}^N$ door

$$\delta(i) = \max_k \delta(i,k) \quad , \; i \in I$$

dan kunnen we het voorgaande als volgt samenvatten

(i) de stationaire strategie $k^{(\infty)}$ ofwel de basisoplossing $(x_1^{k(1)},\ldots,x_N^{k(N)})$ is optimaal als $\delta = 0$

(ii) de strategie $k^{(\infty)}$ kan worden verbeterd tot $\tilde{k}^{(\infty)}$ (de basisoplossing $x_1^{k(1)},\ldots,x_N^{k(N)}$ kan worden verbeterd tot $x_1^{\tilde{k}(1)},\ldots,x_N^{\tilde{k}(N)}$) met

$$r(\tilde{k}) + \beta P(\tilde{k}) \, v_\beta(k^{(\infty)}) = v_\beta(k^{(\infty)}) + \delta$$

als $\delta \neq 0$ (altijd $\delta \geq 0$)

Vergelijken we dit resultaat met paragraaf 4.9, in het bijzonder stelling 4.11., dan zien we dat bovenstaande methode om het lineaire programmeringsprobleem II op te lossen precies overeenkomt met de policy iteration methode.

5. Markov beslissingsproblemen met oneindige tijdshorizon met als criterium gemiddelde opbrengst per tijdseenheid

5.1 Inleiding

In hoofdstuk 4 hebben we een verdisconteringsfactor ingevoerd, omdat in het beslissingsprobleem met oneindige tijdshorizon de totale verwachte opbrengst in het algemeen niet eindig of niet gedefinieerd is. Daarmee kregen we een geschikte criteriumfunctie om strategieën te vergelijken.
Een andere mogelijke criteriumfunctie is de gemiddelde verwachte opbrengst per tijdseenheid.

Definitie 5.1. Zij s een willekeurige strategie en v een willekeurige vector in $\mathbb{R}^n$ dan definiëren we de vector $V_n(s,v)$ door

$$(5.1) \qquad V_n(s,v) := \mathbb{E}_s\left[\sum_{t=0}^{n-1} r(X_t,A_t) + v(X_n)\right].$$

Dus $V_n(s,v)$ is de totale verwachte opbrengst bij gebruik van strategie s in het n-staps beslissingsprobleem met eindopbrengst v.

Definitie 5.2. Zij s een willekeurige strategie dan definiëren we g(s), de gemiddelde opbrengst per tijdseenheid onder strategie s, door

$$(5.2) \qquad g(s) = \liminf_{n\to\infty} n^{-1} V_n(s,0) \ .$$

In de definitie (5.2) kan in plaats van liminf ook limsup genomen worden.
We zijn nu in eerste instantie geïnteresseerd in het bepalen van een strategie s die vector g(s) maximaliseert.

5.2 Stationaire strategieën

In het algemeen geldt dat er al een stationaire strategie is die g(s) maximaliseert. We zullen dit voor een speciaal geval nog bewijzen. Laten we daarom eerst de stationaire strategieën wat nauwkeuriger bekijken. Voor een stationaire strategie $f^{(\infty)}$ kunnen we $V_n(f^{(\infty)},v)$ herschrijven in de vorm

$$(5.3) \qquad V_n(f^{(\infty)},v) = \sum_{t=0}^{n-1} P^t(f)\, r(f) + P^n(f)v$$

Schrijven we verder g(f) in plaats van $g(f^{(\infty)})$ dan geldt dus

$$g(f) = \liminf_{n\to\infty} n^{-1} \sum_{t=0}^{n-1} P^t(f)r(f) = P^*(f)r(f)$$

met

$$P^*(f) := \lim_{n\to\infty} n^{-1} \sum_{t=0}^{n-1} P^t(f).$$

We weten dat per recurrente klasse van P(f) de rijen van $P^*(f)$ gelijk zijn dus ook dat g(f) per klasse van P(f) constant is. In het geval van meerdere klassen zal g(f) in het algemeen geen constante vector zijn. Voor een transiente toestand i zal in dat geval g(f)(i) een gewogen som van de g-waarden op de verschillende klassen zijn met wegingsfactoren gelijk aan de kansen om uiteindelijk in de verschillende recurrente klassen terecht te komen. We zullen ons in het vervolg beperken tot beslissingsproblemen waarin de bij de verschillende stationaire strategieën behorende Markov ketens steeds *aperiodiek* zijn. In paragraaf 5.3 zal worden aangetoond dat dit geen wezenlijke beperking is.
Gevolg:

$$P^*(f) = \lim_{n\to\infty} P^n(f)$$

Verder zegt stelling 2.26.4 in Stochastische Processen I dat de convergentie van $P^n(f)$ naar $P^*(f)$ geometrisch verloopt. Omdat ook de verblijfduur in transiente toestanden exponentieel begrensd is geldt er

Lemma 5.1. Als de bij P(f) behorende Markov keten aperiodiek is, dan bestaan er getallen b en $0 \le \rho < 1$ zodat voor alle i,j en n

$$\left|P^n(f)_{ij} - P^*(f)_{ij}\right| \le b\rho^n.$$

We bewijzen dit lemma hier niet.

Beschouw nu $V_n(f^{(\infty)},0)$:

$$V_n(f^{(\infty)},0) = \sum_{t=0}^{n-1} P^t(f)r(f)$$

$$= \sum_{t=0}^{n-1} P^*(f)r(f) + \sum_{t=0}^{n-1} (P^t(f) - P^*(f))r(f) \qquad (5.4)$$

$$= ng(f) + \sum_{t=0}^{n-1} (P^t(f) - P^*(f))r(f)$$

Definiëren we

$$(5.5) \qquad v(n,f) := \sum_{t=0}^{n-1} (P^t(f) - P^*(f))r(f)$$

dan volgt uit lemma 5.1 dat

$$\lim_{n\to\infty} v(n,f) \text{ bestaat}$$

Definieer nog

$$(5.6) \qquad v(f) = \lim_{n\to\infty} v(n,f)$$

Dan geldt de volgende stelling:

Stelling 5.1

(i) $P(f)g(f) = g(f)$

(ii) $r(f) + P(f)v(f) = v(f) + g(f)$

(iii) $P^*(f)v(f) = 0$

Bewijs. Uit

$$V_t(f^{(\infty)},0) = t\, g(f) + v(t,f) \quad , t = n,n+1$$

en

$$V_{n+1}(f^{(\infty)},0) = r(f) + P(f)\, V_n(f^{(\infty)},0)$$

volgt:

$$(5.7) \qquad r(f) + P(f)(ng(f) + v(n,f)) = (n+1)g(f) + v(n+1,f)$$

Delen we links en rechts door n dan volgt met $n \to \infty$

$$P(f)g(f) = g(f)$$

(Dit is ook direct in te zien met $g(f) = P^*(f)r(f)$ en $P(f)P^*(f) = P^*(f)$).

Hiermee kunnen we (5.7) reduceren tot

$$r(f) + P(f)v(n,f) = g(f) + v(n+1,f)$$

Met (5.6) en $n \to \infty$ volgt hieruit bewering (ii)

De laatste bewering volgt door v(n,f) voor te vermenigvuldigen

met $P^*(f)$:

$$P^*(f)v(n,f) = P^*(f)\sum_{t=0}^{n-1}(P^t(f) - P^*(f))r(f)$$

zodat met

$$P^*(f)P(f) = P^*(f) \text{ en } P^*(f)P^*(f) = P^*(f) \text{ volgt}$$

$$P^*(f)(P^t(f) - P^*(f)) = P^*(f)P^t(f) - P^*(f) =$$

$$= P^*(f)\,P^{t-1}(f) - P^*(f) = \ldots = P^*(f) - P^*(f) = 0$$

Dus

$$P^*(f)v(n,f) = 0$$

Met (5.6) en $n \to \infty$ volgt nu ook (iii). □

Er geldt bovendien dat het stelsel uit stelling 5.1 uniek oplosbaar is.

Stelling 5.2 Het stelsel

(i) $P(f)g = g$

(ii) $r(f) + P(f)v = v + g$

(iii) $P^*(f)v = 0$

heeft de unieke oplossing $(g,v) = (g(f),v(f))$.

Bewijs. Vermenigvuldigen we (ii) met $P^*(f)$ dan volgt met

$P^*(f)P(f) = P^*(f)$

$P^*(f)r(f) = P^*(f)g$

Verder volgt uit (i) dat $P^n(f)g = g$ voor alle n, dus ook

$$P^*(f)g = g$$

zodat

$$P^*(f)r(f) = g$$

Maar $P^*(f)r(f) = g(f)$ dus $g = g(f)$

Laat vervolgens $(g(f),v_1)$ en $(g(f),v_2)$ twee oplossingen van het stelsel zijn, dan geldt dus

$$r(f) + P(f)v_1 = v_1 + g(f)$$

en $$r(f) + P(f)v_2 = v_2 + g(f)$$

Trekken we deze laatste twee vergelijkingen van elkaar af dan vinden we

$$P(f)\ (v_1 - v_2) = (v_1 - v_2)$$

Dus ook weer $P^*(f)\ (v_1 - v_2) = (v_1 - v_2)$.

Maar $P^*(f)v_1 = P^*(f)v_2 = 0$ dus $v_1 = v_2 = v(f)$, waarmee het bewijs is voltooid. □

Opmerking. In het geval dat P(f) periodiek is zal de rij v(n,f) uit (5.5) niet convergeren. Wel convergeert

$$n^{-1} \sum_{t=0}^{n-1} v(t,f)$$

Met $v(f) := \lim_{n\to\infty} n^{-1} \Sigma_{t=0}^{n-1} v(t,f)$ geldt dan opnieuw stelling 5.1.
Ook stelling 5.2 geldt weer.

5.3 De aperiodiciteitstransformatie

Zoals in de vorige paragraaf is opgemerkt is de beperking tot uitsluitend aperiodieke ketens geen wezenlijke beperking.
We zullen dat hier aantonen.
Beschouw eens de volgende data transformatie, $0 < \alpha < 1$

$$\tilde{r}(i,k) = r(i,k)$$

$$\tilde{p}_{ij}^{k} = \alpha\delta_{ij} + (1 - \alpha)p_{ij}^{k}$$

zodat voor alle f

$$\tilde{r}(f) = r(f) \text{ en } \tilde{P}(f) = \alpha I + (1 - \alpha)\ P(f)$$

Het is duidelijk dat deze transformatie opnieuw een Markov beslissingsprobleem geeft (de matrices $\tilde{P}(f)$ zijn weer stochastisch) en dat alle Markov matrices $\tilde{P}(f)$ aperiodiek worden, immers $\tilde{p}_{ii}^{k} \geq \alpha > 0$ voor alle i en k.
Het getransformeerde probleem is ook equivalent met het oorspronkelijke probleem.

Stelling 5.3. Voor de oplossing $\tilde{g}(f)$, $\tilde{v}(f)$ van het stelsel

(i) $\tilde{P}(f)g = g$

(ii) $\tilde{r}(f) + \tilde{P}(f)v = v + g$

(iii) $\tilde{P}^*(f)v = 0$

geldt $\tilde{g}(f) = g(f)$ en $\tilde{v}(f) = (1 - \alpha)^{-1} v(f)$.

Bewijs. De oplossing van bovenstaand stelsel is volgens stelling 5.2 uniek, dus we hoeven slechts te controleren of de oplossing $(g(f), (1 - \alpha)^{-1} v(f))$ voldoet. Met stelling 5.1 volgt:

(i) $\tilde{P}(f)g(f) = \alpha g(f) + (1 - \alpha)P(f)g(f) = g(f)$

(ii) $\tilde{r}(f) + \tilde{P}(f)(1 - \alpha)^{-1} v(f) = r(f) + \alpha(1 - \alpha)^{-1} v(f) + P(f)v(f) =$

$$= v(f) + g(f) + \alpha(1 - \alpha)^{-1} v(f) = (1 - \alpha)^{-1} v(f) + g(f).$$

Om (iii) te bewijzen is het voldoende aan te tonen dat $\tilde{P}^*(f) = P^*(f)$.
Uit $\tilde{P}(f)P^*(f) = [\alpha I + (1 - \alpha)P(f)]P^*(f) = P^*(f)$ volgt ook
$\tilde{P}^*(f)P^*(f) = P^*(f)$
Maar ook $\tilde{P}^*(f)P(f) = \tilde{P}^*(1 - \alpha)^{-1}(\tilde{P} - \alpha I) = \tilde{P}^*(f)$ zodat
$\tilde{P}^*(f)P^*(f) = \tilde{P}^*(f)$.
Dus $P^*(f) = \tilde{P}^*(f)$. □

5.4 Relatie gemiddelde opbrengst - verdisconteerde opbrengst

Met behulp van stelling 5.1 krijgen we nu een mooie relatie tussen de oplossing $g(f), v(f)$ en de totale verwachte verdisconteerde opbrengst $v_\beta(f^{(\infty)})$.

$$v_\beta(f^{(\infty)}) = \sum_{n=0}^{\infty} \beta^n P^n(f) r(f)$$

$$= \sum_{n=0}^{\infty} \beta^n P^n(f) [v(f) + g(f) - P(f)v(f)]$$

$$= (1 - \beta)^{-1} g(f) + \sum_{n=0}^{\infty} \beta^n P^n(f)[v(f) - \beta P(f)v(f) + \tag{5.8}$$

$$- (1 - \beta)P(f)v(f)]$$

$$= (1 - \beta)^{-1} g(f) + v(f) - (1 - \beta) \sum_{n=0}^{\infty} \beta^n P^{n+1}(f)v(f).$$

Omdat $P^n(f)v(f)$ geometrisch naar 0 gaat, is $\Sigma\beta^n P^{n+1}(f)v(f)$ uniform op $0 \leq \beta \leq 1$ begrensd. Dus we krijgen de relatie

$$v_\beta(f) = (1 - \beta)^{-1} g(f) + v(f) + 0(1 - \beta), \quad \beta \uparrow 1.$$

5.5 Successieve approximatie

Evenals in het verdisconteerde geval kunnen we ook nu weer proberen met de methode van de successieve approximaties een (bijna) gemiddeld optimale strategie te bepalen.
Allereerst geven we het volgende equivalent van stelling 4.5.

Stelling 5.4

(i) Als $u \geq r(f) + P(f)w$ dan

$$g(f) \leq \max_i (u - w)(i).e$$

(ii) Als $u \leq r(f) + P(f)w$ dan

$$g(f) \geq \min_i (u - w)(i).e$$

Bewijs.

(i) Uit $r(f) \leq u - P(f)w$ volgt na vermenigvuldiging met $P^*(f)$

$$g(f) = P^*(f)r(f) \leq P^*(f)(u - w)$$

Dus met $u - w \leq \max_i (u - w)(i).e$ en $P^*(f)e = e$

$$g(f) \leq \max_i (u - w)(i).e$$

(ii) Analoog met $u - w \geq \min_i (u - w)(i).e$ □

Een gevolg van deze stelling is

Stelling 5.5

Als $u = \max_f \{r(f) + P(f)w\}$ dan geldt voor alle strategieën s

$$g(s) \leq \max_i (u - w)(i).e$$

Bewijs. Zij s een willekeurige strategie. Beschouw nu het iteratieproces

$$v_n = \max_f \{r(f) + P(f)v_{n-1}\}\ ,\ v_0 = w \text{ (dus } v_1 = u)$$

Dan geldt volgens stelling 3.2

$$V_T(s,w) \leq v_T$$

Ook geldt

$$v_T \leq T \max_i (u - w)(i).e + w$$

Immers

$$v_1 \leq v_0 + \max_i (v_1 - v_0)(i)e$$

zodat

$$v_2 = \max_f\{r(f) + P(f)v_1\} \le \max_f\{r(f) + P(f)(v_0 + \max_i(v_1 - v_0)(i).e)\}$$

$$= \max_f\{r(f) + P(f)v_0\} + \max_i(v_1 - v_0)(i).e$$

$$\le v_0 + 2\max_i(v_1 - v_0)(i).e$$

etc.

Verder $|V_T(s,w) - V_T(s,0)| \le \|w\|e$.

Dus

$$g(s) = \liminf_{T\to\infty} T^{-1}V_T(s,0) = \liminf_{T\to\infty} T^{-1}V_T(s,w)$$

$$\le \liminf_{T\to\infty} T^{-1}v_T \le \max_i (u - w)(i).e.$$

We kunnen de resultaten van stellingen 5.4 en 5.5 combineren.

Gevolg. Als $u = \max_f\{r(f) + P(f)w\} = r(h) + P(h)w$ dan geldt

$$(5.9)\quad \min_i(u - w)(i).e \le g(h) \le \max_s g(s) \le \max_i(u - w)(i).e$$

Merk op dat we $\max_s g(s)$ schrijven terwijl we feitelijk $\sup_s g(s)$ moeten schrijven. Maar zoals al eerder opgemerkt wordt dit supremum ook aangenomen (door een stationaire strategie).

De in bovenstaand gevolg gegeven grenzen zullen in het algemeen weinig waarde hebben tenzij we een iteratieproces v_n kunnen aangeven waarbij het $sp(v_n - v_{n-1})$ (het verschil tussen de maximale en de minimale component van $v_n - v_{n-1}$) klein wordt.

Dit zal zeker niet mogelijk zijn als $\max_s g(s)$ geen constante vector is, wat het geval is als een optimale stationaire strategie f^* meerdere recurrente ketens met verschillende waarden heeft. Ook periodiciteit zal de convergentie van $sp(v_n - v_{n-1})$ kunnen verstoren.

Voorbeeld 5.1. Beschouw het beslissingsprobleem met twee toestanden en in elke toestand maar een actie. Er is dan maar een strategie, zeg f. Zij verder $r(f) = (2,0)^T$, $P(f) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Met $v_0 = (0,0)^T$ volgt dan $v_{2n} = (n,n)^T$, $v_{2n+1} = (n+2,n)^T$.

Zodat $sp(v_{n+1} - v_n) = 2$ voor alle n.

Wel convergeert $sp(v_n - v_{n-1})$ naar nul als $\max_s g(s)$ constant is en alle ketens aperiodiek zijn. We bewijzen dit hier voor een bijzonder

geval.

Laat $\{v_n\}$ een rij successieve approximaties zijn met

$$v_0 = v \text{ , } v \text{ willekeurig}$$

$$v_n = \max_f\{r(f) + P(f)v_{n-1}\}.$$

Dan geldt de volgende stelling. ($A \geq 0 \Longleftrightarrow (A)_{ij} \geq 0$ voor alle i en j).

Stelling 5.6. Als er een matrix $A \geq 0$ bestaat met gelijke rijen en $A \neq 0$ (dan heeft A dus tenminste één strict positieve, constante, kolom) zodanig dat $P(f) - A \geq 0$ voor alle f dan geldt:

$$sp(v_{n+1} - v_n) \leq (1 - \rho)^n \, sp(v_1 - v_0)$$

met ρ gelijk aan de rijsom van A.

In dit geval heeft P(f) slechts een recurrente keten, immers een toestand j waarvoor $(A)_{ij} > 0$ voor alle i is uit elke toestand bereikbaar. Dan is dus ook g(f) constant (evenals $\max_s g(s)$).

Bewijs. Laat f_t en f_{t+1} zodanig zijn dat

$$v_{t+1} = r(f_{t+1}) + P(f_{t+1})v_t \text{ en } v_t = r(f_t) + P(f_t)v_{t-1}$$

Dan geldt dus

$$v_{t+1} - v_t \leq r(f_{t+1}) + P(f_{t+1})v_t - \{r(f_{t+1}) + P(f_{t+1})v_{t-1}\}$$

$$= P(f_{t+1})(v_t - v_{t-1})$$

$$= P((f_{t+1}) - A)(v_t - v_{t-1}) + A(v_t - v_{t-1})$$

$$\leq (1 - \rho) \max_i (v_t - v_{t-1})(i).e + A(v_t - v_{t-1})$$

Analoog

$$v_{t+1} - v_t \geq (1 - \rho) \min_i (v_t - v_{t-1})(i).e + A(v_t - v_{t-1})$$

A heeft gelijke rijen dus $A(v_t - v_{t-1})$ is een constante vector.
Zodat

$$sp(v_{t+1} - v_t) \leq (1 - \rho) \left[\max_i (v_t - v_{t-1})(i) - \min_i (v_t - v_{t-1})(i)\right] =$$

$$= (1 - \rho) sp(v_t - v_{t-1})$$

t is willekeurig, dus

$$sp(v_{n+1} - v_n) \leq (1 - \rho)sp(v_n - v_{n-1}) \leq \ldots \leq (1 - \rho)^n sp(v_1 - v_0)$$ □

Uit het bewijs van stelling 5.6 volgt al dat de voorwaarden verzwakt kunnen worden. Immers in het bewijs spelen slechts twee matrices een rol, nl. $P(f_t)$ en $P(f_{t+1})$. We kunnen de voorwaarden uit stelling 5.6 dus verzwakken tot:

(5.10) voor elk tweetal beslissingsregels f en g is er een matrix $A(f,g) \geq 0$ met gelijke rijen en rijsom tenminste $\rho > 0$ zodanig dat $P(f) - A(f,g) \geq 0$ en $P(g) - A(f,g) \geq 0$.

Een voorbeeld van een beslissingsprobleem waarbij wel aan de bovenstaande voorwaarde maat niet aan de voorwaarde van stelling 5.6 is voldaan is

Voorbeeld 5.2. $I = \{1,2\}$. Er zijn drie mogelijke overgangsmatrices nl.

$$\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \text{ en } \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

Men kan bewijzen dat in het aperiodieke geval $v_n - v_{n-1}$ altijd convergeert naar max $g(s)$, ook als dit maximum niet constant is. We zullen dat hier niet doen.

De belangrijkste consequentie van de in deze paragraaf bewezen resultaten is:

Stelling 5.7. Als voldaan is aan de voorwaarden (5.10) dan bestaat er een stationaire strategie die gemiddeld optimaal is, en levert de methode van successieve approximaties een bijna optimale stationaire strategie en goede onder-en bovengrenzen voor $\max_s g(s)$.

Bewijs. Het bestaan van bijna optimale stationaire strategieën en goede grenzen volgt direct uit (5.9) en stelling 5.6 (met de voorwaarde (5.10)). Dat er ook een optimale stationaire strategie is, volgt uit het feit dat er slechts eindig veel stationaire strategieën zijn. □

5.6 De policy iteration methode

Net als bij het verdisconteerde probleem kan ook hier een policy iteration algoritme ontwikkeld worden. We zullen dat hier alleen doen voor de situatie waarin alle optredende ketens irreducibel zijn. Het algoritme verbetert stationaire strategieën of beslissingsregels (Eng. policies). Dat dit een geschikte aanpak kan zijn volgt uit de vorige paragraaf waar, althans voor een bijzonder geval (vgl. stelling 5.7), is aangetoond dat er een stationaire optimale strategie bestaat.

We beschouwen dus het irreducibele beslissingsprobleem.
Laat $(g(f),v(f))$ de bij f behorende (volgens stelling 5.2 unieke) oplossing zijn van het stelsel uit stelling 5.2. Dan is vanwege de irreducibiliteit van P(f) de vector g(f) constant. Notatie $g(f) = g_f.e$.
Definieer nu voor elke beslissingsregel h de vector $\psi(h,f)$ door

$$r(h) + P(h)v(f) = v(f) + g_f.e + \psi(h,f) \tag{5.11}$$

Dan geldt

Lemma 5.2.

$$g_h.e = g_f.e + P^*(h)\psi(h,f)$$

Bewijs. Vermenigvuldig (5.11) voor met $P^*(h)$. □

Dit lemma vormt de basis voor een algoritme. Immers, uit stelling 5.1 (ii) volgt $\psi(f,f) = 0$, dus

$$\max_h \psi(h,f) \geq 0 \tag{5.12}$$

En dit levert dan de volgende stelling.

Stelling 5.8.

(i) Als $\psi(h,f) \geq 0$ en $\psi(h,f) \neq 0$ terwijl P(h) irreducibel is, dan geldt

$$g_h > g_f$$

(ii) Als voor alle h $\psi(h,f) \leq 0$ dan geldt voor alle h

$$g_h \leq g_f,$$

dus f is gemiddeld optimaal.

Bewijs. (i) Uit P(h) irreducibel volgt $P^*(h)$ $(i,j) > 0$ voor alle i,j dus $P^*(h)\psi(h,f) > 0$, zodat met lemma 5.2 geldt $g_h > g_f$.
(ii) Direct uit lemma 5.2. □

Hiermee verkrijgen we het volgende algoritme.

Policy iteration algoritme (irreducibel)

Laat f de huidige beslissingsregel zijn.

Stap 1. Bepaal de oplossing $(g_f,v(f))$ van het stelsel uit stelling 5.2.

Stap 2. Bepaal een beslissingsregel h zodat

$$r(h) + P(h)v(f) = \max\{r(\cdot) + P(\cdot)v(f)\}$$

Als nu $\psi(h,f) > 0$ voor zekere $i \in I$, dan is h beter dan f.
Vervang f door h en ga terug naar stap 1, etc.
Als $\psi(h,f) \equiv 0$ dan is f (en ook h) gemiddeld optimaal.

Merk op dat het feit dat P(h) geen transiente toestanden bevat essentieel is in stelling 5.8 daar waar uit $\psi(h,f) \geq 0$ en $\psi_j(h,f) > 0$ voor zekere j geconcludeerd wordt dat $g_h > g_f$ is.
Immers als j transient is onder P(h) dan is $P^*(h)(i,j) = 0$ voor alle i.
Dus ook $P^*(h)(i,j)\psi_j(h,f) = 0$,
In het geval dat er ook transiente toestanden zijn werkt het algoritme niet, althans, niet in deze vorm. We kunnen het algoritme echter aanpassen zodat we opnieuw convergentie kunnen aantonen. Ook in de aanzienlijk ingewikkelde situatie van meerdere ketens, dus niet constante g(f), kan een policy iteration algoritme geformuleerd worden. We zullen dat hier niet doen.

Een in numerieke zin minder aantrekkelijk punt van het algoritme is het oplossen van het stelsel uit stelling 5.2. Vergelijking (iii): $P^*(f)v = 0$ is daarbij onplezierig omdat $P^*(f)$ onbekend is en bepaald moet worden uit $P^*(f) = P^*(f)P(f)$. In dit speciale geval (irreducibiliteit) is dit echter overbodig zoals hieronder zal worden aangetoond.
Beschouw nog eens de vergelijkingen

(5.13) (i) $P(f)g = g$
(ii) $r(f) + P(f)v = v + g$

Uit (i) volgt $g = P^*(f)g$ zodat als P(f) irreducibel is g constant moet zijn. Vermenigvuldigen we (ii) nog met $P^*(f)$ dan krijgen we $P^*(f)r(f) = P^*(f)g = g$. Dus

$$g = g(f) = g_f.e$$

We weten al (stelling 5.1) dat $(g_f, v(f))$ een oplossing is van (5.13). En we zien direct dat dan ook $(g_f, v(f) + \alpha e)$, $\alpha \in \mathbb{R}$, een oplossing is. Dat dit ook alle oplossingen zijn wordt als volgt bewezen. Laat (g_f, v) een oplossing zijn.
Dan geldt voor de oplossingen (g_f, v) en $(g_f, v(f))$

$$r(f) + P(f)v = v + g_f.e$$

en $$r(f) + P(f)v(f) = v(f) + g_f.e$$

Zodat $P(f)(v - v(f)) = (v - v(f))$, waaruit met de irreducibiliteit van P(f) volgt

$$v - v(f) = P^*(f)(v - v(f)) = \alpha e,$$

voor zekere $\alpha \in \mathbb{R}$.
Dus alle oplossingen van (5.13) zijn van de vorm $(g_f, v(f) + \alpha e)$.
De vergelijking $P^*(f)v = 0$ legt slechts de constante α vast ($\alpha = 0$).
Een andere manier om de constante vast te leggen is het nulstellen van een van de componenten, b.v.

$$v_N = 0$$

Ook is het paar (5.13) te vervangen door de enkele vergelijking

$$r(f) + P(f)v = v + g.e$$

waarin g nu een scalar is. Ga na.
Hiermee vinden we dus een alternatief voor stap 1 uit de policy iteration methode nl.
stap 1. Bepaal de oplossing (g_f, v_f) van het stelsel

$$\text{(5.14)} \qquad \begin{cases} r(f) + P(f)v = v + g.e \\ v_N = 0 \end{cases}$$

We zien dat het aantal vergelijkingen in (5.14) aanzienlijk minder is dan in het stelsel uit stelling 5.2. Hoewel dat laatste stelsel natuurlijk een groot aantal afhankelijke vergelijkingen bevat.
Men gaat gemakkelijk na dat beide algoritmes equivalent zijn. Dit alternatieve algoritme zullen we in de volgende paragraaf opnieuw tegenkomen.

5.7 Lineaire programmering

We beschouwen in deze paragraaf opnieuw het irreducibele Markov beslissingsprobleem. We zullen het probleem in een aantal stappen vertalen naar een lineair programmeringsprobleem dat dan weer de policy iteration methode uit de vorige paragraaf oplevert.

Laat f een beslissingsregel zijn dan geldt er voor de limietverdeling (eventueel Cesarolimiet) $P^*(f)$ en de gemiddelde opbrengst g_f

$$\text{(5.15)} \qquad \begin{aligned} & p_j^*(f) = \Sigma_i \; p_i^*(f) \; p_{ij}^{f(i)} \\ & \Sigma_i p_i^*(f) = 1 \\ & p_i^*(f) \geq 0, \; i \in I \\ & g_f = \Sigma_i p_i^*(f) r(i, f(i)) \end{aligned}$$

Het doel is nu een mathematisch programmeringsprobleem te formuleren dat g_f maximaliseert over alle beslissingsregels.
We kunnen elke stationaire strategie vastleggen door getallen $d_i^k \in \{0,1\}$, en wel als volgt

$$d_i^k = \begin{cases} 1 \text{ als } f(i) = k \\ 0 \text{ als } f(i) \neq k \end{cases}$$

Omgekeerd legt ook elke collectie d_i^k's met

$$d_i^k \in \{0,1\}$$
$$\Sigma_k d_i^k = 1, \; i \in I$$

een stationaire strategie vast.
Hiermee kunnen we (5.15) gedeeltelijk herschrijven:

$$p_j^*(f) = \sum_{i,k} p_i^*(f) \; d_i^k p_{ij}^k$$

$$g_f = \sum_{i,k} p_i^*(f) \; d_i^k r(i,k)$$

Schrijven we nog

$$p_i^*(f) = \sum_k p_i^*(f) \; d_i^k$$

Dan levert ons dit het volgende mathematisch programmeringsprobleem.

I
$$\begin{cases} \max \sum_{i,k} x_i d_i^k r(i,k) \\ \text{met als nevenvoorwaarden} \\ \Sigma_k x_j d_j^k = \sum_{i,k} x_i d_i^k p_{ij}^k \;, \; j \in I \\ \Sigma_i x_i = 1 \\ x_i \geq 0 \quad , \; i \in I \\ d_i^k \in \{0,1\} \;, \; i \in I, \; k \in K \\ \Sigma_k d_i^k = 1 \quad , \; i \in I \end{cases}$$

Gebruikmakend van de 1 - 1 relatie tussen stationaire strategieën en oplossingen van $d_i^k \in \{0,1\}$, $\Sigma_k d_i^k = 1$, $i \in I$ en van de eenduidigheid van de oplossing van

$$xP(f) = x \ , \ \Sigma x_i = 1 \ , \ x \geq 0$$

wordt nu gemakkelijk aangetoond dat:

(i) Elke toegelaten oplossing van I correspondeert met een (unieke) stationaire strategie.

(ii) De bijbehorende objectwaarde gelijk is aan de gemiddelde opbrengst per tijdseenheid voor die strategie.

Dus het mathematisch programmeringsprobleem I is equivalent met het oorspronkelijke probleem: maximaliseer g_f.

Probleem I is echter zowel niet-lineair als gedeeltelijk geheeltallig. Dit laatste kan niet wezenlijk zijn omdat de geheeltalligheid voortvloeit uit de beperking tot zuivere stationaire strategieën. Laten we deze beperking vallen dan krijgen we het volgende probleem:

$$\text{II} \begin{cases} \max \sum_{i,k} x_i d_i^k r(i,k) \\ \text{met als nevenvoorwaarden} \\ \sum_k x_j d_j^k = \sum_{i,k} x_i d_i^k p_{ij}^k \ , \ j \in I \\ \sum_i x_i = 1 \\ \sum_k d_i^k = 1 \ , \ i \in I \\ x_i \geq 0, \ d_i^k \geq 0, \ i \in I, \ k \in K \end{cases}$$

De nevenvoorwaarden bij probleem II zijn wat zwakker dan die bij I. Op de vraag of beide problemen toch equivalent zijn (wat met de interpretatie via gemengde strategieën ook direct kan worden aangetoond) komen we nog terug.

Eerst zullen we proberen voor het niet-lineaire probleem II een equivalent lineair probleem te formuleren.

Merk op dat de termen die II niet lineair maken steeds van de vorm $x_i d_i^k$ zijn.

Substitueer daarom x_i^k voor $x_i d_i^k$ in II. Met verder:

$$x_i = \sum_k x_i d_i^k = \sum_k x_i^k \ ,$$

zodat

$$\sum_i x_i = \sum_{i,k} x_i^k$$

vinden we dan het lineaire programmeringsprobleem

III
$$\begin{cases} \max \sum_{i,k} x_i^k \, r(i,k) \\ \text{met als nevenvoorwaarden} \\ \sum_k x_j^k = \sum_{i,k} x_i^k p_{ij} \quad , j \in I \\ \sum_{i,k} x_i^k = 1 \\ x_i^k \geq 0, \quad i \in I, \; k \in K \end{cases}$$

De nevenvoorwaarden van III zijn weer zwakker dan die van probleem II. We zullen nu aantonen dat de problemen I, II en III equivalent zijn. Het is duidelijk dat een oplossing van I voldoet aan II en aan III met steeds dezelfde objectwaarde. Analoog voldoen oplossingen van II aan III. Zodat

$$\max \text{ III} \geq \max \text{ II} \geq \max \text{ I}$$

We moeten nog aantonen dat max III = max I.
Eerst max III = max II.
Stel $\{y_i^k\}_{i,k}$ is een toegelaten oplossing van III.
Definieer dan

(5.16)
$$y_i := \sum_k y_i^k$$
$$e_i^k := \begin{cases} y_i^k / y_i \text{ als } y_i > 0 \\ \text{willekeurig als } y_i = 0 \text{ (wel zodanig dat } \sum_k e_i^k = 1) \end{cases}$$

Gevolg hiervan is dat de y_i juist de limietverdeling vormen van de (gemengde) stationaire strategie s met $s(i,k) = e_i^k$, $i \in I$, $k \in K$.
Dus max III = max II.
Maar bovendien is ook de keten met overgangskansen

$$p_{ij} = \sum_k e_i^k p_{ij}$$

weer irreducibel. Dus moet ook gelden $y_i > 0$ voor alle $i \in I$.
Verder zijn de vergelijkingen

$$\sum_k x_j^k = \sum_{i,k} x_i^k p_{ij} \quad , j \in I$$

afhankelijk. Immers sommeren van deze vergelijkingen over j levert de identiteit:

$$\sum_{j,k} x_j^k = \sum_{i,j,k} x_i^k p_{ij}^k = \sum_{i,k} x_i^k \sum_j p_{ij}^k = \sum_{i,k} x_i^k$$

De rang van het stelsel nevenvoorwaarden bij III is dus ten hoogste N. Dus basisoplossingen hebben ten hoogste N niet-nul variabelen. Maar ook geldt $y_i > 0$ voor alle $i \in I$, dus bij elke i is er tenminste één k waarvoor $y_i^k > 0$ is, zodat basisoplossingen minstens N niet-nul variabelen hebben. *Samenvattend: basisoplossingen hebben precies N niet-nul variabelen en bovendien is er bij elke i precies één k waarvoor* $x_i^k > 0$ is.

Laat dus $\{y_i^k\}_{i,k}$ een basisoplossing zijn van III dan zien we dat de corresponderende strategie (vgl. (5.16)) zuiver is: $e_i^k \in \{0,1\}$ voor alle i,k. Dus basisoplossingen van III corresponderen met oplossingen van I, terwijl de bijbehorende objectwaarden gelijk zijn. Er is altijd een basisoplossing die optimaal is dus

$$\max \text{ III} = \max \text{ I}.$$

waarmee is aangetoond dat de problemen I, II en III equivalent zijn.

5.8 Relatie tussen policy iteration en lineaire programmering

Er is opnieuw een duidelijke relatie tussen een bepaalde oplossingsmethode voor het lineaire programmeringsprobleem III en de policy iterationmethode uit paragraaf 5.6.

Beschouw nog eens probleem III

$$\text{III}' \begin{cases} \max \sum_{i,k} x_i^k r(i,k) \\ \text{onder de voorwaarden} \\ \sum_k x_j^k = \sum_{i,k} x_i^k p_{ij}^k, \quad j = 1,\ldots,N-1 \\ \sum_{i,k} x_i^k = 1 \\ x_i^k \geq 0, \quad i \in I, \ k \in K \end{cases}$$

Zoals in de vorige paragraaf is opgemerkt zijn de eerste N vergelijkingen bij probleem III afhankelijk. Omdat een basisoplossing van III correspondeert met een irreducibele Markov keten kunnen we een willekeurige van deze N vergelijkingen weg laten. We hebben hier

gekozen voor het weglaten van de N-de vergelijking.
Voor III' kunnen we nu het volgende tableau opstellen

x_1^1 ... $x_1^{K_0}$	x_2^1 ... $x_2^{K_0}$	... x_{N-1}^1 ... $x_{N-1}^{K_0}$	x_N^1 ... $x_N^{K_0}$	
$1-p_{11}^1$ ... $1-p_{11}^{K_0}$	$-p_{21}^1$ ... $-p_{21}^{K_0}$	... $-p_{N-11}^1$ ... $-p_{N-11}^{K_0}$	$-p_{N1}^1$ ... $-p_{N1}^{K_0}$	0
$-p_{12}^1$ ... $-p_{12}^{K_0}$	$1-p_{22}^1$ ... $1-p_{22}^{K_0}$	... $-p_{N-12}^1$ ... $-p_{N-12}^{K_0}$	$-p_{N2}^1$ ... $-p_{N2}^{K_0}$	0
⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮
$-p_{1N-1}^1$ ... $-p_{1N-1}^{K_0}$	$-p_{2N-1}^1$ ... $-p_{2N-1}^{K_0}$	... $1-p_{N-1N-1}^1$ ... $1-p_{N-1N-1}^{K_0}$	$-p_{N-1N}^1$ ... $-p_{N-1N}^{K_0}$	0
1 ... 1	1 ... 1	... 1 ... 1	1 ... 1	1
$r(1,1)$ ... $r(1,K_0)$	$r(2,1)$ ... $r(2,K_0)$	... $r(N-1,1)$ ... $r(N-1,K_0)$	$r(N,1)$ ... $r(N,K_0)$	

We weten ook dat bij een basisoplossing $\{x_i^k\}_{i,k}$ er voor elke i precies één k(i) is met $x_i^{k(i)} > 0$.
Beschouw dus eens een basis $x_i^{k(i)}$, $i \in I$.
Om na te gaan of deze basisoplossing optimaal is bepalen we de vector van veegconstanten $w = (v_1,\ldots,v_{N-1},g)$. En wel zo dat als de vergelijkingen uit het tableau met deze constanten vermenigvuldigd worden en vervolgens van de laatste rij worden afgetrokken er juist op de plaatsen r(i,k(i)) nullen ontstaan.
Dus de constanten $(v_1,\ldots,v_{N-1},g)$ moeten voldoen aan:

$$v_1(1-p_{11}^{k(1)}) + v_2(-p_{12}^{k(1)}) + \ldots + v_{N-1}(-p_{1N-1}^{k(1)}) + g - r(1,k(1)) = 0$$

$$\begin{matrix} v_1(-p_{21}^{k(2)}) + v_2(1-p_{22}^{k(2)}) + \ldots + v_{N-1}(-p_{2N-1}^{k(2)}) + g - r(2,k(2)) = 0 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ v_1(-p_{N1}^{k(N)}) + v_2(-p_{N2}^{k(N)}) + \ldots + v_{N-1}(-p_{NN-1}^{k(N)}) + g - r(N,k(N)) = 0 \end{matrix}$$

Ofwel in vector-matrix notatie met $v = (v_1,\ldots,v_{N-1},v_N)$

$$\begin{cases} v + g.e - P(k)v - r(k) = 0 \\ v_N = 0 \end{cases}$$

Dus $(g,v) = (g_k, v(f) - v_N(f).e)$

Om na te gaan of deze basisoplossing optimaal is moeten we nagaan of er na het vegen nog negatieve elementen in de laatste rij voorkomen, dat wil zeggen controleren of (met $v_N = 0$)

$$v_i - \sum_j p_{ij}^k v_j + g - r(i,k) \leq 0$$

voor alle $i \in I$ en $k \in K$.

Dus nagaan of voor alle toegelaten basisoplossingen

$$(x_1^{h(1)}, \ldots, x_N^{h(N)} \text{ geldt}$$

$$r(h) + P(h)v \leq v + g.e \tag{5.17}$$

Als (5.17) voor alle h geldt dan is k dus gemiddeld optimaal. Zo niet, dan is er nog verbetering mogelijk. Gebruikmakend van het feit dat we al weten dat er bij een basisoplossing y_i^k voor elke i een k is zodat $y_i^k > 0$ kunnen we de volgende aanpak volgen.

Als

$$\max_k \{r(i,k) + \sum_j p_{ij}^k v_j - g - v_i\} > 0 \tag{5.18}$$

Vervang dan in de basis $(x_1^{k(1)}, \ldots, x_N^{k(N)})$ $x_i^{k(i)}$ door $x_i^{h(i)}$ waarbij h(i) een maximalisator is in (5.18). Doe dit voor alle i, wat neerkomt op het bepalen van een beslissingsregel die

$$r(\cdot) + P(\cdot)v - ge - v$$

maximaliseert.

Precies de verbeterstap uit de policy iteration methode uit paragraaf 5.7. (met $v_N = 0$).

Samenvattend zien we, dat de hierboven gepresenteerde methode om basisoplossingen te verbeteren, waarbij eventueel meerdere basisvariabelen tegelijk vervangen worden, precies de policy iteration methode uit paragraaf 5.7, met $v_N = 0$ oplevert.

5.9 Benadering van de policy iteration methode

Een nadeel van de policy iteration methode is de noodzaak steeds een stelsel vergelijkingen op te lossen.

In plaats van een exacte oplossing (g_f, v) van de vergelijking

$$r(f) + P(f)v = v + g.e$$

(we beschouwen weer het geval dat alle ketens irreducibel zijn) bepalen we een benadering.

En wel als volgt.

Zij v_0 een willekeurige vector en zij f een stationaire strategie

Bepaal vervolgens

$$v_n = r(f) + P(f)v_{n-1},\ n = 1,2,\ldots$$

Dan geldt

$$\begin{aligned} v_n - v_{n-1} &= r(f) + P(f)v_{n-1} - (r(f) + P(f)v_{n-2}) \\ &= P(f)\ (v_n - v_{n-1}) = \ldots = P^{n-1}(f)\ (v_1 - v_0) \\ &= P^{n-1}(f)\ (r(f) + P(f)\ v_0 - v_0) \\ &= P^*(f)\ r(f) + (P^{n-1}(f) - P^*(f))\ (r(f) + P(f)\ v_0 - v_0) \\ &= g_f.e + (P^{n-1}(f) - P^*(f))\ (v_1 - v_0) \end{aligned}$$

Veronderstellen we alle ketens weer aperiodiek dan convergeert $P^n(f) - P^*(f)$ volgens lemma 5.1 exponentieel naar nul. Bovendien geldt

$$P^*(f)\ (P^{n-1}(f) - P^*(f))\ (v_1 - v_0) = 0,$$

zodat $(P^{n-1}(f) - P^*(f))(v_1 - v_0)$ zowel niet negatieve als niet positieve componenten bevat.
Dus $v_n - v_{n-1}$ levert een goede benadering voor g_f

$$\min_i\ (v_n - v_{n-1})(i) \le g_f \le \max\ (v_n - v_{n-1})(i). \tag{5.19}$$

en $\quad sp(v_n - v_{n-1}) \to 0$ als $n \to \infty$.

Bovendien convergeert $v_n - ng_f e$ naar een oplossing van

$$r(f) + P(f)v = v + g_f.e \tag{5.20}$$

Immers

$$\begin{aligned} v_n - v(f) &= r(f) + P(f)v_{n-1} - (r(f) + P(f)v(f) - g_f.e) \\ &= P(f)(v_{n-1} - v(f)) + g_f.e \\ &= \ldots = P^n(f)\ (v_0 - v(f)) + ng_f.e \end{aligned}$$

Dus $v_n - ng_f.e$ convergeert naar $v(f) + P^*(f)(v_0 - v(f))$.

$P^*(f)\ (v_0 - v(f)) = ce$ voor zeker $c \in \mathbb{R}$ dus (vgl. paragraaf 5.5)

$v_n - g_f.e$ convergeert naar een oplossing van (5.20).

Dit leidt tot het volgende alternatief voor stap 1 van het policy iteration algoritme

stap 1" . Zij $v_0 \in \mathbb{R}^n$ bepaal $v_n = r(f) + P(f)v_{n-1}$, $n = 1,2,...$
tot $sp\ (v_n - v_{n-1}) \leq \varepsilon$

Het feit dat we nu niet beschikken over een exacte oplossing van (5.20) dwingt ons ook stap 2 aan te passen. Bijvoorbeeld tot

stap 2" . Zij v_n (of $v_n - v_n(N).e$) de in stap 1^{11} gevonden benaderde oplossing van (5.19).
Bepaal een beslissingsregel h zodat

$$r(h) + P(h)\ v_{n-1} = \max\ \{r(\cdot) + P(\cdot)v_{n-1}\}$$

Als $r(h) + P(h)v_{n-1} \leq v_n + \alpha e$, dan is f bijna optimaal.
Zo niet vervang dan f door h.
Neem vervolgens v_n als eerste benadering voor een oplossing van $r(h) + P(h)v = v + g_h.e$. etc.

In het resterende deel van deze paragraaf zullen we het convergentiegedrag van deze benaderde policy iteration methode bestuderen.
Dat stap 1" eindigt, is reeds aangetoond. Beschouw nu eerst het in stap 2" gegeven criterium voor bijna-optimaliteit.

Lemma 5.3. Als $sp(v_n - v_{n-1}) \leq \varepsilon$ en $r(h) + P(h)v_{n-1} \leq v_n + \alpha e$ dan geldt er $g_h \leq g_f + \alpha + \varepsilon$

Bewijs. Uit $sp(v_n - v_{n-1}) \leq \varepsilon$ en (5.19) volgt

$$v_n \leq v_{n-1} + g_f.e + \varepsilon e$$

Dit geeft

$$r(h) + P(h)v_{n-1} \leq v_n + \alpha e \leq v_{n-1} + (g_f + \alpha + \varepsilon)e$$

Vermenigvuldigen we deze vergelijking voor met $P^*(h)$ dan vinden we $P^*(h)r(h) = g_n.e \leq (g_f + \alpha + \varepsilon)e$. □

Dus door α en ε voldoende klein te kiezen vinden we inderdaad een bijna optimale strategie. Mits echter het algoritme eindigt. Een belangrijk punt daarbij is de vraag of de verbeterstap ook echt een betere strategie oplevert.
Zij h de verbetering van f volgens stap 2" .

Definieer γ door

$$r(h) + P(h)v_{n-1} = v_n + \gamma$$

Dan geldt dus $\gamma \geq 0$ en $\gamma(i) > \alpha$ voor tenminste één $i \in I$.
Zodat met

$$r(h) = v_n - v_{n-1} + \gamma + v_{n-1} - P(h)v_{n-1}$$

volgt

$$g_h.e = P^*(h)r(h) = P^*(h)(v_n - v_{n-1}) + P^*(h)\gamma$$

$$\geq g_f.e - \varepsilon e + P^*(h)\gamma$$

Definieer θ door

$$\theta = \min_{i,j,h} P^*(h)(i,j) \quad \text{(natuurlijk } \theta > 0\text{)}$$

Dan geldt het volgende lemma

Lemma 5.4. Als $\alpha\theta \geq \varepsilon$ dan $g_h > g_f$

Bewijs. Laat $j \in I$ een component van γ zijn met $\gamma(j) > \alpha$ (zo'n j bestaat). Dan $(P^*(h)\gamma)(i) \geq P^*(h)(i,j)\gamma(j) > \alpha\theta$
Dus $g_h > g_f - \varepsilon + \alpha\theta \geq g_f$. □

θ is echter niet bekend. We kunnen dit probleem omzeilen door ε niet constant te nemen. Dit levert dan het volgende alternatieve algoritme.

Benaderde Policy Iteration

Kies $\varepsilon_t > 0$, $t = 0,1,\ldots$ met $\varepsilon_t \downarrow 0$ en $v_0^0 \in \mathbb{R}^n$

Stap 1. Zij f_t de huidige beslissingsregel.
Bepaal $v_n^t = r(f_t) + P(f_t)v_{n-1}^t$, $n = 1,2,..$ tot
$sp(v_{n_t}^t - v_{n_t-1}^t) \leq \varepsilon_t$

Stap 2. Bepaal f_{t+1} zodanig dat

$$r(f_{t+1}) + P(f_{t+1})v_{n_t-1} = \max\{r(f) + P(f)v_{n_t-1}\}$$

Als

$$r(f_{t+1}) + P(f_{t+1})v_{n_t-1} \leq v_{n_t} + \alpha e$$

dan is f_t $(\alpha + \varepsilon_t)$-optimaal.

Zo niet vervang dan f_t door f_{t+1}, definieer $v_0^{t+1} = v_{n_t-1}^t$

en ga verder met stap 1.

Dit algoritme convergeert nu. Immers $\varepsilon_t \downarrow 0$ zodat voor t voldoende groot $\varepsilon_t < \alpha\theta$. Op den duur geldt dus als f_t verbeterd wordt tot f_{t+1} ook $g_{f_{t+1}} > g_{f_t}$.

Er zijn maar eindig veel verschillende strategieën dus moet het algoritme convergeren.

6. Markov spelen

6.1 Inleiding

In de voorgaande hoofdstukken hebben we steeds gekeken naar beslissingsproblemen met één beslisser. Er zijn echter ook beslissingsproblemen met meerdere beslissers, denk aan spelen. In dit hoofdstuk zullen we wat uitgebreider kijken naar Markov beslissingsproblemen met twee spelers. Het grootste deel van dit hoofdstuk is daarbij gewijd aan het geval dat de belangen van beide spelers volstrekt tegengesteld zijn.
We kunnen het spel dan als volgt beschrijven.
Er is een dynamisch systeem met eindige toestandsruimte $I = \{1,2,\ldots,N\}$ dat op discrete tijdstippen, $t = 0,1,2,\ldots$, wordt waargenomen. Het systeem wordt bestuurd door twee spelers, P_1 en P_2. Beide spelers hebben een eindige verzameling van mogelijke beslissingen, zeg $K = \{1,\ldots,K_0\}$ voor P_1 en $L = \{1,\ldots,L_0\}$ voor P_2. Als op zeker moment het systeem in toestand i zit kiezen beide spelers, gelijktijdig, een beslissing. Laat P_1 beslissing a nemen en P_2 beslissing b dan gaat het systeem met kans p_{ij}^{ab} naar toestand j. Bovendien ontvangt P_1 dan een bedrag $r(i,a,b)$ van P_2.
Het bovenstaande spel noemen we het twee personen nulsom Markov spel (ook wel stochastische spel).
Beide spelers zijn geïnteresseerd in het vinden van een strategie die hen een in zekere zin maximale opbrengst geeft. Merk op dat het nulsom karakter van het spel tot gevolg heeft dat er geen enkele reden is voor samenwerking tussen de beide spelers.
In het algemeen is een strategie s, voor P_1, een rij functies $s = (s_0,s_1,\ldots)$, met

s_0: $I \to K$, waarin K de verzameling kansverdelingen op K is.

s_n: $(I \times K \times L)^n \times I \to K$. Dus s_n legt voor elke historie $h_n = (i_0, a_0, b_0, i_1,\ldots,i_{n-1},a_{n-1},b_{n-1})$, de opeenvolging van vroegere toestanden en acties (van beide spelers), de kans $s_n(h_n,i,a)$ vast waarmee, als het systeem op $t = n$ in i zit bij verleden h_n, speler 1 actie a kiest.

Een Markov strategie is een strategie waarbij de kansen $s_n(h_n,i,a)$ niet van h_n afhangen. Notatie $s = (f_0,f_1,\ldots)$ waarbij $f_n(i)$ de op tijdstip n voorgeschreven gemengde beslissing is. $f_n(i,a)$ is dan de kans dat actie a gekozen wordt. Een stationaire strategie is een Markov strategie waarbij alle functies f_n gelijk zijn. Notatie $s = f^{(\infty)}$.
Merk op dat we hier, in tegenstelling tot de voorgaande hoofdstukken, te

maken hebben met gemengde strategieën. We zitten nu in een spelsituatie zodat niet te verwachten is dat zuivere strategieën al optimaal zullen zijn.

Analoog definiëren we voor P_2 strategieën $\sigma = (\sigma_0, \sigma_1 \ldots)$ met

σ_0: $I \times \mathcal{L}$, waarin $\mathcal{L}$ de verzameling kansverdelingen op L is

σ_n: $(I \times K \times L)^n \times I \to \mathcal{L}$,

Markov strategieën $\sigma = (g_0, g_1 \ldots)$ en stationaire strategieën $\sigma = g^{(\infty)}$.

Bij elk paar strategieën (s,σ) kunnen we weer een stochastisch proces definiëren. Laat i de starttoestand zijn en laten de stochastische variabelen X_n, A_n en B_n de toestand, beslissing van P_1 respectievelijk de beslissing van P_2 op tijdstip n zijn.

Dan definiëren we het bij s en σ behorende stochastische proces $(X_0, A_0, B_0, X_1, \ldots)$ door

$$\mathbb{P}_{i,s,\sigma}(X_0 = i_0) = \delta_{ii_0} \qquad (\delta_{ij} = 1 \text{ als } i = j, \text{ anders } 0)$$

$$\mathbb{P}_{i,s,\sigma}(X_0 = i_0,\ A_0 = a_0) = \delta_{ii_0} s_0(i_0,a_0)$$

$$\mathbb{P}_{i,s,\sigma}(X_0 = i_0,\ A_0 = a_0,\ B_0 = b_0) = \delta_{ii_0}\, s_0(i_0,a_0)\sigma_0(i_0,b_0)$$

$$\mathbb{P}_{i,s,\sigma}(X_0{=}i_0, A_0{=}a_0, B_0{=}b_0 X_1{=}i_1) = \delta_{ii_0} s_0(i_0,a_0)\sigma_0(i_0,b_0) p_{i_0 i_1}^{a_0 b_0}$$

etc.

Verwachtingen met betrekking tot dit proces noteren we $\mathbb{E}_{i,s,\sigma}$. Of in vectornotatie: $\mathbb{E}_{s,\sigma}$ is de vector met i-de component $\mathbb{E}_{i,s,\sigma}$.

6.2 Het matrixspel

Matrixspelen zijn al uitgebreid geanalyseerd in het college Besliskunde. We herhalen hier nog even de belangrijkste resultaten.

Een matrixspel is een spel tussen twee personen dat gekarakteriseerd wordt door een matrix

$$A = \begin{pmatrix} a_{11} & \cdots\cdots & a_{1n} \\ & & \\ a_{m1} & \cdots\cdots & a_{mn} \end{pmatrix}$$

en als volgt wordt gespeeld. Speler 1 kiest een rij uit de matrix en speler 2 een kolom. Daarmee wordt een element a_{ij} vastgelegd dat de uitbetaling

representeert die speler 1 van speler 2 ontvangt.
P_1 tracht dit bedrag te maximaliseren terwijl P_2 het natuurlijk wil minimaliseren.
In het algemeen geldt

$$\max_i \min_j a_{ij} \leq \min_j \max_i a_{ij} \tag{6.1}$$

Het gelijkteken geldt slechts als de matrix A een zadelpunt heeft (d.w.z. een element $a_{i_0 j_0}$ dat maximaal is in de j_0-de kolom en minimaal in de i_0-de rij). We kunnen in dat geval zeggen dat het spel een waarde heeft, n.l. $a_{i_0 j_0} = \max_i \min_j a_{ij}$.
Meestal zal echter het kleiner dan teken gelden. Om dan toch een waarde te kunnen definiëren beschouwen we de gemengde uitbreiding van dit spel. Een speler kiest nu niet meer een rij of kolom maar een kansverdeling over de rijen respectievelijk kolommen.
Laat $p = (p_1,\ldots,p_m)$ en $q = (q_1,\ldots,q_n)$ zulke kansverdelingen zijn dan geldt er:

$$\max_p \min_q \sum_{i,j} p_i q_j a_{ij} = \min_q \max_p \sum_{i,j} p_i q_j a_{ij} =: v(A)$$

$v(A)$ heet dan de waarde van het matrixspel en gemengde beslissingen p^* en q^*, die voldoen aan

$$\sum_{i,j} p_i q_j^* a_{ij} \leq \sum_{i,j} p_i^* q_j^* a_{ij} \leq \sum_{i,j} p_i^* q_j a_{ij}$$

voor alle p en q, noemen we optimaal. Natuurlijk geldt ook $\sum_{i,j} p_i^* q_j^* a_{ij} = v(A)$.

6.3 Het M-staps Markov spel

We zullen in deze paragraaf het Markov spel met eindige tijdshorizon beschouwen. En daarbij resultaten afleiden die volstrekt analoog zijn aan de resultaten uit hoofdstuk 3.
Beschouw eens het M-stapsspel met einduitkering $q \in \mathbb{R}^N$ van P_2 aan P_1.
Laat s en σ een willekeurige strategie voor P_1 respectievelijk P_2 zijn, dan definiëren we $V_M(s,\sigma)$ door

$$V_M(s,\sigma) = \mathbb{E}_{s,\sigma}\left[\sum_{n=0}^{M-1} r(X_n,A_n,B_n) + q(X_M)\right] \tag{6.2}$$

$V_M(s,\sigma)$ is dus de totale verwachte opbrengst voor P_1 in het M-stapsspel met einduitkering q (van P_2 aan P_1) als P_1 strategie s en P_2 strategie σ speelt.

P_1 is dus geïnteresseerd in het maximaliseren van $V_M(s,\sigma)$ maar P_2 die deze opbrengst aan P_1 moet betalen is geïnteresseerd in het minimaliseren van $V_M(s,\sigma)$

We zullen laten zien dat dit spel een waarde heeft, d.w.z. dat

$$\max_s \min_\sigma V_M(s,\sigma) = \min_\sigma \max_s V_M(s,\sigma).$$

De aanpak die we zullen volgen om te laten zien dat dit spel inderdaad een waarde heeft is dezelfde als die in hoofdstuk 3.

Laat f en g beslissingsregels voor P_1 respectievelijk P_2 zijn, dan definiëren we

$r(f,g)$: de vector in $\mathbb{R}^N$ met i-de component $\sum_{a,b} f(i,a)g(i,b)\, r(i,a,b)$

$P(f,g)$: de $N \times N$ matrix met i,j-de element $\sum_{a,b} f(i,a)g(i,b)p_{ij}^{ab}$

En definieer

$$v_0 := q$$

$$v_n := \max_f \min_g \{r(f,g) + P(f,g)v_{n-1}\},\ n = 1,2,\ldots,M \tag{6.3}$$

waarbij maxmin weer componentsgewijs wordt genomen. In de n-de iteratiestap in toestand i moet volgens (6.3) bepaald worden

$$\max_{f(i)} \min_{g(i)} \sum_{a,b} f(i,a)g(i,b)[r(i,a,b) + \sum_j p_{ij}^{ab}\, v_{n-1}(j)]$$

Dat is juist de waarde van het $K_0 \times L_0$ matrix spel met elementen

$r(i,a,b) + \sum_j p_{ij}^{ab}\, v_{n-1}(j)$, $a \in K$, $b \in L$.

Er bestaan dus gemengde beslissingen $f_n(i)$ en $g_n(i)$ die optimaal zijn in dit matrix spel. Zulke gemengde beslissing bestaan natuurlijk voor alle i en vormen tezamen beslissingsregels f_n en g_n die voldoen aan

$$r(f,g_n) + P(f,g_n)v_{n-1} \le v_n = r(f_n,g_n) + P(f_n,g_n)v_{n-1} \le \tag{6.4}$$

$$\le r(f_n,g) + P(f_n,g)v_{n-1}\ ,$$

voor alle f en g.

Definieer de Markov strategieën $s^* = (f_M,\ldots,f_1)$ en $\sigma^* = (g_M,\ldots,g_1)$ waarbij f_n en g_n voldoen aan (6.4) voor $n = 1,2,\ldots,M$.
Dan geldt de volgende stelling:

Stelling 6.1. Voor alle s en alle σ geldt

$$V_M(s,\sigma^*) \leq V_M(s^*,\sigma^*) = v_M \leq V_M(s^*,\sigma).$$

Dus het M-stapsspel heeft een waarde, nl. v_M en de strategieën s^* en σ^* zijn optimaal.

Bewijs. Het bewijs verloopt volkomen analoog aan het bewijs van stelling 3.1.
Nummer eerst beslissingstijdstippen weer in omgekeerde volgorde.
En definieer $v_m(h_m,i,s,\sigma^*)$ als de verwachte opbrengst voor P_1 vanaf tijdstip m (omgekeerde tijd), als het systeem zich op tijdstip m in i bevindt, de historie h_m is, P_1 strategie s speelt en P_2 strategie σ^* volgt.
Met inductie kunnen we nu bewijzen dat

(6.5) $$v_m(h_m,i,s,\sigma^*) \leq v_m(i) \quad \text{voor alle } m.$$

$m = 0$ is triviaal.
Aannemend dat (6.5) geldt voor n, bewijzen we dat (6.5) ook geldt voor n+1.

$$v_{n+1}(h_{n+1},i,s,\sigma^*) = \sum_{a,b} s_{n+1}(h_{n+1},i,a)\,\sigma^*_{n+1}(i,b)$$

$$[r(i,a,b) + \sum_j p_{ij}^{ab} v_n(h_{n+1},i,a,b,j,s,\sigma^*)]$$

$$\leq \sum_{a,b} s_{n+1}(h_{n+1},i,a)\,\sigma^*_{n+1}(i,b)\,[r(i,a,b) + \sum_j p_{ij}^{ab} v_n(j)]$$

$$\leq v_{n+1}(i)$$

s was willekeurig zodat daarmee vgl. (6.5) voor $m = n+1$ bewezen is.
Daarmee volgt

$$V_M(i,s,\sigma^*) = v_M(h_M,i,s,\sigma^*) \leq v_M(i)$$

Analoog bewijzen we

$$V_M(i,s^*,\sigma) \geq v_M(i)$$

Gecombineerd levert dit dus

$$V_M(s,\sigma^*) \leq v_M = V_M(s^*,\sigma^*) \leq V_M(s^*,\sigma)$$

voor alle s en σ □

Gevolg van deze stelling is dat we ons opnieuw kunnen beperken tot Markov strategieën, hier echter zijn dat wel gemengde strategieën.
Om de waarde van een M-staps probleem te bepalen moeten we M × N lineaire programmeringsproblemen oplossen, een voor elke toestand op elk tijdstip. Dit zijn echter steeds betrekkelijk kleine problemen van de vorm

$$\begin{cases} \max v \\ \sum_i p_i a_{ij} - v \geq 0, \quad j = 1,\ldots,L_0 \\ \sum_i p_i = 1 \end{cases}$$

Het M-stapsprobleem is ook als één lineair programmeringsprobleem te formuleren. Maar zelfs al gebruiken we daarbij dat we alleen Markov strategieën hoeven te beschouwen dan nog zal de omvang van het probleem immens zijn.

6.4 Markov spelen over oneindige horizon met verdiscontering

In deze paragraaf zullen we een aantal resultaten afleiden die we ook voor het verdisconteerde Markov beslissingsprobleem hebben gevonden. We verdisconteren de uitbetalingen van P_2 aan P_1 weer met een factor $0 \leq \beta < 1$.
Laat s en σ willekeurige strategieën voor beide spelers zijn dan definiëren we $V_\beta(s,\sigma)$ door

$$V_\beta(s,\sigma) := \mathbb{E}_{s,\sigma} \sum_{n=0}^{\infty} \beta^n r(X_n,A_n,B_n) \tag{6.6}$$

$V_\beta(s,\sigma)$ is dus de totale verwachte verdisconteerde opbrengst voor P_1 en het verlies voor P_2.
We zullen nu eerst aantonen dat dit spel een waarde v_β heeft en dat er optimale stationaire strategieën bestaan; d.w.z. er bestaan strategieën $f^{*(\infty)}$ en $g^{*(\infty)}$ zdd voor alle s en σ

$$V_\beta(s,g^{*(\infty)}) \leq V_\beta(f^{*(\infty)},g^{*(\infty)}) = v_\beta \leq V_\beta(f^{*(\infty)},\sigma) \tag{6.7}$$

Merk allereerst op dat voor elk paar strategieën s, de opbrengst vanaf tijdstip t weer begrensd is door

(6.8) $$A\beta^t/(1-\beta)$$

waarin $A := \max_{i,a,b} r(i,a,b)$.

Definieer vervolgens de operatoren $L_\beta(f,g)$ en U_β op $\mathbb{R}^N$ door

$$L_\beta(f,g)v = r(f,g) + \beta P(f,g)v$$

$$U_\beta v = \max_f \min_g L_\beta(f,g)v$$

U_β is dus een operator die aan v toevoegt de waarde van het 1-staps verdisconteerde Markov spel met eindopbrengst v.

Lemma 6.1. Voor alle $v, w \in \mathbb{R}^N$ en alle f en g geldt

(i) $$\beta \min_i (v-w)(i)e \le L_\beta(f,g)v - L_\beta(f,g)w \le \beta \max_i (v-w)(i).e$$

(ii) $$\beta \min_i (v-w)(i).e \le U_\beta v - U_\beta w \le \beta \max_i (v-w)(i).e$$

(iii) $$\|U_\beta v - U_\beta w\| \le \beta \|v-w\|.$$

Bewijs. Het bewijs van (i) is volstrekt analoog aan soortgelijke bewijzen voor $L_\beta(f)$ in hoofdstuk 4.

(ii). Laat f_v, g_v en f_w, g_w beslissingsregels zijn die voldoen aan

$$L_\beta(f,g_v)v \le L_\beta(f_v,g_v)v \le L_\beta(f_v,g)v \quad , \text{ en}$$

$$L_\beta(f,g_w)w \le L_\beta(f_w,g_w)w \le L_\beta(f_w,g)w \quad , \text{ voor alle f en g}$$

Zulke beslissingsregels bestaan. Dan geldt dus

$$\begin{aligned} U_\beta v - U_\beta w &= L_\beta(f_v,g_v)v - L_\beta(f_w,g_w)w \\ &\le L_\beta(f_v,g_w)v - L_\beta(f_v,g_w)w = \beta P(f_v,g_w)(v-w) \\ &\le \beta \max_i (v-w)(i).e. \end{aligned}$$

Analoog bewijzen we de linkerongelijkheid in (ii). Daaruit volgt dan direct (iii). □

Lemma 6.1 (iii) zegt dat U_β een contractie is met betrekking tot de maximumnorm in $\mathbb{R}^N$. Met deze maximum norm is $\mathbb{R}^N$ ook een Banach ruimte. Dus er bestaat een unieke vector $v_\beta \in \mathbb{R}^N$ waarvoor geldt

$$U_\beta v_\beta = v_\beta$$

Stelling 6.2. Het ∞-horizon Markovspel met verdiscontering heeft de waarde v_β en strategieën $f^{*(\infty)}$ en $g^{*(\infty)}$ die voldoen aan

$$L_\beta(f,g^*)v_\beta \le L_\beta(f^*,g^*)v_\beta (= v_\beta) \le L_\beta(f^*,g)v_\beta \tag{6.9}$$

zijn optimaal.

Bewijs. Beschouw het M-stapsspel met einduitkering v_β. Dit spel heeft de waarde v_β. Immers, de waarde is gelijk aan $U^M v_\beta$ en met $Uv_\beta = v_\beta$ volgt

$$U_\beta^M v_\beta = U_\beta^{M-1} U_\beta v_\beta = U_\beta^{M-1} v_\beta = \ldots = v_\beta$$

bovendien, vgl. stelling 6.1, zijn de strategieën $(f^*,\ldots,f^*)$ en $(g^*,\ldots,g^*)$ optimaal in dit spel. Dus voor alle s en σ geldt, met $W_M(s,\sigma)$ de opbrengst voor P_1 in het M-stapsspel met einduitkering v_β;

$$W_M(s,g^{*(\infty)}) \le W_M(f^{*(\infty)},g^{*(\infty)}) = v_\beta \le W_M(f^{*(\infty)},\sigma) \tag{6.10}$$

Met (6.8) en $\beta^M v_\beta \to 0$ $(M \to \infty)$ volgt

$$W_M(s,\sigma) \to V_\beta(s,\sigma) \text{ als } M \to \infty.$$

Dus volgt uit (6.19) voor alle s en σ

$$V_\beta(s,g^{*(\infty)}) \le V_\beta(f^{*(\infty)},g^{*(\infty)}) = v_\beta \le V_\beta(f^{*(\infty)},\sigma)$$ □

Ook kunnen we met de methode van de successieve approximaties weer grenzen voor v_β en ε-optimale stationaire strategieen bepalen.
Beschouw het iteratieproces

$$v_0 = v,\ v_n = U_\beta v_{n-1}\ ,\ n = 1,2,\ldots$$

en laat f_n, g_n beslissingsregels zijn die voldoen aan

$$L_\beta(f,g_n)v_{n-1} \le L_\beta(f_n,g_n)v_{n-1} = v_n \le L_\beta(f_n,g)v_{n-1} \text{ voor alle f en g.}$$

Dan geldt de volgende stelling

Stelling 6.3

(i) $\lim_{n\to\infty} v_n = v_\beta$

(ii) $V_\beta(s,g_n^{(\infty)}) \le v_n + \beta(1-\beta)^{-1} \max_i (v_n - v_{n-1})(i).e$

(iii) $V(f_n^{(\infty)},\sigma) \ge v_n + \beta(1-\beta)^{-1} \min_i (v_n - v_{n-1})(i).e$

(iv) $v_n + \beta(1-\beta)^{-1} \min_i (v_n - v_{n-1})(i)e \le v_\beta \le v_n + \beta(1-\beta)^{-1} \max_i (v_n - v_{n-1})(i).e$

Bewijs. (i) $\|v_n - v_\beta\| = \|U_\beta v_{n-1} - U_\beta v_\beta\| \le \beta\|v_{n-1} - v_\beta\| \le \ldots \le \beta^n \|v - v_\beta\|$

Dus $\|v_n - v\| \to 0$ als $n \to \infty$.

(ii). Feitelijk moeten we (ii) bewijzen voor alle strategieën s, maar we zullen dat hier alleen doen voor stationaire strategieën. We zouden dit kunnen rechtvaardigen door te zeggen dat zodra P_2 een stationaire strategie vastlegt, het voor P_1 resterende probleem een gewoon Markov beslissingsprobleem wordt. We moeten daarbij echter voorzichtig zijn omdat P_1 hier acties kan kiezen afhankelijk van de acties die P_2 in het verleden heeft gekozen. Men toont echter gemakkelijk aan dat het toepassen van zulke strategieën niet zinvol is.

We bewijzen dus $V_\beta(f^{(\infty)},g_n^{(\infty)}) \le v_n + \beta(1-\beta)^{-1} \max_i (v_n - v_{n-1})(i)$

Er geldt

$$V_\beta(f^{(\infty)},g_n^{(\infty)}) = \lim_{t\to\infty} L_\beta^t(f,g_n)0,$$

maar ook

$$\|L_\beta^t(f,g_n)v_{n-1} - L_\beta^t(f,g_n)0\| \le \beta^t \|v_{n-1}\| \to 0 \quad (t \to \infty)$$

Dus beschouwen we

$$\begin{aligned} L_\beta^t(f,g_n)v_{n-1} &= L_\beta^{t-1}(f,g_n)L_\beta(f,g_n)v_{n-1} \le L_\beta^{t-1}(f,g_n)v_n \\ &\le L_\beta^{t-1}(f,g_n)[v_{n-1} + \max_i (v_n - v_{n-1})(i).e] \\ &= L_\beta^{t-1}(f,g_n)v_{n-1} + \beta^{t-1} \max_i (v_n - v_{n-1})(i).e \\ &\le \ldots \le v_n + (\beta + \beta^2 + \ldots + \beta^{t-1}) \max_i (v_n - v_{n-1})(i) \end{aligned}$$

Met $t \to \infty$ volgt dus voor alle f (f was willekeurig)

$$V_\beta(f^{(\infty)},g_n^{(\infty)}) \le v_n + \beta(1-\beta)^{-1} \max_i (v_n - v_{n-1})(i)$$

(iii). Analoog aan (ii). (iv) volgt uit (ii) en (iii) samen. □

Tot nu toe zagen we steeds dat de resultaten volkomen analoog waren aan die bij het verdisconteerde Markov beslissingsprobleem. We moeten echter voorzichtig zijn. De volgende rechtsreekse generalisatie van de policy iteration methode convergeert niet altijd.

Policy iteration methode A

$v_0 = 0$.

Bepaal f_n en g_n, zodanig dat

$$L_\beta(f,g_n)v_{n-1} \leq L_\beta(f_n,g_n)v_{n-1} \leq L_\beta(f_n,g)v_{n-1} \text{ voor alle } f \text{ en } g$$

en definieer $v_n := V_\beta(f_n,g_n)$, $n = 1,2,\ldots$

Voorbeeld 6.1. Beschouw het volgende probleem. $I = \{1,2\}$, $K_0 = L_0 = 2$

$r(1,a,b)$	1	2
1	3	6
2	2	1

$r(2,a,b)$	1	2
1	0	0
2	0	0

p_{11}^{ab}	1	2
1	1	1/3
2	1	1

p_{21}^{ab}	1	2
1	0	0
2	0	0

Verder $p_{i2}^{ab} = 1 - p_{i1}^{ab}$ en $\beta = 3/4$

Passen we het bovenstaande algoritme op dit probleem toe dan vinden we $v_{2n-1} = (12,0)^T$, $v_{2n} = (4,0)^T$, $n = 1,2\ldots$. Met bijbehorende strategieën $f_{2n-1}(1,1) = g_{2n-1}(1,1) = 1$, $f_{2n}(1,1) = g_{2n}(1,1) = 0$.
Ga na.

Wel convergeert het volgende algoritme, dat ook als een generalisatie van de policy iteration methode is te beschouwen.

Policy iteration methode B

Kies v_0 zodanig dat $U_\beta v_0 \leq v_0$

Bepaal g_n zodanig dat

$L_\beta(f,g_n)v_{n-1} \leq U_\beta v_{n-1}$ voor alle f

en bepaal v_n zodanig dat

$$v_n = \max_f V_\beta(f^{(\infty)},g_n^{(\infty)}),$$

$n = 1,2,\ldots$.

Met inductie kan men nu het volgende resultaat bewijzen.

Lemma 6.2. Voor de policy iteration methode B geldt

(6.11) $$v_\beta \le v_n \le U_\beta^n v_0$$

Bewijs. Gebruikmakend van de monotonie van U_β en $L_\beta(f,g)$, die direct te bewijzen is, en het feit dat

$$L_\beta^n(f,g)v \to V(f^{(\infty)},g^{(\infty)})$$

voor alle f,g en v bewijzen we (6.11) met inductie.
Duidelijk is dat $v_\beta \le v_n$. Immers laat f^* voldoen aan (6.9) dan geldt

$$v_n = \max_f V_\beta(f^{(\infty)},g_n^{(\infty)}) \ge V_\beta(f^{*(\infty)},g_n^{(\infty)}) \ge v_\beta.$$

Resteert te bewijzen $v_n \le U_\beta^n v_0$.
Zij $\bar{f}_n$ zodanig dat $\max_f V_\beta(f^{(\infty)},g_n^{(\infty)}) = V_\beta(\bar{f}_n^{(\infty)},g_n^{(\infty)}) = v_n$

$n = 1$. $v_1 = \lim_{m\to\infty} L_\beta^m(\bar{f}_1,g_1)v_0$. Verder

$$L_\beta(\bar{f}_1,g_1)v_0 \le U_\beta v_0 \le v_0$$

Dus ook met de monotonie van $L_\beta(\bar{f}_1,g_1)$

$$L_\beta^m(\bar{f}_1,g_1)v_0 \le L_\beta(\bar{f}_1,g_1)v_0 \le U_\beta v_0 \text{ voor alle } m$$

zodat met $m \to \infty$

$$V_\beta(\bar{f}_1^{(\infty)}, g_1^{(\infty)}) = v_1 \le U_\beta v_0$$

Voor willekeurige n verloopt het bewijs analoog.

Uit $v_n = \max_f V_\beta(f^{(\infty)}, g_n^{(\infty)})$ volgt ook (vgl. de funktionaalvergelijking in hoofdstuk 4)

$$v_n = \max_f L_\beta(f,g_n)v_n$$

Dus $$U_\beta v_n = \max_f \min_g L_\beta(f,g)v_n \le \max_f L_\beta(f,g_n)v_n = v_n$$

Hieruit volgt als in het geval $n = 1$

$$v_{n+1} \le U_\beta v_n$$

Zodat met de monotonie van U_β

$$v_n \leq U_\beta v_{n-1} \leq U_\beta U_\beta v_{n-2} \leq \ldots \leq U_\beta^n v_0 \,. \qquad \square$$

Hieruit volgt direct

Stelling 6.4. De policy iteration methode B convergeert, d.w.z. levert goede onder en bovengrenzen voor v_β en ε-optimale strategieën voor beide spelers.

Bewijs. Uit $v_\beta \leq v_n \leq U_\beta^n v_0$ volgt met $U_\beta^n v_0 \to v_\beta$ ook $v_n \to v_\beta$. Dus $U_\beta v_n - v_n \to 0$. Met stelling 6.3 vinden we dan op den duur goede onder- en bovengrenzen voor v_β en ε-optimale strategieën. $\square$

6.5 Niet nul-som Markov spelen

In de voorgaande paragrafen hebben we ons uitsluitend bezig gehouden met nul-som spelen.

In deze paragraaf willen we nog wat aspecten van niet nul-som spelen bekijken.

In de niet nul-som situatie hebben we niet langer te maken met slechts één opbrengstfunctie (betalingen van P_2 aan P_1) maar met twee opbrengstfuncties

$r_1(i,a,b)$ voor P_1, en

$r_2(i,a,b)$ voor P_2.

Beschouwen we allereerst het bimatrixspel. Het bimatrixspel wordt gekarakteriseerd door twee m × n matrices, zeg A en B.

$$\begin{pmatrix} (a_{11},b_{11}) & \cdots & (a_{1n},b_{1n}) \\ \vdots & & \vdots \\ (a_{m1},b_{m1}) & \cdots & (a_{mn},b_{mn}) \end{pmatrix}$$

Het spel verloopt als volgt. Speler 1 kiest een rij (of een kansverdeling over de rijen) en speler 2 een kolom (of kansverdeling over de kolommen).

Is het resultaat (eventueel na loting) het paar (i,j) dan ontvangt P_1 het bedrag a_{ij} en P_2 een bedrag b_{ij}.

Een evenwichtspunt voor dit spel is een paar gemengde beslissingen $p^* = (p_1^*,\ldots,p_m^*)$, $q^* = (q_1^*,\ldots,q_n^*)$ met de eigenschap:

$$\sum_{i,j} p_i q_j^* a_{ij} \leq \sum_{i,j} p_i^* q_j^* a_{ij} \quad \text{voor alle } p$$

$$\sum_{i,j} p_i^* q_j b_{ij} \leq \sum_{i,j} p_i^* q_j^* b_{ij} \quad \text{voor alle } q.$$

Een dergelijk evenwichtspaar heet een Nash evenwichtspunt.
Elke bimatrix heeft tenminste één Nash-punt. In het algemeen kunnen er meerdere Nash punten zijn die ook, in tegenstelling tot de situatie bij matrix spelen, verschillende waarden kunnen hebben.
Voorbeeld 6.2. Het bimatrix spel

$$\begin{pmatrix} (1,3) & (0,0) \\ (0,0) & (3,1) \end{pmatrix}$$

heeft drie evenwichtswaarden, n.l. (1,3), (3,1) en (3/4,3/4). De laatste behoort bij de gemengde beslissingen p = (1/4,3/4), q = (3/4,1/4).

Keren we nu terug naar het Markov spel.
Het niet-nul som Markov spel verloopt volstrekt analoog aan het nul-som Markov spel; het enige verschil is dat we nu twee objectfuncties hebben; in het M-stapsspel met einduitkeringen q_1 voor P_1 en q_2 voor P_2:

$$V_{1,M}(s,\sigma) := \mathbb{E}_{s,\sigma}\left[\sum_{n=0}^{M-1} r_1(X_n,A_n,B_n) + q_1(X_M)\right]$$

$$V_{2,M}(s,\sigma) := \mathbb{E}_{s,\sigma}\left[\sum_{n=0}^{M-1} r_2(X_n,A_n,B_n) + q_2(X_M)\right]$$

Door het M-stapsspel in (bi-)normaal vorm te brengen kunnen we inzien dat dit spel ook een Nash evenwichtspunt bezit. Een andere, wat constructievere methode, levert de aanpak met dynamische programmering.
Definieer

$$v_{1,0} = q_1 \ , \quad v_{2,0} = q_2$$

en bepaal $v_{1,n}$, $v_{2,n}$, f_n en g_n zodanig dat voor alle f en g

(6.12) $$r_1(f,g_n) + \beta P(f,g_n)v_{1,n-1} \leq r_1(f_n,g_n) + \beta P(f_n,g_n)v_{1,n-1} = v_{1,n}$$

(6.13) $$r_2(f_n,g) + \beta P(f_n,g)v_{2,n-1} \leq r_2(f_n,g_n) + \beta P(f_n,g_n)v_{2,n-1} = v_{2,n}$$

Componentsgewijs vormt dus $(v_{1,n},v_{2,n})$ een Nash evenwichtspunt van het bimatrixspel $(r_1(.,.) + \beta P(.,.)v_{1,n-1},\ r_2(.,.) + \beta P(.,.)v_{2,n-1})$ met corresponderende evenwichtsstrategieën f_n en g_n.

Stelling 6.5. Voor het M-staps niet-nul som Markov spel vormen de strategieën $s^* = (f_M,\ldots,f_1)$ en $\sigma^* = (g_M,\ldots,g_1)$, gedefinieerd door (6.12) en (6.13) een Nash evenwichtspunt met bijbehorende evenwichtswaarde $(v_{1,M},v_{2,M})$, dat wil zeggen voor alle s en σ geldt

$$V_{1,M}(s,\sigma^*) \leq V_{1,M}(s^*,\sigma^*) = v_{1,M}$$

$$V_{2,M}(s^*,\sigma) \leq V_{2,M}(s^*,\sigma^*) = v_{2,M}$$

Bewijs. Het bewijs verloopt volstrekt analoog aan de bewijzen van de stellingen 3.1 en 6.3. □

De strategieën die we op deze wijze vinden zijn altijd Markov strategieën. Dat was in het nulsom spel geen probleem. Alle paren evenwichtsstrategieën hadden dezelfde evenwichtswaarde.
Dus ook de evenwichtsparen van niet-Markov strategieën. Hier echter kunnen, doordat de evenwichtswaarde niet voor elk paar Nash evenwichtsstrategieën gelijk is, belangrijke evenwichtspunten niet met deze dynamische programmeringsaanpak ontdekt worden.

Voorbeeld 6.3. Beschouw eens het ∞-horizon niet-nul som Markov spel met verdisconteringsfactor β, met slechts één toestand. Feitelijk hebben we dus te maken met één bimatrix spel dat steeds herhaald wordt. Het spel is het volgende

$$\begin{pmatrix} (5,5) & (8,0) \\ (0,8) & (7,7) \end{pmatrix}$$

We gaan gemakkelijk na, dat, hoewel (7.7) veel aantrekkelijker is dan (5,5), het spel alleen (5,5) als Nash evenwichtspunt bezit. In het ∞-horizon spel is het paar (7,7) wel een Nash punt, maar het behoort niet bij de stationaire strategieën $f^{(\infty)}$ en $g^{(\infty)}$ met $f(1,2)\ ;\ g(1,2) = 1$. Dan immers levert het voordeel om om te switchen naar $\hat{f}$ resp. $\hat{g}$ met $\hat{f}(1,1) = 1$ en $\hat{g}(1,1) = 1$, als tenminste de ander zijn strategie vast houdt. De strategieën die dit evenwichtspunt vormen zijn de niet-Markov strategieën $\tilde{s}$ en $\tilde{\sigma}$ die in woorden als volgt luiden speel actie 2 net zo lang tot je tegenstander een keer 1 speelt, zodra hij een keer

beslissing 1 neemt swich je zelf definitief naar beslissing 1. Als β voldoende dicht bij 1 ligt, $\beta \geq 1/3$ (ga na), is bedriegen niet zinvol dus vormen $\tilde{s}$ en $\tilde{\sigma}$ een paar evenwichtsstrategieën die niet Markov zijn en dat voor beiden aantrekkelijker is dan het evenwichtspaar in de Markov strategieën.

In het ∞-horizon niet nul-som Markov spel met verdiscontering treden nog wel meer problemen op. Zo zal in het algemeen $(v_{1,n}, v_{2,n})$ niet uniek bepaald worden door $(v_{1,n-1}, v_{2,n-1})$. Zelfs al is $(v_{1,n}, v_{2,n})$ uniek bepaald (hetgeen kan door het opleggen van een aantal extra eisen) dan nog zal de operator, die uit $(v_{1,n-1}, v_{2,n-1})$ het paar $(v_{1,n}, v_{2,n})$ bepaalt, in het algemeen geen contractie zijn.
Het is dan ook onduidelijk of de benadering van het ∞-horizon spel door eindig-horizon problemen enig nut heeft.

Overigens ook bij eindig -staps spelen doet zich al het probleem voor dat we ons niet zonder meer tot Markov strategieën kunnen beperken.
Voorbeeld 6.4. Beschouw het volgende bimatrixspel dat tweemaal achtereen gespeeld wordt.

$$\begin{pmatrix} (0,5) & (5,5) & (12,0) \\ (0,5) & (5,5) & (10,10) \end{pmatrix}$$

In Markov strategieën is het beste dat P_2 kan krijgen 5 + 5 = 10 door steeds actie 1 of 2 te kiezen. Immers P_1 zal actie 1 prefereren.
In niet-Markov strategieën vinden we een extra evenwichtspunt, nl. P_1 speelt de eerste keer actie 2 de tweede keer actie 1 en P_2 speelt als strategie eerst actie 3 daarna als P_1 de eerste keer actie 1 koos ook actie 1 en als P_1 actie 2 nam actie 2. Dus, als P_1 de eerste keer meewerkte wordt hij de tweede keer beloond met opbrengst 5, anders wordt hij bestraft met opbrengst 0.
Dit evenwichtspunt heeft de waarde (15,15) en is dus superieur aan de evenwichtspunten in Markov strategieën die waarden hebben tussen (0,10) en (10,10).

7. Het algemene totale kosten model

7.1. Inleiding

In dit hoofdstuk willen we het model van het verdisconteerde Markov beslissingsprobleem met eindige toestands- en beslissingsruimten generaliseren. En we zullen van een aantal resultaten uit hoofdstuk 4 nagaan of die ook in dit model nog gelden.

We beschouwen het volgende model

$I = \{1,2,\ldots\}$ de aftelbare toestandsruimte

A = de willekeurige actieruimte met daarop een σ-algebra $\mathcal{A}$ die alle 1-puntsverzamelingen omvat

$r(i,a)$ = de opbrengststructuur waarbij we eisen dat $r(i,a)$ voor elke i meetbaar is in a

p_{ij}^a = het overgangsmechanisme; meetbaar in a voor alle i en j, met verder $p_{ij}^a \geq 0$, $\sum_{j\in I} p_{ij}^a \leq 1$.

De twee belangrijkste Markov beslissingsproblemen die aan dit model voldoen zijn

Voorbeeld 7.1. Het verdisconteerde Markov beslissingsprobleem met overgangskansen q_{ij}^a en verdisconteringsfactor β. Definieer namelijk $p_{ij}^a = \beta q_{ij}^a$.

Voorbeeld 7.2. Het semi-Markov beslissingsprobleem met verdiscontering (factor β). In de semi-Markov situatie is de tijd benodigd om bij actie a van i naar j te gaan niet langer constant (= 1) maar stochastisch met verdelingsfunctie $F_{ij}^a(t)$. We nemen verder aan dat de opbrengsten op twee manieren ontstaan; nl. een directe opbrengst $r_1(i,a)$ bij het nemen van de actie en een opbrengst per tijdseenheid $r_2(i,a)$ in de periode tot de nieuwe toestand is bereikt. Zij verder q_{ij}^a de kans dat, als in i actie a genomen wordt, het systeem naar toestand j gaat. Na de transformaties

$$r(i,a) = r_1(i,a) + \sum_j \int_0^\infty \beta^t r_2(i,a)\, dF_{ij}(t)$$

$$p_{ij}^a = q_{ij}^a \int_0^\infty \beta^t dF_{ij}(t)$$

past ook dit probleem in het algemene model van dit hoofdstuk.

Een strategie voor dit algemene beslissingsprobleem is weer een rij $s = (s_0, s_1, \ldots)$ waarbij nu $s_t(i_0, a_0, i_1, \ldots, i_{t-1}, a_{t-1}, i_t, \cdot)$ een kansmaat is op A. Verder is voor elke $B \in A$ de functie $s_t(\cdot, B)$ meetbaar in $(h_t, i_t) = (i_0, a_0, \ldots, i_{t-1}, a_{t-1}, i_t)$ met betrekking tot de product σ-algebra op $I \times A \times \ldots \times I = (I \times A)^t \times I$; op I nemen we de σ-algebra van alle deelverzamelingen. Bij iedere starttoestand $i \in I$ en iedere strategie s kunnen we weer een stochastisch proces $\{(I_t, A_t),\ t = 0,1,\ldots\}$ definiëren als volgt

$$\mathbb{P}_{i,s}(I_0 = i_0,\ A_0 \in B_0,\ I_1 = i_1, \ldots, I_t = i_t) =$$

$$\delta_{ii_0} \int_{a_0 \in B_0} s_0(i_0, da_0) p_{i_0 i_1}^{a_0} \int_{a_1 \in B_1} s_1(i_0, a_0, i_1, da_1) p_{i_1 i_2}^{a_1}$$

$$\ldots \int_{a_{t-1} \in B_{t-1}} s_{t-1}(i_0, a_0, i_1, \ldots, i_{t-1}, da_{t-1}) p_{i_{t-1} i_t}^{a_{t-1}}$$

voor elke $B_0, B_1, \ldots \in A$, $i_0, \ldots, i_t \in I$.
In het algemeen zullen de maten $\mathbb{P}_{i,s}(.)$ geen kansmaten zijn, immers $\sum_j p_{ij}^a$ kan kleiner dan 1 zijn. Om dit te ondervangen zullen we een extra toestand invoeren, zeg toestand 0, met

$$\text{(7.1)} \qquad \begin{cases} p_{i0}^a = 1 - \sum_{j \in I} p_{ij}^a, & a \in A,\ i \in I \\ r(0,a) = 0, & a \in A \\ p_{00}^a = 1, & a \in A\ . \end{cases}$$

Hierna sommeren alle kansen weer tot 1.
We moeten dit doen om te kunnen spreken over een kansmaat $\mathbb{P}_{i,s}$ op $(I \times A)^\infty$ en over de totale verwachte opbrengst. Het is echter zuiver formeel en speelt verder geen enkele rol.
We definiëren nu de totale verwachte opbrengst bij strategie s en start in toestand i door

$$\text{(7.2)} \qquad v(i,s) := \mathbb{E}_{i,s} \sum_{n=0}^{\infty} r(I_n, A_n)\ ,$$

mits natuurlijk het positieve of het negatieve deel van de integraal eindig is.
Om dat te garanderen eisen we

$$\text{(7.3)} \qquad \mathbb{E}_{i,s} \sum_{n=0}^{\infty} r^+(I_n, A_n) < \infty \qquad \text{voor alle } i \text{ en } s\ ,$$

waarin

$$r^+(i,a) = \max\{0, r(i,a)\} .$$

Als gevolg van (7.3) geldt er voor alle i en s

$$-\infty \leq v(i,s) < \infty .$$

Bovendien stelt (7.3) ons in staat om de sommatievolgorde in (7.2) te verwisselen

$$\text{(7.4)} \qquad \mathbb{E}_{i,s} \sum_{n=0}^{\infty} r(I_n, A_n) = \sum_{n=0}^{\infty} \mathbb{E}_{i,s}\, r(I_n, A_n) .$$

In het vervolg zullen drie verzamelingen van strategieën een rol spelen.

S = verzameling van alle strategieën (in de zin van dit hoofdstuk)

RM = verzameling van alle gemengde Markov strategieën (randomized Markov); RM $\subset$ S

M = verzameling van zuivere Markov strategieën; M $\subset$ RM .

(Het Markov zijn van een strategie houdt weer in dat de functies $s_t(i_0, a_0, \ldots, a_{t-1}, i_t, B)$ niet van $i_0, a_0, \ldots, a_{t-1}$ afhangen.)
We zien nu ook waarom geëist is dat $\mathcal{A}$ de 1-puntsverzamelingen bevat. Dit maakt immers dat elke zuivere beslissingsfunctie s die de Markov structuur heeft voldoet aan de meetbaarheidseisen die in dit hoofdstuk aan strategieën zijn opgelegd en dus een zuivere Markov strategie is, dus $s \in M$.

In de volgende paragrafen willen we onder meer laten zien dat

$$\text{(7.5)} \qquad \sup_{s \in S} v(i,s) = \sup_{s \in RM} v(i,s) = \sup_{s \in M} v(i,s) .$$

Definieer nog $v(i) = \sup_{s \in S} v(i,s)$.

7.2. Beperking tot gemengde Markov strategieën

In deze paragraaf zullen we bewijzen dat de eerste gelijkheid in (7.5) inderdaad geldt, dus

$$\sup_{s \in S} v(i,s) = \sup_{s \in RM} v(i,s) .$$

We zullen eerst aantonen dat het mogelijk is bij elke strategie $s \in S$ een strategie $\bar{s} \in RM$ te bepalen die bij starttoestand i dezelfde totale verwachte opbrengst heeft. Het is in het algemeen niet mogelijk een strategie $\bar{s} \in RM$ te construeren die het zelfde stochastisch proces voortbrengt zodat bv.

$$\mathbb{P}_{i,s}(I_0 = i_0, A_0 \in B_0, \ldots, A_n \in B_n) = \mathbb{P}_{i,\bar{s}}(I_0 = i_0, A_0 \in B_0, \ldots, A_n \in B_n) \ .$$

Voorbeeld 7.3. Zij $I = \{1\}$ en $A = \{1,2\}$. Laat strategie s als volgt luiden: Loot op $t = 0$ met gelijke kansen tussen de acties 1 en 2 en kies daarna steeds dezelfde actie. Ga na dat een strategie uit RM nooit dezelfde simultane kansen kan opleveren (veronderstel $p_{11}^a > 0$ voor $a \in A$).

Wel geldt het volgende lemma.

Lemma 7.1. Voor elke $i \in I$ en $s \in S$ bestaat er een $\bar{s} \in RM$ zodat voor de marginale verdelingen van s en $\bar{s}$ geldt:

$$\mathbb{P}_{i,\bar{s}}(I_n = i_n, A_n \in B_n) = \mathbb{P}_{i,s}(I_n = i_n, A_n \in B_n) \ ,$$

voor alle $i_n \in I$, $B_n \in \mathcal{A}$, $n = 0,1,\ldots$.

Bewijs. Bij gegeven i en s construeren we $\bar{s}$ als volgt. Definieer $\bar{s}_0 = s_0$. Dan geldt ook

$$\mathbb{P}_{i,\bar{s}}(I_1 = i_1) = \mathbb{P}_{i,s}(I_1 = i_1) \ .$$

Vervolgens definiëren we $\bar{s}_1$ door

$$\bar{s}_1(i_1, B_1) = \mathbb{P}_{i,s}(A_1 \in B_1 \mid I_1 = i_1), \ B_1 \in \mathcal{A} \ ,$$

indien $\mathbb{P}_{i,s}(I_1 = i_1) > 0$, en willekeurig (maar wel gemengd Markov) als $\mathbb{P}_{i,s}(I_1 = i_1) = 0$.
Dan geldt als $\mathbb{P}_{i,s}(I_1 = i_1) > 0$

$$\mathbb{P}_{i,\bar{s}}(I_1 = i_1, A_1 \in B_1) = \mathbb{P}_{i,\bar{s}}(I_1 = i_1)\mathbb{P}_{i,\bar{s}}(A_1 \in B_1 \mid I_1 = i_1)$$

$$= \mathbb{P}_{i,s}(I_1 = i_1)\mathbb{P}_{i,s}(A_1 \in B_1 \mid I_1 = i_1) = \mathbb{P}_{i,s}(I_1 = i_1, A_1 \in B_1) \ .$$

Als $\mathbb{P}_{i,s}(I_1 = i_1) = 0$ zijn beide kansen 0.
En er geldt

$$\mathbb{P}_{i,\bar{s}}(I_2 = i_2) = \mathbb{P}_{i,s}(I_2 = i_2) \ .$$

Zo voortgaand definiëren we $\bar{s}_n$ voor alle n door

$$\bar{s}_n(i_n, B_n) = \mathbb{P}_{i,s}(A_n \in B_n \mid I_n = i_n), \ B_n \in \mathcal{A} \ .$$

Dan leveren $\bar{s}$ en s bij starttoestand i dezelfde marginale verdelingen

$$\mathbb{P}_{i,\bar{s}}(I_n = i_n, A_n \in B_n) = \mathbb{P}_{i,s}(I_n = i_n, A_n \in B_n) \ ,$$

voor alle $i_n \in I$, $B_n \in \mathcal{A}$ en $n = 0,1,\ldots$.
$\bar{s}$ is gedefinieerd als gemengde Markov strategie en voldoet aan de meetbaarheidscondities voor strategieën.
Daarmee is het lemma bewezen. □

Hieruit volgt direct

Stelling 7.1. Voor elke $i \in I$ en $s \in S$ bestaat er een $\bar{s} \in RM$ zodat

$$v(i,\bar{s}) = v(i,s) \ .$$

Bewijs. Voor de in lemma 7.1 geconstrueerde $\bar{s} \in RM$ geldt

$$\mathbb{P}_{i,\bar{s}}(I_n = i_n, A_n \in B_n) = \mathbb{P}_{i,s}(I_n = i_n, A_n \in B_n) \ .$$

En dus ook

$$\mathbb{E}_{i,\bar{s}} r(I_n,A_n) = \mathbb{E}_{i,s} r(I_n,A_n) \ .$$

Met vergelijkingen (7.4) en (7.2) volgt dan

$$v(i,\bar{s}) = v(i,s) \ .$$ □

Hiermee vinden we

Stelling 7.2. Onder de voorwaarde (7.3) geldt

$$\sup_{s \in S} v(i,s) = \sup_{s \in RM} v(i,s) \ .$$

In het algemeen geldt niet dat er bij elke strategie $s \in S$ een $\bar{s} \in RM$ bestaat zodat

$$v(i,\bar{s}) = v(i,s), \quad \text{voor alle } i \text{ tegelijk (ga na)} \ .$$

Ook bestaat er niet in het algemeen een strategie $\bar{s} \in RM$ die voldoet aan

$$v(i,\bar{s}) \geq v(i) - \varepsilon \ ,$$

voor alle $i \in I$ tegelijk.

Uit lemma 7.1 en stelling 7.1 volgt ook dat de voorwaarde (7.3) equivalent is met

$$\mathbb{E}_{i,s} \sum_{n=0}^{\infty} r^+(I_n,A_n) < \infty \quad \text{voor alle } i \text{ en } s \in RM \ . \tag{7.6}$$

7.3. Beperking tot zuivere Markov strategieën voor het eindig-staps probleem

We beschouwen in deze paragraaf het T-staps probleem met einduitkering $w \leq 0$. Uit lemma 7.1 volgt reeds dat we ons ook in dit geval kunnen beperken tot gemengde Markov strategieën. We zullen laten zien dat we in dit eindig-staps probleem iedere gemengde Markov strategie kunnen vervangen door een zuivere Markov strategie met minstens dezelfde verwachte opbrengst.
In het bewijs hiervan maken we gebruik van het volgende resultaat.

Stelling 7.3. Zij $q(.)$ een kansmaat op $A,\mathcal{A}$, zij $u(.)$ meetbaar en $\int_A q(da)u^+(a) < \infty$, dan bestaat er een $a_0 \in A$ zodanig dat

$$\int_A q(da)u(a) \leq u(a_0) \ .$$

Bewijs. Zelf. □

Met behulp hiervan bewijzen we

Stelling 7.4. Zij $s \in RM$ willekeurig, $w \leq 0$ en zij verder voldaan aan (7.6) dan bestaat er een $\bar{s} \in M$ zodanig dat

$$\mathbb{E}_{i,s}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)] \leq \mathbb{E}_{i,\bar{s}}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)] \ .$$

Bewijs. De aanpak is feitelijk dezelfde als in het bewijs van stelling 3.1. Met als enig probleem het controleren of bepaalde optredende functies integreerbaar zijn. En dat blijkt op grond van (7.6) en $w \leq 0$ steeds het geval te zijn.
Beschouw eerst eens de verwachte opbrengst voor strategie s vanaf tijdstip $T-1$. Als het systeem op $t = T-1$ in j zit, bedraagt deze

$$\int_A s_{T-1}(j,da)[r(j,a) + \sum_k p_{jk}^a w(k)] \ , \tag{7.7}$$

(de integraal in goed gedefinieerd immers

$$\int_A s_{T-1}(j,da)[r(j,a) + \sum_k p_{jk}^a w(k)]^+ \leq \int_A s_{T-1}(j,da)r^+(j,a) < \infty) \ .$$

Volgens stelling 7.3 bestaat er nu een $f_{T-1}(j) \in A$ waarvoor geldt

$$\int_A s_{T-1}(j,da)[r(j,a) + \sum_k p_{jk}^a w(k)] \leq r(j,f_{T-1}(j)) + \sum_k p_{jk}^{f_{T-1}(j)} w(k) .$$

Doen we dit voor alle $j \in I$ dan vinden we een beslissingsregel f_{T-1} waarvoor met $s^{(T-1)} = (s_0,s_1,\ldots,s_{T-2},f_{T-1})$ geldt

$$\mathbb{E}_{i,s}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)] \leq \mathbb{E}_{i,s^{(T-1)}}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)] .$$

Vervolgens kunnen we s_{T-2} verbeteren. Met

$$u_{T-1}(k) = r(k,f_{T-1}(k)) + \sum_\ell p_{k\ell}^{f_{T-1}(k)} w(\ell)$$

bedraagt de opbrengst onder strategie $s^{(T-1)}$ vanaf T - 2

$$\int_A s_{T-2}(j,a)[r(j,a) + \sum_k p_{jk}^a u_{T-1}(k)] .$$

We kunnen dit herschrijven in de vorm

$$\mathbb{E}_{j,(s_{T-2},f_{T-1})}[\sum_{n=0}^{1} r(I_n,A_n) + w(I_2)] ,$$

waaruit met (7.6) en $w \leq 0$ volgt dat de integraal weer goed gedefinieerd is. En er bestaat volgens stelling 7.3 dus een $f_{T-2}(j)$ zodat

$$\int_A s_{T-2}(j,a)[r(j,a) + \sum_k p_{jk}^a u_{T-1}(k)] \leq r(j,f_{T-2}(j) + \sum_k p_{jk}^{f_{T-2}(j)} u_{T-1}(k),$$

en wel voor alle $j \in I$.
Voor de strategie $s^{(T-2)} = (s_0,\ldots,s_{T-3},f_{T-2},f_{T-1})$ geldt nu

$$\mathbb{E}_{i,s}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)] \leq \mathbb{E}_{i,s^{(T-2)}}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)] .$$

Zo voortgaand construeren we een strategie $\bar{s} = (f_0,f_1,\ldots,f_{T-1})$ die in het T-staps probleem tenminste dezelfde opbrengst geeft als s. □

Men kan bewijzen dat de voorwaarde $w \leq 0$ kan worden verzwakt tot $w^+ \leq v^+ < \infty$. Dan geldt namelijk ook

(7.8) $\mathbb{E}_{i,s} w^+(I_t)$ voor alle i, s en t .

En zoals we nog zullen zien (in corollarium 7.2) geldt onder voorwaarde (7.6) inderdaad $v^+ < \infty$.

Uit stelling 7.4 en lemma 7.1 volgt nu dat we ons tot zuivere Markov strategieën kunnen beperken.

Stelling 7.5. Voor het T-staps probleem met einduitkering $w \leq 0$ geldt

$$\sup_{s\in S} \mathbb{E}_{i,s}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)] = \sup_{s\in M} \mathbb{E}_{i,s}[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)],$$

voor alle $i \in I$.

7.4. Positieve dynamische programmering

In deze paragraaf beschouwen we de situatie dat alle opbrengsten niet negatief zijn: $r(i,a) \geq 0$, voor alle $i \in I$, $a \in A$. Zodat ook $r(i,a) = r^+(i,a)$. Een aantal van de resultaten in deze paragraaf zijn ook van belang voor het algemene probleem met zowel positieve als negatieve opbrengsten.
Allereerst bewijzen we het volgende lemma.

Lemma 7.2. Uit voorwaarde (7.6) volgt

$$\sup_{s\in M} \mathbb{E}_{i,s} \sum_{n=0}^{\infty} r(I_n,A_n) < \infty \quad \text{voor alle } i \in I . \tag{7.9}$$

Bewijs. Stel dat het supremum in (7.9) niet eindig is, dan bestaan er dus strategieën $s^1,s^2,\ldots \in M$ met

$$\mathbb{E}_{i,s^m} \sum_{n=0}^{\infty} r(I_n,A_n) \geq 2^m, \quad m = 1,2,\ldots .$$

Beschouw nu de strategie $\hat{s}$ die een loting is tussen de strategieën $s^1,s^2,\ldots$ en wel zo dat strategie s^m met kans 2^{-m} wordt gekozen. Nu is $\hat{s}$ weliswaar geen strategie volgens de definitie van strategie in paragraaf 7.1 (de acties op latere tijdstippen hangen niet af van een lotingsexperiment vooraf) maar er bestaat wel een $\hat{\hat{s}} \in RM$ die dezelfde marginale verdelingen heeft als $\hat{s}$. Ga na dat $\hat{\hat{s}}$ aan de meetbaarheidseisen voldoet. Voor $\hat{\hat{s}}$ geldt dan dat

$$\mathbb{E}_{i,\hat{\hat{s}}} \sum_{n=0}^{\infty} r(I_n,A_n) \geq \sum_{m=1}^{\infty} 2^{-m}2^m = \infty .$$

Maar dit is in tegenspraak met (7.6). Daarmee is het lemma bewezen. □

Met behulp van dit lemma vinden we nu de volgende twee belangrijke resultaten.

Stelling 7.6. Onder de voorwaarde (7.9) geldt

i) $$\sup_{s\in S} \mathbb{E}_{i,s} \sum_{n=0}^{\infty} r(I_n,A_n) < \infty, \quad i \in I$$

ii) $$\sup_{s\in S} \mathbb{E}_{i,s} \sum_{n=0}^{\infty} r(I_n,A_n) = \sup_{s\in M} \mathbb{E}_{i,s} \sum_{n=0}^{\infty} r(I_n,A_n), \quad i \in I .$$

Bewijs.

i) Stel dat het supremum in i) voor zekere j niet eindig is, dan bestaat er voor elke N een $s \in S$, en met stelling 7.1 dus ook een $\bar{s} \in RM$, waarvoor geldt

$$\mathbb{E}_{j,\bar{s}} \sum_{n=0}^{\infty} r(I_n,A_n) > N .$$

Voor T voldoende groot geldt dan

$$\mathbb{E}_{j,\bar{s}} \sum_{n=0}^{T-1} r(I_n,A_n) > N .$$

Volgens stelling 7.4 bestaat er dan ook een zuivere Markov strategie $\hat{s}$ waarvoor geldt

$$\mathbb{E}_{j,\hat{s}} \sum_{n=0}^{T-1} r(I_n,A_n) > N .$$

We kunnen $\hat{s}$ natuurlijk uitbreiden tot een zuivere Markov strategie s^* voor alle t waarvoor geldt

$$\mathbb{E}_{j,s^*} \sum_{n=0}^{\infty} r(I_n,A_n) > N .$$

Dit zou gelden voor alle N maar dat is in tegenspraak met (7.9). Daarmee is i) bewezen.

ii) Zij $s \in S$ willekeurig, dan bestaat er volgens stelling 7.1 een equivalente $\bar{s} \in RM$. Het totaal van de opbrengsten onder strategie $\bar{s}$ kunnen we als volgt splitsen

$$\mathbb{E}_{i,\bar{s}} \sum_{n=0}^{\infty} r(I_n,A_n) = \mathbb{E}_{i,\bar{s}} \sum_{n=0}^{T-1} r(I_n,A_n) + \mathbb{E}_{i,\bar{s}} \sum_{n=T}^{\infty} r(I_n,A_n) .$$

Uit i) volgt dat de tweede term in het rechterlid voor T voldoende groot kleiner is dan ε. Voor het eerste stuk kunnen we volgens stelling 7.4

een equivalente zuivere Markov strategie vinden, die uitgebouwd tot een zuivere Markov strategie voor het ∞-horizon probleem op het stuk vanaf T hoogstens ε ten opzichte van $\bar{s}$ verspeelt. En dus ook in totaal hoogstens ε slechter in dan $\bar{s}$ en s. ε en s zijn willekeurig waarmee ii) is bewezen. □

Uit stelling 7.6 volgt nu direct ($r(i,a) \geq 0$ voor alle i en a)

Corollarium 7.1. De voorwaarden (7.3), (7.6) en (7.9) zijn equivalent.

Bewijs. Direct met lemma 7.2 en stelling 7.6 i). □

En

Corollarium 7.2. Onder de voorwaarde (7.3), (7.6) of (7.9) geldt voor de waardevector v

$$v < \infty .$$

7.5. De functionaalvergelijking

We keren nu weer terug naar het algemene geval met zowel positieve als negatieve opbrengsten. Merk allereerst op dat de corollaria 7.1 en 7.2 niet slechts gelden in het positieve dynamische programmeringsmodel maar algemeen. Immers het feit er ook negatieve opbrengsten zijn speelt bij die twee uitspraken geen enkele rol, omdat we ons daarbij volledig tot het probleem met i.p.v. $r(i,a)$ alleen de positieve delen, $r^+(i,a)$, kunnen beperken. Dus er geldt

Corollarium 7.3.

i) De voorwaarden (7.3), (7.6) en (7.9) zijn equivalent.

ii) Onder de voorwaarde (7.3), (7.6) of (7.9) geldt voor de waardevector v

$$v < \infty .$$

We zullen in deze paragraaf de functionaalvergelijking

$$w = \sup_f \{r(f) + P(f)w\} \tag{7.10}$$

beschouwen.

In paragraaf 4.3 is het verdisconteerde Markov beslissingsproces beschouwd en daar is bewezen dat de waarde v_β de unieke oplossing is van de functionaalvergelijking

$$w = \sup_f\{r(f) + \beta P(f)w\} .$$

Hoe zit dat nu hier? Beschouw eens het volgende triviale voorbeeld.

Voorbeeld 7.4. $I = \{1\}$, $A = \{1\}$, $r(1,1) = 0$, $p_{11}^1 = 1$.
Duidelijk is dat $v = 0$. De functionaalvergelijking (7.10) luidt hier $w = 0 + w$. We zien dus dat v voldoet aan de functionaalvergelijking maar ook dat de oplossing bepaald niet uniek is.

We zullen nu eerst bewijzen dat v altijd aan de functionaalvergelijking voldoet.

Stelling 7.7. Onder de voorwaarde (7.9) (dus ook (7.3) of (7.6)) voldoet v aan de functionaalvergelijking

(7.11) $$w = \sup_f\{r(f) + P(f)w\} .$$

Bewijs. Eerst $v \geq \sup_f\{r(f) + P(f)v\}$. Zij s een uniforme ε-optimale strategie: $v(i,s) \geq v(i) - \varepsilon$ voor alle $i \in I$ (dat zo een s bestaat volgt uit $v < \infty$), en zij f een beslissingsregel. Dan is ook (f,s) (eerst f daarna s) een strategie en er geldt voor alle $i \in I$

$$v(i) \geq v(i,(f,s)) = r(i,f(i)) + \sum_j p_{ij}^{f(i)} v(j,s)$$

$$\geq r(i,f(i)) + \sum_j p_{ij}^{f(i)} v(j) - \varepsilon .$$

Omdat (f,s) een strategie is en v(i,s) en v(i) voor alle i hoogstens ε verschillen bestaat de som $\sum_j p_{ij}^{f(i)} v(j)$. ε en f waren willekeurig, dus $v \geq \sup_f\{r(f) + P(f)v\}$.
Nu $v \leq \sup_f\{r(f) + P(f)v\}$. Zij f weer een willekeurige beslissingsregel en $s = (s_1, s_2, \ldots)$ een gemengde Markov strategie dan geldt voor de strategie (f,s) voor elke i

$$v(i,(f,s)) = r(i,f(i)) + \sum_j p_{ij}^{f(i)} v(j,s)$$

$$\leq r(i,f(i)) + \sum_j p_{ij}^{f(i)} v(j)$$

$$\leq \sup_a\{r(i,a) + \sum_j p_{ij}^a v(j)\} .$$

(De som $\sum p_{ij}^a v(j)$ is voor alle a weer goed gedefinieerd.)

Er geldt dus ook

$$\sup_{(f,s)} v(i,(f,s)) \leq \sup_{a}\{r(i,a) + \sum_{j} p_{ij}v(j)\} .$$

Nu is met s_0 een "gemengde beslissingsregel" ook

$$\sup_{f} v(i,(f,s)) = \sup_{s_0} v(i,(s_0,s))$$

op grond van stelling 7.4. Zodat

$$\sup_{(f,s)} v(i,(f,s)) = \sup_{(s_0,s)} v(i,s_0,s)) = v(i) .$$

En dus

$$v \leq \sup_{f}\{r(f) + P(f)v\} .$$

Daarmee is het bewijs voltooid. □

Merk op dat $v(i) = -\infty$ voor een of meer i hier niet hoeft te worden uitgesloten.

Voor het positieve dynamische programmeringsprobleem vinden we nu het volgende resultaat.

Stelling 7.8. Als $r(i,a) \geq 0$ voor alle i en a en (7.9) geldt, dan is de waarde v de kleinste niet negatieve oplossing van de functionaalvergelijking.

Bewijs. Stel $w \geq 0$ is een oplossing van (7.10). Zij $s = (f_0,f_1,\ldots)$ een willekeurige Markov strategie, dan geldt voor alle T

$$\mathbb{E}_s \sum_{n=0}^{T-1} r(I_n,A_n) \leq \mathbb{E}_s[\sum_{n=0}^{T-1} r(I_n,A_n) + w(I_T)]$$

$$= r(f_0) + P(f_0)r(f_1) + \ldots + P(f_0)\ldots P(f_{T-2})[r(f_{T-1}) + P(f_{T-1})w]$$

$$\leq r(f_0) + P(f_0)r(f_1) + \ldots + P(f_0)\ldots P(f_{T-3})[r(f_{T-2}) + P(f_{T-2})w]$$

$$\leq \ldots \leq w .$$

Dus ook

$$v(s) = \lim_{T\to\infty} \mathbb{E}_s \sum_{n=0}^{T-1} r(I_n,A_n) \leq w$$

$s \in M$ was willekeurig zodat met stelling 7.6 ii) volgt

$$v(i) = \sup_{s \in M} v(i,s) \leq w(i) \quad \text{voor alle } i \in I . \qquad \square$$

7.6. Negatieve dynamische programmering

In deze paragraaf beschouwen we nu het geval dat $r(i,a) \leq 0$ is voor alle i en a: het zogenaamde negatieve dynamische programmeringsprobleem (hiervoor is triviaal aan (7.3), (7.6) en (7.9) voldaan). Het feit dat v aan de functionaalvergelijking voldoet stelt ons in staat de volgende stelling te bewijzen.

Stelling 7.9. Als $r(i,a) \leq 0$ voor alle i en a dan bestaat er een zuivere Markov strategie, $s \in M$, zodanig dat

$$v(s) \geq v - \varepsilon e ,$$

d.w.z. s is uniform ε-optimaal.

Bewijs. Op grond van stelling 7.7 bestaan er beslissingsregels f_n, $n = 0,1,..$ zodanig dat

$$r(f_n) + P(f_n)v \geq v - \varepsilon 2^{-(n+1)} e .$$

Voor de strategie $s = (f_0, f_1, \ldots)$ geldt dan met $v \leq 0$ (vgl. het bewijs van stelling 7.8)

$$\mathbb{E}_s \sum_{n=0}^{T-1} r(I_n,A_n) \geq \mathbb{E}_s[\sum_{n=0}^{T-1} r(I_n,A_n) + v(I_T)]$$

$$= r(f_0) + P(f_0)r(f_1) + \ldots + P(f_0)\ldots P(f_{T-2})[r(f_{T-1}) + P(f_{T-1})v]$$

$$\geq r(f_0) + P(f_0)r(f_1) + \ldots + P(f_0)\ldots P(f_{T-2})[v - \varepsilon 2^{-T} e]$$

$$\geq \ldots \geq v - \varepsilon(2^{-1} + \ldots + 2^{-T})e \geq v - \varepsilon e$$

zodat ook

$$v(s) = \lim_{T \to \infty} \mathbb{E}_s \sum_{n=0}^{T-1} r(I_n,A_n) \geq v - \varepsilon e . \qquad \square$$

Door de bewijzen van de stellingen 7.8 en 7.9 te combineren kunnen we ook het volgende "equivalent" van stelling 7.8 bewijzen.

Stelling 7.10. Als $r(i,a) \leq 0$ voor alle i en a dan is v de grootste niet positieve oplossing van de functionaalvergelijking (7.10).

7.7. De beperking tot zuivere Markov strategieën in het ∞-horizon probleem met zowel positieve als negatieve opbrengsten

We hebben nu alle resultaten afgeleid die we nodig hebben om te bewijzen dat we ons ook in het geval van zowel positieve als negatieve opbrengsten tot zuivere Markov strategieën kunnen beperken.

Stelling 7.11. Onder de voorwaarde (7.9) geldt

$$\sup_{s\in M} v(i,s) = \sup_{s\in S} v(i,s) \quad \text{voor alle } i \in I \ .$$

Bewijs. Zij $\hat{s} \in S$ willekeurig, dan bestaat er volgens stelling 7.1 een $\bar{s} \in RM$ zodat $v(i,\hat{s}) = v(i,\bar{s})$. We kunnen $v(i,\bar{s})$ nu in drie delen splitsen:

$$v(i,\bar{s}) = \mathbb{E}_{i,\bar{s}} \sum_{n=0}^{T-1} r(I_n,A_n) + \mathbb{E}_{i,\bar{s}} \sum_{n=T}^{\infty} r^+(I_n,A_n) + \mathbb{E}_{i,\bar{s}} \sum_{n=T}^{\infty} r^-(I_n,A_n)$$

waarbij $r^-(i,a) := \min\{0,r(i,a)\}$.
Voor T voldoende groot is de middelste term kleiner dan ε. Definiëren we verder $z(s)$ en z door

$$z(s) = \mathbb{E}_s \sum_{n=0}^{\infty} r^-(I_n,A_n) \text{ en } z = \sup_{s\in S} z(s) \ .$$

Volgens stelling 7.9 bestaat er een $\tilde{s} \in M$ zodanig dat

$$z(\tilde{s}) \geq z - \varepsilon e \ .$$

Door nu $\bar{s}$ vanaf T te vervangen door $\tilde{s} = (\tilde{f}_0,\tilde{f}_1,\ldots)$ krijgen we een strategie $s^* = (\bar{s}_0,\bar{s}_1,\ldots,\bar{s}_{T-1},\tilde{f}_0,\tilde{f}_1,\ldots)$ waarvoor geldt

$$\mathbb{E}_{i,s^*} \sum_{n=T}^{\infty} r^-(I_n,A_n) = \mathbb{E}_{i,\bar{s}}\, z(I_T,\tilde{s})$$

$$\geq \mathbb{E}_{i,\bar{s}}\, z(I_T,(\bar{s}_T,\bar{s}_{T+1},\ldots)) - \varepsilon = \mathbb{E}_{i,\bar{s}} \sum_{n=T}^{\infty} r^-(I_n,A_n) - \varepsilon \ .$$

(Merk op dat we inderdaad de uniforme ε-optimaliteit van $\tilde{s}$ gebruikt hebben.)

Beschouw nu nog het T-stapsprobleem met einduitkering $z(\tilde{s})$. Daarvoor bestaat er een zuivere Markov strategie die minstens even goed is als $\bar{s}$, zeg $(f_0,f_1,\ldots,f_{T-1})$.
Voor de zuivere Markov strategie $s^{**} = (f_0,f_1,\ldots,f_{T-1},\tilde{f}_0,\tilde{f}_1,\ldots)$ voor het oneindig horizon probleem geldt dan

$$\mathbb{E}_{i,s^{**}} \sum_{n=0}^{\infty} r(I_n,A_n) = \mathbb{E}_{i,s^{**}}[\sum_{n=0}^{T-1} r(I_n,A_n) + z(I_T,\tilde{s})]$$

$$\geq \mathbb{E}_{i,\tilde{s}}[\sum_{n=0}^{T-1} r(I_n,A_n) + z(I_T,\tilde{s})]$$

$$\geq \mathbb{E}_{i,\tilde{s}}[\sum_{n=0}^{T-1} r(I_n,A_n) + \sum_{n=T}^{\infty} r^{-}(I_n,A_n)] - \varepsilon \geq v(i,\bar{s}) - 2\varepsilon \ .$$

ε is willekeurig evenals s, zodat

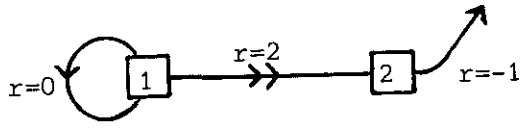
$$\sup_{s\in M} v(i,s) = \sup_{s\in S} v(i,s) \ .$$ □

Dus voor elke toestand i en $\varepsilon > 0$ bestaat er een zuivere Markov strategie die ε-optimaal is voor starttoestand i. Het is niet algemeen zo dat er een zuivere, of zelfs maar gemengde Markov strategie bestaat die voor alle starttoestanden ε-optimaal is.

7.8. Successieve approximaties

Beschouw eens het volgende voorbeeld

Voorbeeld 7.5. $I = \{1,2\}$, $A = \{1,2\}$, $p^1_{11} = p^2_{12} = 1$, alle andere p^a_{ij} zijn 0, $r(1,1) = 0$, $r(1,2) = 1$, $r(2,1) = r(2,2) = -1$. Je kunt dus kiezen tussen in 1 blijven en niets krijgen en naar toestand 2 gaan en 2 ontvangen. Zodra je in 2 zit moet je weer 1 terugbetalen en verlaat je het systeem.



Duidelijk is dat

$$\sup_{s\in S} \mathbb{E}_{1,s} \sum_{n=0}^{T-1} r(I_n,A_n) = 2 \quad \text{voor alle } T \ .$$

(Kies namelijk de strategie "T - 1 keer blijven zitten en dan naar 2"). En dat

$$\sup_{s\in S} \mathbb{E}_{1,s} \sum_{n=0}^{\infty} r(I_n,A_n) = 1 \ .$$

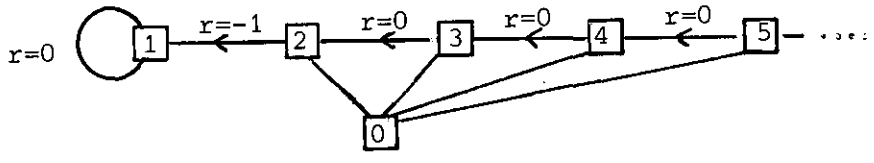
Dus de methode van de successieve approximaties werkt niet algemeen. Wel echter in de positieve dynamische programmeringssituatie.

Stelling 7.12. Als $r(i,a) \geq 0$ voor alle i en a dan convergeert de methode van de successieve approximaties voor elke startvector v_0 met $0 \leq v_0 \leq v$.

Bewijs. Zelf. □

In het negatieve dynamische programmeringsprobleem zal de methode van successieve approximaties echter weer niet noodzakelijk convergeren.

Voorbeeld 7.6. $I = \{0,1,\ldots\}$. Alle $r(i,a)$ zijn nul, behalve $r(2,a) = -1$.



Van toestand n ga je naar n - 1, enz. tot je in toestand 1 komt daar blijf je dan verder. In toestand 0 kun je kiezen naar welke toestand je gaat: 2 of 3 of 4 of ... . Je kunt niet naar 0 en 1. Je kunt je bezoek aan 2 dus willekeurig lang uitstellen maar niet vermijden zodat $v(0) = -1$, maar

$$\sup_{s\in M} \mathbb{E}_{0,s} \sum_{n=0}^{T-1} r(I_n,A_n) = 0 \ .$$