Stochastische ∞-staps beslissingsproblemen *)

door G. de Leve **)

U.D.C. 65.011:51

Summary:

In this paper a survey of known methods for solving stochastic ∞-stage decision problems is given. In addition a generalization for solving problems in which the state of the system is changing continuously in the course of time is discussed.

§ 1. Inleiding

De studie, welke betrekking heeft op het ontwerpen en toepassen van oplossingsmethoden voor ∞-staps beslissingsproblemen vormt een wezenlijk deel van de besliskunde.

In dit artikel zullen wij nagaan hoe het wiskundige ∞-staps beslissingsprobleem kan worden gedefinieerd en opgelost. Wij zullen daartoe een aantal veronderstellingen maken, welke, naar wij menen, aansluiten bij de physische structuur van het beslissingsprobleem.

Bij een ∞-staps beslissingsprobleem wordt de keuze van een beslissing o.a. bepaald door de toestand, waarin de beslissing wordt genomen. Spreekt men echter van „de toestand", dan moet er ook iets bestaan waarop deze toestand betrekking heeft. In de wiskundige formulering van het ∞-staps beslissingsprobleem spreekt men van de *toestand van het systeem.*

Bij een vervangingsprobleem kan het systeem b.v. de meer abstracte benaming zijn van de betreffende machine, terwijl bij een voorraadprobleem met het systeem veelal de voorraad wordt bedoeld.

Aangenomen wordt dat het systeem zich in verschillende toestanden kan bevinden en dat deze toestanden kunnen worden beschreven met behulp van kwantitatieve grootheden.

Indien de toestand van het systeem kan worden beschreven met behulp van N kwantitatieve grootheden, dan kan iedere toestand worden weergegeven door een punt ψ uit een N-dimensionale cartesische ruimte Ψ. Deze ruimte Ψ zullen wij de *toestandsruimte* noemen.

Omgekeerd behoeft niet met ieder punt uit deze N-dimensionale ruimte een mogelijke toestand van het systeem te corresponderen.

In het hierna volgende zullen wij steeds aannemen, dat de toestand van het

---

*) Rapport S 301 (SP 79) van de Afdeling Mathematische Statistiek van het Mathematisch Centrum; de auteur hield over dit onderwerp een voordracht op de Statistische Dag 1962.

**) Medewerker Afdeling Mathematische Statistiek van het Mathematisch Centrum, Amsterdam.

systeem kan worden weergegeven door een punt uit een N-dimensionale ruimte. Wij zullen niet ingaan op de vraag hoe men aan de hand van een gegeven physisch beslissingsprobleem de toestand van het systeem definieert. Wij volstaan slechts met vast te stellen, dat in de toestand van het systeem dat deel van de aanwezige informatie over het systeem is verwerkt, dat kenmerkend is voor de situatie waarin het systeem op het beschouwde tijdstip verkeert. Het verdient uiteraard aanbeveling om bij het vaststellen van de toestandsgrootheden zo „zuinig" mogelijk te werk te gaan.

In dit overzicht zullen wij ons beperken tot die ∞-staps beslissingsproblemen, waarbij de toestand van het systeem in de loop van de tijd verandert op een *stochastische* wijze. Verder zullen wij aannemen, dat in een eindig tijdsinterval slechts een eindig aantal beslissingen wordt genomen. Bijgevolg wordt slechts op *discrete* tijdstippen beslist.

Indien de beslisser geen beslissingen neemt en als in de loop van de tijd toch toestandsveranderingen plaats vinden, dan spreekt men van een *natuurlijk proces* in de toestandsruimte.

Het natuurlijke proces dwingt het systeem tot het maken van een stochastische wandeling door de toestandsruimte. Het natuurlijke proces zelf wordt gedefinieerd door een stochastisch proces.

Veronderstelling no. 1.

Wij zullen aannemen, dat het natuurlijke proces in de toestandsruimte kan worden beschreven met behulp van een *stationair sterk* M a r k o v *proces*[1]).

De natuurlijke processen zijn dus voor iedere begintoestand van het systeem gedefinieerd.

Uiteraard spelen ook bij ∞-staps beslissingsproblemen opbrengsten en kosten een belangrijke rol.

Zonder de algemeenheid te schaden zullen wij in het vervolg slechts spreken van kosten. Hoe brengen wij deze kosten tot uitdrukking in het mathematische model?

In het algemeen treft men bij beslissingsproblemen twee soorten van kosten aan. In de eerste plaats kosten, die optreden op discrete tijdstippen (b.v. bij breuk) en vervolgens kosten, welke ontstaan gedurende een tijdsinterval (b.v

---

[1]) Zie voor de definitie van een sterk M a r k o v proces: M. L o è v e, Probability theory, 2de druk, bl. 577.

Verder wordt verondersteld, dat de stochastische wandeling, welke een gevolg is van het natuurlijke proces, met kans 1 slechts een eindig aantal discontinuïteiten heeft in een eindig tijdsinterval en in de tijdsparameter continu is van rechts.

rente derving). In dit artikel zullen wij niet ingaan op de vraag hoe beide soorten van kosten worden gedefinieerd in het mathematische model, maar wel zullen wij aangeven aan welke eigenschappen zij zullen voldoen.

Veronderstelling no. 2.

Als in een tijdsinterval geen beslissingen worden genomen dan:

a) zijn voor iedere mogelijke stochastische wandeling door de toestandsruimte in dat tijdsinterval de totale kosten gedefinieerd.
b) zullen de gemaakte kosten slechts afhangen van de toestanden, welke het systeem in dat tijdsinterval heeft doorlopen.
c) geldt voor een volledige en disjuncte onderverdeling van het interval in deelintervallen, dat de totale kosten gelijk zijn aan de som van de kosten gemaakt in die deelintervallen.

Men merke op, dat de toestand van het systeem voldoende uitgebreid gedefinieerd moet zijn, wil in een concreet beslissingsprobleem aan deze veronderstellingen zijn voldaan.

Aangezien aan iedere stochastische wandeling door de toestandsruimte in een tijdsinterval op ondubbelzinnige wijze de bijbehorende kosten zijn toegevoegd, zal de beslisser, om hoge onkosten in de toekomst zoveel mogelijk te vermijden, trachten bepaalde wandelingen van het systeem te voorkomen, of, als dit onmogelijk is, minder waarschijnlijk te maken.

Indien men bij een beslissingsprobleem nagaat wat het effect is van een beslissing, dan constateert men steeds dat het effect correspondeert met een toestandsverandering in de wiskundige formulering. Als bij een vervangingsprobleem de toestand van het systeem o.a. gegeven wordt door de leeftijd van de beschouwde machine, dan zal de beslissing, welke de vervanging van de oude door een nieuwe machine inhoudt, een toestandsverandering veroorzaken. Ook bij een voorraadprobleem zal de beslissing, welke leidt tot het bestellen van goederen een toestandsverandering te weeg brengen. Immers de omstandigheid, dat een bepaalde hoeveelheid goederen op zeker tijdstip is besteld en nog niet is afgeleverd, bepaalt mede de te nemen beslissing en maakt derhalve deel uit van de toestand van het systeem.

Veronderstelling no. 3.

In de wiskundige formulering van een ∞-staps beslissingsprobleem vloeien uit beslissingen toestandsveranderingen van het systeem voort.

Dit zijn toestandsveranderingen die niet worden veroorzaakt door het natuurlijke proces. Bij vervangingsproblemen zullen de „begintoestanden" van

de nieuwe machines niet altijd identiek zijn. Aangezien men meestal een aselecte keuze doet uit de verzameling van nieuwe machines, brengt een beslissing een stochastische toestandsverandering te weeg. Daar deze toestandsveranderingen kunnen worden beschreven met behulp van kansverdelingen maken wij de volgende veronderstelling:

Veronderstelling no. 4.

In de mathematische formulering van het ∞-staps beslissingsprobleem wordt een beslissing geidentificeerd met een kansverdeling (eventueel ontaard) van de toestand, waarin het systeem ten gevolge van de „beslissing" zal komen.

Naar analogie van de toestandsruimte Ψ zullen beslissingen worden weergegeven door punten d uit de z.g. *beslissingsruimte* D.

Indien men zich beperkt tot kansverdelingen, waarvan alle momenten bestaan dan kan voor D een ∞-dimensionale Cartesische ruimte gekozen worden[2])

Het nemen van een beslissing brengt doorgaans ook kosten met zich mede (inruilen van een machine, kopen van goederen etc.). Met betrekking tot deze kosten maken wij de volgende veronderstelling.

Veronderstelling no. 5.

De kosten, welke uit een beslissing voortvloeien hangen behalve van de beslissing alleen af

a) van de toestand waarin de beslissing is genomen, en
b) van de toestand waarin het systeem tengevolge van de beslissing is terecht gekomen.

Op grond van de veronderstellingen no. 2 en no. 5 kunnen wij vaststellen:

Conclusie no. 1.

De kosten, die bij een ∞-staps beslissingsprobleem in elk tijdsinterval worden gemaakt, hangen alleen af van de in dat tijdsinterval doorlopen toestanden en de in dat tijdsinterval genomen beslissingen.

Wij weten, dat als geen beslissingen worden genomen het systeem onderworpen is aan een natuurlijk proces. Wij zullen thans dienen vast te stellen wat er gebeurt nadat een beslissing is genomen.

[2]) Voor het geval, dat de toestandsruimte een twee-dimensionale ruimte (ψ_1, ψ_2) is en de beslissingen normale verdelingen zijn, dan kunnen alle beslissingen (verdelingen) worden weergegeven door punten uit een 5-dimensionale ruimte. Immers, iedere 2- dimensionale normale verdeling wordt door de parameterwaarden $\mu_{\psi_1}, \mu_{\psi_2}, \sigma_{\psi_1}, \sigma_{\psi_2}$ en $\varrho_{\psi_1\psi_2}$ volledig bepaald.

Veronderstelling no. 6.

In dit artikel wordt aangenomen, dat het gedrag van het systeem tussen twee opeenvolgende beslissingen kan worden beschreven met behulp van het natuurlijke stationaire sterke Markov proces, waarvan de begintoestand gelijk is aan de toestand, waarin het systeem ten gevolge van de eerste beslissing is terecht gekomen[3]).

In het begin van deze paragraaf hebben wij aangenomen, dat slechts op discrete tijdstippen een beslissing kan worden genomen. De formulering van het ∞-staps beslissingsprobleem wordt echter eenvoudiger als men aanneemt, dat op *ieder* tijdstip een beslissing wordt genomen maar dat deze beslissingen slechts op discrete tijdstippen aanleiding kunnen geven tot „kunstmatige" toestandsveranderingen. Op de overige tijdstippen wordt het systeem met kans 1 „geplaatst" in de toestand, waarin het zich reeds bevond. Wij dienen nu de volgende veronderstelling in te voeren, opdat ook dit soort van beslissingen (z.g. ontaarde beslissingen) gedefinieerd kan worden.

Veronderstelling no. 7.

De beslissingsruimte D bevat alle ontaarde kansverdelingen, welke geconcentreerd zijn in één punt van de toestandsruimte.

Uit de structuur van menig beslissingsprobleem volgt, dat op zeker beslissingstijdstip niet altijd alle beslissingen zijn toegelaten. Wij maken nu de volgende veronderstelling:

Veronderstelling no. 8.

In dit artikel wordt aangenomen, dat de omstandigheid of een beslissing is toegelaten alleen wordt bepaald door de toestand van het systeem op het beslissingstijdstip.

In de wiskundige formulering is bijgevolg aan iedere toestand $\psi \epsilon \Psi$ een verzameling van toegelaten beslissingen in D toegevoegd.

De *oplossing* van het ∞-staps beslissingsprobleem zal worden gegeven in de vorm van een *beslissingsvoorschrift* (strategie), met behulp waarvan voor elk tijdstip, op grond van de beschikbare informatie, ondubbelzinnig kan worden vastgesteld welke beslissing moet worden genomen.

[3]) Het beeld van het door beslissingen gestoorde proces in de toestandsruimte is compleet. Het systeem is vanaf de begintoestand tot aan het eerste beslissingstijdstip onderworpen aan een natuurlijk proces. Op het eerste beslissingstijdstip vindt een toestandsverandering plaats ten gevolge van de beslissing en daarna gaat het proces verder als een natuurlijk proces met als begintoestand de toestand, waarin het systeem ten gevolge van de beslissing is terecht gekomen.

Indien een beslissingsvoorschrift wordt toegepast dan is het duidelijk, dat door de extra toestandsveranderingen, welke een gevolg zijn van de genomen beslissingen, het natuurlijke proces niet meer geëigend is om het gebeuren in de toestandsruimte te beschrijven.

Wij zullen ons nu beperken tot die beslissingsvoorschriften, waarvoor geldt, dat een waarschijnlijkheidsveld geconstrueerd kan worden met behulp waarvan het gebeuren in de toestandsruimte *wel* kan worden beschreven.

Voor het bepalen van de *optimale* strategie dient men de beschikking te hebben over een criterium.

Als aan het criterium een aantal „vanzelfsprekende" eigenschappen wordt opgelegd, dan kan men bewijzen, dat men slechts de toestand van het systeem op het beslissingstijdstip behoeft te kennen om het optimale beslissingsvoorschrift te kunnen toepassen. Informatie over het gedrag van het systeem in het verleden stelt de beslisser dus niet in staat om tot betere beslissingen te geraken. Dit resultaat behoeft niemand te verwonderen. Immers, op de beslissingstijdstippen wordt men door het „verleden van het systeem" niet beperkt in de keuze van de beslissing (veronderstelling no. 8). Het „verleden"heeft ook geen invloed op de ontwikkeling na de beslissing (veronderstelling no. 6) en tenslotte hangen ook de kosten, welke na een beslissing worden gemaakt niet af van wat vroeger is geschied (veronderstellingen no. 2 en no. 5). Wij kunnen nu de volgende conclusie trekken:

Conclusie no. 2.

Het optimale beslissingsvoorschrift voegt aan iedere toestand van het systeem (punt ψ in Ψ) een toegelaten beslissing (punt d in D) toe.

In het hierna volgende zullen wij ons beperken tot beslissingsvoorschriften z van de volgende vorm:

$$d = z(\psi) \tag{1.1}$$

Met ieder beslissingsvoorschrift z van dit type correspondeert in de toestandsruimte een verzameling A_z (*storingszone*) van toestanden met de volgende eigenschappen:

a) zodra het systeem een toestand aanneemt uit de verzameling A_z dan volgt uit de genomen beslissing met kans 1 een toestandsverandering.
b) indien een toestand in A_z op verschillende tijdstippen door het systeem wordt aangenomen, dan zullen op die tijdstippen ook dezelfde beslissingen worden genomen.

Verder merken wij nog op, dat voor ieder reëel beslissingsvoorschrift z de

verzameling A_z gesloten moet zijn. Dat deze voorwaarde noodzakelijk is kan men eenvoudig inzien door na te gaan wat er gebeurt als A_z open is en het systeem A_z binnen wandelt langs een continu pad. Omdat A_z open is kan men geen tijdstip aanwijzen, waarop het systeem voor het eerst in A_z kwam.

Indien men het systeem slechts aanschouwt op de tijdstippen, waarop het een toestand aanneemt uit de verzameling A_z, dan kunnen de overgangen tussen deze toestanden worden beschreven door een stationair Markov proces met discrete tijdsparameter[4]).

De tijdstippen, waarop het systeem een toestand aanneemt uit A_z, zullen wij in het vervolg reële beslissingstijdstippen noemen.

In de paragrafen 2 en 3 zullen nu een tweetal bekende oplossingsmethoden worden besproken, waarna in § 4 een generalisatie zal worden behandeld.

§ 2. Dynamische programmering

Een klasse van ∞-staps beslissingsproblemen kan worden opgelost met behulp van *dynamische programmering*.

In een dynamisch programmeringsprobleem mogen slechts op van te voren vastgestelde, equidistante tijdstippen niet-ontaarde beslissingen worden genomen. Voor de eenvoud zullen wij aannemen, dat de lengte van het tijdsinterval tussen twee opeenvolgende tijdstippen gelijk is aan de tijdseenheid. Om er voor te zorgen, dat slechts op equidistante tijdstippen niet-ontaarde beslissingen worden genomen, voegt men aan het begrip toestand een toestandsgrootheid toe, die aangeeft hoeveel tijd reeds verstreken is sinds het laatste voorafgaande gegeven beslissingstijdstip. Deze toestandsgrootheid doorloopt dus het interval [0,1]. Wij kunnen nu dankzij veronderstelling no. 8 voorkomen, dat tussentijds ook niet-ontaarde beslissingen worden genomen. Uit de definitie van de toegevoegde toestandsgrootheid volgt, dat op de equidistante tijdstippen altijd een toestandsverandering plaats vindt. Men kan deze toestandsverandering toeschrijven aan het natuurlijke proces, maar het past meer in de lijn van het betoog om deze toestandsverandering te zien als een gevolg van een beslissing op het desbetreffende tijdstip. Bijgevolg worden op de gegeven tijdstippen nooit ontaarde beslissingen genomen. Met andere woorden: het systeem neemt op die tijdstippen onverschillig de toegepaste strategie z altijd een toestand uit A_z aan. Deze gekunstelde beschrijving dient slechts om de problemen, welke

[4]) Van het beslissingsvoorschrift z wordt geëist, dat met kans 1 slechts een eindig aantal beslissingen wordt genomen in een eindig tijdsinterval. In [3]) hebben wij geconstateerd, dat het beeld van het gestoorde proces compleet is. Uit het gestelde volgt, dat op het beslissingstijdstip het systeem twee toestanden aanneemt. Men kan nu het gestoorde proces „ontbinden" in het Markov proces met discrete tijdsparameter en een stationair sterk Markov proces met een continue tijdsparameter. Dit laatste proces bezit dan dezelfde eigenschappen als het natuurlijke proces, maar toestanden uit de verzameling A_z worden niet doorlopen.

worden besproken in deze en de volgende paragrafen, te kunnen behandelen vanuit één gezichtspunt.

In het ∞-staps dynamische programmeringsprobleem worden onkosten verdisconteerd en wel op de volgende wijze: onkosten gemaakt in de periode tussen het $k$de en het $(k+1)$ste reële beslissingstijdstip worden vermenigvuldigd met een factor α^{k-1}, waarbij α voldoet aan:

$$0 \leqq \alpha < 1. \tag{2.1}$$

De onkosten, welke worden gemaakt in deze periode zijn stochastisch. Indien I_k de toestand aangeeft op het $k$de reële beslissingstijdstip en d_k de genomen beslissing dan wordt de voorwaardelijke verwachting van deze onkosten bij gegeven I_k en d_k, aangeduid door de functie $h(I_k; d_k)$.

Indien de beslissingen worden genomen volgens een beslissingsvoorschrift z dan is op A_z een functie $C(z; I)$ gedefinieerd, die de verwachting van de totale verdisconteerde onkosten aangeeft als in de toestand I en op de volgende reële beslissingstijdstippen wordt beslist volgens z.

Aangezien na de eerste beslissing in een ∞-staps beslissingsprobleem een nieuw ∞-staps beslissingsprobleem ontstaat van de zelfde vorm, maar wellicht met een andere begintoestand, geldt:

$$C(z; I_1) = h(I_1; z(I_1)) + \alpha \mathscr{E}\{C(z; \underline{I}_2) \mid I_1; z(I_1)\} \tag{2.2}$$

Indien D_{I_1} de klasse van toegelaten beslissingen voorstelt als het systeem zich bevindt in I_1 dan geldt voor de optimale strategie z_0:

$$C(z_0; I_1) = \min_{d_1 \in D_{I_1}} [h(I_1; d_1) + \alpha \mathscr{E}\{C(z_0; I_2) \mid I_1; d_1\}]. \tag{2.3}$$

In een dynamisch programmeringsprobleem wordt de oplossing verkregen door de functionaalvergelijking (2.3) op te lossen.

Beschouwen wij nogmaals de relatie (2.3) dan constateren wij dat bij de bepaling van de optimale strategie de functie $h(I; d)$ een zeer belangrijke rol speelt. Men kan $h(I; d)$ interpreteren als het verlies, dat men zal lijden als men in de toestand I een beslissing d neemt. In werkelijkheid echter is $h(I; d)$ de verwachting van de onkosten gemaakt in een tijdsinterval.

§ 3 Markovian Decision processes

Ook nu beschouwen wij eerst de situatie, welke ontstaat als slechts niet-ontaarde beslissingen mogen worden genomen op van te voren vastgestelde tijdstippen.

Indien steeds wordt beslist volgens een beslissingsvoorschrift z en als de toestand op het eerste gegeven beslissingstijdstip wordt gegeven door I_1 dan wordt

de kans op het aannemen van een toestand in een verzameling B op het $j$de gegeven beslissingstijdstip gedefinieerd door $Q_z^{j-1}(B; I_1)$; wij vinden derhalve voor de verwachting van de onkosten voor de periode tussen het $j$de en het $(j+1)$ste tijdstip:

$$\int_{A_z} Q_z^{j-1}(dI_j; I_1)\, h(I_j; z(I_j)), \qquad j \geqq 2 \tag{3.1}$$

waarbij A_z wederom de verzameling is van toestanden in de toestandsruimte, waarin alleen niet-ontaarde beslissingen worden genomen (zie par. 2).

De verwachting van de onkosten voor de eerste m perioden wordt gegeven door:

$$h(I_1; z(I_1)) + \sum_{j=2}^{m} \int_{A_z} Q_z^{j-1}(dI_j; I_1)\, h(I_j; z(I_j)) \tag{3.2}$$

en dus gemiddeld per periode:

$$\frac{h(I_1, z(I_1))}{m} + \frac{1}{m} \sum_{j=2}^{m} \int_{A_z} Q_z^{j-1}(dI_j; I_1)\, h(I_j; z(I_j)) \tag{3.3}$$

Beschouwen wij nu de rij van toestandsgrootheden $\underline{I}_j$.

Men kan bewijzen, dat onder zekere voorwaarden de toestandsveranderingen $\underline{I}_j \to \underline{I}_{j+1}$ kunnen worden beschreven door een stationair Markovproces met discrete tijdsparameter en met één of meer invariante verdelingen. Als I_1 de begintoestand aangeeft, dan geldt voor de invariante verdeling $Q_z(B; I_1)$:

$$Q_z(B; I_1) = \lim_{m\to\infty} \frac{1}{m} \sum_{j=2}^{m+1} Q_z^{j-1}(B; I_1) \tag{3.4}$$

en derhalve is het plausibel, dat als $m \to \infty$ (3.3) overgaat in:

$$\int_{A_z} Q_z(dI; I_1)\, h(I; z(I)) \tag{3.5}$$

Het ligt voor de hand om als *criterium* voor de optimale strategie te kiezen:

$$C(z; I_1) = \int_{A_z} Q_z(dI; I_1)\, h(I; z(I)) \tag{3.6}$$

Men bedenke echter wel, dat dit criterium slechts bruikbaar is voor die strategieën waarvoor het gedefinieerd is.

Ook nu blijkt, dat de uitdrukking $h(I; d)$ een belangrijke rol speelt bij de bepaling van de optimale strategie. Tevens zien wij dat de invariante kansverdeling van het Markovproces met discrete tijdsparameter, ingevoerd om de toestandsveranderingen $\underline{I}_j \to \underline{I}_{j+1}$ te beschrijven, van groot nut is.

In plaats van (3.6) mogen wij ook schrijven:

$$C(z; I_1) = \mathscr{E}\{h(\underline{I}; z(\underline{I})) \mid I_1\} \tag{3.7}$$

waarbij de verwachting dan betrekking heeft op de invariante verdeling van het M a r k o v proces met discrete tijdsparameter.

De verwachting van de kosten in de stationaire toestand kan ook in andere situaties dienen als criterium voor de optimale strategie.

Wij zullen nu een situatie beschouwen, waarin reële beslissingen kunnen worden genomen op momenten, waarop het systeem van toestand verandert. Stel dat, als de strategie z wordt toegepast, de kans op een toestandsverandering van I naar een toestand uit de verzameling B binnen een periode van de lengte dt wordt gegeven door:

$$\varrho_z(B; I)\, dt \tag{3.8}$$

Verder zij gegeven, dat de kans op meer dan één toestandsverandering in die periode van de orde $(dt)^2$ is.

Indien het systeem zich gedurende een tijdseenheid in de toestand I bevindt dan zullen de kosten $l(I; z)$ bedragen, terwijl bij een toestandsverandering van $I \to I'$ een bedrag gelijk aan $l(I; I'; z)$ aan kosten wordt gemaakt.

Als criterium voor de optimale strategie kiezen wij de verwachting van de kosten per tijdseenheid in de stationaire toestand. Dit criterium wordt dan gegeven door:

$$C(z; I_1) = \int_{\Psi} Q_z(dI; I_1)\ \{l(I; z) + \int_{\Psi} \varrho_z(dI'; I)\, l(I; I'; z)\} \tag{3.9}$$

waarbij $Q_z(B; I_1)$ de invariante kansverdeling aangeeft van het proces, dat ontstaat bij toepassing van de strategie z.

§ 4 Een generalisatie

Indien één van de toestandsgrootheden de tijd aangeeft verstreken sinds de laatste reële beslissing en als men vrij wil zijn in de keuze van de tijdstippen waarop niet-ontaarde beslissingen mogen worden genomen, dan kunnen de methoden, besproken in de paragrafen 2 en 3, niet worden gebruikt. Immers bij de problemen in deze paragrafen zijn òf de reële beslissingstijdstippen van te voren gegeven òf zij vallen samen met de momenten, waarop toestandsveranderingen plaatsvinden. Indien het laatste het geval is dan wordt verondersteld, dat de kans op meer dan één toestandsverandering in een periode dt van de orde $(dt)^2$ is. De hierboven gegeven toestandsgrootheid verandert echter continu in de tijd en aan de gestelde voorwaarde wordt dus zeker niet voldaan. Bij

vervangings- en voorraadproblemen kan men b.v. toestandsgrootheden van dit type aantreffen.

Ook in de situatie, welke wij in deze paragraaf zullen behandelen maakt het systeem een stochastische wandeling door de toestandsruimte Ψ.

Zoals reeds in § 1 is vastgesteld wordt de omstandigheid of de beslisser al of niet een toestandsverandering van het systeem zal veroorzaken bepaald door de toestand van het systeem op dat tijdstip.

Indien een strategie z wordt toegepast, dan zal men pas ingrijpen in het natuurlijke proces, zodra het systeem een toestand aanneemt uit de verzameling A_z.

Indien beslissingen op discrete tijdstippen worden genomen, dan kunnen onder zekere voorwaarden de toestandsveranderingen $\underline{I}_j \to \underline{I}_{j+1}$ worden beschreven met behulp van een stochastisch proces met discrete tijdsparameter. Men kan bewijzen, dat dit proces in de verzameling A_z een stationair Markov proces is met discrete tijdsparameter. In onze verdere beschouwingen speelt dit proces een belangrijke rol.

Stel dat Z_0 een klasse is van strategieën met een aantal speciale eigenschappen. Een van deze eigenschappen houdt in dat het hierboven genoemde Markov proces één invariante kansverdeling bezit, terwijl bovendien de doorsnede van alle verzamelingen A_z ($z \epsilon Z_0$) niet leeg is[5]). Zij A_0 een niet lege deelverzameling van deze doorsnede. Tenslotte zullen wij aannemen, dat bij een natuurlijk proces met kans 1 vanuit iedere toestand $\psi \epsilon \Psi$ het systeem na een eindige tijd een toestand aanneemt uit de verzameling A_0 (de z.g. *stopzone*).

Wij beschouwen vervolgens een rij van stochastische wandelingen $\{w_n; n = 0,1,2, \ldots\}$ in de toestandsruimte met de volgende eigenschappen:

a) w_n begint in de toestand $\psi_0 \epsilon \Psi$.
b) er wordt achtereenvolgens n maal volgens het beslissingsvoorschrift z een niet-ontaarde beslissing genomen.
c) na de $n$de niet-ontaarde beslissing wordt er geen reële beslissing meer genomen, maar wordt de wandeling beëindigd zodra een toestand uit de verzameling A_0 wordt aangenomen.

Stel dat de verwachting van de kosten, tijdens de wandeling w_n gemaakt, bestaat en wordt aangegeven door $k_n(\psi_0; z)$. Stel vervolgens dat tevens de verwachting van de tijdsduur van de wandeling w_n bestaat en wordt aangeduid met $t_n(\psi_0; z)$.

We maken nu het volgende gedachten experiment:

De beslisser heeft in de begintoestand ψ_0 de keuze tussen:

---

[5]) Voor meer fuiken zie [4].

a) het zelf dragen van de kosten gemaakt door het systeem tijdens de wandeling w_n

en

b) het uitbesteden van het systeem tegen een vast bedrag λ per tijdseenheid.

Indien wij nu de periode beschouwen, waarin de wandeling w_n wordt gemaakt en als de beslisser zijn keuze zal doen op grond van de verwachting van de kosten, die in deze periode worden gemaakt, dan is het duidelijk, dat hij aan a) de voorkeur geeft als geldt:

$$k_n(\psi_0; z) < \lambda t_n(\psi_0; z) \tag{4.1}$$

of

$$\frac{k_n(\psi_0; z)}{t_n(\psi_0; z)} < \lambda \tag{4.2}$$

Thans voeren wij in de functies $k(\psi; d)$ en $t(\psi; d)$ gedefinieerd voor iedere $\psi \epsilon \Psi$ en iedere toegelaten d in D.

Beschouwen wij hiertoe twee stochastische wandelingen vanuit de toestand ψ. In de eerste wandeling begint het systeem in ψ direct al met een toestandsverandering, te weeg gebracht door een beslissing d, en daarna gaat het verder tot dat het voor het eerst een toestand aanneemt uit A_0. In de tweede wandeling gaat het systeem regelrecht van ψ naar A_0.

Als de beslissing d ontaard is dan zijn beide stochastische wandelingen identiek.

De functiewaarde van $k(\psi; d)$ geeft het verschil aan tussen de verwachtingswaarden van de te maken kosten in de eerste en de tweede wandeling.

De functiewaarde van $t(\psi; d)$ geeft het verschil aan tussen de verwachtingswaarden van de tijdsduren van de eerste en de tweede wandeling.

Uit deze beide definities volgt:

a) $k(\psi; d) = 0$ (4.3)

$t(\psi; d) = 0$ (4.4)

als d ontaard is,

en, als d niet ontaard is en de strategie z voldoet bij de gegeven waarden van d en ψ aan $d = z(\psi)$:

b) $k(\psi; d) = k_1(\psi; z) - k_0(\psi; z)$ (4.5)

$t(\psi; d) = t_1(\psi; z) - t_0(\psi; z)$ (4.6)

Men kan nu bewijzen dat geldt:

a) $k_n(\psi_0; z) = k_0(\psi_0; z) + \sum_{j=1}^{n} \mathscr{E}\{k(\underline{I}_j; z(\underline{I}_j) \mid \psi_0\}$ (4.7)

b) $t_n(\psi_0; z) = t_0(\psi_0; z) + \sum_{j=1}^{n} \mathscr{E}\{t(\underline{I}_j; z(\underline{I}_j) \mid \psi_0\}$ (4.8)

terwijl onder zekere voorwaarden tevens kan worden aangetoond dat:

$$\lim_{n\to\infty} \frac{k_n(\psi_0; z)}{t_n(\psi_0; z)} = \frac{\mathscr{E}\{k(\underline{I}; z(\underline{I})) \mid \psi_0\}}{\mathscr{E}\{t(\underline{I}; z(\underline{I})) \mid \psi_0\}}, \tag{4.9}$$

waarbij de verwachting is genomen met betrekking tot de bij ψ_0 behorende invariante verdeling van het stationaire Markovproces met discrete tijdsparameter, dat ontstaat in de verzameling A_z bij voortdurende toepassing van het beslissingsvoorschrift z.

Uit (4.2) en (4.9) volgt voor zeer hoge waarden van n dat de beslisser de voorkeur geeft aan eigen beheer als geldt:

$$\frac{\mathscr{E}\{k(\underline{I}; z(\underline{I})) \mid \psi_0\}}{\mathscr{E}\{t(\underline{I}; z(\underline{I})) \mid \psi_0\}} \leqq \lambda \tag{4.10}$$

Door in het gedachtenexperiment voor de situatie b) verschillende waarden van λ te beschouwen kunnen wij de strategieën van Z_0 ordenen. Indien een optimale strategie $z_0 \epsilon Z_0$ bestaat, dan bereikt het linkerlid van (4.10) voor $z=z_0$ het minimum. Het optimaliteitscriterium wordt derhalve gegeven door:

$$C(z; \psi_0) = \frac{\mathscr{E}\{k(\underline{I}; z(\underline{I})) \mid \psi_0\}}{\mathscr{E}\{t(\underline{I}; z(\underline{I})) \mid \psi_0\}} \tag{4.11}$$

Men kan $k(\psi,d)$ en $t(\psi;d)$ als volgt interpreteren: $k(\psi,d)$ is het verlies, dat men zal lijden als in ψ een beslissing d wordt genomen, terwijl $t(\psi, d)$ de lengte van de periode is waarin dit verlies wordt geleden.

In werkelijkheid hebben deze functies een andere betekenis. Belangrijk is evenwel, dat beide functies bij gegeven ψ en d onafhankelijk zijn van de gevolgde strategie z.

Om eenvoudig te kunnen inzien hoe men met behulp van (4.11) de optimale strategie z_0 kan bepalen maken wij wederom een gedachten experiment. Stel, dat de beslisser een kostenvergoeding krijgt gedurende de periode, dat hij het systeem beheert.

Stel vervolgens, dat deze kostenvergoeding is gebaseerd op het toepassen van de strategie z vanuit de begintoestand ψ_1 en dat zij derhalve gelijk is aan $C(z;\psi_1)$ per tijdseenheid. Wij beschouwen nu de situatie, waarin de beslisser n maal zal beslissen volgens strategie z en daarna de stochastische wandeling zal beëindigen zodra het systeem een toestand aanneemt uit de verzameling A_0.

Wij weten uit het voorafgaande, dat, als de begintoestand van de wandeling ψ_2 is, de verwachting van de te maken kosten en die van de tijdsduur gegeven worden door $k_n(\psi_2; z)$ resp. $t_n(\psi_2; z)$.

De verwachting van het bedrag, dat de beslisser uit eigen zak moet bijbetalen, wordt dan gegeven door:

$$g_n(z;\psi_2;\psi_1) = k_n(\psi_2;z) - C(z;\psi_1)\, t_n(\psi_2;z) \tag{4.12}$$

Indien het aantal beslissingen n$\to\infty$ dan vinden wij voor de verwachting van het uit eigen zak te betalen bedrag:

$$g(z;\psi_2;\psi_1) = \lim_{n\to\infty} \{k_n(\psi_2;z) - C(z;\psi_1)\, t_n(\psi_2;z)\} \tag{4.13}$$

waarbij stilzwijgend wordt aangenomen dat kringfuiken afwezig zijn.

Stel, dat de beslisser in de begintoestand ook mag weigeren het systeem te beheren. Hij is echter dan wel verplicht de kosten te dragen zolang het systeem geen toestand aanneemt uit de verzameling A_0. De verwachting van het bedrag dat dan uit eigen zak moet worden bijbetaald, wordt gegeven door:

$$k_0(\psi_2;z) - C(z;\psi_1)\, t_0(\psi_2;z) \tag{4.14}$$

De verwachting van het bedrag ,dat de beslisser vrijwillig voor zijn rekening neemt, is dus gelijk aan:[6])

$$\hat{g}(z;\psi_2;\psi_1) = g(z;\psi_2;\psi_1) - [k_0(\psi_2;z) - C(z;\psi_1)\, t_0(\psi_2;z)] \tag{4.15}$$

uit de relaties (4.13) en (4.14) volgt:

$$\hat{g}(z;\psi_2;\psi_1) = k(\psi_1;z(\psi_1)) - C(z;\psi_1)\, t(\psi_1;z(\psi_1)) + \\ + \mathscr{E}\{\hat{g}(z;\underline{I};\psi_1) \mid \underline{I}\in A_z;\psi_1;z(\psi_1)\} \tag{4.16}$$

Uiteraard zal de beslisser trachten zijn aandeel in de kosten te minimaliseren.

Indien echter de kostenvergoeding gebaseerd is op de optimale strategie z_0 dan kan voor iedere $\psi\in\Psi$ worden aangetoond:

$$\hat{g}(z_0;\psi_1;\psi_1) = \min_{d_1} [k(\psi_1;d_1) - C(z_0;\psi_1)\, t(\psi_1;d_1) + \\ + \mathscr{E}\{\hat{g}(z_0;\underline{I};\psi_1) \mid \underline{I}\in A_{z_0};\psi_1;d_1\}] \tag{4.17}$$

Uit (4.17) volgt dus, dat de beslisser geen positieve winstverwachting heeft, wanneer hij in ψ_1 anders beslist dan volgens z_0.

Wij zullen nu de kans, dat het systeem in het natuurlijke proces vanuit de toestand ψ in een periode van de lengte t

a) een toestand aanneemt uit A_0 en wel de eerste keer een toestand uit $A_0.B$ of

b) geen toestand aanneemt uit A_0 maar aan het eind van die periode in een toestand van B is,

aanduiden met $\tilde{p}^t(B;\psi)$.

---

[6]) Als dit bedrag negatief is, verdient de beslisser aan het beheer.

Het is plausibel, dat als de kostenvergoeding gebaseerd is op de optimale strategie z_0, het voor de beslisser ook nadelig is om in toestanden van A_{z_0}, buiten A_0, ontaarde beslissingen te nemen. Indien ψ_1 een dergelijke toestand is, dan kan bovenstaande bewering ook als volgt worden uitgedrukt:

$$\int_{A_{z_0}} \tilde{p}^t(dI; \psi_1)\,[\hat{g}(z_0; I; \psi_1) - \hat{g}(z_0; \psi_1; \psi_1)] \geqq 0. \qquad (4.18)$$

voor $t \geqq 0$.

In plaats van met behulp van (4.11) alleen, kan men de optimale strategie z_0 dus ook bepalen met (4.11), (4.17) en (4.18) tezamen.
Men gaat dan als volgt te werk:

1) kies een willekeurig beslissingsvoorschrift $z^{(1)} \epsilon Z_0$
2) bepaal met behulp van (4.11) de ψ-functie $C(z^{(1)}; \psi)$
3) los de functionaalvergelijking (4.16) op
4) minimaliseer voor iedere $\psi \notin A_{z^{(1)}}$ met betrekking tot toegelaten d en voor iedere $\psi \epsilon A_{z^{(1)}}$ met betrekking tot toegelaten en niet-ontaarde d, de uitdrukking:

$$k(\psi; d) - C(z^{(1)}; \psi)\, t(\psi; d) + \mathscr{E}\{\hat{g}(z^{(1)}; \underline{I}; \psi) \mid \underline{I} \epsilon A_{z^{(1)}}; \psi; d\} \qquad (4.19)$$

5) Uit 4) vinden wij voor iedere $\psi \epsilon \Psi$ een beslissing d en dus een beslissingsvoorschrift.
   Dit beslissingsvoorschrift zullen wij aangeven met $\hat{z}^{(1)}$
6) bepaal met behulp van (4.11) de ψ-functie $C(\hat{z}^{(1)}; \psi)$
7) los de functionaalvergelijking (4.16) op voor de strategie $\hat{z}^{(1)}$
8) verwijder uit $A_{\hat{z}^{(1)}}$ de toestanden ψ, waarvoor, voor $t > 0$, geldt:

$$\int_{A_{\hat{z}^{(1)}}} \tilde{p}^t(dI; \psi)\,[\hat{g}(z_0; I; \psi) - \hat{g}(z_0; \psi; \psi)] < 0 \qquad (4.20)$$

9) Door de overige toestanden in $A_{\hat{z}^{(1)}}$ het beslissingsvoorschrift te handhaven, verkrijgt men een nieuw beslissings voorschrift $z^{(2)}$.
10) herhaal de gehele procedure, maar nu met $z^{(2)}$ in plaats van $z^{(1)}$.

Men kan bewijzen, dat onder zekere voorwaarden langs deze iteratieve weg de optimale strategie kan worden gevonden.

§ 5 Conclusie

De in § 4 behandelde methode kan worden beschouwd als een generalisatie van de eerste van de in § 3 besproken methoden.

Wij zullen tot slot nagaan hoe men het eerste probleem van § 3 oplost met

behulp van de methode ontwikkeld in § 4. Doordat in het eerste probleem van § 3 voor alle strategieën z de verzamelingen A_z identiek zijn, mogen wij A_0 gelijk kiezen aan A_z. Bijgevolg geldt voor iedere $\psi \epsilon A_z$:

$$t(\psi; d) = 1 \tag{5.1}$$

$$k(\psi; d) = h(\psi; d) \tag{5.2}$$

Uit (5.1) en (5.2) volgt dan, dat de criteria (3.7) en (4.11) identiek zijn.

Men kan tevens aantonen dat de iteratieprocedure gegeven in § 4 overgaat in die van R. A. H o w a r d, welke speciaal ontwikkeld is voor de problemen, besproken in § 3.

Literatuur:

[1] R. B e l l m a n, Dynamic Programming, Princeton 1957

[2] R. A. H o w a r d, Dynamic Programming and Markov processes, John Wiley, New York 1960

[3] G. d e L e v e, Decision Rules for adjusting Markovian Processes, Math. Centr. Rapport S 282a, 1961.

[4] G. d e L e v e, Stochastic ∞–stage Decisionproblems Math. Centr. Rapport S 302, (SP 81), 1962.