CWI Syllabi

CWI
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands
Telephone 31 -20 592 9333, telex 12571 (mactr nl),
telefax 31 -20 592 4199

CWI is the nationally funded Dutch institute for research in Mathematics and Computer Science.

CWI Syllabus 31

Vakantiecursus 1992
Systeemtheorie

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

1991 Mathematics Subject Classification: 93-01.
ISBN 90 6196 409 1
NUGI-code: 811

Errata bij "Inleiding gewone differentiaalvergelijkingen"

- Blz. 54, 3^+: "$t \to \infty$" wordt "$t \to +\infty$".
- Blz. 60, $11^+, 12^+$: Vervang "snelle eigenvector" door "langzame eigenvector" en "langzame eigenvector" door "snelle eigenvector".
- Blz. 66: Stelling 3.3 wordt:

  Stelling 3.3 *Beschouw het lineaire tweedimensionale stelsel (10). Geef de eigenwaarden van A aan met λ_1, λ_2. Het stelsel is*

  (i) asymptotisch stabiel dan en slechts dan als $\mathrm{Re}(\lambda_i) < 0$ $(i = 1, 2)$.

  (ii) stabiel dan en slechts dan als $\mathrm{Re}(\lambda_i) \leq 0$ $(i = 1, 2)$ *én als* $\lambda_1 = \lambda_2 = 0$, *dan* $A = 0$.

- Blz. 68, 12^+: vervang "annemelijk" door "aannemelijk".
- Blz. 80: Stelling 4.1 wordt:

  Stelling 4.1 *Beschouw het lineaire stelsel (85) en laat N, M, n_i, λ_i $(i = 1, \cdots, N)$, m_i, μ_i $(i = 1, \cdots, M)$ gedefinieerd zijn als hierboven. Laat $\mu_i = \alpha_i + i\beta_i$ $(i = 1, \cdots, M)$.*

  (i) Dan zijn alle oplossingen van (85) van de vorm

$$x(t) = \sum_{i=1}^{N} \sum_{j=0}^{n_i - 1} B_{ij} t^j e^{\lambda_i t} + \sum_{i=1}^{M} \sum_{j=0}^{m_i - 1} e^{\alpha_i t} [R_{ij} \cos \beta_i t + S_{ij} \sin \beta_i t]$$

  waar $B_{ij}, R_{ij}, S_{ij} \in \mathbb{R}^n$ voldoen aan

$$AB_{in_i - 1} = \lambda_i B_{in_i - 1}$$

$$AB_{ij} = \lambda_i B_{ij} + (j+1) B_{ij+1}$$

voor $j = n_i - 2, \cdots, 0$ *en*

$$\begin{aligned} AR_{in_i-1} &= \alpha_i R_{in_i-1} - \beta_i S_{in_i-1} \\ AS_{in_i-1} &= \beta_i R_{in_i-1} + \alpha_i S_{in_i-1} \\ AR_{ij} &= \alpha_i R_{ij} - \beta_i S_{ij} + (j+1) R_{ij+1} \\ AS_{ij} &= \beta_i R_{ij} + \alpha_i S_{ij} + (j+1) S_{ij+1} \end{aligned}$$

voor $j = m_i - 2, \cdots, 0$.

(ii) Er bestaan n lineair onafhankelijke oplossingen van (85) van de vorm (95).

- Blz. 82: vergelijking (102) wordt:

$$x(t) = e^{At} x_0$$

- Blz. 84: de laatste regel wordt:

$n_i = g_i$ respectievelijk $m_i = g_i$

- Blz. 86, 2^+: de matrix A wordt:

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(0) & \cdots & \frac{\partial f_1}{\partial x_n}(0) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(0) & \cdots & \frac{\partial f_n}{\partial x_n}(0) \end{pmatrix}$$

Inhoud

TEN GELEIDE

Het onderwerp voor de vacantiecursus 1992 is nu eens gekozen uit het rijk geschakeerde gebied van de toepassingen van de wiskunde: de systeemtheorie. Deze theorie kan — zeer ruw gezegd — omschreven worden als de bestudering van de interactie tussen een welomschreven gedeelte van de ons omringende wereld en de omgeving daarvan.
Hoewel dit onderwerp duidelijk valt buiten de stof die bij het VWO centraal staat, leek het de organisatoren van deze vacantiecursus toch een goede gedachte aandacht te vragen voor juist deze toepassing van de wiskunde vanwege het grote belang daarvan voor het dagelijkse leven; maar ook het toenemende besef dat de wiskunde zich steeds meer manifesteert in onze samenleving sprak een woord mee.

Deze keuze heeft een duidelijke consequentie voor de structuur van de cursus. Na een algemene inleiding zullen de voornaamste mathematische hulpmiddelen bijeengezet worden die dienen als voorbereiding op en voor een goed begrip van het eigenlijke onderwerp dat in een viertal voordrachten uiteengezet zal worden, waarvan de laatste verdere perspectieven zal schetsen.

Deze opbouw kan op de toehoorders tweeërlei effect hebben. In de eerste plaats kan men gestimuleerd worden door deze cursus om zich verder te verdiepen in de systeemtheorie, maar het is ook denkbaar dat men door de inleidende voordrachten over lineaire algebra, differentiaalvergelijkingen en stochastiek gemotiveerd raakt deze onderwerpen weer eens op te halen en verder uit te diepen.

In beide gevallen zouden de organisatoren van mening zijn dat de cursus zijn doel bereikt heeft: aanzet tot het zelfstandig beoefenen van een van de facetten van de altijd boeiende "mathematikè technè".

Evenals in de vorige jaren mag een ten geleide niet afgesloten worden zonder een woord van zeer hartelijke dank aan de medewerksters en medewerkers van het CWI, die met zoveel inzet en zorgvuldigheid deze zo fraai uitgevoerde syllabus produceerden en ervoor zorgden dat deze ook ruim op tijd beschikbaar was.

A.W. Grootendorst

Introduction to mathematical system theory

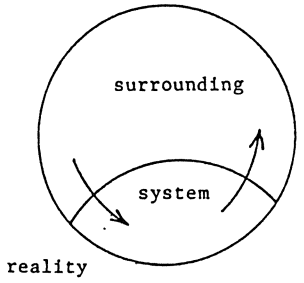
G.J. Olsder
Delft University of Technology

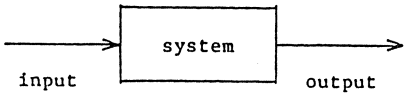
Abstract

In this introduction we will talk about systems. By means of examples, various concepts will be introduced which then will be formalized mathematically.

1 What is mathematical system theory?

A system is part of reality which we think to be a separated unit within this reality. The reality outside the system is called the surrounding. The interaction between system and surrounding is realized via quantities, quite often functions of time, which are called input and output. The system is influenced by the surrounding via the input(-functions) and the system has an influence on the surrounding by means of the output(-functions).





Three examples:

(1.1) How to fly an aeroplane; the position of the control wheel (the input) has an influence on the course (the output).

(1.2) In the economy: the interest rate (the input) has an influence on the investmentbehaviour (the output).

(1.3) Rainfall (the input) has an influence on the height of the water in a river (the output).

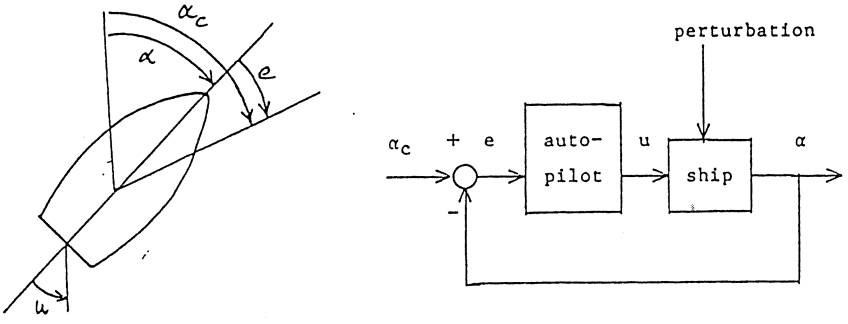
In many fields of study, a phenomenon is not studied directly but indirectly through a model of the phenomenon. A model is a representation, often in mathematical terms, of what are felt to be the important features of the object or system under study. By the manipulation of the representation, it is hoped that new knowledge about the modelled phenomenon can be obtained without the danger, cost, or inconvenience of manipulating the real phenomenon itself. In mathematical system theory we only work with models and when talking about a system we mean a modelled version of the system as part of reality.
Most modeling uses mathematics. The important features of many physical phenomena can be described numerically and the relations between these features described by equations or inequalities. Particulary in the natural sciences and engineering, properties such as mass, acceleration and forces are describable by mathematical equations. To successfully utilize the modeling approach, however, requires a knowledge of both the modeled phenomena and properties of the modeling technique. The development of high-speed computers has greatly increased the use and usefulness of modeling. By representing a system as a mathematical model, converting that model into instructions for a computer, and running the computer, it is possible to model larger and more complex systems than ever before.

Mathematical system theory is concerned with the study and control of input/output phenomena. The emphasis is on the dynamic behaviour of these phenomona, i.e. how do characteristic features (such as input and output) change in time and what are the relationships, also as functions of time. Mahematical system theory found its feet around 1950, the (classic) control theory played a simulating role. Initially system theory was more or less a collection of concepts and techniques from the theory of differential equations, linear algebra, matrix theory, probability theory, statistics, and, to a lesser extent, complex function theory. Later on (around 1960) system theory got its own face; "own" results were obtained which were especially related to the "structure" of the "box" between input and output. It is important to realize that nowadays the applications of known mathematical techniques is not the most important issue in mathematical system theory; it is rather the development of own original mathematical concepts and algorithms.

Mathematical system theory forms the mathematical base for technical areas such as automatic control and networks. It is also the starting point for other mathematical subjects such as optimal control theory (here one tries to find an input function which yields an output function that must satisfy a certain requirement as well as possible) and filter theory (the interpretation of the input function is here measurements with measurement errors, the system tries to realize an input which equals the "ideal" measurements, that is, without measurement errors). Mathematical system theory also plays a role in economics (specially in macro-economic control theory and time series analysis), theoretical computer science (via automation theory, Petry-networks) and management science (models of firms and other organizations). At last mathematical system theory forms the hard, mathematical, core of more philosophically oriented areas such as general systems theory and cybernetics.

Example of a system with feedback

Autopilot of a boat. An autopilot is a device which receives as input the present heading α of a boat (measured by an instrument such as a magnetic compass or a gyrocompass) and the heading α_c desired (reference point) by the navigator. Using this information, the device automatically outputs, as a funcion of time, the positioning command u of the rudder so as to achieve the smallest possible heading error $e = \alpha_c - \alpha$.



Given the dynamics of the boat and the external perturbations (wind, swell, etc.) the theory of automatic control helps to determine a control input command $u = f(e)$ that meets the imposed technical specifications (stability, accuracy, response time, etc.). For example, this control might be bang-bang:

$$u = \begin{cases} +u_{\max} & \text{if} \quad e > 0, \\ -u_{\max} & \text{if} \quad e < 0. \end{cases}$$

It might be proportional:

$$u = K.e.$$

It might be proportional, integrating and differentiating (PID-control):

$$u = K.e + K' \int^{t} e(s)ds + K'' \frac{de}{dt}$$

Automatic control theory aids in the choice of the best control law. If the ship itself is considered as a system, then the input to the ship is the rudder setting u (and possibly perturbations) and the output is the course α. The autopilot is another system; its input is the error signal e and output is the desired rudder setting u. Thus we see that the output of one system can be the input of another system. The combination of ship, autopilot and the connection from α to α_c (see the figure) can also be considered as a system; the input is the desired course α_c and the output is the real course α.

Example of an optimal control problem

The notion of a ship is described by

$$\dot{x}(t) = f(x, u, t),$$

where the state $x = (x_1, x_2) \in I\!R^2$ represents the ship's position with respect to a fixed coordinate system. The vector $u = (u_1, u_2) \in I\!R^2$ represents the control and t is the time. De notation $\dot{x}$ refers to the time derivatives of the two state components. One control variable to be chosen is the ship's heading u_1; the other one, u_2, is the ship's velocity. The problem now is to chosse u_1 and u_2 in such a way that the ship uses as little fuel as possible such that, if it leaves Rotterdam at a certain time, it should have reached New York not more than 10 days later. The function u_1 and u_2 may depend on available information such as time, wether forecast, ocean streams, et cetera. Formally, $u = (u_1, u_2)$ must be chosen such that

$$\int_{t_0}^{t_f} g(x, u, t)dt$$

is minimized. This criterion describes the fuel used. The function g is the amount of fuel used per time unit; t_0 is the departure time and t_f is the arrival time.

Example of a filter problem

NAVSAT is the acronym for NAVigation by means of SATellites. It refers to a worldwide navigation system being studed by the European Space Agency ESA. The NAVSAT system is still in the development phase with feasibility studies currently being performed by several European aerospace research institutes. At the National Aerospace Laboratory NLR, Amsterdam, the Netherlands, for instance, a simulation tool has been developed with which various alternative NAVSAT concepts and scenarios can be evaluated.
The central idea of satellite based navigation systems is the following. A user (such as an airplane or a ship) receives messages from satellites, from which he can estimate his own position. Such a satellite broadcasts its own coordinates (in some known reference frame) and the time instant at which this message is broadcast. The user measures the time instant at which he receives this message on his own clock. Thus he knows the time difference between sending and receiving the message which yields the distance between the position of the satellite and the user. If the user can calculate these distances with respect to at least three different satellites, he can in principle calculate his own position. Complicating factors in these calculations are: i) different satellites send messages at different time instants while the user moves in the meantime, ii) there are several different sources of error present in the data, e.g. unknown ionospheric and tropospheric delays, the clocks of the satellites and of the user not running exactly synchronously, the satellite position being broadcast with only limited accuracy.
The problem to be solved by the user is to calculate his position as accurately as possible if he gets the information of the satellites and if he knows the stochastic characteristics of the errors or uncertainties mentioned above. As the satellites broadcast the information periodically, the user can update the estimate of his position, which is a function of time, periodically.

2 A short history

Feedback -the key concept of system theory- is found in many places such as nature and living organisms. An example is the control of the body temperature. Also, social and economic processes are controlled by feedback mechanisms. In most technical equipment use is made of control mechanisms.
In the old times feedback was already applied in for instance, the Babylonic waterwheels and for the control of waterheights in Roman aquaducts. Historian Otto Mayr describes the first explicit use of a feedback mechanism as having been designed by Cornelis Drebbel [1572-1633], both an engineer and alchemist. He designed the "Athanor", an oven with which he optimistically hoped to change lead into gold. The control of the temperature in this oven was rather complex and could be viewed as a feedback design.
Drebbel's invention was then used for commercial purposes by his son in law, Augustus Kuffler [1595-1677]. Kuffler was a temporary of Christian Huygens [1629-1695], who himself designed a fly-wheel for the control of the rotational speed of windmills. This idea was refined by R. Hooke [1635-1703] and J. Watt [1736-1819], the latter being the inventor of the steam-engine. In the middle of the 19th century more than 75.000 flyball governors of James Watt were in use. Soon it was realized that these contraptions gave problems if the control was too rigid. Nowadays one realizes that that behaviour was a form of instability due to a high gain in the feedback loop. This problem of bad behaviour was given to J.C. Maxwell [1831-1879] -the Maxwell of the electromagnetism- who was the first to study the mathematical analysis of stability problems. His paper "On Governors" can be viewed as the first mathematical article devoted to control theory.
The next important development started in the period before the second world war in the Bell Labs in the USA. The invention of the electronic amplification by means of feedback started the design and use of feedback controllers in communication-devices. In the theoretical area frequency-domain techniques were developed for the analysis of stability and sensitivity. H. Nyquist [1889-1976] and H.W. Bode [1905-1982] are the most important representants of this direction.
Norbert Wiener [1894-1964] worked on the fire-control of anti-aircraft defence during the second world war. He also advocated control theory as some kind of artificial intelligence as an independent discipline which he called "Cybernetics" (this word was already used by A.M. Ampere [1775-1836]).
System theory and automatic control, as known nowadays, found their feet in the years sixty of the current century. Two developments contributed to that. First there were fundamental theoretical developments in the fifties. Names attached to these developments are R. Bellman (dynamic programming), L.S. Pontryagin (optimal control) and R.E. Kalman (state models and recursive filtering). Secondly there was the invention of the chip at the end of the sixties and the subsequent development of the micro-electronics. This has led to cheap and fast computers by means of which control algorithms with a high degree of complexity can be really used.

3 Mathematical descriptions of dynamic systems

The input quantities at time t will be denoted by $u(t)$ and the output quantities by $y(t)$. For the input function, resp. output function, as functions of time we write $u(.)$ and $y(.)$. If no misunderstanding is possible these functions are sometimes simply written as u and y. The time will either be continuous ($t \in T$ with $T = (-\infty, +\infty)$ or $T = [t_0, \infty)$) or be discrete ($t \in T$ with $T = \mathbb{Z}$ or $T = \{t_1, t_2, ..., t_n, ...\}$). If $T = \mathbb{R}$ we talk about time-continuous systems, if $T = \mathbb{Z}$ we talk about time-discrete systems.

Two ways exist in order to describe the dynamic behaviour of systems; an external and an internal description. The external description considers the system as an input/output map, i.e. $y(t) = f(u(.), t)$. If a system is described by means of the internal or state space form description, another quantity, the state $x(t)$, is introduced. Later on in this section we will see the usefulness of this concept.

Definition 3.1 of the external description.

A system in input/output form is defined as

$$\Sigma_{I/O} = \{T, U, \underline{U}, Y, \underline{Y}, F\},$$

whereby

i) T is the time axis (i.e. $T = \mathbb{R}$ or $\mathbb{Z}$ or a subset of $\mathbb{R}$ or $\mathbb{Z}$)

ii) U is the set of input values; this set is called the input space. Quite often $U = \mathbb{R}^m$, or U is a subset of $\mathbb{R}^m$.

iii) $\underline{U}$ is a set of functions from $T \rightarrow U$; $\underline{U}$ is the set of admissible input functions; $\underline{U} \subset \{f | f : T \rightarrow U\}$.

iv) Y is the set of output values. Usually $Y = \mathbb{R}^p$; Y is called the output space.

v) $\underline{Y}$ is the set of functions from $T \rightarrow Y$.

vi) F is a mapping from $\underline{U}$ to $\underline{Y}$: $F : \underline{U} \rightarrow \underline{Y}$
F defines the relation between input- and output functions. If $u \in \underline{U}$, then Fu is the resulting output function. Its value at time t is denoted by $(Fu)(t)$. The mapping F is called the input/output caon or the system function. It is assumed that F is i.e. if $u_1, u_2 \in \underline{U}$ and $u_1(t) = u_2(t)$ for $t \leq t'$ with $t' \in T$, then $(Fu_1)(t') = (Fu_2)(t')$ and therefore also $(Fu_1(t) = (Fu_2)(t)$ for all $t \leq t'$.

Definition 3.2

The system $\Sigma_{I/O}$ is called linear if $U, Y, \underline{U}$ and $\underline{Y}$ are linear vectorspaces (for example $U = \mathbb{R}^m$, $Y = \mathbb{R}^p$) and if $F : \underline{U} \rightarrow \underline{Y}$ is a linear mapping. The latter requirement means that if $u_1, u_2 \in \underline{U}$, -then also $u_1 + u_2 \in \underline{U}$ and $\lambda u_1 \in \underline{U}$ whereby λ is an arbitrary scalar- then $F(u_1 + u_2) = Fu_1 + Fu_2$ and $F(\lambda u_1) = \lambda F u_1$.

Definition 3.3

The system $\Sigma_{I/O}$ is called time-invariant (or, equivalently, stationary) if

i) T is closed with respect to addition, i.e. if $t_1, t_2 \in T$ then also $t_1 + t_2 \in T$,

ii) $\underline{U}$ and $\underline{Y}$ are invariant with respect to the shift operator S_τ defined by $(S_\tau u)(t) = u(t+\tau)$, $(S_\tau y)(t) = y(t+\tau)$, i.e. $S_\tau \underline{U} \subset \underline{U}$ and $S_\tau \underline{Y} \subset \underline{Y}$ for all $\tau \in T$.

iii) $S_\tau F = F S_\tau$ for all $\tau \in T$.

To say it in a simple way: a system is time-invariant if a shift in the time axis yields an equivalent system. If $t \to u(t)$ leads to an output $t \to y(t)$; then $t \to u(t-\tau)$ should result in $t \to y(t-\tau)$. If a signal is applied one hour later, we get the same response, expect for a delay of one hour.

Definition 3.4

The sysem $\Sigma_{I/O}$ is called memoryless or static if s function f exists, $f : U \times T \to Y$ such that $(Fu)(t) = f(u(t), t)$. This means that Fu at time t only depends on $u(t)$ and not on the past (or future) of u.

Example Population dynamics

We want to express the population N (the output) as a function of the number of births per time unit or, equivalently, the birth rate (the input B). If $P(x,t)$ is the probability that somebody, born at time $t-x$, is still alive at time t (at which time he/she has an age of x), then

$$N(t) = \int_{-\infty}^{t} P(t-s,t)B(s)ds \tag{1}$$

If this integral is well defined (depending on the functions P and B), then (1) describes a system in input/output form. A reasonable assumption is that a quantity L exists such that $P(x,t) = 0$ for $x > L$. Then

$$N(t) = \int_{t-L}^{t} P(t-s,t)B(s)ds$$

If $P(.,.)$ is continuous in its arguments and if $B(.)$ is piecewise continuous (i.e. on each finite interval $B(.)$ it has at most a finite number of discontinuities and at points of discontinuity, left and right limit of $B(.)$ must exist), then this integral exists. Returning to (1) and assuming that a function $g(.)$ exists such that $P(t-s,t) = g(t-s)$; we can write (1) as

$$y(t) = \int_{-\infty}^{t} g(t-s)u(s)ds \tag{2}$$

If this integral exists for all $u \in \underline{U}$ then it can be interpreted as a time-invariant, strictly causal input/output system. For such a system the probability that somebody is still alive at age x is determined by x only and not by the date of birth.

Example

An important class of input/output systems is of the form

$$y(t) = \int_{-\infty}^{t} g(t-s)u(s)ds$$

For reason of simplicity we assume that $y(t) \in \mathbb{R}$, $u(t) \in \mathbb{R}$, though the extention to the vector case is straightforward (then g becomes a matrix). If for instance

$$g(t) = \begin{cases} e^{-t} & \text{for} \quad t > 0 \\ 0 & \text{for} \quad t \leq 0 \end{cases}$$

then we can write

$$y(t) = \int_{-\infty}^{+\infty} g(t-s)u(s)ds = \int_{-\infty}^{+\infty} g(\tau)u(t-\tau)d\tau.$$

A reasonable class of input functions for which these latter integrals exist is the class of piecewise continuous functions, which are zero for $t \leq t_0$ for some t_0. This class is indicated by PC_+. In the later sections we will tacitly assume that $\underline{U} = PC_+$ unless $\underline{U}$ is explicitly defined differently.
Another interpretation (other than the population dynamics) of this example is that $y(t)$ denotes the waterheight in a lake and $u(t)$ is the water input (rivers, rain) per unit of time. The function g symbolizes what remains after evaporation. Intuitively it is clear that if we know the waterheight at time $t'(t' < t)$, then we only need $u(s)$, $t' \leq s \leq t$, in order to calculate the waterheight at time t. Thus we do not need the whole past of $u(.)$ as in (2). We now, however, introduced another quantity, $y(t')$, in order to specify the input/output behaviour. If $y(t')$ is known and also $u(t)$ for $t \geq t'$, then the height of the water is completely determined for $t \geq t'$.

In the example above a quantity x (there denoted by y) was introduced with the property that if $x(t_1)$ is known and if also $u(t)$, $t \geq t_1$ is known, then the behaviour of the system is completely specified for $t \geq t_1$. In the example above we had $x = y$, but this is not true in general. We call $x(t)$ the state of the system at time t. It turns out that for many systems such a state can be found. The symbol y will be reserved for the output (therefore not necessarily $x = y$ in general). The past of the system (for $t < t_1$) is "projected" on (or is summarized by) the state at time t_1; $x(t_1)$ symbolizes all knowledge of the past which we need in order to describe the future behaviour of the system.

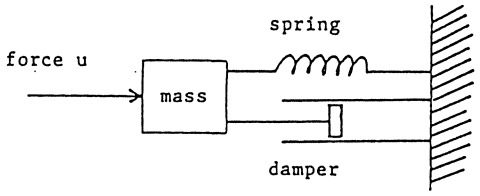
Example

A mass m moves along a straight line and is connected with a spring with characteristic

constant k. There is friction which is a function of the speed of the mass. An external force $u(t)$ acts on the mass. Classical mechanics tells us that if we want to describe the motion of the mass from a time instant t_1 onwards while the force $u(t)$, $t \geq t_1$ is being exerted, that the position and velocity of the mass at time t_1 should be known. The state of this system therefore is the vector

$$x(t) = \begin{pmatrix} q(t) \\ v(t) \end{pmatrix},$$

where q denotes the position and v the velocity.



Example

Two persons play the game of goose. As time variable we denote the number of times n that both persons have thrown the die (n is increased by 1 after both persons had a turn). This is a time discrete system. As input at time n, $u(n)$, we define

$$u(n) = \begin{pmatrix} \text{number of spots on the die at } n\text{-th throw,} & \text{first person} \\ \text{"} \quad \text{"} \quad \text{"} \quad \text{"}, & \text{second person} \end{pmatrix}$$

The state can be defined as

$$x(n) = \begin{pmatrix} \text{position of first person's marker on the board} \\ \text{" second " " " " " } \end{pmatrix}$$

For simplicity we have assumed that the rule "pass your turn" does not exist. If this rule would be allowed, what could then be defined as the state of the system?

Definition 3.5 of the internal description of a system (or, equivalently, of a system in state space form).

A system in state space form is defined as

$$\Sigma_M = \{T, U, \underline{U}, Y, \underline{Y}, X, \phi, r\},$$

where:

i) $T, U, \underline{U}, Y$ and $\underline{Y}$ are the same as in definition of the external description

ii) X is the state space; $x(t) \in X$. Quite often $X = \mathbb{R}^n$ or X is a subset of $\mathbb{R}^n$.

iii) $\phi : T_+^2 \times X \times \underline{U} \rightarrow X$
whereby $T_+^2 = \{(t_1, t_0) \in T^2 \text{ with } t_1 \geq t_0\}$. The mapping ϕ is called the state evolution function. The quantity $\phi(t_1, t_0, x_0, u)$ denotes the state at time t_1, which was obtained by applying the input $u \in \underline{U}$ and starting from the state x_0 at time t_0. The function ϕ must

a) be consistent, i.e. $\phi(t, t, x, u) = x$

b) satisfy the semi-group property:
$\phi(t_2, t_1, \phi(t_1, t_0, x_0, u), u) = \phi(t_2, t_0, x_0, u)$

c) be determinate; if
$u_1, u_2 \in \underline{U}$ and $u_1(t) = u_2(t)$, $t_0 \leq t \leq t_1$, then
$\phi(t_1, t_0, x_0, u_1) = \phi(t_1, t_0, x_0, u_2)$.

iv) $r : T \times X \times U \rightarrow Y$ is the output function (or measurement function or observation function) $y(t) = r(x(t), u(t), t)$. It is the value of the output at time t if the system is in state $x(t)$ and $u(t)$ is the input at time t. The function $r(., x(.), u(.))$ must belong to $\underline{Y}$.

Example

The starting point is -see before-

$$y(t) = \int_{-\infty}^{t} e^{-t+s} u(s) ds,$$

with y denoting the height of the water and u the water input in the lake. This internal description can be written as a differential equation:

$$\frac{dy(t)}{dt} = -y(t) + u(t), \text{ equivalently, } \dot{y}(t) = -y(t) + u(t).$$

If the initial value is: $y(t_0) = y_0$, then the solution to this differential equation can be written as

$$y(t) = e^{-(t-t_0)} y_0 + \int_{t_0}^{t} e^{-(t-s)} u(s) ds.$$

The quantity y can be interpreted as the state;

$$y(t) = r(t, x(t), u(t)) = x(t),$$

whereby

$$x(t) = \phi(t, t_0, x_0, u) = e^{-(t-t_0} x_0 + \int_{t_0)}^{t} e^{-(t-s)} u(s) ds, \quad (x_0 = y_0).$$

It is easily seen that r and ϕ satisfy all the conditions.

The choice of the state is not unique. In the above example is $100x$ also a state (if the height is measured in centimeters instead of meters). In the example with the mass sprong and friction, one can take (q,p) as state instead of (q,v), where p is the impulse of the mass. Even a linear combination $(a_{11}q + a_{12}v,\ a_{21}q + a_{22}v)$ can be chosen as the state provided that the matrix $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is nonsingular. From the physical point of view, however, there are usually only a few natural choices for the state. The vector (q,v,a), where a is the acceleration, is also a state. Usually we try to keep the state as small as possible in the sense that the dimension of X should be as small as possible. As a trivial state (but not the smallest one!) one could take the whole past of the input function. Then $x(t)$ is the function $u : (-\infty,t) \to U$; for each time t the state is a function! This state is very large and will not be of great use.

Definition 3.6

Σ_M is called linear if $U, Y, \underline{U}, \underline{Y}$ and X are linear vectorspaces and if

i) the mapping $\phi(t_1,t_0,.,.) : X \times \underline{U} \to X$ is jointly linear in both arguments (i.e. if $\phi(t_1,t_0,x_0,u) = x$ and $\phi(t_1,t_0,\tilde{x_0},\tilde{u}) = \tilde{x}$ then $\phi(t_1,t_0,\lambda x_0,\lambda u) = \lambda x$ and $\phi(t_1,t_0,x_0 + \tilde{x}_0, u + \tilde{u}) = x + \tilde{x}$).

ii) the mapping $r(t,.,.) : X \times U \to$ is jointly linear in both arguments.

Definition 3.7

Σ_M is called time invariant if $t_1, t_2 \in T$ then $t_1 + t_2 \in T$, $S_t\underline{U} \subset \underline{U}$, $S_t\underline{Y} \subset \underline{Y}$ for all $t \in T$ and if moreover

i) $\phi(t_1 + t,\ t_0 + t,\ x_0,\ u) = \phi(t_1,t_0,x_0,S_tu)$ for all $t \in T$

ii) $r(t,x,u)$ is independent of t and therefore written as $r(x,u)$.

Definition 3.8

Σ_M is called strictly causal if $r(x,u,t)$ does not depend on u.

Definition 3.9

Σ_M is called autonomous if U consists of only one element.
(Therefore no control is possible).

So far we talked about the external description and the state space form description of a system. Some words will be devoted now as to how one description can be derived from the other. Suppose $\Sigma_M = \{T, U, \underline{U}, Y, \underline{Y}, X, \phi, r\}$ is a description in state space form. In order to obtain the corresponding $\Sigma_{I/O}$ the essential idea is to eliminate x from the ϕ- and r-relations. Suppose for simplicity that Σ_M is time invariant. Choose a $t_0 \in T$ and a $x_0 \in X$ (think of initial time and initial state) and define

$$(Fu)(t) = r(\phi(t,t_0,x_0,u),u(t)) \quad \text{for} \quad t \geq t_0.$$

Thus we obtained a system

$$\Sigma_{I/O} = \{T \cap [t_0, \infty), U, \underline{U}, Y, \underline{Y}, F\}$$

The time axis can be extended to the whole T by defining

$$x(t) = x_0, \;\; u(t) = u_0, \;\; y(t) = y_0, \;\; t < t_0,$$

where u_0 and y_0 are constants in resp. U and Y. For every choice we get in principle another F. The state x_0 will usually be interpreted as an equilibrium for the system. A natural choice for x_0 is the zero-element of X. Similarly choices for u_0 and y_0 are the zero-elements of U resp. Y. If in addition t_0 is chosen close to $-\infty$ (if $T = (-\infty, +\infty)$) then we say that the system is in equilibrium or at rest at "$t = -\infty$". The reverse problem as how to obtain $\sum_M$ from $\sum_{I/O}$ is far more difficult. Now one has to create a space X instead of eliminate X. For linear systems this problem has been solved satisfactory. A whole theory has been built around the "creation" of the state space X and it is called realization theory.

4 Differential and difference systems

Differential systems are a subclass of continuous time systems, whereas difference systems are a subclass of discrete time systems. We start with the former. We assume that U, Y and X are linear vectorspaces, and more specifically, assume $U = \mathbb{R}^m$, $Y = \mathbb{R}^p$ and $X = \mathbb{R}^n$. Consider

$$\dot{x}(t) = f(x(t), u(t), t) \tag{3}$$
$$y(t) = g(x(t), u(t), t) \tag{4}$$

where $\dot{}$ denotes the derivative with respect to time. Relations (3) and (4) are vector relations. Componentswise they can be written as

$$\begin{aligned}
\dot{x}_1(t) &= f_1(x_1(t), ..., x_n(t), u_1(t), ..., u_m(t), t) \\
&\vdots \\
\dot{x}_n(t) &= f_n(x_1(t), ..., x_n(t), u_1(t), ..., u_m(t), t) \\
y_1(t) &= g_1(x_1(t), ..., x_n(t), u_1(t), ..., u_m(t), t) \\
&\vdots \\
y_p(t) &= g_p(x_1(t), ..., x_n(t), u_1(t), ..., u_m(t), t.
\end{aligned}$$

Suppose we are given a certain input function $\overline{u} \in \underline{U}$ and define the following (vector-) differential equation

$$\dot{x}(t) = f(x(t), \overline{u}(t), t) = \overline{f}(x(t), t) \tag{5}$$

The theory of ordinary differential equations gives conditions on $\overline{f}$ such that (5) has a unique solution on $[t_0, \infty)$ for an initial value $x(t_0) = x_0$(x_0 is also an n-vector). A sufficiency condition for the existence of a unique solution are the following points:

i) $\overline{f}$ is piecewise continuous in t.

ii) $\overline{f}$ satisfies the so-called Lipschitz-condition: there is a continuous time function $K(t)$ such that

$$\|\overline{f}(x_2, t) - \overline{f}(x_1, t)\| \leq K(t)\|x_2 - x_1\|$$

for all $t \in [t_0, \infty)$ and for all $x_1, x_2 \in I\!R^n$. The norm $\|.\|$ denotes the Euclidian norm. If a is a vector with components $a_1, ..., a_n$ then $\|a\| = \sqrt{\sum_{i=1}^{n} a_i^2}$.

The solution of (5), function $x(.)$, will be differentiable in t (the value of $\dot{x}(t)$ equals $\overline{f}(x(t), t)$) except in the points where $\overline{f}$ is discontinuous in t; at those discontinuity points the solution $x(t)$ will be continuous but not continuously differentiable. With respect to the conditions on $f(x, u, t)$, such that $\dot{x} = f(x, u, t)$ has a unique solution on $[t_0, +\infty)$ for each $u \in \underline{U}$ we assume that:

i) the elements of $\underline{U}$ are piecewise continuous functions: $I\!R \rightarrow I\!R^m$.

ii) f is continuous in x and u, piecewise continuous in t. For all $\overline{u} \in \underline{U}$ a function $K(t)$ exists such that for all $t \in [t, \infty)$:

$$\|f(x_1, \overline{u}, t) - f(x_2, \overline{u}, t)\| \leq K(t) \,.\, \|x_1 - x_2\|,$$

for all $x_1, x_2 \in I\!R$. The function K must be continuous with respect to t.

Thus a mapping $\phi_{\overline{u}}$ has been defined with

$$\phi_{\overline{u}} : T_+^2 \times X \rightarrow X,$$

where $T_+^2 = \{(t_1, t_0) \in T^2 | t_1 \geq t_0\}$ and $X = I\!R^n$, by the fact that $\phi_{\overline{u}}(t_1, t_0, x_0)$ is the solution of (5) at time t_1. If we write $\phi(t_1, t_0, x_0, \overline{u}) = \phi_{\overline{u}}(t_1, t_0, x_0)$, then it is easily verified that ϕ satisfies all conditions of a state evolution function. If we define the observation function $r : X \times U \times T \rightarrow Y$ as $g : X \times U \times T \rightarrow Y$, then we have obtained a system in state space form. For g and $\underline{Y}$ we require that

iii) g is continuous in x and u , and is piecewise continuous in t,

iv) $\underline{Y}$ is the set of piecewise continuous functions: $I\!R \rightarrow Y = I\!R^p$.

Unless explicitly stated, we will always assume that the above conditions i) - iv) are satisfied. The system (3), (4) is then called a differential system.

A differential system is a generalisation of a set of first order differential equations. If a system is described by means of a differential equation $\dot{x} = f(x, t)$ then the state x evolves in time

without an influence from outside. If we define the output function as $y = r(x,t)$ -a special case is $y = x$- then we obtain an autonomous system

$$\begin{aligned} \dot{x} &= f(x,t), \\ y &= g(x,t). \end{aligned}$$

If in the model $\dot{x} = f(x,u)$ we substitute for u a given input function $u_1(.)$, a time dependent (not time-invariant) differential equation is obtained:

$$\dot{x}(t) = f(x(t), u_1(t)) = \overline{f}(x(t), t).$$

Consider the following set of equations

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{6}$$
$$y(t) = C(t)x(t) + D(t)u(t) \tag{7}$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, A, B, C and D are matrices of sizes $n \times n$, $n \times m$, $p \times n$ and $p \times m$ respectively. The elements of these matrices are piecewise continuous functions of time. It can be verified that (5), (6) is a differential system (the proof will not be given here). From the theory of ordinary differential equations it is known that the solution $x(t)$ of (5) with initial condition $x(t_0) = x_0$ can be written as

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t,s)B(s)u(s)ds,$$

where $\Phi(t, t_0)$ is the transition matrix beloning to $\dot{x} = A(t)x$, the $n \times n$ matrix Φ satisfies

$$\frac{d}{dt}\Phi(t, t_0) = A(t)\Phi(t, t_0), \quad \Phi(t_0, t_0) = I,$$

where I is the identity-matrix. The state evolution function is given by

$$\phi(t_1, t_0, x_0, u) = \Phi(t_1, t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_1, s)B(s)u(s)ds$$

The output function r is here defined by

$$r(x, u, t) = C(t)x + D(t)u.$$

It is easily verified that ϕ is linear in x_0 and u and that r is linear in x and u. Therefore (6), (7) defines a linear system in state space form; it is called a linear differential system.
Consider next the equations (3) and (4) again and assume that the functions f and g do not explicitly depend on t, such that we can write (with an abuse of notation):

$$\begin{aligned} \dot{x} &= f(x,u) \\ y &= g(x,u) \end{aligned}$$

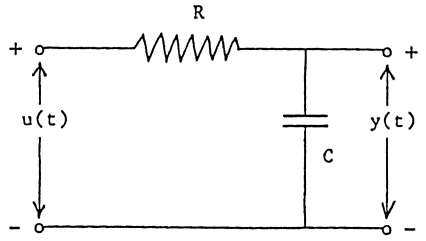
It can be verified that this differential system is time-invariant. If in the eqs. (6), (7) the matrices do not depend on time (they are constants), then we have a time-invariant linear differential system:

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned}$$

The matrices A, B, C and D are constants. If D equals the zero-matrix, then this system is strictly causal.

Example

Consider the resistor-capacitor network shown in the figure. An experiment is performed by applying a voltage $u(t)$, the input, and measuring a voltage $y(t)$, the output.



If Q is the electrical charge on the capacitor, then

$$y(t) = \frac{1}{C}Q(t)$$

and, by Kirchov's laws,

$$\frac{dQ(t)}{dt} = \frac{-1}{RC}Q(t) + \frac{1}{R}u(t)$$

By identifying Q with x, we have obtained a linear, time-invariant differential system. If R would be time dependent (for example R increases if it gets warmer), then the time-invariance is lost.

Example of a prey-predicator system

Suppose x_1 denotes the amount of preys (anchovy) and x_2 the amount of predators (salmon), u_1 is the fraction of anchovy caught by fishermen per unit of time and similarly, u_2 is the fraction of salmon caught per unit of time. The equations describing the evolution of x_1 and x_2 are, according to Volterra (1860-1940);

$$\begin{aligned} \dot{x}_1 &= ax_1 - bx_1x_2 - u_1x_1 \\ \dot{x}_2 &= cx_1x_2 - dx_2 - u_2x_2 \end{aligned}$$

where a, b, c and d are positive constants. The term ax_1 is due to birth, the term $-dx_2$ is due to (natural) death. The terms $-bx_1x_2$ and cx_1x_2 are due to the fact that salmon eat anchovy. As output function we can for instance take the amount of anchovy;

$$y + (1\ 0)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Thus a time-invariant, strictly causal system is defined which is not linear.

If instead of a differential equation we start with a difference equation

$$x(k+1) = f(x(k), u(k), k), \quad k \in \mathbb{Z}$$

and define the output function as

$$y(k) = g(x(k), u(k), k), \quad k \in \mathbb{Z},$$

a time discrete system in state space form has been defined. If we assume $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, $Y = \mathbb{R}^p$, then

$$\begin{aligned} f: &\quad \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{Z} \to \mathbb{R}^n \\ g: &\quad \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{Z} \to \mathbb{R}^p. \end{aligned}$$

In contrast to the differential equations describing the time continuous system in state space form, we now need, in the time discrete case, not impose any smoothness conditions on f and g. The system is linear if we can write

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned}$$

where A, B, C and D are matrices of appropiate sizes. If these matrices do not depend on k, the system is time-invariant.
If the function g does not explicitly depend on u, then the system is strictly causal.

Example of a national economy.

Let $y(k)$ be the total national income in year k,
$c(k)$ be the consumer expediture in year k,
$i(k)$ be the investments in year k,
$u(k)$ be the government expediture in year k.

We will make the following assumptions

i) $y(k) = c(k) + i(k) + u(k)$;

ii) the consumer expediture is a fixed fraction of the total income of the previous year;
$c(k) = my(k-1), \ 0 \leq m \leq 1$;

iii) the investment in year k depends on the increase in consumer expediture from year $k-1$ to year k;
$i(k) = \mu(c(k) - c(k-1))$,
where μ is a positive constant.

We want to write the evolution of the national economy in state-space form. We have

$$\begin{aligned} i(k+1) - \mu c(k+1) &= -\mu c(k) \\ c(k+1) = my &= m\{i(k) - \mu c(k) + (1+\mu)c(k) + u(k)\} \\ &= m\{i(k) - \mu c(k)\} + m(1+\mu)c(k) + mu(k) \end{aligned}$$

If a state vector is defined as $x(k) = \begin{pmatrix} x_1(k) \\ x_2(k) \end{pmatrix}$, where

$$\begin{aligned} x_1(k) &= i(k) - \mu c(k), \\ x_2(k) &= c(k), \end{aligned}$$

then the state evolution equation is

$$\begin{pmatrix} x_1(k+1) \\ x_2(k+1) \end{pmatrix} = \begin{pmatrix} 0 & -\mu \\ m & m(1+\mu) \end{pmatrix} \begin{pmatrix} x_1(k) \\ x_2(k) \end{pmatrix} + \begin{pmatrix} 0 \\ m \end{pmatrix} u(k)$$

and the output function is

$$y(k) = (1 \quad 1+\mu) \begin{pmatrix} x_1(k) \\ x_2(k) \end{pmatrix} + u(k)$$

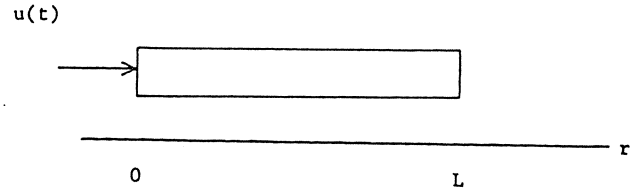
Thus a linear, time-invariant time discrete system has been defined.

5 Distributed and stochastic systems

In this section we will briefly talk about two classes of systems which are (also) important from a practical point of view, but which will not be discussed in these notes (apart from this section). In all examples so far the state space X was either finite dimensional ($\mathbb{R}^n$) or even finite. In the physical examples a finite dimensional state space could be constructed because physical quantities as mass, velocity, electric charge, temperature were thought to be concentrated in one point. For some problems such a simplification may lead to inadmissible conclusions and therefore the dependence of electric charge, temperature, etc., are not only time, but also location (spatial) dependent. Such quantities are then elements of a function space and the state space is infinite dimensional. Such systems are called distributed systems (this in contrast to systems with finite dimensional state spaces which are called lumped systems).

Example

Consider a metal bar of length L which is insulated from its environement, except at the left side where the bar is heated by a jet with heat transfer $u(t)$.



The temperature of the bar at position r, with $0 \leq r \leq L$, is denoted by $T(t,r)$; r is the spatial variable. In order to be able to determine the thermal behaviour of the bar one must know $T(t_0,r)$, $0 \leq r \leq 1$, the initial temperature distribution and $u(t)$, $t \geq t_0$. The state of the system is $T(t,.) : [0,L] \rightarrow I\!R$. From physics it is known that T satisfies a partial differential equation:

$$\frac{\partial T(t,r)}{\partial t} = c\frac{\partial^2 T(t,r)}{\partial r^2}, \tag{8}$$

where c is a carasteristic constant of the bar. At the left side we have

$$-A\frac{\partial T(t,r)}{\partial r}|_{r=0} = u(t) \tag{9}$$

where A is the surface of the cross section of the bar. At the right hand side of the bar we have

$$\frac{\partial T(t,r)}{\partial r}|_{r=L} = 0 \tag{10}$$

because of the insulation there. The evolution of the state is described by the partial differential equation (8), with boundary conditions (9) and (10). In this example the input enters the problem only via the boundary conditions. In other problems the input can also be distributed. Can you give an interpretation of the partial differential equation

$$\frac{\partial T(t,r)}{\partial t} = c\frac{\partial^2 T(t,r)}{\partial r^2} + u(t,r)?$$

The system which we have considered sofar are all deterministic. Once the initial condition and input function are known, the future behaviour is uniquely determined. There are many systems in practice in which the future is (partly) determined by processes of a stochastic, probabilistic nature. The winner of the game of goose is not determined at the outset of the game; the evolution of the game depends on the outcomes of the die, which usually are modelled in probabilistic way. In principle it may be possible to describe the throwing of a die in a deterministic way, but such a model would be extremely complicated and it is preferred to describe the outcome of a die probabilistically. If random influences determine the future of a system, it is called a stochastic system. A quantity x, within a stochastic system, could be interpreted as the state if, given $x(t)$ and $u(t), s \geq t$, all quantities within the system are determined in a probabilistic way. That is for instance that the probability distribution

functions are uniquely determined by $x(t)$ and $u(s)$, $s \geq t$. The future behaviour is then characterized by probabilistic laws, but the actual outcome of the system (who will win the game of goose) is not known before the evolution has really taken place.

Example

An industrial area can be in two situations: the atmosphere is good (G) or the atmosphere is bad (B). In both situtions two possible actions exist: start the alarm phase ($u = 1$) or not ($u = 0$). Depending on the atmospheric condition and the action, the atmosphere of the next day will be good or bad according to the following probabilistic rule;

initial condition \ condition tomorrow	G	B
G	0.8	0.2
B	0.4	0.6

$u = 0$

initial condition \ condition tomorrow	G	B
G	0.9	0.1
B	0.6	0.4

$u = 1$

The numbers in these tabular forms denote trasformation probabilities. If it is assumed that the transition probabilities are independent (i.e. there is no correlation with respect to time), then the state of this stochastic system is the atmospheric situation; $X = \{G, B\}$

6 Why linear system and how they arise

In this section we will be concerned with linear differential systems

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{11}$$
$$y(t) = C(t)x(t) + D(t)u(t) \tag{12}$$

The dot in (11) denotes the derivative with respect to time. In both the continuous and discrete time case we assume that $x \in X = \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in Y = \mathbb{R}^p$. The matrices A, B, C and D have sizes $n \times n, n \times m, p \times n$ and $p \times m$ respectively.. These matrices will be piecewise continuous in time. The classes of input and output functions $\underline{U}$ and $\underline{Y}$ are also assumed to be piecewise continuous.

There are two reasons for the importance of linear systems. The first one is that they are analytically attractable. These systems can be analyzed much better than nonlinear systems. This is particularly true if the matrices in (11), (12) are constant with respect to time. In this case the solution, expressed in an initial conditon and the input function, can be written down explicitly as we will see later on. The second reason is that many systems are "almost" linear or can, at least, be approximated by linear systems. Even non-linear systems may locally be linearized i.e. in the neighbourhood of a solution small perturbations will behave as a linear systems. We will now make the idea of linearization more precise.

Consider a non-linear differential equation

$$\dot{x} = f(x,u),\ x \in \mathbb{R}^n,\ u \in \mathbb{R}^m \tag{13}$$

(The restriction to a time-invariant system is not essential). Given a solution $\tilde{x}(t)$ of (13) with initial condition $\tilde{x}(0) = \tilde{x}_0$ and input function $\tilde{u}(t)$, consider another solution $\tilde{x}(t) + z(t)$, with initial condition $\tilde{x}_0 + z_0$ and input function $\tilde{u}(t) + v(t)$, where $\tilde{x} + z$ is "in the neighbourhood of" $\tilde{x}$ and $\tilde{u} + v$ "in the neighbourhood of" $\tilde{u}$. This will be made more precise later on. We have

$$\dot{\tilde{x}} = f(\tilde{x},\tilde{u}),\ \tilde{x}(0) = \tilde{x}_0 \tag{14}$$

$$\frac{d}{dt}(\tilde{x}+z) = f(\tilde{x}+z,\ \tilde{u}+v),\ \tilde{x}(0)+z(0) = \tilde{x}_0 + z_0. \tag{15}$$

We assume z and v to be small such that the right hand side of (15) can be expanded into a Taylor series, where the expansion up to linear terms yields a good approximation:

$$\frac{d}{dt}(\tilde{x}+z) = f(\tilde{x},\tilde{u}) + \frac{\partial f}{\partial x}(\tilde{x},\tilde{u})z + \frac{\partial f}{\partial u}(\tilde{x},\tilde{u})v + \text{ higher order terms.} \tag{16}$$

This is a vector equation. Written out in components the terms are

$$\frac{d}{dt}(\tilde{x}) = \begin{pmatrix} \frac{d\tilde{x}_1}{dt} \\ \vdots \\ \frac{d\tilde{x}_n}{dt} \end{pmatrix},\quad f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix},\quad \frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}.$$

If (14) is subtracted from (16), we get

$$\dot{z} = \frac{\partial f}{\partial x}(\tilde{x},\tilde{u})z + \frac{\partial f}{\partial u}(\tilde{x},\tilde{u})v \tag{17}$$

for the approximated system. This differential equation is linear, since the coefficients $\frac{\partial f}{\partial x}(\tilde{x},\tilde{u})$ and $\frac{\partial f}{\partial u}(\tilde{x},\tilde{u})$ are given matrix-functions of time. Therefore we write for (17)

$$\dot{z} = A(t)z + B(t)v \tag{18}$$

The output function $y = g(x,u)$ can also be linearized around the pair $(\tilde{x},\tilde{u})$. If $\tilde{y} = g(\tilde{x},\tilde{u})$ and $\tilde{y} + w = g(\tilde{x}+z,\ \tilde{u}+v)$, then

$$\tilde{y} + w = g(\tilde{x},\tilde{u}) + \frac{\partial g(\tilde{x},\tilde{u})}{\partial x}z + \frac{\partial g(\tilde{z},\tilde{u})}{\partial u}v + \text{ higher order terms}$$

and therefore, as an approximation,

$$w(t) = \frac{\partial g(\tilde{x},\tilde{u})}{\partial x}z + \frac{\partial g(\tilde{x},\tilde{u})}{\partial u}v$$

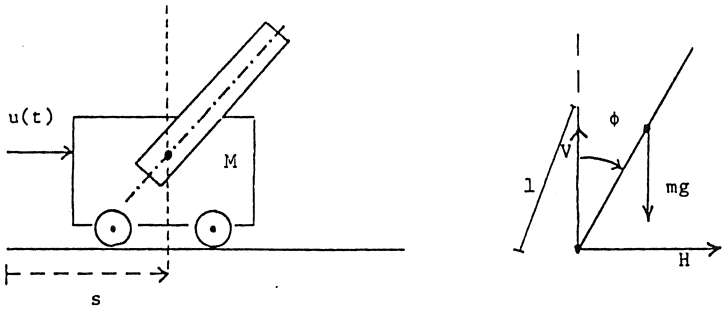
which we write as

$$w(t) = C(t)z + D(t)v \tag{19}$$

Equations (18) and (19) together form the linearized system, linearized around the solution $(\tilde{x}(t), \tilde{u}(t))$.

Example of the inverted pendulum.

Consider the inverted pendulum in the following figure. The pivot of the pendulum is mounted on a carriage which can move in horizontal direction. The carriage is driven by a small motor that at time t exerts a force $u(t)$ on the carriage. This force is the input variable to the system.



The mass of the carriage will be indicated by M, that of the pendulum by m. The distance of the pendulum between pivot and center of gravity is l. In the figure $H(t)$ denotes the horizontal reaction force and $V(t)$ is the vertical reaction force in the pivot. The angle that the pendulum makes with the vertical is indicated by $\phi(t)$. For the center of gravity of the pendulum we have

$$m\frac{d^2}{dt^2}(s + l\sin\phi) = H; \qquad m\frac{d^2}{dt^2}(l\cos\phi) = V - mg; \tag{20}$$

$$J\frac{d^2\phi}{dt^2} = Vl\sin\phi - Hl\cos\phi. \tag{21}$$

The function $s(t)$ denotes the position of the carriage and J is the moment of inertia with respect to the center of gravity. If the pendulum has a uniform mass distribution, then

$$J = \frac{1}{3}ml^2.$$

The equation which describes the movement of the carriage is

$$M\frac{d^2s}{dt^2} = u - H \tag{22}$$

Substitution of H and V in (20) into (21) and (22) leads to

$$\left.\begin{array}{l} \frac{4l}{3}\ddot{\phi} - g\sin\phi + \ddot{s}\cos\phi = 0 \\ (M+m)\ddot{s} + ml(\ddot{\phi}\cos\phi - \dot{\phi}^2\sin\phi) = u. \end{array}\right\} \tag{23}$$

This system can be written as a set of four first order differential equations where the state vector is defined as $x = (\phi, \dot{\phi}, s, \dot{s})^T$, where T denotes the transpose. We either can linearize that vector differential equation, or we can linearize (23) directly and then afterwards we will construct a set of linear differential equations. We will continue with the latter method (the reader should do the first method himzelf and convince himzelf that the answers will be the same). Linearization of (23) will be shown around the solution

$$\tilde{\phi}(t) = \dot{\tilde{\phi}}(t) = \tilde{s}(t) = \dot{\tilde{s}}(t) = u(t) = 0$$

and leads to (i.e. the nonlinear terms in (23) are replaced by Taylor series expansion up to the linear term)

$$\frac{4l}{3}\ddot{\phi} - g\phi + \ddot{s} = 0, \quad (M+M)\ddot{s} + ml\ddot{\phi} = u,$$

which can be written as

$$\frac{dx}{dt} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ a_{41} & 0 & 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ b_2 \\ 0 \\ b_4 \end{pmatrix} u, \tag{24}$$

where

$$a_{21} = \frac{3g(M+M)}{1(4M+m)}, \quad a_{41} = \frac{-3gm}{4M+m}, \quad b_2 = \frac{-3}{1(4M+m)}, \quad b_4 = \frac{4}{4M+m}.$$

If we take $M = .98kg$, $m = .08kg$, $l = .312m$ and $g = 10m/sec^2$, then (24) becomes

$$\dot{x} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -.6 & 0 & 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ -2.4 \\ 0 \\ 1 \end{pmatrix} u$$

If s and ϕ are measured quantitis, then the output function is

$$y = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} x$$

Hulpmiddelen uit de Lineaire Algebra

A.W. Grootendorst
Aardbeistraat 11
2564 TM Den Haag

0. In deze voordracht, getiteld "hulpmiddelen uit de Lineaire Algebra", zullen enkele onderwerpen aan de orde gesteld worden die voor een goed begrip van de hoofdvoordrachten over systeemtheorie nodig zijn. Uiteraard is het niet wel mogelijk een volledige inleiding in de lineaire algebra te geven; dat zou een boek op zichzelf worden. In deze inleiding is er van uitgegaan dat de toehoorder-lezer de meest elementaire begrippen uit de lineaire algebra tot zijn beschikking heeft. Dit is op zichzelf al een zeer subjectief standpunt. Toch zullen enkele van deze begrippen zoals rang van een matrix, basistransformatie, verband tussen lineaire afbeelding en matrix-representatie daarvan, etc. nog even naar voren gehaald worden.
Daarnaast worden in deze inleiding enkele stellingen en begrippen behandeld die wellicht niet aan iedereen bekend zijn, zoals de stelling van Cayley-Hamilton, diagonalisatie en de Jordan normaalvorm, en wel omdat deze door de andere sprekers gebruikt worden. Genoemde stelling zal exact bewezen worden. Van de Jordan normaalvorm wordt alleen de definitie gegeven en aan voorbeelden toegelicht. Voor nadere details wordt daarbij verwezen naar de opgegeven literatuur. Hiervan wordt vooral het Schaum dictaat van Seymour Lipschutz aanbevolen (litt. (4)). Met zijn heldere tekst en de 600 vraagstukken, die voor een deel uitwerkingen van eerder genoemde stellingen omvatten, is dit boek een uitmuntende handleiding bij het zelfstandig bestuderen van de stof.

1. Lineaire vectorruimten

In de lineaire algebra staat het begrip lineaire vectorruimte (ook wel lineaire ruimte of vectorruimte genoemd) centraal. Daarom beginnen we met de volgende

Definitie. Een lineaire vectorruimte bestaat uit een additief geschreven Abelse* groep V waarvan de elementen vectoren worden genoemd en een lichaam K waarvan de elementen scalairen worden genoemd. Behalve de gebruikelijke operaties die inherent zijn aan een groep en een lichaam, is er ook een vermenigvuldiging gedefinieerd van de scalairen van K met de vectoren van V, d.w.z. aan elk paar (α, a) met $\alpha \in K$ en $a \in V$ is een element van $V-$genoteerd als αa – toegevoegd. Hierbij gelden de volgende regels voor alle $\alpha, \beta \in K$ en alle $a, b \in V$:

*Niels Henrik Abel (1802-1829)

$$\begin{aligned} \alpha(a+b) &= \alpha a + \alpha b \\ (\alpha+\beta)a &= \alpha a + \beta a \\ (\alpha\beta)a &= \alpha(\beta a) \\ 1.a &= a \end{aligned}$$

waarbij 1 het één-element van K is. Voor de volledigheid nog even de definities van Abelse groep en van lichaam.

DEFINITIES. Een Abelse groep V (formeler $(V,+)$) is een niet-lege verzameling elementen waarbij aan elk tweetal $a \in V$, $b \in V$ een derde element, de som van a en b genaamd, is toegevoegd. Deze som noteren we als $a+b$ en daarbij geldt:

$$\begin{aligned} a+b &= b+a \\ (a+b)+c &= a+(b+c) \end{aligned}$$

Er is een nulelement $0 \in V$ zodanig dat

$$a+0 = 0+a = a \quad \text{voor alle} \quad a \in V.$$

Bij elke $a \in V$ is er een element $(-a)$ (tegengestelde van a genaamd) zodat

$$a+(-a) = (-a)+a = 0.$$

Een lichaam L (formeler $(L,+,\times)$) is een niet-lege verzameling elementen die voor de operatie $+$ (de optelling) een Abelse groep vormt. De elementen die van 0 verschillen vormen voor de vermenigvuldiging een Abelse groep. Optelling en vermenigvuldiging zijn verbonden door de *associatieve* wet:

$$a \times (b+c) = a \times b + a \times c$$

en natuurlijk vanwege de commutativiteit van de vermenigvuldiging:

$$(b+c) \times a = a \times (b+c) = a \times b + a \times c = b \times a + c \times a.$$

In deze voordrachtenserie kiezen we voor K steeds $\mathbb{R}$ (het lichaam van de reële getallen) of $\mathbb{C}$ (het lichaam van de complexe getallen).
Het standaardvoorbeeld van een vectorruimte over $\mathbb{R}$ is het volgende. V bestaat daarbij uit alle rijen (of kolommen) met n elementen uit $\mathbb{R}$ (dus $V = \mathbb{R}^n$). De optelling gebeurt elementsgewijs, d.w.z.

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)$$

en de vermenigvuldiging met een scalair α is gedefinieerd door

$$\alpha(a_1, a_2, \ldots, a_n) = (\alpha a_1, \alpha a_2, \ldots, \alpha a_n).$$

Dit voorbeeld heeft een bijzondere eigenschap en wij zullen ons uitsluitend beperken tot vectorruimten met die eigenschap. Wij zullen ons nl. uitsluitend bezighouden met *eindig-dimensionale vectorruimten.* Hierbij geldt de

DEFINITIE. Een eindig-dimensionale vectorruimte met dimensie n heeft een lineair onafhankelijke basis bestaande uit n (lineair onafhankelijke) vectoren. Daarbij geldt de

DEFINITIE. De elementen $a_1, a_2, \ldots, a_n \in V$ heten lineair onafhankelijk over K als uit

$$\alpha_1 a_1 + \alpha_2 a_2 + \ldots + \alpha_n a_n = 0 \quad (\alpha_i \in K)$$

volgt $\alpha_1 = \alpha_2 = \ldots = a_n = 0$.

Zijn deze elementen niet onafhankelijk over K dan noemen wij ze lineair afhankelijk over K. In dat geval bestaan er dus elementen α_i $(i = 1, 2, \ldots, n)$ die *niet alle nul* zijn zodanig

dat $\alpha_1 a_1 + \alpha_2 \alpha_2 + \ldots \alpha_n a_n = 0$.

Verder hebben we de

DEFINITIE. Een stel elementen $b_1, \ldots, b_n$ heet een basis voor een vectorruimte indien zij lineair onafhankelijk zijn en ieder element $a \in V$ geschreven kan worden in de gedaante

$$a = \alpha_1 b_1 + \alpha_2 b_2 + \ldots + \alpha_n b_n. \quad (\alpha_i \in K)$$

Uit de onafhankelijkheid van de vectoren $b_1, \ldots, b_n$ volgt dat de α_i eenduidig bepaald zijn door a.

OPMERKING 1. Een *basis* is dus altijd een lineair onafhankelijke basis. Onder een stel *voortbrengenden* (eventueel oneindig veel), verstaan we een verzameling S van vectoren zodanig dat elke a in V geschreven kan worden als een lineaire combinatie van eindig veel elementen van S met coëfficiënten in K.

OPMERKING 2. Reeds nu wijzen we erop dat in een n-dimensionale vectorruimte de vectoren op één-één duidige wijze corresponderen met n-tupels, rijtjes elementen van de scalairen in K. Het eerder genoemde voorbeeld is dus zeer wezenlijk.

OPMERKING 3. Het getal n, de dimensie van de vectorruimte, is onafhankelijk van de keuze van de basisvectoren.

2. OVERGANG OP EEN ANDERE BASIS

Laat V een eindig-dimensionale vectorruimte zijn over K met basis $\{e_1, e_2, \ldots, e_n\}$. Elke vector $v \in V$ is dan op één, doch slechts één, wijze te schrijven als

$$v = v_1 e_1 + v_2 e_2 + \ldots + v_n e_n$$

met $v_i \in K$, de zgn. kentallen van V t.o.v. deze basis. Bij deze basiskeuze behoort dus bij $v \in V$ een kolomvector (element van K^n), $\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ die we aangeven met $[v]_e$. Bij een andere keuze van een basis zal dezelfde vector v i.h.a. andere kentallen hebben. Indien we overgaan op de basis $\{f_1, f_2, \ldots, f_n\}$ dan geldt bijv.

$$v = v_1' f_1 + v_2' f_2 + \ldots + v_n' f_n.$$

De bijbehorende kolomvector van de kentallen noteren we nu als $[v]_f$.
Een voor de hand liggende vraag is: wat is het verband tussen $[v]_e$ en $[v]_f$?
Welnu, men kan bewijzen dat er een niet-singuliere $n \times n$-matrix P bestaat, d.w.z. met determinant waarde $\neq 0$ zodanig dat $[v]_e = P[v]_f$ voor alle $v \in V$ en dus ook $[v]_f = P^{-1}[v]_e$. Deze matrix noemt men de *overgangsmatrix*. Natuurlijk hangt P af van de oude e-basis en de nieuwe f-basis. De notatie P_e^f zou exacter zijn, maar omdat we hier een vaste e-basis en een vaste f-basis beschouwen, schrijven we kortheidshalve P, gedachtig aan de uitspraak die wel wordt toegeschreven aan Luther: "Jede Konsequenz führt zum Teufel". De matrix P kunnen we als volgt beschrijven: Indien voor de nieuwe basisvectoren $f_1, f_2, \ldots, f_n$ geldt:

$$\begin{aligned} f_1 &= p_{11}e_1 + p_{12}e_2 + \ldots + p_{1n}e_n \\ f_2 &= p_{21}e_1 + p_{22}e_2 + \ldots + p_{2n}e_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ f_n &= p_{n1}e_1 + p_{n2}e_2 + \ldots + p_{nn}e_n \end{aligned}$$

dan is P de getransponeerde van de coëfficiënten-matrix, dus

$$P = \begin{pmatrix} p_{11} & p_{21} \cdots p_{n1} \\ p_{12} & p_{22} \cdots p_{n2} \\ \cdots\cdots & \cdots\cdots\cdots \\ p_{1n} & p_{2n} \cdots p_{nn} \end{pmatrix}.$$

In de kolommen staan dus de kentallen van de nieuwe basisvectoren (t.o.v. de oude basis). Voor elke $v \in V$ geldt dan, zoals gezegd:

$$[v]_e = P[v]_f \quad \text{en} \quad [v]_f = P^{-1}[v]_e.$$

Een voorbeeld. $V = \mathbb{R}^2$, $K = \mathbb{R}$ met als e-basis $\{e_1, e_2\}$ met $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Als f basis kiezen we $\{f_1, f_2\} = \{2e_1 + 3e_2, 4e_1 + 5e_2\}$. Hierbij geldt

$$f_1 = 2e_1 + 3e_2$$

$$f_2 = 4e_1 + 5e_2$$

dus $P = \begin{pmatrix} 2 & 4 \\ 3 & 5 \end{pmatrix}$ en $P^{-1} = \frac{1}{2}\begin{pmatrix} -5 & 4 \\ 3 & -2 \end{pmatrix}$.

Voor de vector v met $[v]_e = \begin{pmatrix} a \\ b \end{pmatrix}$ geldt dan

$$[v]_f = \frac{1}{2}\begin{pmatrix} -5 & 4 \\ 3 & -2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix},$$

v heeft dus als kentallen t.o.v. de nieuwe-basis:

$$\frac{1}{2}(-5a + 4b) \quad \text{en} \quad \frac{1}{2}(3a - 2b).$$

3. Lineaire afbeeldingen en matrices

A. *Lineaire afbeeldingen*

Onder een lineaire afbeelding A van de vectorruimte V over K in zichzelf verstaan we een afbeelding van V in zichzelf waarbij aan elke $v \in V$ op één-duidige wijze een element $Av \in V$ wordt toegevoegd en wel zó dat voor alle $v, w \in V$ en $\alpha, \beta \in K$ geldt:

$$A(\alpha v + \beta w) = \alpha Av + \beta Aw.$$

De verzameling $\{Av | v \in V\}$ heet de beeldruimte van A, notatie ImA. De verzameling $\{v \in V | Av = 0\}$ heet de kern van A, notatie Ker$A$. Beide zijn lineaire deelruimten van V, d.w.z. deelverzamelingen van V die op zichzelf een lineaire ruimte zijn. De som van hun dimensies is de dimensie van V.

Voorbeeld. Als V de drie dimensionale Euclidische ruimte is en A de loodrechte projectie op het XOY-vlak dan geldt Ker$A = Z$ - as; Im$A = XOY$-vlak.

B *Lineaire afbeeldingen door matrices gerepresenteerd*

Op zeer natuurlijke wijze kan men bij een n-dimensionale vectorruimte aan een lineaire afbeelding een matrix koppelen. Laat V weer een n-dimensionale vectorruimte over K zijn, met basis $\{e_1, e_2, \ldots, e_n\}$ die we voorlopig vasthouden. Een lineaire afbeelding A voert dan de eenheidsvectoren e_i over in Ae_i, $(i = 1, 2, \ldots, n)$ waarvoor dan geldt:

$$\begin{aligned} Ae_1 &= a_{11}a_1 + a_{12}e_2 + \ldots + a_{1n}e_n \\ Ae_2 &= a_{21}e_1 + a_{22}e_2 + \ldots + a_{2n}e_n \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Ae_n &= a_{n1}e_1 + a_{n2}e_2 + \ldots + a_{nn}e_n. \end{aligned}$$

De getransponeerde van de coëfficientenmatrix, dus de matrix

$$\begin{pmatrix} a_{11} & a_{21} & \ldots & a_{n1} \\ a_{12} & a_{22} & & a_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ a_{1n} & a_{2n} & & a_{nn} \end{pmatrix}$$

noemen wij de matrix die behoort bij deze lineaire afbeelding en de gekozen e-basis. Formeel schrijven we daarvoor $[A]_e$, maar als de basis vastligt en er geen verdere verwarring dreigt, dan schrijven we gewoon A. Het belang ervan schuilt in het volgende. Als $v \in V$ dan voert A deze vector over in Av en men rekent nu gemakkelijk na dat

$$[Av]_e = [A]_e[v]_e.$$

Uitvoeriger

$$\text{als} \quad v = v_1e_1 + v_2e_2 + \ldots + v_ne_n$$

$$\text{en} \quad Av = v_1'e_1 + v_2'e_2 + \ldots + v_n'e_n$$

dan geldt:

$$\begin{pmatrix} v_1' \\ \vdots \\ v_n' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \ldots & a_{n1} \\ a_{12} & a_{22} & \ldots & a_{n2} \\ \vdots & & & \\ a_{1n} & a_{2n} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

Deze matrix beschrijft dus het effect van de lineaire afbeelding A op de kolomvectoren $[v]_e$.

3. Lineaire afbeeldingen en basistransformatie

De vraag doet zich nu direct voor: hoe wordt de lineaire afbeelding A gerepresenteerd bij een andere keuze van de basis?
Hiertoe verwijzen we allereerst naar wat wij hierboven vonden. Bij overgang van e-basis naar f-basis geldt:

$$[v]_e = P[v]_f \quad \text{en} \quad [v]_f = P^{-1}[v]_e \quad (P = [P]_e^f)$$

dus

$$[Av]_e = [A]_e[v]_e = [A]_eP[v]_f$$

en dus $\quad P^{-1}[Av]_e = P^{-1}[A]_eP[v]_f$

maar

$$P^{-1}[Av]_e = [Av]_f$$

dus

$$[Av]_f = P^{-1}[A]_eP[v]_f$$

Indien we uitgaan van de f-basis wordt het effect van de lineaire afbeelding A dus beschreven door de matrix

$$[A]_f = P^{-1}[A]_eP$$

waarin P de overgangsmatrix $[P]_e^f$ is, behorende bij de overgang van de e-basis naar de f-basis.
We zeggen dan dat $[A]_f$ en $[A]_e$ equivalent zijn op grond van de

DEFINITIE. De matrix A is equivalent met de matrix B indien er een niet-singuliere matrix P bestaat zodanig dat

$$B = P^{-1}AP.$$

Niet singulier betekent dat de determinant $|P| \neq 0$, dus P heeft een inverse P^{-1} zodanig dat $P^{-1}P = PP^{-1} = E$.
Men ziet onmiddellijk in dat deze equivalentierelatie voldoet aan reflexiviteit, symmetrie en transitiviteit.

4. EIGENVECTOREN EN EIGENWAARDEN

Wanneer A een lineaire afbeelding is van een n-dimensionale vectorruimte V over het lichaam K, in zichzelf dan kan men vragen naar de vectoren v ($v \neq 0$) waarvoor geldt:

$$Av = tv \qquad (t \in K).$$

Deze gaan dus over in een scalair veelvoud van zichzelf.
Men noemt zo'n vector een eigenvector van A en t de bijbehorende eigenwaarde.
Kiest men in V een basis $\{e_1, e_2, \ldots, e_n\}$ dan wordt A gerepresenteerd door een matrix $[A]_e$ die werkt op de kolomvector $[v]_e$ gevormd door de kentallen $v_1, v_2, \ldots, v_n$ van v t.o.v. deze basis. Er geldt dan

$$[A]_e[v]_e = t[v]_e.$$

Uitgeschreven:

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \\ a_{n1} & \ldots & A_{nn} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} tv_1 \\ \ldots \\ tv_n \end{pmatrix}.$$

De kentallen $v_1, v_2, \ldots, v_n$ en t voldoen dus aan het stelsel vergelijkingen:

$$\begin{aligned} (t - a_{11})v_1 - a_{12}v_2 - a_{13}v_3 \ldots - a_{1n}v_n &= 0 \\ -a_{21}v_1 + (t - a_{22})v_2 \ldots - a_{2n}v_n &= 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ -a_{n1}v_1 - a_{n2}v_2 \ldots + (t - a_{nn})v_n &= 0. \end{aligned}$$

N.e.v. voorwaarde voor het bestaan van een oplossing $(v_1, v_2, \ldots, v_n)$ verschillend van de nulvector is dat de determinant

$$\begin{vmatrix} t - a_{11} & -a_{12} \ldots - a_{1n} \\ -a_{21} & t - a_{22} \ldots - a_{2n} \\ \ldots\ldots\ldots\ldots\ldots & \\ -a_{n1} & -a_{n2} \ldots t - a_{nn} \end{vmatrix} = 0.$$

Hiervoor schrijven we

$|tE - A| = 0$.*

Dit is een vergelijking van de graad n in t, de zgn. *karakteristieke vergelijking* van A. Indien we deze uitschrijven in de gedaante

$$t^n + a_{n-1}t^{n-1} + a_{n-2}t^{n-2} + \ldots + a_0 = 0,$$

dan vinden we onmiddellijk

$$a_{n-1} = -(a_{11} + a_{22} + \ldots + a_{nn}),$$

dus het tegengestelde van het spoor S van de matrix A en voor a_n geldt: $a_n = (-1)^n|A|$. Het aantal wortels t hangt uiteraard af van het lichaam K. Voor $K = \mathbb{C}$ heeft de vergelijking n (niet noodzakelijk verschillende) wortels. De multipliciteit van een wortel t_i noemt men de *algebraïsche multipliciteit* daarvan. Het aantal lineair onafhankelijke vectoren dat bij één wortel t_i behoort, noemt men de *meetkundige multipliciteit* (zie voorbeelden). Men kan bewijzen dat voor elke wortel t_i de meetkundige multipliciteit niet groter is dan de algebraïsche multipliciteit. (Zie. bijv. litt. [4]). Zou men in V een andere basis gekozen hebben, dan zou de matrix A vervangen moeten worden door de matrix $P^{-1}AP$ waarin P de overgangsmatrix is (zie bldz. 4). Dit maakt echter geen verschil voor de karakteristieke vergelijking,

daar $|tE - P^{-1}AP| = |tP^{-1}P - P^{-1}AP| =$

$$|P^{-1}| \cdot |tE - A|\,|P| = |tE - A|.$$

Van belang is de

STELLING. Eigenvectoren, behorende bij onderling verschillende eigenwaarden, zijn lineair onafhankelijk.

BEWIJS. Stel dat $v_1, v_2, \ldots, v_k$ eigenvectoren zijn van A die behoren bij de onderling verschillende eigenwaarden $t_1, t_2, \ldots, t_k$. We geven het bewijs via volledige inductie.

$k = 1$. Uit $a_1v_1 = 0$ volgt $a_1 = 0$ daar $v_1 \neq 0$.

Stel dat de bewering juist is voor $v_1, v_2, \ldots, v_p$; we moeten dan de juistheid aantonen voor $v_1, v_2, \ldots v_p, v_{p+1}$.

Laat $a_1v_1 + a_2v_2 + \ldots + a_pv_p + a_{p+1}v_{p+1} = 0,$ (4.1)
dan geldt

$$a_1Tv_1 + a_2Tv_2 + \ldots + a_pTv_p + a_{p+1}Tv_{p+1} = T(0)$$

*Informeel schrijven we A i.p.v. $[A]_e$. Dit leidt niet tot misverstand zoals we zullen zien.

dus $a_1t_1v_1 + a_2t_2v_2 + \ldots + a_pt_pv_p + a_{p+1}t_{p+1}v_{p+1} = 0.$ (4.2)

Vermenigvuldig (4.1) met t_{p+1} en trek (4.2) er vanaf, dan krijgen we:

$$a_1(t_{p+1} - t_1)v_1 + a_2(t_{p+1} - t_2)v_2 + \ldots + a_p(t_{p+1} - t_p)v_p = 0.$$

Volgens de inductieveronderstelling zijn hier alle coëfficiënten $a_i(t_{p+1} - t_i)$ gelijk aan 0 $(i = 1, 2, \ldots p)$ dus, daar $t_{p+1} - t_i \neq 0$, geldt $a_i = 0$ $(i = 1, 2, \ldots, p)$. Van (4.1) resteert dus alleen nog $a_{p+1}v_{p+1} = 0$ en dus volgt ook $a_{p+1} = 0$, daar $v_{p+1} \neq 0$.

N.B. Bij een meervoudige wortel (dus gelijke eigenwaarden) kunnen lineair onafhankelijke vectoren behoren (zie voorbeeld 2). De genoemde stelling is dus niet omkeerbaar!

TWEE VOORBEELDEN

1. Bepaald de eigenwaarden en de bijbehorende eigenvectoren

van de matrix $A = \begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & 3 \\ 1 & 1 & 1 \end{pmatrix}$.

De karakteristieke vergelijking is

$$|tE - A| = 0$$

d.w.z. $\begin{vmatrix} (t-1) & 1 & -1 \\ -2 & t & -3 \\ -1 & -1 & (t-1) \end{vmatrix} = 0$

dus

$$t^3 - 2t^3 - t + 2 = 0$$

met wortels $t_1 = 1$, $t_2 = -1$, $t_3 = 2$, die dus elk de algebraïsche multipliciteit 1 hebben. Bij $t_1 = 1$ geldt voor de bijbehorende eigenvectoren

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} : \qquad \begin{aligned} v_2 - v_3 &= 0 \\ -2v_1 + v_2 - 3v_3 &= 0 \\ -v_1 + v_2 &= 0 \end{aligned}$$

dus $\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \alpha \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \qquad (\alpha \in K).$

Men rekent gemakkelijk na:
Bij $t_2 = -1$ en $t_3 = 2$ behoren resp. de eigenvectoren

$$\beta\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \quad \text{en} \quad \gamma\begin{pmatrix} -1 \\ 5 \\ 4 \end{pmatrix} \qquad (\beta, \gamma \in K).$$

2. Bepaal de eigenwaarden en de eigenvectoren van de matrix

$$A = \begin{pmatrix} 1 & -2 & 2 \\ -2 & -2 & 4 \\ 2 & 4 & -2 \end{pmatrix}.$$

Nu gaat de karakteristieke vergelijking

$$|tE - A| = 0$$

over in $\begin{vmatrix} (t-1) & 2 & -2 \\ 2 & (t+2) & -4 \\ -2 & -4 & (t+2) \end{vmatrix} = 0$

d.w.z. $t^3 + 3t^2 - 24t + 28 = 0.$

Hiervan zijn de wortels $t_1 = -7$, $t_2 = t_3 = 2$. Bij t_1 behoren de eigenvectoren die we vinden uit:

$$\begin{aligned} -8v_1 + 2v_2 - 2v_3 &= 0 \\ 2v_1 - 5v_2 - 4v_3 &= 0 \\ -2v_1 - 4v_2 - 5v_3 &= 0 \end{aligned}$$

dus $\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \alpha\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \qquad (\alpha \in K).$

Bij $t_2 = t_3 = 2$ vinden we:

$$\begin{aligned} v_1 + 2v_2 - 2v_3 &= 0 \\ 2v_1 + 4v_1 - 4v_3 &= 0 \\ -2v_1 - 4v_2 + 4v_3 &= 0 \end{aligned}$$

dus $v_1 + 2v_2 - 2v_3 = 0.$

Hieraan voldoen alle vectoren die lineaire combinaties zijn van de twee onafhankelijk vectoren

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{en} \quad \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

De meetkundige multipliciteit is dus 2 en deze eigenvectoren spannen een twee dimensionale *eigenruimte* op:

$$\beta \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \gamma \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \qquad (\beta, \gamma \in K).$$

5 Diagonaliseerbaarheid en Jordanvorm

A. Diagonaliseerbaarheid

Men noemt een $n \times n$-matrix diagonaliseerbaar indien hij equivalent is met een diagonaalmatrix, d.w.z. een matrix waarin alle elementen buiten de hoofddiagonaal nul zijn. Verder noemen we de matrix A equivalent met de matrix B (zoals we al eerder opmerkten) indien er een niet-singuliere matrix P bestaat zodanig dat $P^{-1}AP = B$.

Een voorbeeld van een diagonaaliseerbare matrix is $\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$, immers

$$\begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix} = \begin{pmatrix} -20 & 0 \\ 0 & 5 \end{pmatrix}.$$

Er rijzen dan onmiddellijk twee vragen.

1. Wat zijn de nodige en voldoende voorwaarden opdat een matrix diagonaliseerbaar is?

2. Hoe voert men een eventueel mogelijke diagonalisatie uit?

Allereerst de

Stelling. Een $n \times n$-matrix is dan en slechts dan diagonaliseerbaar indien hij n lineair onafhankelijke eigenvectoren heeft.

We stellen ons voor dat de beschouwde matrices een n-dimensionale vectorruimte, bestaande uit kolommen met lengte n en met elementen in K, in zichzelf afbeeldt.

Bewijs.

a. Stel dat de matrix A diagonaliseerbaar is.
Dan is er dus een niet-singuliere matrix T zodanig dat

$$T^{-1}AT = \begin{pmatrix} \lambda_1 & \dots & \dots & 0 \\ \vdots & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \lambda_n \end{pmatrix}$$

Voor de eenheidsvectoren $e_i = [e_i] = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i^e$plaats geldt dan

$$\begin{aligned} & T^{-1}ATe_i && = \lambda_i e_i \\ \text{en dus} \quad & A(Te_i) && = \lambda_i Te_i. \end{aligned}$$

Dit betekent dat de vectoren $Te_i (i = 1, 2 \ldots, n)$ eigenvectoren zijn van A met eigenwaarde λ_i.
De onafhankelijkheid van deze vectoren blijkt aldus: stel dat $\alpha_1 Te_1 + \alpha_2 Te_2 + \cdots + \alpha_n Te_n = 0$ dan volgt, door vermenigvuldiging met T^{-1}

$$\alpha_1 e_1 + \alpha_2 e_2 + \cdots \alpha_n e_n = 0$$

en dus, omdat $e_1, e_2, \ldots, e_n$ lineair onafhankelijk zijn: $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$.

b. Stel dat de matrix A n lineair onafhankelijke eigenvectoren $a_1, a_2, \ldots, a_n$ heeft met eigenwaarden $\lambda_1, \lambda_2, \ldots, \lambda_n$. We kunnen deze dan in de genoemde vectorruimte als basis kiezen. Zoals eerder werd aangetoond, wordt het effect van A — als afbeelding — bij deze nieuwe basis beschreven door de matrix $P^{-1}AP$ waarin P de niet-singuliere matrix is met in de kolommen de kentallen van de als basisvectoren gekozen eigenvectoren (kentallen t.o.v. de oude basis).
Anderzijds zal bij deze basiskeuze de matrix $P^{-1}AP$ de diagonaalvorm hebben daar de nieuwe eenheidsvectoren

$$\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i - \text{de plaats}$$

eigenvectoren zijn en dus overgaan in een veelvoud van zichzelf.

Dus

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix}.$$

Hiermee is tevens de tweede vraag beantwoord, daar dit bewijs constructief is.

Voorbeelden.

1. $A = \begin{pmatrix} 5 & -1 \\ 6 & -2 \end{pmatrix}$. Karakteristieke vergelijking

$$\begin{vmatrix} t-5 & 1 \\ -6 & t+2 \end{vmatrix} = 0, \quad t^2 - 3t - 4 = 0 \Rightarrow t_1 = -1, t_2 = 4;$$

bijbehorende eigenvectoren resp. $\begin{pmatrix} 1 \\ 6 \end{pmatrix}$ en $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

dus

$$P = \begin{pmatrix} 1 & 1 \\ 6 & 1 \end{pmatrix}, \; P^{-1} = \frac{1}{5}\begin{pmatrix} -1 & 1 \\ 6 & -1 \end{pmatrix} \text{ en}$$

$$P^{-1}AP = \frac{1}{5}\begin{pmatrix} -1 & 1 \\ 6 & -1 \end{pmatrix}\begin{pmatrix} 5 & -1 \\ 6 & -2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 6 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 4 \end{pmatrix}.$$

2. Met de resultaten van de voorbeelden op bldz. 9 kunnen we ook

$$\begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & 3 \\ 1 & 1 & 1 \end{pmatrix}$$

diagonaliseren.
Op grond van de theorie zal moeten gelden

$$\begin{pmatrix} -1 & 1 & -1 \\ 1 & 1 & 5 \\ 1 & -1 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & -1 \\ 1 & 1 & 5 \\ 1 & -1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

en

$$\begin{pmatrix} 1 & 1 & 0 \\ 2 & 0 & 1 \\ -2 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -2 & 2 \\ -2 & -2 & 4 \\ 2 & 4 & -2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 2 & 0 & 1 \\ -2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -7 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

OPMERKING 1. Aangezien bij verschillende eigenwaarden lineair onafhankelijke eigenvectoren behoren, is een $n \times n$-matrix met n verschillende eigenwaarden diagonaliseerbaar. Deze voorwaarde is dus voldoende, maar niet noodzakelijk (zie tweede voorbeeld).

OPMERKING 2. Onder het minimumpolynoom van een $n \times n$-matrix A verstaat men het monieke polynoom van de laagste graad met de gedaante

$$t^m + a_{m-1}t^{m-1} + a_{m-2}t^{m-2} + \cdots + a_0 E \quad (a_i \in k)$$

waarvan A nulpunt is. Hierin is E de $n \times n$-eenheidsmatrix.
Men kan dan bewijzen dat een matrix A dan en slechts dan diagonaliseerbaar is als zijn minimumpolynoom (met 1 i.p.v. E) een produkt is van onderling verschillende lineaire factoren.
Tussen het minimumpolynoom en het karakteristieke polynoom (met weer 1 i.p.v. E) bestaat dit verband dat zij dezelfde irreducibele factoren hebben (maar niet noodzakelijk in hetzelfde aantal!)

OPMERKING 3. Het diagonaliseren van een matrix heeft in de meetkunde een toepassing bij het klassificeren van 2de graadskrommen en 2de-graadsoppervlakken.
Bij het berekenen van machten van een matrix kan diagonalisatie veel werk besparen, immers als $P^{-1}AP = D$ (diagonaalmatrix), dan geldt:

$$\begin{aligned} P^{-1}A^nP &= (P^{-1}AP)^n = D^n, \text{ dus} \\ A^n &= PD^nP^{-1}, \end{aligned}$$

maar de macht van een diagonaalmatrix is direct op te schrijven. Dit is van belang bij de berekening van e^A, waarbij A een matrix is.
De formele definitie daarvan luidt:

$$e^A = E + \frac{A}{1!} + \frac{A^2}{2!} + \cdots + \frac{A^n}{n!} + \cdots$$

E is weer de $n \times n$-eenheidsmatrix. Uiteraard wordt convergentie ondersteld.

B. Jordan normaalvorm *

Wanneer een matrix niet diagonaliseerbaar is, dan kan hij toch vaak in een handelbare vorm gebracht worden. Met dit laatste bedoelen we dat de matrix equivalent is met een matrix in de Jordan normaalvorm.
Een matrix heeft de Jordan normaalvorm indien deze is opgebouwd uit vierkante blokken rond de hoofddiagonaal die de volgende gedaante (maar niet noodzakelijk dezelfde afmetingen) hebben.

$$\begin{pmatrix} \alpha & 1 & 0 & 0 & 0 \\ 0 & \alpha & 1 & 0 & 0 \\ 0 & 0 & \alpha & 1 & 0 \\ 0 & 0 & 0 & \alpha & 1 \\ 0 & 0 & 0 & 0 & \alpha \end{pmatrix}$$

dus overal nullen behalve op de hoofddiagonaal (die onderling gelijke elementen bevat) en op de nevendiagonaal "boven" de hoofddiagonaal louter enen.
Een concreet voorbeeld:

$$\left(\begin{array}{cc|cc|c|c|cc} 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 8 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \end{array}\right)$$

*C. Jordan (1838-1921)

Indien een matrix A equivalent is met een Jordan matrix J dan bestaat er een niet-singuliere T matrix T zodanig dat $T^{-1}AT = J$ en dus

$$AT = TJ = T \left(\begin{array}{ccc|cccccc|c} \alpha_1 & 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \alpha_1 & 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \alpha_1 & 1 & 0 & & & & & \\ \hline \cdot & & 0 & \alpha_2 & 1 & 0 & & & & \\ \cdot & & & 0 & \ddots & 1 & 0 & & & \\ \cdot & & & & & \alpha_2 & 1 & 0 & & \\ \cdot & & & & & 0 & \alpha_2 & & 0 & \\ \hline \cdot & & & & & & 0 & \ddots & 1 & 0 \\ 0 & & & & & & & & \alpha_n & 1 \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & \alpha_n \end{array}\right).$$

We noteren de kolomvectoren van T als $k_1, \ldots, k_n$.
Het is dan duidelijk dat

$$Ak_i = \alpha_i k_i + \beta_i k_{i-1}$$

waarbij $\beta_1 = 0$ en $\beta_2 = \ldots = \beta_n = 1$ en α_i de collectie $\{\alpha_i\}$ doorloopt (gelijke α_i meervoudig geteld). Indien $\beta_i = 0$, dan is k_i een eigenvector van A.
Indien $\beta_i = 1$, dan noemen we k_i een gegeneraliseerde eigenvector van A.
Aangezien $A^n = T^{-1}J^nT$ en J^n vrij eenvoudig is te berekenen, kunnen we hiermee zonder al te veel moeite e^A berekenen.
De vraag is nu natuurlijk "onder welke voorwaarde kunnen we een gegeven matrix in de Jordan normaalvorm brengen en hoe ziet die er dan uit?"
We volstaan met het antwoord en verwijzen voor de bewijzen naar de literatuur (bijv. [4]).
Er geldt de

STELLING. Een matrix A kan in de Jordan normaalvorm gebracht worden indien de karakteristieke veelterm (en dus ook de minimumveelterm) in lineaire niet noodzakelijk verschillende factoren te ontbinden is.
Dit is dus steeds het geval als de elementen van A tot $\mathbb{C}$ behoren, maar men kan natuurlijk het coëfficiëntenlichaam altijd uitbreiden door het in te bedden in zijn algebraische afsluiting.
Wat betreft de gedaante van de bijbehorende Jordanvorm het volgende:
Laat het karakteristieke polynoom van A zijn:

$$(t - \alpha_1)^{n_1}(t - \alpha_2)^{n_2} \cdots (t - \alpha_r)^{n_r}$$

en het minimum polynoom:

$$(t - \alpha_1)^{m_1}(t - \alpha_2)^{m_2} \cdots (t - \alpha_r)^{m_r}$$

met $\alpha_i \neq \alpha_j$ als $i \neq j$.
De normaalvorm bestaat dan uit blokken van de gedaante:

$$\longleftarrow j \longrightarrow$$

$$J_{ij} \begin{pmatrix} \alpha_i & 1 & 0 & 0 & \dots & 0 \\ 0 & \alpha_i & 1 & 0 & & \cdot \\ 0 & 0 & \alpha_i & 1 & & \cdot \\ \vdots & & & \ddots & 1 & \cdot \\ \cdot & & & & \alpha_i & 1 \\ 0 & 0 & 0 & & 0 & \alpha_i \end{pmatrix} \begin{matrix} \uparrow \\ j \\ \downarrow \end{matrix}$$

en wel zo dat

1. Er hoort bij elke $i(i = 1, 2, \dots r)$ minstens één blok met afmeting $m_i \times m_i$. Bij deze i hebben alle andere J_{ij} kleinere of gelijke afmetingen.
2. De som van de lengten (= breedten) van alle J_{ij} die bij een vaste i behoren is n_i.
3. Bij vaste i is het aantal blokken J_{ij} gelijk aan de meetkundige multipliceert van α_i
4. Het aantal blokken J_{ij} bij elke i is geheel bepaald door de gegeven matrix.

Tot slot een voorbeeld
Zij van een matrix het karakteristieke polynoom $(t-3)^6(t-5)^4$ en het minimumpolynoom $(t-3)^3(t-5)^2$ dan geldt met de gebruikte notaties:

$$\alpha_1 = 3, \alpha_2 = 5\,;\; m_1 = 3, m_2 = 2\,;\; n_1 = 6, n_2 = 4.$$

Het blok waarin 3 optreedt is een 6×6-matrix, het blok met 5 is een 4×4 matrix.
We geven hier alle mogelijkheden voor het 6×6-resp. resp. 4×4-blok. Door alle combinaties van deze blokken te nemen krijgt men alle bijbehorende (echt verschillende) Jordan matrices van 10×10.

$$\left(\begin{array}{ccc|ccc} 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{array}\right) \quad \left(\begin{array}{ccc|ccc} 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 3 \end{array}\right) \quad \left(\begin{array}{ccc|ccc} 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 3 \end{array}\right)$$

$$\left(\begin{array}{cc|cc} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{array}\right) \quad \left(\begin{array}{cc|cc} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{array}\right)$$

6. Enkele opmerkingen over matrices

A. Rang van een matrix

Bij een gegeven $m \times n$ matrix kunnen we de ruimte beschouwen die opgespannen wordt de rijvectoren (de zgn. rijruimte)
De dimensie daarvan noemen we de rij-rang, hier te noteren als r.
Evenzo kunnen we de ruimte beschouwen die door de kolommen wordt opgespannen (de zgn. kolomruimte). De dimensie daarvan noemen we de kolomrang, hier te noteren als k.
Het is direct duidelijk dat geldt:

$$r \leq \min(m, n) \quad \text{en} \quad k \leq \min(m, n).$$

Door op de bekende wijze met rijen resp. kolommen te "vegen" kan men r, resp. k berekenen.
Als voorbeeld berekenen we de rijrang van

$$\begin{pmatrix} 1 & 1 & 2 & 3 & -2 \\ 2 & 4 & 0 & 0 & -8 \\ 0 & -2 & 4 & 6 & 4 \end{pmatrix}$$

Door op de bekende manier te vegen, d.w.z. handige combinaties van de rijvectoren te kiezen, vinden we achtereenvolgens

$$\begin{pmatrix} 1 & 1 & 2 & 3 & -2 \\ 1 & 2 & 0 & 0 & -4 \\ 0 & -1 & 2 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 2 & 3 & -2 \\ 0 & 1 & -2 & -3 & -2 \\ 0 & -1 & 2 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 4 & 6 & 0 \\ 0 & 1 & -2 & -3 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

De eerste en tweede rij zijn duidelijk lineair onafhankelijk. De rijrang is dus 2.
Van fundamenteel belang is de stelling die wij hier niet bewijzen:
In iedere $m \times n$-matrix zijn rijrang en kolomrang aan elkaar gelijk. Deze gemeenschappelijke waarde noemen we de rang van de matrix. Een vierkante $n \times n$-matrix met rang kleiner dan n noemen we singulier.

B. Geadjungeerde matrix

Zoals bekend kan men de waarde van een determinant berekenen door "te ontwikkelen naar een rij of een kolom", d.w.z. bij een gegeven $n \times n$-matrix A vindt men de determinant als volgt:

$$|A| = \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{vmatrix} = \sum_{j=1}^{n} a_{ij} A_{ij} = \sum_{i=1}^{n} a_{ij} A_{ij} \quad (j = 1, \dots n) \tag{6.1}$$

Hierin zijn de "cofactoren" A_{ij} de van een teken voorziene onderdeterminanten $|M_{ij}|$, de zgn. minoren die men verkrijgt door in $|A|$ de i-de rij en de j-de kolom te schrappen. Het teken dat men toevoegt is $(-1)^{i+j}$.
Ook wordt bekend verondersteld

$$\sum_{i=1}^{n} a_{ij} A_{it} = \sum_{j=1}^{n} a_{ij} A_{tj} = 0 \quad (i \neq t) \tag{6.2}$$

VOORBEELD. $A = \begin{pmatrix} 1 & 2 & -3 \\ 4 & 5 & 6 \\ 3 & 2 & 1 \end{pmatrix}$ dan geldt, bij ontwikkeling naar de tweede kolom:

$$|A| = -2\begin{vmatrix} 4 & 6 \\ 3 & 1 \end{vmatrix} + 5\begin{vmatrix} 1 & -3 \\ 3 & 1 \end{vmatrix} - 2\begin{vmatrix} 1 & -3 \\ 4 & 6 \end{vmatrix} = 42.$$

Deze cofactoren A_{ij} spelen een belangrijke rol bij de definitie van de klassieke geadjungeerde A^* van een $n \times n$-matrix A. Hieronder verstaan we nl. de matrix die op de plaats van a_{ij} heeft staan de cofactor A_{ji} van a_{ji} (let op de volgorde van de indices!), dus

$$A^* = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & & & A_{n2} \\ \vdots & & & \vdots \\ A_{1n} & \dots & \dots & A_{nn} \end{pmatrix}.$$

In ons voorbeeld:

$$A^* = \begin{pmatrix} \begin{vmatrix} 5 & 6 \\ 2 & 1 \end{vmatrix} & -\begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} & \begin{vmatrix} 2 & -3 \\ 5 & 6 \end{vmatrix} \\ -\begin{vmatrix} 4 & 6 \\ 3 & 1 \end{vmatrix} & \begin{vmatrix} 1 & -3 \\ 3 & 1 \end{vmatrix} & -\begin{vmatrix} 1 & -3 \\ 4 & 6 \end{vmatrix} \\ \begin{vmatrix} 4 & 5 \\ 3 & 2 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} \end{pmatrix}. \tag{6.3}$$

Uit (6.1) en (6.2) volgt:

$$A^*A = AA^* = \begin{pmatrix} |A| & \dots & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & |A| \end{pmatrix} = |A|E \tag{6.4}$$

waarbij E de $n \times n$ eenheidsmatrix is.

N.B. Deze klassieke geadjungeerde van een matrix is duidelijk verschillend van wat bij inproduktruimten bekend staat als geadjungeerde matrix (zonder meer) die men uit een matrix verkrijgt door rijen en kolommen te verwisselen en tevens de elementen te vervangen door hun toegevoegd complexen.
Voorbeeld:

$$\begin{pmatrix} i & 2+i & 3 \\ 5 & 7 & 1-i \\ 6 & 4 & 3-i \end{pmatrix}^* = \begin{pmatrix} -i & 5 & 6 \\ 2-i & 7 & 4 \\ 3 & 1+i & 3+i \end{pmatrix}.$$

7. DE STELLING VAN CAYLEY-HAMILTON *

Deze stelling speelt een belangrijke rol in de toepassingen van de matrix-theorie

*Arthur Cayley (1821-1865); William Rowan Hamilton (1805-1865)

en luidt als volgt:
Iedere vierkante matrix is nulpunt van zijn eigen karakteristieke polynoom.
Hierbij is allereerst een toelichting vereist; er wordt het volgende bedoeld:
Indien de $n \times n$-matrix A als karakteristieke vergelijking heeft

$$t^n + a_{n-1}t^{n-1} - \cdots + a_1 t + a_0 = 0$$

waarbij de a_i scalairen uit het grondlichaam zijn, dan voldoet de matrix A aan de betrekking (tussen matrices):

$$A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 E = N.$$

Hierbij zijn E en N resp. de $n \times n$-eenheidsmatrix en de $n \times n$-nulmatrix.
Zo geldt bijv. dat het karakteristieke polynoom van

$$\begin{pmatrix} 4 & 1 \\ -2 & 7 \end{pmatrix} \quad \text{is} \quad t^2 - 11t + 30$$

en dus voldoet deze matrix aan:

$$\begin{pmatrix} 4 & 1 \\ -2 & 7 \end{pmatrix}^2 - 11\begin{pmatrix} 4 & 1 \\ -2 & 7 \end{pmatrix} + 30\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Algemener, als

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

dan is, zoals we zagen, het karakteristieke polynoom van A:

$$\begin{vmatrix} t - a_{11} & -a_{12} & \dots & -a_{1n} \\ -a_{21} & \dots & \dots & -a_{2n} \\ \vdots & & & \vdots \\ -a_{n1} & \dots & \dots & t - a_{nn} \end{vmatrix} \tag{7.1}$$

Kort geschreven: $|tE - A|$

A voldoet dan aan: $|tE - A| = N$.

waarin N weer de $n \times n$-nulmatrix is.
We schrijven (7.1) weer uit als:

$$t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0. \tag{7.2}$$

Nu het bewijs.
Laat $B(t)$ de (klassieke) geadjungeerde zijn van de matrix $(tE - A)$. Er geldt dan (zie (6.4)

$$B(t) \cdot (tE - A) = (tE - A)B(t) = |tE - A| \cdot E \tag{7.3}$$

Hierin is $|tE - A|$ de determinantwaarde van de matrix $(tE - A)$ en E is weer de $n \times n$-eenheidsmatrix. Zoals uit de definitie van de klassieke geadjungeerde van een matrix blijkt, bestaan de elementen van de matrix $B(t)$ uit de van geschikte tekens voorziene onderdeterminanten van $(tE - A)$ en dit zijn dus polynomen in t met hoogstens de graad $n - 1$ zie (6.3) en (7.1) ; één van hen is bijv.

$$\begin{vmatrix} t - a_{22} & \dots & \dots & a_{2n} \\ a_{32} & t - a_{33} & & \\ \vdots & & \ddots & a_{n-1,n} \\ a_{n2} & \dots & \dots & t - a_{nn} \end{vmatrix}.$$

Deze onderdeterminanten zijn dus veeltermen in t met hoogstens de graad $(n - 1)$. Dit betekent echter dat $B(t)$ te schrijven is in de gedaante

$$B(t) = B_{n-1}t^{n-1} + B_{n-2}t^{n-2} + \cdots + B_1 t + B_0$$

waarbij de $B_i (i = 0, 1, \dots, n - 1)$ $n \times n$-matrices zijn waarin t niet voorkomt. Ter verduidelijking een voorbeeld. Stel $n = 3$ en

$$B(t) = \begin{pmatrix} t^2 + 4 & t^2 + 3t + 1 & t + 1 \\ 5t^2 + 2t + 3 & 2t^2 - 4 & 5 \\ t^2 + 2t & 3t^2 - 3t + 1 & 4t^2 + 4t + 5 \end{pmatrix}$$

Er geldt dan:

$$B(t) = t^2 \begin{pmatrix} 1 & 1 & 0 \\ 5 & 2 & 0 \\ 1 & 3 & 4 \end{pmatrix} + t \begin{pmatrix} 0 & 3 & 1 \\ 2 & 0 & 0 \\ 2 & -3 & 4 \end{pmatrix} + \begin{pmatrix} 4 & 1 & 1 \\ 3 & -4 & 5 \\ 0 & 1 & 5 \end{pmatrix}.$$

Nu het vervolg van het bewijs.
Op grond van (7.3) geldt:

$$B(t) \cdot (tE - A) = |tE - A| \cdot E.$$

Het linkerlid hiervan laat zich schrijven als

$$(B_{n-1}t^{n-1} + B_{n-2}t^{n-2} + \cdots + B_1 t + B_0)(tE - A) =$$

$$= B_{n-1}t^n + (B_{n-2} - B_{n-1}A)t^{n-1} +$$

$$(B_{n-3} - B_{n-2}A)t^{n-2} + \cdots (B_2 - B_1 A)t - B_0 A.$$

Het rechterlid wordt op grond van (7.2)

$$t^n E + a_{n-1}t^{n-1}E + \cdots a_1 tE + a_0 E,$$

waarin de $a_i (i = 0, 1, 2, \dots, n - 1)$ scalairen uit het grondlichaam zijn.
Gelijkstelling van overeenkomstige coëfficiënten (links en rechts) geeft:

$$\begin{aligned} B_{n-1} &= E \\ B_{n-2} - B_{n-1}A &= a_{n-1}E \\ B_{n-3} - B_{n-2}A &= a_{n-2}E \\ \dots\dots\dots\dots\dots\dots & \quad \dots\dots\dots\dots\dots \\ B_0 - B_1 A &= a_1 E \\ -B_0 A &= a_0 E \end{aligned}$$

We vermenigvuldigen elk van deze gelijkheden in linker- en rechterlid (rechts) met resp. $A^n, A^{n-1}, \ldots, A, E$.
Er ontstaat dan:

$$\begin{aligned} B_{n-1}A^n &= A^n \\ B_{n-2}A^{n-1} - B_{n-1}A^n &= a_{n-1}A^{n-1} \\ B_{n-3}A^{n-2} - B_{n-2}A^{n-1} &= a_{n-2}A^{n-2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \quad \ldots\ldots\ldots\ldots\ldots\ldots \\ B_0A - B_1A^2 &= a_1A \\ -B_0A &= a_0E \end{aligned}$$

Optelling geeft dan

$$0 = A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \cdots a_1A + a_0E$$

d.w.z. A voldoet aan de karakteristieke vergelijking (in matrixvorm)

$$t^n + a_{n-1}t^{n-1} + a_{n-2}t^{n-2} + \cdots + a_1t + a_0E = N$$

Een belangrijk gevolg van deze stelling is:
Elke positieve gehele macht van de $n \times n$-matrix A kan geschreven worden als een lineaire combinatie van de matrices $E, A, A^2, \ldots, A^{n-1}$ met scalaire coëfficiënten.
Immers

$$A^n = -a_{n-1}A^{n-1} - a_{n-2}A^{n-2} - \cdots - a_0E \tag{7.4}$$

$$A^{n+1} = -a_{n-1}A^n - a_{n-2}A^{n-1} - \cdots - a_0A,$$

maar hierin laat A^n zich via (7.4) weer uitdrukken in lagere machten van A etc.

8. Vectorruimten met een inwendig produkt

A. Een bijzondere plaats wordt ingenomen door die vectorruimten waarop een zgn. inwendig produkt is gedefinieerd, ook wel inprodukt-ruimte genoemd.
We beperken ons hier weer tot vectorruimten over $\mathbb{R}$ (het lichaam van de reële getallen) of $\mathbb{C}$ (het lichaam van de complexe getallen).
Onder een inwendig produkt op een vectorruimte V over K verstaan we een afbeelding van $V \times V$ in K, d.w.z. dat aan ieder paar vectoren u, v in V een scalar in K is toegevoegd, genoteerd als $< u, v >$.
Hierbij wordt geeist:

1. $< u, v >= \overline{< v, u >}^*$ (8.1)

2. $< u, u > \geq 0$ en $< u, u >= 0$ *d.e.s.d.* als $u = 0$. (8.2)

   N.B $< u, u >= \overline{< u, u >}$ en dus reëel!

3. $< \alpha u + \beta v, w >= \alpha < u, w > + \beta < v, w >$ (8.3)

$^*\overline{< v, u >}$ is de toegevoegd complexe van $< v, u >$.
Voor een reële vectorruimte betekent dat dus symmetrie.

Deze laatste eis heeft een consequentie voor $< u, \alpha v + \beta w >$. Hiervoor geldt volgens (1) en (3):

$$< u, \alpha v + \beta w >= \overline{< \alpha v + \beta w, u >} = \overline{< \alpha v, u > + < \beta w, u >} =$$

$$\overline{\alpha < v, u >} + \overline{\beta < w, u >} = \overline{\alpha} < u, v > + \overline{\beta} < u, w > .$$

I.h.b. geldt dus: $< u, \alpha v >= \overline{\alpha} < u, v >$.

Een reële inproduktruimte noemt men een euclidische ruimte, een complexe inproduct-ruimte noemt men een unitaire ruimte.
Men kent aan een vector u een norm (lengte) toe, genoteerd als $||u||$ en gedefinieerd door

$$||u|| = \sqrt{< u, u >}.$$

Met behulp hiervan definieert men ook de afstand $d(u, v)$ tussen twee vectoren u en v, aldus:

$$d(u, v) = ||u - v||.$$

Hiervoor geldt:

$$d(u, v) \geq 0 \text{ en } d(u, v) = 0 \; d.e.s.d. \text{ als } u = v$$

$$d(u, v) = d(v, u)$$

$$d(u, v) \leq d(u, w) + d(w, v)$$

voor alle v, u en w in V.
Op grond hiervan is een inprodukt-ruimte een metrische ruimte.

Van groot belang is de ongelijkheid van Cauchy-Schwartz*

$$| < u, v > | \leq ||u||. ||v||.$$

Deze stelt ons in staat de hoek Θ tussen twee van nul verschillende vectoren te definiëren door

$$\cos \Theta = \frac{| < u, v > |}{||u||.||v||}.$$

Inderdaad is dan Θ te bepalen daar

$$-1 \leq \cos \Theta \leq 1$$

Indien $< u, v >= 0$ dan is $\cos \Theta = 0$ en men zegt dan u en v loodrecht op elkaar staan.

B. Orthonormale stelsels
Men noemt de verzameling $\{u_1, u_2, \ldots, u_t\}$ van vectoren in V een orthogonaal stelsel indien

*A.L. Cauchy (1789-1857); H.A. Schwartz (1843-1921)

$$< u_i, u_j >= 0 \qquad \text{voor } i \neq j.$$

Heeft elke vector bovendien de lengte 1, dus

$$< u_i, u_i >= 1 \qquad i = 1, 2, \ldots t$$

dan heet dit stel vectoren orthonormaal. Met behulp van het Kroneckersymbool δ_{ij} ($\delta_{ij} = 0$ als $i \neq j$ en $\delta_{ii} = 1$) kan men orthonormaliteit karakteriseren door $< u_i, u_j >= \delta_{ij}$.

Van groot belang is het feit dat men in een eindig dimensionale inproduktruimte altijd een orthonormale basis kan kiezen en wel op grond van het *orthogonalisatie-proces van Gram-Schmidt.**
Dit geeft n.l. een constructieve methode om uit een basis van n onafhankelijke vectoren een basis van n onafhankelijke orthonormale vectoren te construeren. De overgangsmatrix P blijkt daar de volgende driehoeks vorm te hebben:

$$P = \begin{pmatrix} p_{11} & 0 & \cdot & \cdots & 0 \\ p_{21} & p_{22} & 0 & & \cdot \\ p_{31} & p_{32} & p_{33} & & \cdot \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{pmatrix}.$$

Wanneer men in een inproduktruimte een orthonormale basis gekozen heeft, dan krijgt het inwendig produkt van de vectoren u en v een bijzondere gedaante, de zgn. standaardgedaante, immers als $\{e_1, e_2, \ldots, e_n\}$ een orthonormale basis is en

$$u = u_1 e_1 + u_2 e_2 + \ldots + u_n e_n$$

$$v = v_1 e_1 + v_2 e_2 + \ldots + v_n e_n$$

met $u_i, v_i \in K$ $(i = 1, 2, \ldots, n)$ dan geldt op grond van (8.3) en omdat $< e_i, e_j >= \delta_{ij}$:

$$< u, v >= u_1 \overline{v}_1 + u_2 \overline{v}_2 + \ldots + u_n \overline{v}_n.$$

In het reële geval:

$$< u, v >= u_1 v_1 + u_2 v_2 + \ldots + u_n v_n.$$

C. Enkele bijzondere typen matrices
Allereerst enkele definities:
Onder de *getransponeerde* A^t van een matrix A verstaat men de matrix die uit A ontstaat door spiegeling om de hoofddiagonaal.
Onder de *geadjungeerde* A^* van matrix A verstaat men de matrix die uit A ontstaat door spiegeling om de hoofddiagonaal en dan alle elementen te vervangen door hun toegevoegd complexen.
Zie ook het voorbeeld op bldz. 18.
Een *symmetrische* matrix is een matrix A waarvoor geldt $A^t = A$, dus bijv.

$$\begin{pmatrix} 1 & 4 & 5 \\ 4 & 2 & 6 \\ 5 & 6 & 3 \end{pmatrix}.$$

*J.P. Gram (1850-1916); E. Schmidt (1876-1959).

Een zelfgeadjungeerde matrix is een matrix A waarvoor geldt $A^* = A$, dus bijv.
$\begin{pmatrix} 1 & 4i & 5+i \\ -4i & 2 & 6-2i \\ 5-i & 6+2i & 3 \end{pmatrix}$.

Een *normale* matrix is een matrix A waarvoor geldt

$$AA^* = A^*A$$

Een *orthogonale matrix* is een (reële) matrix A waarvoor geldt $A^t = A^{-1}$, dus $A^tA = AA^t = E$. Gelijkwaardig met deze definitie is de eigenschap dat zowel de n rij-vectoren als de n kolomvectoren een orthonormaal stel vectoren vormen. Dus als

$$A = (a_{ij}),$$

dan geldt:

$$\sum_{i=1}^{n} a_{ij}a_{ik} = \delta_{jk} \text{ en } \sum_{i=1}^{n} a_{ji}a_{ki} = \delta_{jk}.$$

Een *unitaire matrix* is een (complexe) matrix A waarvoor geldt $A^* = A^{-1}$, dus $A^*A = AA^* = E$. Hiermee is gelijkwaardig dat zowel de n rij vectoren als de n kolomvectoren een orthonormaal stelsel vormen, hetgeen inhoudt:

$$\sum_{i=1}^{n} a_{ij}\overline{a}_{ik} = \delta_{jk} \text{ en } \sum_{i=1}^{n} a_{ji}\overline{a}_{ki} = \delta_{jk}.$$

D. Overgang van de ene orthonormale basis naar de andere orthonormale basis, nogmaals diagonalisatie.
Hierbij geldt de stelling dat de optredende overgangsmatrix (zie bldz. 4) orthogonaal resp. unitair is, al naar dat $K = \mathbb{R}$ of $K = \mathbb{C}$. Omgekeerd geldt ook dat, wanneer men uitgaande van een orthonormale basis op een nieuwe basis overgaat via een orthogonale, resp. unitaire overgangsmatrix, de nieuwe basis eveneens orthonormaal is.
Tot slot een tweetal stellingen die betrekking hebben op diagonalisatie van matrices met reële, resp. complexe coëfficiënten.

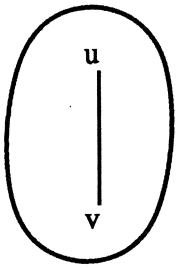
1. Als A een *reële symmetrische* matrix is dan bestaat er een *orthogonale* matrix P zodanig dat $P^{-1}AP = P^tAP$ een diagonaalmatrix is
2. Als A een *complexe normale* matrix is dan bestaat er een *unitaire* matrix P zodanig dat $P^{-1}AP = P^*AP$ een diagonaalmatrix is.

E. Convexe verzamelingen
In één van de voordrachten zal het begrip "convexe verzameling" genoemd en gebruikt worden.
Dit begrip kan gemakkelijk aanschouwelijk duidelijk gemaakt worden:

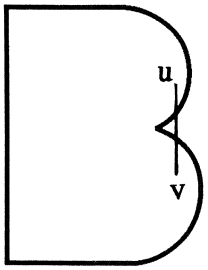
Exacter en voor uitbreiding vatbaar: Een deelruimte W van een vectorruimte V noemt men convex indien bij willekeurige u en v behorende tot W, geldt dat het segment, gedefinieerd door

$$w = u + \lambda\ (v - u) \qquad 0 \leq \lambda \leq 1$$

geheel tot W behoort.



convex



niet convex

LITERATUUR

1. JOHN W. DETTMAN, Introduction to Linear Algebra and Differential Equations. Mc Graw-Hill, New York etc., 1974.

2. P.R. HALMOS, Finite Dimensional Vectorspaces, D. Van Nostrand Company Inc., Princeton etc., 1958.

3. N.H. KUIPER, Analytische meetkunde verklaard met Lineaire Algebra, Noord-Hollandsche Uitg. Mij, Amsterdam 1959.

4. SEYMOUR LIPSCHUTZ, Linear Algebra, Schaum's Outline Series, McGraw-Hill, New York etc., 1968.

5. L. MIRSKY, An Introduction to Linear Algebra, Clarendon Press, Oxford, 1963.

6. G.R. VELDKAMP-F. SCHUH, Lineaire Algebra en Analytische Meetkunde I, N.V.W.J. Thieme & C. Zuthphen, 1961.

Inleiding gewone differentiaalvergelijkingen

Henri Huijberts
Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven
Postbus 513
5600 MB Eindhoven
Telefoon: 040-474230

20 juli 1992

1 Inleiding

Dit dictaat is bedoeld als een korte inleiding op de theorie van gewone differentiaalvergelijkingen. De aandacht zal voornamelijk worden gericht op lineaire differentiaalvergelijkingen en stabiliteitstheorie. Bewijzen worden achterwege gelaten. Hiervoor wordt verwezen naar de referenties aan het eind van het dictaat.

De lezer wordt geacht bekend te zijn met elementaire begrippen uit de matrixtheorie (matrix en vector, matrixvermenigvuldiging, determinant en inverse van een matrix), lineaire algebra (lineaire (on)afhankelijkheid, basis), calculus (functie, continuïteit, differentieerbaarheid, (partiële) afgeleide, kromme, raakvector), complexe getallen (definitie van complex getal, optellen, vermenigvuldigen, complex toegevoegde) en de theorie van ééndimensionale gewone differentiaalvergelijkingen.

2 Gewone differentiaalvergelijkingen

Een *gewone differentiaalvergelijking (DV)* is een stelsel vergelijkingen van de vorm

$$\begin{aligned} \frac{dx_1}{dt} = \dot{x}_1(t) &= f_1(x_1(t), \cdots, x_n(t)) \\ &\vdots \\ \frac{dx_n}{dt} = \dot{x}_n(t) &= f_n(x_1(t), \cdots, x_n(t)) \end{aligned} \tag{1}$$

waar $x_1, \cdots, x_n \in \mathbb{R}$. De onafhankelijke variabele t wordt vaak geïnterpreteerd als de tijd. We zullen in het vervolg aannemen dat voor $i = 1, \cdots, n$ en $j = 1, \cdots, n$ de partiële afgeleiden $(\partial f_i / \partial x_j)$ bestaan en continu zijn (ofwel: de functies $f_1, \cdots, f_n : \mathbb{R}^n \mapsto \mathbb{R}$ zijn *continu differentieerbaar*). Definieer de vectoren

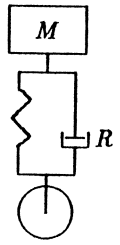
$$x := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad f(x) := \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix}$$

Dan kunnen we de DV (1) schrijven in de verkorte vorm

$$\dot{x} = f(x) \tag{2}$$

met $x \in \mathbb{R}^n$ en f een afbeelding van $\mathbb{R}^n$ naar $\mathbb{R}^n$.

We kunnen ons bij (1) de volgende voorstelling maken. Beschouw een (fysisch) systeem waarvan de eigenschappen op een bepaald tijdstip t



Figuur 1: Vering en schokdemping van een auto

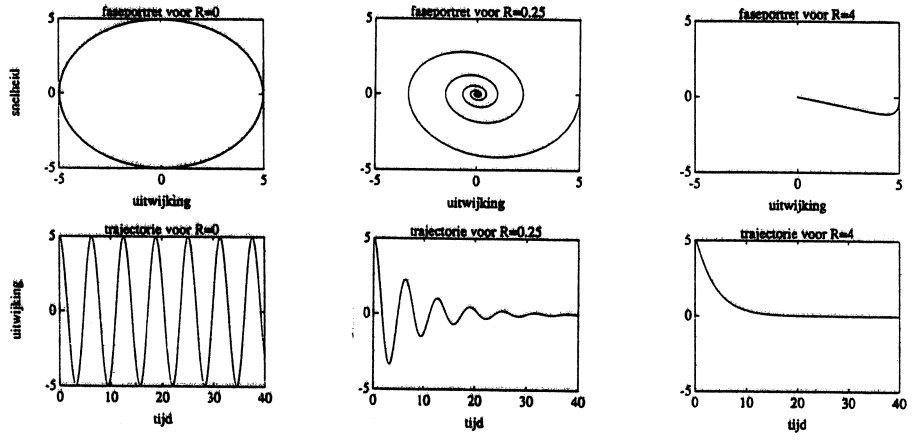
worden gekenmerkt door de variabelen $x_1(t), \cdots, x_n(t) \in \mathbb{R}$. De vector $x(t) = \text{col}(x_1(t), \cdots, x_n(t)) \in \mathbb{R}^n$ wordt de *toestand* van het systeem op tijdstip t genoemd, en $\mathbb{R}^n$ wordt de *toestandsruimte* (ook wel *faseruimte*) genoemd. Door de DV wordt in elk punt van $\mathbb{R}^n$ een richting gegeven, namelijk de vector $f(x)$. Deze vector geeft aan in welke richting het systeem "beweegt", ofwel wat de snelheid van het systeem is. De familie van alle richtingen vormt het *richtingsveld* van de DV. Een oplossing van de DV is dan een kromme in $\mathbb{R}^n$ die "past" in het richtingsveld, d.w.z. een kromme waarvan in elk punt de raakvector de richting $f(x)$ heeft. Formeel levert dit:

Definitie 2.1 Beschouw de DV (1). Een continu differentieerbare functie $\xi : \mathbb{R} \mapsto \mathbb{R}^n$ heet een *oplossing* van de DV als

$$\forall t \in \mathbb{R} : \quad \frac{d\xi}{dt}(t) = f(\xi(t)) \tag{3}$$

Voorbeeld 2.2 Beschouw het (sterk vereenvoudigde) model van de vering en schokdemping van een auto dat wordt gegeven in Figuur 1.
We nemen even aan dat de auto slechts één wiel heeft en een massa $M = 1$ kg. Neem verder aan dat de veerconstante gelijk is aan 1 N/m en de dempingsconstante van de schokdemper gelijk aan R $N/(m/s)$. Verder vergeten we even voor het gemak dat er ook nog zoiets als zwaartekracht bestaat. Uit de mechanica weten we dat het gedrag van een mechanisch systeem geheel wordt beschreven door de positie en de (verticale) snelheid van M. We nemen dus

x_1 = verticale positie van M
x_2 = verticale snelheid van M



Figuur 2: Gedrag van de oplossingen voor $R = 0$, $R = 0.25$ en $R = 4$

Gebruik makend van de Tweede Wet van Newton levert dit het volgende stelsel differentiaalvergelijkingen:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 - Rx_2 \end{aligned} \tag{4}$$

Het resulterende gedrag van de oplossingen voor $R = 0$, $R = 0.25$ en $R = 4$ wordt gegeven in Figuur 2.
Hieruit blijkt dat de situaties $R = 0$ (zeeziekte) en $R = 4$ (deuken in het plafond) niet echt comfortabel zijn.

Belangrijke problemen in de theorie van differentiaalvergelijkingen zijn zogenaamde *beginwaardeproblemen (BWP)*. Bij dit soort problemen zoeken we een oplossing van de DV als gegeven is dat de toestand $x(t_0)$ op een begintijdstip $t_0 \in \mathbb{R}$ gelijk is aan $x_0 \in \mathbb{R}^n$. We noteren dit als volgt:

$$\text{BWP}(f, t_0, x_0) \begin{cases} \dot{x} = f(x) \\ x(t_0) = x_0 \end{cases} \tag{5}$$

De voorwaarde dat voor $i = 1, \cdots, n$ de functies $f_i(x)$ in (1) continu differentieerbaar zijn, garandeert dat voor iedere $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$ het BWP(f, t_0, x_0) precies één oplossing heeft. In het algemeen zal het echter niet mogelijk zijn de oplossingen expliciet te geven. In dit geval kunnen we twee dingen doen. In de eerste plaats kunnen we proberen met behulp van methoden uit de numerieke wiskunde benaderende oplossingen te genereren. Ten tweede kunnen we proberen zo veel mogelijk informatie over het gedrag van de oplossingen te verkrijgen zonder deze oplossingen ook expliciet te berekenen. Men spreekt dan van *kwalitatieve eigenschappen* van de oplossingen. Eén zo'n kwalitatieve eigenschap is de stabiliteit van oplossingen (ofwel het gedrag van de oplossingen als $t \to +\infty$). Om dit concept te definiëren, hebben we nog het volgende nodig. We noemen een punt $x_0 \in \mathbb{R}^n$ een *evenwichtspunt* voor de DV (1) als $f(x_0) = 0$. Merk op dat dit betekent dat de oplosing van het BWP(f, t_0, x_0) wordt gegeven door $x(t) = x_0$ $(\forall t \geq 0)$.

Definitie 2.3 Stabiliteit in de zin van Lyapunov
Beschouw de DV (1) en laat $x_0 \in \mathbb{R}^n$ een evenwichtspunt zijn, ofwel $f(x_0) = 0$. Geef voor $\bar{x} \in \mathbb{R}^n$ de oplossing van het BWP$(f, 0, \bar{x})$ aan met $\xi(t, \bar{x})$. Dan heet x_0

(i) stabiel als

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall \bar{x}, \|\bar{x} - x_0\| < \delta)$$

$$(\xi(t, \bar{x}) \text{ bestaat voor alle } t \geq 0 \text{ en}$$

$$\|\xi(t, \bar{x}) - x_0\| < \epsilon \ (\forall t \geq 0))$$

(ii) asymptotisch stabiel als x_0 stabiel is en bovendien

$$(\exists \delta_1 > 0)(\forall \bar{x}, \|\bar{x} - x_0\| < \delta_1)(\|\xi(t, \bar{x}) - x_0\| \to 0 \ (t \to +\infty))$$

(iii) instabiel als x_0 niet stabiel is, ofwel als

$$(\exists \epsilon > 0)(\forall \delta > 0)(\exists \bar{x}, \|\bar{x} - x_0\| < \delta)$$

$$(\xi(t, \bar{x}) \text{ bestaat niet voor alle } t \geq 0 \text{ of}$$

$$\exists t \geq 0 \text{ waarvoor } \|\xi(t, \bar{x}) - x_0\| > \epsilon)$$

Stabiliteit betekent dus ruwweg dat oplossingen die starten in de buurt van het evenwichtspunt ook in de buurt van dit evenwichtspunt blijven. Bij asymptotische stabiliteit geldt bovendien dat voor $t \to \infty$ de oplossing willekeurig dicht bij het evenwichtspunt komt. Het volgende voorbeeld illustreert het één en ander.

Voorbeeld 2.4 *(i)* Beschouw de DV $\dot{x} = x^2$. Het is duidelijk dat $x = 0$ een evenwichtspunt is. Kies nu $x(0) = \epsilon > 0$. De oplossing wordt dan gegeven door

$$x(t) = \frac{1}{\epsilon - t}$$

We zien dat $x(t)$ niet bestaat voor $t = \epsilon$ (de oplossing "ontploft" als $t \to \epsilon$). We concluderen dat $x = 0$ een instabiel evenwichtspunt is.

(ii) Beschouw de DV $\dot{x} = x$. In dit geval is $x = 0$ weer een evenwichtspunt. Laat weer $x(0) = \epsilon > 0$. Dan wordt de oplossing gegeven door

$$x(t) = \epsilon e^t$$

Alhoewel nu de oplossing niet ontploft voor eindige t, geldt wel dat $x(t) \to +\infty$ $(t \to +\infty)$, en dus is $x = 0$ een instabiel evenwichtspunt.

(iii) Beschouw nu $\dot{x} = -x$. $x = 0$ is weer een evenwichtspunt en met $x(0) = \epsilon > 0$ vinden we de oplossing

$$x(t) = \epsilon e^{-t}$$

We zien nu dat $x(t) \leq \epsilon$ $(\forall t \geq 0)$ en dat $x(t) \to 0$ $(t \to +\infty)$. In dit geval is $x = 0$ dus een asymptotisch stabiel evenwichtspunt.

3 Tweedimensionale stelsels

We beschouwen in deze sectie DV's van de vorm

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2) \\ \dot{x}_2 &= f_2(x_1, x_2) \end{aligned} \tag{6}$$

of, met $x = (x_1\ x_2)^T \in I\!R^2$, $f = (f_1\ f_2)^T$,

$$\dot{x} = f(x) \tag{7}$$

We zullen altijd $t_0 = 0$ nemen en aannemen dat $x = 0$ een evenwichtspunt is, ofwel dat $f(0) = 0$.

Opmerking 3.1 Beide aannamen zijn zonder verlies van algemeenheid om de volgende redenen.

(i) Stel dat $t_0 \neq 0$ en laat $x(t)$ de oplossing zijn van het BWP$(f, t_0, \bar{x})$. Definieer nu $\xi(t) = x(t + t_0)$. Dan:

$$\begin{aligned} \dot{\xi}(t) &= \dot{x}(t + t_0) = f(x(t + t_0)) = f(\xi(t)) \\ \xi(0) &= x(t_0) = \bar{x} \end{aligned}$$

Dus $\xi(t)$ is de oplossing van het BWP$(f, 0, \bar{x})$.

(ii) Stel dat $x_0 \neq 0$ een evenwichtspunt is van (7). Definieer $\xi := x - x_0$. Dan

$$\dot{\xi} = \dot{x} = f(x) = f(\xi + x_0) =: \bar{f}(\xi) \tag{8}$$

en $\bar{f}(0) = f(x_0) = 0$. $\xi = 0$ is dus een evenwichtspunt van (8) en uit de definitie van ξ volgt dat de stabiliteitseigenschappen van $\xi = 0$ voor (8) dezelfde zijn als die van $x = x_0$ voor (7).

Verder nemen we weer aan dat f in (7) continu differentieerbaar is. Uit de Stelling van Taylor volgt dan dat we (6) kunnen schrijven als

$$\dot{x} = Ax + g(x) \tag{9}$$

waarbij de (2,2)-matrix A wordt gegeven door

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(0) & \frac{\partial f_1}{\partial x_2}(0) \\ \\ \frac{\partial f_2}{\partial x_1}(0) & \frac{\partial f_1}{\partial x_1}(0) \end{pmatrix}$$

en $g(x)$ voldoet aan

- $g(x)$ is continu.
- $g(0) = 0$.
- $\lim\limits_{\|x\| \to 0} \dfrac{\|g(x)\|}{\|x\|} = 0$ (ofwel $g(x) = o(x)$).

De laatste eis betekent dat $g(x)$ niet kan worden geschreven als $g(x) = Bx + \tilde{g}(x)$ voor de één of andere (2,2)-matrix $B \neq 0$, ofwel dat alle lineaire termen van f worden vertegenwoordigd in de term Ax in (9). In paragraaf 3.2 zal blijken dat de stabiliteitseigenschappen van (6) in belangrijke mate worden bepaald door de *gelineariseerde vergelijking*

$$\dot{x} = Ax \tag{10}$$

Daarom bestuderen we in paragraaf 3.1 eerst tweedimensionale stelsels van de vorm (10). In paragraaf 3.3 behandelen we een methode om het gedrag van de oplossingen van (10) grafisch weer te geven. In paragraaf 3.4 tenslotte, komen Lyapunovfuncties aan de orde. Deze functies geven een andere manier om stabiliteit na te gaan.

3.1 Lineaire tweedimensionale stelsels

Beschouw het lineaire tweedimensionale stelsel (10) en schrijf het in de vorm

$$\begin{aligned} \dot{x}_1 &= ax_1 + bx_2 \\ \dot{x}_2 &= cx_1 + dx_2 \end{aligned} \tag{11}$$

Vóór we in paragraaf 3.1.2 kunnen kijken naar het gedrag van de oplossingen van (11), zullen we eerst in paragraaf 3.1.1 wat concepten uit de lineaire algebra en de matrixtheorie introduceren.

3.1.1 Matrixtheorie en lineaire algebra

Beschouw de (2,2)-matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Het *karakteristieke polynoom* $p(\lambda)$ van A is gedefinieerd als (met I de (2,2)-eenheidsmatrix)

$$\begin{aligned} p(\lambda) = \det(\lambda I - A) = \det \begin{pmatrix} \lambda - a & -b \\ -c & \lambda - d \end{pmatrix} = \\ \lambda^2 - (a+d)\lambda + (ad - bc) \end{aligned} \tag{12}$$

De *eigenwaarden* van A zijn de oplossingen van de *karakteristieke vergelijking* $p(\lambda) = 0$. Met behulp van de (a, b, c)-formule vinden we dus dat de eigenwaarden van A worden gegeven door

$$\lambda_{1,2} = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad - bc)}}{2} \tag{13}$$

Merk op dat de eigenwaarden in principe complexe getallen zijn. Laat $\lambda_i \in \mathbb{C}$ een eigenwaarde van A zijn. Omdat $\det(\lambda_i I - A) = 0$, bestaat er een vector $v_i \in \mathbb{C}^2$, $v_i \neq 0$, met de eigenschap dat

$$(\lambda_i I - A)v_i = 0 \tag{14}$$

Zo'n vector wordt een *eigenvector* van A bij de eigenwaarde λ_i genoemd. Merk op dat als $v_i \in \mathbb{C}^2$ een eigenvector bij λ_i is, dat dan ook αv_i $(0 \neq \alpha \in \mathbb{C})$ een eigenvector bij λ_i is.

Propositie 3.2 *Beschouw een (2,2)-matrix A met eigenwaarden λ_1, λ_2.*

(i) Laat $\lambda_1, \lambda_2 \in \mathbb{R}$ met $\lambda_1 \neq \lambda_2$ en laat $v_1, v_2 \in \mathbb{R}^2$ eigenvectoren bij respectievelijk λ_1 en λ_2 zijn. Dan zijn v_1 en v_2 lineair onafhankelijk.

(ii) Als $\lambda_1 \in \mathbb{C}\backslash\mathbb{R}$, zeg $\lambda_1 = \alpha + i\beta$ met $\beta \neq 0$, dan geldt $\lambda_2 = \alpha - i\beta = \overline{\lambda}_1$. Verder bestaan er lineair onafhankelijke vectoren $r, s \in \mathbb{R}^2$ zodanig dat $v_1 = r + is$ een eigenvector bij λ_1 en $v_2 = r - is = \overline{v}_1$ een eigenvector bij λ_2 is. Merk op dat dit tegelijk betekent dat ook v_1 en v_2 lineair onafhankelijk zijn.

(iii) Als $\lambda_1 = \lambda_2 = \mu$, dan $\mu \in \mathbb{R}$. Verder:

a. Als $(\mu I - A) \neq 0$, dan bestaat er geen tweetal lineair onafhankelijke eigenvectoren bij μ. Echter, laat v een eigenvector bij μ zijn. Dan bestaat er een vector $w \in \mathbb{R}^2$ zodanig dat v en w lineair onafhankelijk zijn en bovendien

$$(\mu I - A)w + v = w \tag{15}$$

w wordt een gegeneraliseerde eigenvector *bij λ genoemd.*

b. Als $(\mu I - A) = 0$, dan bestaan er twee lineair onafhankelijke eigenvectoren bij μ. ■

We zullen in paragraaf 3.1.2 het stelsel (11) oplossen m.b.v. een zogenaamde *basistransformatie*. Als we ons een vector in $\mathbb{R}^2$ voorstellen als een punt, dan associeert een basistransformatie op een eenduidige manier een nieuw punt met ieder punt in $\mathbb{R}^2$. Een simpel voorbeeld van een basistransformatie is een rotatie van, zeg, 90° tegen de klok in om een punt. Een basistransformatie in $\mathbb{R}^2$ wordt gerepresenteerd door een inverteerbare (2,2)-matrix (zeg T). Als een vector t.o.v. de standaardbasis wordt gerepresenteerd door $(x_1\ x_2)^T$, dan representeert de nieuwe vector

$$\bar{x} = Tx \tag{16}$$

het resultaat van de transformatie. Een rotatie van 90° tegen de klok in rond de oorsprong wordt bijvoorbeeld gerepresenteerd door

$$T = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{17}$$

Beschouw de DV (10). We leiden nu een DV af voor $\bar{x}$ als gedefinieerd in (16). Gebruik makend van de kettingregel en het feit dat uit (16) volgt dat $x = T^{-1}\bar{x}$, vinden we

$$\dot{\bar{x}} = T\dot{x} = TAx = TAT^{-1}\bar{x} \tag{18}$$

3.1.2 Gedrag van oplossingen van lineaire tweedimensionale stelsels

Beschouw het lineaire tweedimensionale stelsel (10). We onderzoeken het gedrag van de oplossingen van dit stelsel. Hierbij onderscheiden we verschillende gevallen. We geven telkens de eigenwaarden van A aan met λ_1, λ_2.

Geval 1: $\lambda_1, \lambda_2 \neq 0$, $\lambda_1 \neq \lambda_2$ (dus ook $\lambda_1, \lambda_2 \in \mathbb{R}$).

Laat v_1, v_2 eigenvectoren bij respectievelijk λ_1 en λ_2 zijn. Uit Propositie 3.2 volgt dan dat v_1 en v_2 lineair onafhankelijk zijn. Daarom is de matrix $V = (v_1\ v_2)$ (de matrix met v_1 en v_2 als kolommen) inverteerbaar. Definieer nu

$$\bar{x} = V^{-1}x \tag{19}$$

Dan volgt uit (18) dat voor $\bar{x}$ de DV (10) overgaat in de DV

$$\dot{\bar{x}} = V^{-1}AV\bar{x} \tag{20}$$

Merk nu op dat uit $Av_i = \lambda_i v_i$ $(i = 1, 2)$ volgt dat

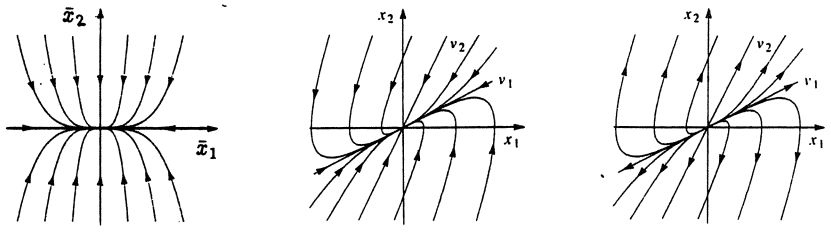
$$AV = V\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{21}$$

en dus

$$V^{-1}AV = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{22}$$

De DV (10) gaat voor $\bar{x}$ dus over in

$$\begin{aligned} \dot{\bar{x}}_1 &= \lambda_1\bar{x}_1 \\ \dot{\bar{x}}_2 &= \lambda_2\bar{x}_2 \end{aligned} \tag{23}$$



(a) (b) (c)

Figuur 3: Faseportretten voor stabiele knoop in $(\bar{x}_1, \bar{x}_2)$-vlak (a), stabiele knoop in (x_1, x_2)-vlak (b), instabiele knoop in (x_1, x_2)-vlak (c)

De oplossing van (23) kunnen we makkelijk geven. Kies beginwaarden $\bar{x}_1(0) = \bar{x}_{10}$, $\bar{x}_2(0) = \bar{x}_{20}$. Dan wordt de oplossing gegeven door:

$$\begin{aligned} \bar{x}_1(t) &= \bar{x}_{10} e^{\lambda_1 t} \\ \bar{x}_2(t) &= \bar{x}_{20} e^{\lambda_2 t} \end{aligned} \tag{24}$$

Door t uit (24) te elimineren krijgen we

$$\bar{x}_2 = c\bar{x}_1^{(\lambda_2/\lambda_1)} \tag{25}$$

waar $c = \bar{x}_{20}/\bar{x}_{10}^{(\lambda_2/\lambda_1)}$. Het *faseportret* van (23) bestaat uit de familie van alle oplossingskrommen van (23) in het $(\bar{x}_1, \bar{x}_2)$-vlak. Dit faseportret kan worden verkregen door in (25) de constante c willekeurige waarden in $\mathbb{R}$ aan te laten nemen. De vorm van het faseportret is afhankelijk van het teken van λ_1 en λ_2. Onderscheid daarom:

a. Beide eigenwaarden zijn negatief

Neem zonder verlies van algemeenheid aan dat $\lambda_2 < \lambda_1 < 0$. In dit geval gaan de termen $e^{\lambda_1 t}$ en $e^{\lambda_2 t}$ in (24) naar nul voor $t \to +\infty$. Bovendien zal de term $e^{\lambda_2 t}$ sneller naar nul gaan dan de term $e^{\lambda_1 t}$. Daarom zullen we λ_2 de *snelle* eigenwaarde en λ_1 de *langzame* eigenwaarde noemen. Voor later gebruik noemen we v_2 de *snelle* eigenvector en v_1 de *langzame* eigenvector. De oplossingstrajectoriën naderen nu dus de oorsprong van het $(\bar{x}_1, \bar{x}_2)$-vlak langs de curve (25). De helling van de curve wordt gegeven door

$$\frac{d\bar{x}_1}{d\bar{x}_2} = c\frac{\lambda_2}{\lambda_1}\bar{x}_1^{(\lambda_2/\lambda_1 - 1)} \tag{26}$$

Omdat $(\lambda_2/\lambda_1 - 1)$ positief is, gaat de helling naar nul als $\bar{x}_1 \to 0$ en naar ∞ als $\bar{x}_1 \to \infty$. Daarom raken de trajectoriën de $\bar{x}_1$-as

(a) (b)

Figuur 4: Faseportretten voor zadelpunt in $(\bar{x}_1, \bar{x}_2)$-vlak (a) en in (x_1, x_2)-vlak (b)

als ze naar de oorsprong gaan en worden ze parallel aan de $\bar{x}_2$-as als ze naar oneindig gaan. Uit deze overwegingen kunnen we in het $(\bar{x}_1, \bar{x}_2)$-vlak het typische faseportret tekenen dat in Figuur 3.a wordt gegeven. In Figuur 3.b is het faseportret in het (x_1, x_2)-vlak gegeven. Dit faseportret verkrijgen we door het faseportret in het $(\bar{x}_1, \bar{x}_2)$-vlak terug te transformeren via $x = V\bar{x}$. Dit betekent dat de lijn opgespannen door v_1 de rol overneemt van de $\bar{x}_1$-as en de lijn opgespannen door v_2 de rol overneemt van de $\bar{x}_2$-as: Figuur 3.a wordt op sommige plaatsen "uitgerekt" en op andere plaatsen "ingedrukt". Merk verder op dat de trajectoriën raken aan de snelle eigenvector v_1 als ze de oorsprong naderen en parallel zijn aan de langzame eigenvector v_2 als ze ver van de oorsprong zijn. Het evenwichtspunt $x = 0$ wordt in dit geval een *stabiele knoop* genoemd.

b. Beide eigenwaarden zijn negatief

In dit geval behoudt het faseportret het karakter van Figuur 3.b, maar met omgekeerde richtingen omdat de termen $e^{\lambda_1 t}$, $e^{\lambda_2 t}$ nu exponentieel groeien als t groeit. Figuur 3.c geeft het faseportret in het (x_1, x_2)-vlak voor het geval dat $\lambda_2 > \lambda_1 > 0$. Het evenwichtspunt wordt nu een *instabiele knoop* genoemd.

c. De eigenwaarden hebben tegengesteld teken

Neem zonder verlies van algemeenheid aan dat $\lambda_2 < 0 < \lambda_1$. In dit geval hebben we dat $e^{\lambda_1 t} \to \infty$, $e^{\lambda_2 t} \to 0$ als $t \to +\infty$. De trajectorievergelijking (25) heeft nu een negatieve exponent (λ_2/λ_1). Daarom heeft het faseportret in het $(\bar{x}_1, \bar{x}_2)$-vlak de typische vorm die wordt getoond in Figuur 4.a. Het faseportret in het (x_1, x_2)-vlak wordt gegeven in Figuur 4.b. In dit geval wordt het

evenwichtspunt een *zadelpunt* genoemd.

Geval 2: Complexe eigenwaarden.

Laat $\lambda_1 = \alpha + i\beta$, $\lambda_2 = \alpha - i\beta$ met $\beta \neq 0$. Uit Propositie 3.2 volgt dat er lineair onafhankelijke vectoren $r, s \in \mathbb{R}^2$ bestaan zodanig dat $v_1 = r + is$, $v_2 = r - is$ eigenvectoren bij respectievelijk λ_1 en λ_2 zijn. Vorm de matrix $V = (r\ s)$ en definieer $\bar{x} = V^{-1}x$. Dan gaat de DV (3.4) voor $\bar{x}$ over in

$$\dot{\bar{x}} = V^{-1}AV\bar{x} \tag{27}$$

Nu volgt uit

$$Av_1 = A(r + is) = Ar + iAs$$

en

$$Av_1 = (\alpha + i\beta)(r + is) = (\alpha r - \beta s) + i(\beta r + \alpha s)$$

dat

$$\begin{array}{rcl} Ar & = & (\alpha r - \beta s) \\ As & = & (\beta r + \alpha s) \end{array} \tag{28}$$

Uit (28) volgt

$$AV = V\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \Rightarrow V^{-1}AV = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \tag{29}$$

en dus gaat voor $\bar{x}$ de DV (10) over in de DV

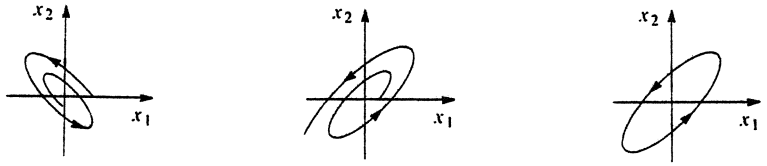
$$\dot{\bar{x}} = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}\bar{x} \tag{30}$$

Beschouw nu een punt $\bar{x} = (\bar{x}_1, \bar{x}_2) \in \mathbb{R}^2$. Laat r de afstand van $\bar{x}$ tot de oorsprong zijn en ϕ de hoek die de lijn door de oorsprong en $\bar{x}$ met de positieve $\bar{x}_1$-as maakt. Dan:

$$\begin{array}{rcl} \bar{x}_1 & = & r\cos\phi \\ \bar{x}_2 & = & r\sin\phi \end{array} \tag{31}$$

De nieuwe coördinaten (r, ϕ) worden *poolcoördinaten* genoemd. We zullen nu uit (28) een DV voor r en ϕ afleiden. Met behulp van de kettingregel vinden we

$$\begin{pmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \end{pmatrix} = \begin{pmatrix} \dot{r}\cos\phi - r\dot{\phi}\sin\phi \\ \dot{r}\sin\phi + r\dot{\phi}\cos\phi \end{pmatrix} = \begin{pmatrix} \cos\phi & r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix}\begin{pmatrix} \dot{r} \\ \dot{\phi} \end{pmatrix} \tag{32}$$



(a) (b) (c)

Figuur 5: Faseportretten voor stabiele spiraal (a), instabiele spiraal (b), centrum (c) in (x_1, x_2)-vlak

en dus met (28)

$$\begin{pmatrix} \dot{r} \\ \dot{\phi} \end{pmatrix} = \begin{pmatrix} \cos\phi & r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix}^{-1} \begin{pmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \end{pmatrix} =$$

$$\frac{1}{r}\begin{pmatrix} r\cos\phi & r\sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}\begin{pmatrix} r\cos\phi \\ r\sin\phi \end{pmatrix} =$$

$$= \cdots = \begin{pmatrix} \alpha r \\ \beta \end{pmatrix}$$

Kies beginwaarden $r(0) = r_0$ en $\phi(0) = \phi_0$. Dan volgt uit het bovenstaande:

$$\begin{aligned} r(t) &= r_0 e^{\alpha t} \\ \phi(t) &= \phi_0 + \beta t \end{aligned} \tag{33}$$

(33) definieert een spiraal in het $(\bar{x}_1, \bar{x}_2)$-vlak. Afhankelijk van de waarde van α zullen de trajectorieën één van de drie vormen die worden getoond in Figuur 5 aannemen. Als $\alpha < 0$ convergeert de spiraal naar de oorsprong, als $\alpha > 0$ beweegt de spiraal zich van de oorsprong af, en als $\alpha = 0$, dan zijn de trajectoriën cirkels met straal r_0. Het evenwichtspunt heet een *stabiel spiraalpunt* als $\alpha < 0$, een *instabiel spiraalpunt* als $\alpha > 0$ en een *centrum* als $\alpha = 0$.

Geval 3: Meervoudige eigenwaarden ongelijk aan nul, $\lambda_1 = \lambda_2 = \mu \neq 0$.

Onderscheid de volgende gevallen.

a. $(\mu I - A) = 0$.

(a) (b)

Figuur 6: Faseportretten voor stabiele ster (a) en instabiele ster (b) in $(\bar{x}_1, \bar{x}_2)$-vlak

De DV (10) wordt nu

$$\begin{aligned} \dot{x}_1 &= \mu x_1 \\ \dot{x}_2 &= \mu x_2 \end{aligned} \tag{34}$$

met als oplossing (met $x_1(0) = x_{10}, x_2(0) = x_{20}$)

$$\begin{aligned} x_1(t) &= x_{10} e^{\mu t} \\ x_2(t) &= x_{20} e^{\mu t} \end{aligned} \tag{35}$$

Elimineren van t uit (35) geeft

$$x_1 = x_2 \frac{x_{10}}{x_{20}} \tag{36}$$

Dit levert de faseportretten in Figuur 6. Het evenwichtspunt wordt voor $\mu < 0$ een *stabiele ster* genoemd en voor $\mu > 0$ een *instabiele ster.*

b. $(\mu I - A) \neq 0$.

Volgens Propositie 3.2 bestaat er een eigenvector v bij μ en een vector w zodanig dat v en w lineair onafhankelijk zijn en

$$Aw = \mu w + v \tag{37}$$

Vorm de matrix $V = (v\ w)$. Dan volgt uit het feit dat $Av = \mu v$ en uit (37) dat

$$AV = (\mu v\ \mu w + v) = V \begin{pmatrix} \mu & 1 \\ 0 & \mu \end{pmatrix} \tag{38}$$

(a) (b)

Figuur 7: Faseportretten voor stabiele oneigenlijke knoop (a) en instabiele oneigenlijke knoop (b) in $(\bar{x}_1, \bar{x}_2)$-vlak

en dus

$$V^{-1}AV = \begin{pmatrix} \mu & 1 \\ 0 & \mu \end{pmatrix} \tag{39}$$

Definieer $\bar{x} = V^{-1}x$. Dan gaat (10) voor $\bar{x}$ over in

$$\begin{aligned} \dot{\bar{x}}_1 &= \mu\bar{x}_1 + \bar{x}_2 \\ \dot{\bar{x}}_2 &= \mu\bar{x}_2 \end{aligned} \tag{40}$$

Kies $\bar{x}_1(0) = \bar{x}_{10}$, $\bar{x}_2(0) = \bar{x}_{20}$. De oplossing van (40) wordt dan gegeven door

$$\begin{aligned} \bar{x}_1(t) &= (\bar{x}_{10} + \bar{x}_{20}t)e^{\mu t} \\ \bar{x}_2(t) &= \bar{x}_{20}e^{\mu t} \end{aligned} \tag{41}$$

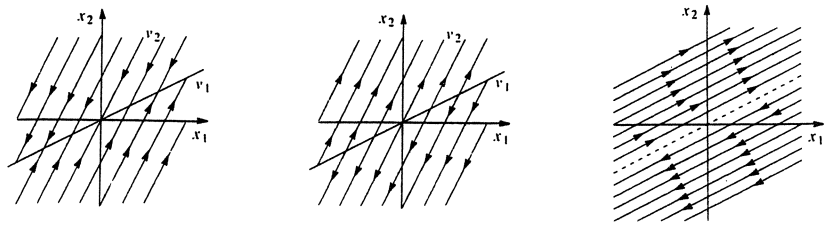
Als we t uit (41) elimineren krijgen we de vergelijking

$$\bar{x}_1 = \bar{x}_2\left[\frac{\bar{x}_{10}}{\bar{x}_{20}} + \frac{1}{\mu}\ln\left(\frac{\bar{x}_2}{\bar{x}_{20}}\right)\right] \tag{42}$$

Figuur 7 geeft de bijbehorende faseportretten. Het evenwichtspunt $x = 0$ wordt voor $\mu < 0$ een *stabiele oneigenlijke knoop* genoemd en voor $\mu > 0$ een *instabiele oneigenlijke knoop.*

Geval 4: Eén of beide eigenwaarden gelijk aan nul.

In dit geval is het faseportret gedegenereerd. We onderscheiden:



Figuur 8: Faseportretten als één eigenwaarde gelijk aan nul is (a,b) en als beide eigenwaarden gelijk aan nul zijn (c)

a. $\lambda_1 = 0, \lambda_2 \neq 0$

Laat v_1 een eigenvector zijn bij λ_1 en v_2 een eigenvector bij λ_2. Merk op dat nu voor alle $\alpha \in \mathbb{R}$ geldt dat

$$A(\alpha v_1) = 0$$

Dit betekent dat het stelsel (10) nu niet meer één evenwichtspunt heeft, maar een hele lijn van evenwichtspunten. Definieer weer de matrix $V = (v_1\ v_2)$ en $\bar{x} = V^{-1}x$. Net als in Geval 1 leiden we de volgende DV voor $\bar{x}$ af:

$$\begin{aligned} \dot{\bar{x}}_1 &= 0 \\ \dot{\bar{x}}_2 &= \lambda_2 \bar{x}_2 \end{aligned} \tag{43}$$

met als oplossingen

$$\begin{aligned} \bar{x}_1(t) &= \bar{x}_{10} \\ \bar{x}_2(t) &= \bar{x}_{20} e^{\lambda_2 t} \end{aligned} \tag{44}$$

Het faseportret in het $(\bar{x}_1, \bar{x}_2)$-vlak bestaat dus uit rechte lijnen evenwijdig aan de $\bar{x}_2$-as. De richting waarin de trajectoriën worden doorlopen hangt af van het teken van λ_2. In Figuur 8 zijn de faseportretten in het (x_1, x_2)-vlak getekend. Merk op dat voor $\lambda_2 < 0$ alle punten op de lijn opgespannen door v_1 stabiele (maar niet asymptotisch stabiele) evenwichtspunten voor (10) zijn en dat deze punten voor $\lambda_2 > 0$ instabiele evenwichtspunten zijn.

b. $\lambda_1 = \lambda_2 = 0$

Net als in Geval 3 kunnen we hier weer onderscheid maken tussen $(\lambda_1 I - A) = 0$ en $(\lambda_1 I - A) \neq 0$. In het eerste geval is er weinig

te beleven: $A = 0$ en het hele (x_1, x_2)-vlak bestaat uit evenwichtspunten. Daarom bekijken we verder het tweede geval (dus $A \neq 0$). Laat v een eigenvector bij $\lambda_1 = 0$ zijn en w een vector die voldoet aan $Aw = v$. Vorm weer $V = (v\ w)$ en definieer $\bar{x} = V^{-1}x$. Dan vinden we net als in Geval 3 de volgende DV voor $\bar{x}$:

$$\begin{aligned} \dot{\bar{x}}_1 &= \bar{x}_2 \\ \dot{\bar{x}}_2 &= 0 \end{aligned} \tag{45}$$

met als oplossing

$$\begin{aligned} \bar{x}_1(t) &= \bar{x}_{10} + \bar{x}_{20}t \\ \bar{x}_2(t) &= \bar{x}_{20} \end{aligned} \tag{46}$$

Dit levert in het (x_1, x_2)-vlak het faseportret in Figuur 8.c. Merk nu op dat alle punten op de lijn opgespannen door v instabiele evenwichtspunten zijn.

Uit het bovenstaande kunnen we met betrekking tot stabiliteit de volgende conclusie trekken.

Stelling 3.3 *Beschouw het lineaire tweedimensionale stelsel (10). Geef de eigenwaarden van A aan met* λ_1, λ_2. *Het stelsel is*

(i) asymptotisch stabiel dan en slechts dan als $\mathrm{Re}(\lambda_i) < 0$ $(i = 1, 2)$.

(ii) stabiel dan en slechts dan als $\mathrm{Re}(\lambda_i) \leq 0$ $(i = 1, 2)$ *en ten minste één van de eigenwaarden van A ongelijk aan nul is.* ■

3.2 Gedrag van oplossingen van tweedimensionale stelsels in de buurt van een evenwichtspunt

We keren nu terug naar de bestudering van het stelsel

$$\dot{x} = Ax + g(x) \tag{47}$$

waar $x \in \mathbb{R}^2$, $g(0) = 0$, g is continu en $g(x) = o(x)$. Beschouw verder het gelineariseerde stelsel

$$\dot{x} = Ax \tag{48}$$

We zullen proberen de volgende twee stellingen aannemelijk te maken (een exact bewijs vergt wat meer moeite, maar is in feite op dezelfde argumenten gebaseerd).

Stelling 3.4 *Beschouw de onderstaande tabel. Als $x = 0$ voor het gelineariseerde stelsel (48) van het type in de eerste kolom is, dan is $x = 0$ voor (47) van het type in de tweede kolom.*

lineair stelsel	niet-lineair stelsel
knoop	(oneigenlijke) knoop
ster	ster eigenlijke knoop oneigenlijke knoop spiraal
zadelpunt	zadelpunt
centrum	centrum stabiele spiraal instabiele spiraal
oneigenlijke knoop	oneigenlijke knoop eigenlijke knoop spiraal

■

Stelling 3.5 *Beschouw het tweedimensionale stelsel (47) en het gelineariseerde stelsel (48). Laat λ_1, λ_2 de eigenwaarden van A zijn.*

(i) Als $Re(\lambda_i) < 0$ $(i = 1, 2)$, dan is de oorsprong een asymptotisch stabiel evenwichtspunt voor (47).

(ii) Als ten minste één van de eigenwaarden van A een reëel deel groter dan nul heeft, dan is de oorsprong een instabiel evenwichtspunt voor (47). ■

Zoals gezegd, zullen we nu proberen deze stellingen aannemelijk te maken. Het feit dat $g(x) = o(x)$ impliceert dat er voor iedere $\epsilon > 0$ een $\delta > 0$ bestaat zodanig dat

$$\|g(x)\| < \epsilon\|x\| \quad \text{als } \|x\| < \delta \tag{49}$$

Dit betekent op z'n beurt dat we in de buurt van $x = 0$ de DV (47) kunnen interpreteren als een (kleine) verstoring van de gelineariseerde DV (48) en dat we zouden kunnen vermoeden dat het faseportret van (47) in de buurt van de oorsprong er (ongeveer) hetzelfde uitziet als dat van (48). Of dit vermoeden gerechtvaardigd is, hangt af van de vraag hoe het faseportret van (48) verandert onder kleine verstoringen.

Om hier een gevoel voor te krijgen, bekijken we nu het gedrag van de oplossingen van het verstoorde lineaire stelsel

$$\dot{x} = (A + \Delta A)x \tag{50}$$

waar ΔA een (2,2)-matrix met willekeurig kleine elementen, zeg

$$\Delta A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \tag{51}$$

De eigenwaarden van $(A + \Delta A)$ worden dan gegeven door

$$\lambda^{\Delta}_{1,2} = \frac{(a+\alpha+d+\delta)}{2} \pm \frac{\sqrt{(a+\alpha+d+\delta)^2 - 4((a+\alpha)(d+\delta) - (b+\beta)(c+\gamma))}}{2}$$

Hieruit volgt dat de eigenwaarden van $(A+\Delta A)$ continu van $\alpha, \beta, \gamma, \delta$ afhangen. Dit betekent dat als $\mathrm{Re}(\lambda_i) < 0$ ($\mathrm{Re}(\lambda_i) > 0$), we voor $\alpha, \beta, \gamma, \delta$ klein genoeg zullen hebben dat $\mathrm{Re}(\lambda^{\Delta}_i) < 0$ ($\mathrm{Re}(\lambda^{\Delta}_i) > 0$). Dit maakt tegelijk bovenstaande stellingen annemelijk.

Stelling 3.5 doet geen uitspraak voor het geval dat er eigenwaarden van A op de imaginaire as liggen. In dit geval is de situatie ook geheel verschillend. Om dit in te zien, beschouwen we

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{52}$$

met

$$\Delta A = \begin{pmatrix} \mu & 0 \\ 0 & \mu \end{pmatrix} \tag{53}$$

De eigenwaarden van A zijn $\lambda_{1,2} = \pm i$ en de oorsprong is dus een centrum voor (48). De eigenwaarden van $(A + \Delta A)$ zijn

$$\lambda^{\Delta}_{1,2} = \mu \pm i \tag{54}$$

We zien dus dat het evenwichtspunt van het verstoorde systeem een stabiele knoop wordt als $\mu < 0$ en een instabiele knoop als $\mu > 0$. Het faseportret verandert in dit geval dus drastisch. Oók als één of beide eigenwaarden gelijk aan nul zijn, zal het faseportret veranderen, waarbij het verstoorde stelsel zowel stabiel, asymptotisch stabiel als instabiel kan worden. Om het één en ander nog verder te illustreren, beschouwen we het volgende voorbeeld.

Voorbeeld 3.6 Beschouw het stelsel

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + kx_2^3 \end{aligned} \tag{55}$$

met $k \in \mathbb{R}$. Het is duidelijk dat de oorsprong een evenwichtspunt is. Het gelineariseerde stelsel wordt gegeven door

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 \end{aligned} \tag{56}$$

De oorsprong is dus een centrum voor het gelineariseerde stelsel (en daarmee ook een centrum voor (55) met $k = 0$). Definieer de functie

$$V(x_1, x_2) = x_1^2 + x_2^2 \tag{57}$$

Merk op dat V het kwadraat van de afstand van (x_1, x_2) tot de oorsprong is. Kies een beginpunt $(x_{10}, x_{20}) \neq (0,0)$ en geef de oplossing aan met $(x_1(t), x_2(t))$. We bekijken nu de functie

$$F(t) = V(x_1(t), x_2(t)) \tag{58}$$

Deze functie geeft dus een maat voor de afstand van de oplossing $(x_1(t), x_2(t))$ tot de oorsprong op een bepaald tijdstip. Dit betekent dat de oorsprong een stabiel evenwichtspunt is voor (55) als $F(t)$ begrensd is voor alle $t \geq 0$ en $\lim_{t\to+\infty} F(t) < \infty$, een asymptotisch stabiel evenwichtspunt als bovendien $\lim_{t\to+\infty} F(t) = 0$ en dat de oorsprong een instabiel evenwichtspunt is als $F(t) = \infty$ voor eindige t of als $\lim_{t\to+\infty} F(t) = \infty$. Met behulp van de kettingregel vinden we:

$$\begin{aligned} \frac{dF(t)}{dt} = & \frac{d}{dt}(V(x_1(t), x_2(t))) = \\ & \frac{\partial V}{\partial x_1}(x_1(t), x_2(t))\dot{x}_1(t) + \frac{\partial V}{\partial x_2}(x_1(t), x_2(t))\dot{x}_2(t) = \\ & 2x_1(t)x_2(t) + 2x_2(t)(-x_1(t) + kx_2^3(t)) = 2kx_2^4(t) \end{aligned}$$

Beschouw nu het geval dat $k > 0$. Dan zien we dat

$$\frac{dF(t)}{dt} \begin{cases} > 0 & \text{als } x_2(t) \neq 0 \\ = 0 & \text{als } x_2(t) = 0 \end{cases} \tag{59}$$

Dus de afstand van $(x_1(t), x_2(t))$ tot de oorsprong groeit zolang $x_2(t) \neq 0$. Dit betekent dat de oorsprong een instabiel evenwichtspunt is *tenzij* er een $T > 0$ bestaat zodanig dat $x_2(t) = 0$ voor alle $t \geq T$. Dit zou in het bijzonder impliceren dat $x_1(t) = \dot{x}_2(t) = 0$ voor alle $t \geq T$. Dus: $(x_1(t), x_2(t)) = (0,0)$ $(\forall t \geq T)$. Dit is echter in strijd met het feit dat de

oplossingen uniek zijn. De conclusie is dus dat als $k > 0$, de oorsprong een instabiel evenwichtspunt voor (55) is. Op een analoge manier kan worden afgeleid dat voor $k < 0$ de oorsprong een asymptotisch stabiel evenwichtspunt voor (55) is.

3.3 Globaal faseportret

In de vorige paragraaf hebben we gezien dat we voor een DV (47) met de oorsprong als evenwichtspunt in veel gevallen het faseportret in de buurt van de oorsprong kunnen tekenen. We zullen nu een methode geven voor het tekenen van het faseportret in het hele (x_1, x_2)-vlak. Merk in de eerste plaats op dat een stelsel van de vorm (47) in het algemeen meerdere (geïsoleerde) evenwichtspunten zal hebben. Met behulp van de theorie uit de vorige paragraaf en Opmerking 3.1 kan het type van elk van deze evenwichtspunten worden bepaald. Vervolgens maken we gebruik van de zogenaamde *isoclienenmethode*. Een *isoclien* is een kromme in het (x_1, x_2)-vlak die voldoet aan de vergelijking

$$\frac{f_2(x_1, x_2)}{f_1(x_1, x_2)} = c \tag{60}$$

waar $c \in \mathbb{R} \cup \infty$. Merk op dat

$$\frac{dx_2}{dx_1} = \frac{f_2(x_1, x_2)}{f_1(x_1, x_2)} \tag{61}$$

De isoclien bij een constante $c \in \mathbb{R} \cup \infty$ bestaat dus uit alle punten waarin de oplossingen van (47) een helling c hebben. Belangrijke isoclienen zijn de horizontale isoclien ($c = 0$), gegeven door

$$f_2(x_1, x_2) = 0 \tag{62}$$

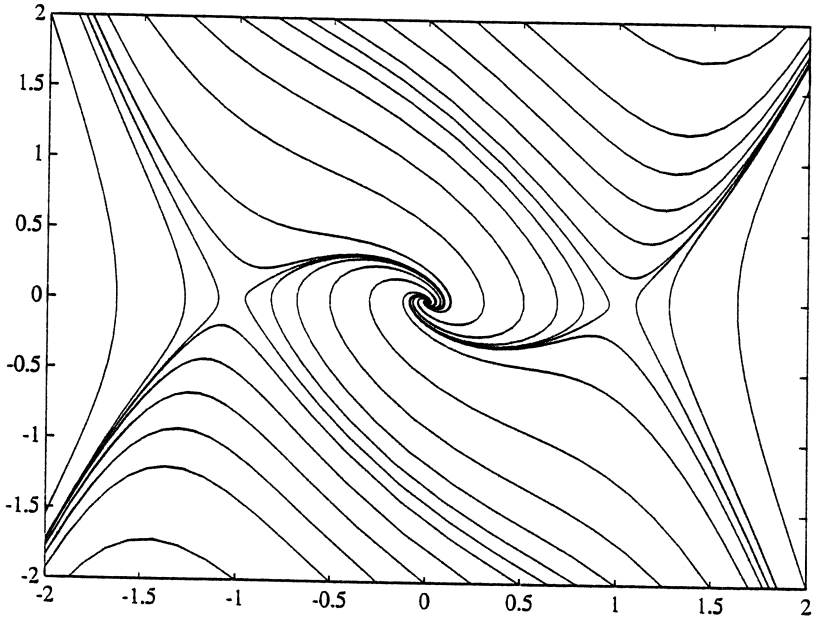
en de verticale isoclien ($c = \infty$), gegeven door

$$f_1(x_1, x_2) = 0 \tag{63}$$

Merk op dat de evenwichtspunten precies de snijpunten van de horizontale en de verticale isoclien zijn. Door een aantal isoclienen in het (x_1, x_2)-vlak te tekenen en rekening te houden met het gedrag in de buurt van de evenwichtspunten kunnen we een schets van het faseportret maken. We zullen dit illustreren aan de hand van een voorbeeld.

Voorbeeld 3.7 Beschouw het stelsel

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + x_1^3 - x_2 \end{aligned} \tag{64}$$



Figuur 9: Globaal faseportret

De evenwichtspunten van (64) worden gevonden door het volgende stelsel vergelijkingen op te lossen:

$$\begin{aligned} x_2 &= 0 \\ -x_1 + x_1^3 - x_2 &= 0 \end{aligned} \tag{65}$$

Dit levert de evenwichtspunten

$$\begin{aligned} x_A &= (0,0) \\ x_B &= (1,0) \\ x_C &= (-1,0) \end{aligned} \tag{66}$$

Voor $x_A = (0,0)$ vinden we

$$A = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}$$

De eigenwaarden van A worden gegeven door

$$\lambda_{1,2} = -\frac{1}{2} \pm \frac{1}{2} i\sqrt{3}$$

en dus is x_A een stabiele spiraal voor (64). Voor x_B en x_C vinden we

$$A = \begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix}$$

met als eigenwaarden

$$\lambda_1 = -2,\ \lambda_2 = 1$$

en dus zijn x_B en x_C zadelpunten voor (64). Eigenvectoren v_1 en v_2 bij λ_1 en λ_2 worden gegeven door

$$v_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix},\ v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

De horizontale isoclien wordt gegeven door

$$x_2 = x_1(x_1^2 - 1)$$

en de verticale isoclien door

$$x_2 = 0$$

Voor $c = -1$ wordt de isoclien gegeven door

$$x_1 = -1,\ x_1 = 0,\ x_1 = 1$$

Voor $c \notin \{-1, 0, \infty\}$ vinden we als isoclien

$$x_2 = \frac{x_1(x_1^2 - 1)}{c + 1}$$

Rekening houdend met het gedrag in de buurt van de evenwichtspunten en de isoclienen, vinden we het globale faseportret dat wordt gegeven in Figuur 9.

3.4 Lyapunovfuncties

In Paragraaf 3.2 hebben we gezien dat we voor een stelsel (47) stabiliteitsuitspraken kunnen doen op grond van het gelineariseerde stelsel (48) als A geen eigenwaarden heeft op de imaginaire as. In Voorbeeld 3.6 behandelden we een stelsel waarvoor A alle eigenwaarden op de imaginaire as had en waar we toch stabiliteitsuitspraken konden doen op grond van de functie $V(x_1, x_2) = x_1^2 + x_2^2$. Deze functie is een voorbeeld van een zogenaamde *Lyapunovfunctie.* We zullen in deze paragraaf de aanpak van Voorbeeld 3.6 generaliseren.

Beschouw weer een stelsel

$$\dot{x} = f(x) \tag{67}$$

waar $x \in \mathbb{R}^2$, f is continu differentieerbaar en $f(0) = 0$. Beschouw verder een continu differentieerbare functie $V : \mathbb{R}^2 \mapsto \mathbb{R}$. De *gradiënt* van V is de kolomvector

$$\nabla V := \begin{pmatrix} \frac{\partial V}{\partial x_1} \\ \\ \frac{\partial V}{\partial x_2} \end{pmatrix} \tag{68}$$

Laat $\langle \cdot, \cdot \rangle$ het *inproduct* op $\mathbb{R}^2$ aangeven, dus voor $x, y \in \mathbb{R}^2$:

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 \tag{69}$$

De afgeleide van V langs oplossingen van (67) wordt gegeven door

$$\begin{aligned} \frac{dV}{dt} = & \frac{\partial V}{\partial x_1}(x)\dot{x}_1 + \frac{\partial V}{\partial x_2}(x)\dot{x}_2 = \\ & \frac{\partial V}{\partial x_1}(x)f_1(x) + \frac{\partial V}{\partial x_2}(x)f_2(x) = \\ & \langle \nabla V, f \rangle(x) =: \mathcal{L}_f V(x) \end{aligned} \tag{70}$$

Laat nu $\epsilon > 0$ gegeven zijn en definieer

$$B_\epsilon := \{x \in \mathbb{R}^2 \mid \|x\| < \epsilon\} \tag{71}$$

B_ϵ bestaat dus uit alle $x \in \mathbb{R}^2$ waarvoor de afstand tot de oorsprong kleiner dan ϵ is.

Definitie 3.8 Een continue fuctie $V : \mathbb{R}^2 \mapsto \mathbb{R}$ die voldoet aan $V(0) = 0$ heet

(i) *positief semi-definiet* als er een $\epsilon > 0$ is zodat

$$\forall x \in B_\epsilon : V(x) \geq 0$$

(ii) *positief definiet* als er een $\epsilon > 0$ is zodat

$$\forall x \in B_\epsilon \backslash \{0\} : V(x) > 0$$

(iii) *negatief semi-definiet* als er een $\epsilon > 0$ is zodat

$$\forall x \in B_\epsilon : V(x) \leq 0$$

(iv) *negatief definiet* als er een $\epsilon > 0$ is zodat

$$\forall x \in B_\epsilon \backslash \{0\} : V(x) < 0$$

(v) indefiniet als V niet (semi-)definiet is

Definitie 3.9 Een continu differentieerbare functie $V : \mathbb{R}^2 \mapsto \mathbb{R}$ die voldoet aan $V(0) = 0$ heet

(i) een *zwakke Lyapunovfunctie* voor (67) als $\mathcal{L}_f V$ negatief semi-definiet is.

(ii) een *sterke Lyapunovfunctie* voor (67) als $\mathcal{L}_f V$ negatief definiet is.

Voorbeeld 3.6 Encore De functie $V(x_1, x_2) = x_1^2 + x_2^2$ is positief definiet en voor $k < 0$ geldt

$$\mathcal{L}_f V = 2kx_2^4 \leq 0 \text{ op } B_\epsilon \ (\forall \epsilon > 0) \tag{72}$$

Dit betekent dat V voor $k > 0$ een zwakke Lyapunovfunctie is. Beschouw voor $k > 0$ de functie $W(x) = -V(x)$. Dan

$$\mathcal{L}_f W = -\mathcal{L}_f V = -2kx_2^4 \leq 0 \text{ op } B_\epsilon \ (\forall \epsilon > 0) \tag{73}$$

en dus is W voor $k < 0$ een zwakke Lyapunovfunctie. ■

Als er een Lyapunovfunctie bestaat voor (67), dan kunnen we de volgende stabiliteitsuitspraken doen.

Stelling 3.10 *Beschouw de DV (67) en neem aan dat* $f(0) = 0$.

(i) Zij V een zwakke Lyapunovfunctie voor (67). Als V positief definiet is, dan is de oorsprong een stabiel evenwichtspunt voor (67). Als V niet positief semi-definiet is, dan is de oorsprong geen asymptotisch stabiel evenwichtspunt voor (67).

(ii) Zij V een sterke Lyapunovfunctie voor (67). Als V positief definiet is, dan is de oorsprong een asymptotisch stabiel evenwichtspunt voor (67). Als V niet positief definiet is, dan is de oorsprong een instabiel evenwichtspunt voor (67). ■

Voorbeeld 3.6 Nog weer een keer Voor $k > 0$ is $V(x)$ een positief definiete zwakke Lyapunovfunctie. Uit Stelling 3.10 volgt dan dat voor $k > 0$ de oorsprong een stabiel evenwichtspunt is. Verder geldt voor $k < 0$ dat $W(x)$ een negatief definiete zwakke Lyapunovfunctie is. Uit

Stelling 3.10 volgt dan dat voor $k > 0$ de oorsprong geen asymptotisch stabiel evenwichtspunt is.

Als we de bovenstaande uitspraken vergelijken met de uitspraken die in eerste instantie in dit voorbeeld werden gedaan, dan zien we dat bovenstaand uitspraken zwakker zijn: op grond van Stelling 3.10 en de functies V en W kunnen we slechts besluiten tot stabiliteit (in plaats van asymptotische stabiliteit) respectievelijk geen asymptotische stabiliteit (in plaats van instabiliteit). Dit komt doordat we eerder behalve de argumenten van Stelling 3.10 ook nog impliciet een argument hebben gebruikt dat bekend staat als *LaSalle's Invariantieprincipe*. Het voert te ver om dit principe hier uitputtend te behandelen. Kort gezegd komt het principe op het volgende neer:

LaSalle's Invariantieprincipe

Beschouw de DV (67) en neem aan dat $f(0) = 0$. Laat V een positief definiete zwakke Lyapunovfunctie voor (67) zijn. Als er een $\epsilon > 0$ bestaat zodanig dat de enige oplossing van (67) in de verzameling $B_\epsilon \cap \{x \mid \mathcal{L}_f V(x) = 0\}$ de nuloplossing $(x(t) = 0, \forall t)$ is, dan is de oorsprong een asymptotisch stabiel evenwichtspunt voor (67).

We kunnen wel tot asymptotische stabiliteit (instabiliteit) besluiten door gebruik te maken van alleen Stelling 3.10 (en dus niet van LaSalle's Invariantieprincipe) met behulp van een andere Lyapunovfunctie.

Laat $k < 0$ en beschouw

$$\tilde{V}(x) = x_1^2 + x_2^2 + ax_1^3x_2 = V(x) + ax_1^3x_2$$

waar $0 < a < -\frac{8}{9}k$. Als $x_1 \to 0$, $x_2 \to 0$, dan gaat $ax_1^3x_2$ sneller naar nul dan $V(x_1, x_2)$. Dit betekent dat $\tilde{V}$ positief definiet is. We vinden

$$\mathcal{L}_f\tilde{V} = \begin{pmatrix} 2x_1 + 3ax_1^2x_2 & 2x_2 + ax_1^3 \end{pmatrix} \begin{pmatrix} x_2 \\ -x_1 + kx_2^3 \end{pmatrix} =$$

$$-P(x_1, x_2) + kax_1^3x_2^3$$

waar $P(x_1, x_2) = ax_1^4 - 3ax_1^2 - 2kx_2^4$. Als $x_1 \to 0$, $x_2 \to 0$, dan gaat de term $kax_1^3x_2^3$ sneller naar nul dan P. Dit betekent dat $\mathcal{L}_f\tilde{V}$ negatief definiet is als P positief definiet is. Kwadraat afsplitsen levert

$$P(x_1, x_2) = a(x_1^4 - 3x_1^2x_2^2) - 2kx_2^4 = a(x_1^2 - \frac{3}{2}x_2^2)^2 - (\frac{9}{4}a + 2k)x_2^4$$

Omdat $a < -\frac{8}{9}k$, hebben we

$$\alpha := -(\frac{9}{4}a + 2k) > (\frac{9 \cdot 8}{4 \cdot 9}k - 2k) = 0$$

en dus

$$P(x_1, x_2) = a(x_1^2 - \frac{3}{2}x_2^2)^2 + \alpha x_2^4$$

waar $\alpha > 0$. Het is nu makkelijk in te zien dat P positief definiet is. Dit betekent dat $\tilde{V}$ een sterke Lyapunovfunctie is en dus dat de oorsprong een asymptotisch stabiel evenwichtspunt voor (55) is. Op dezelfde manier kunnen we voor $k > 0$ instabiliteit aantonen met behulp van $\tilde{W}(x) = -\tilde{V}(x)$, waar $a < -\frac{8}{9}k$. ∎

Zoals we zien, kunnen met behulp van Stelling 3.10 stabiliteitsuitspraken worden gedaan *als* we een Lyapunovfunctie voor (67) hebben. Er bestaat echter geen algemene methode voor de constructie van Lyapunovfuncties, alhoewel wel kan worden aangetoond dat als de oorsprong een (asymptotisch) stabiel evenwichtspunt is, er een Lyapunovfunctie bestaat. Er bestaat wel een aantal *ad hoc* methoden voor de constructie van Lyapunovfuncties. We zullen hier enige van deze methoden aanstippen.

Soms gaat men uit van een willekeurige positief definiete functie $V(x)$ en wordt $\mathcal{L}_f V(x)$ berekend. Vaak wordt $V(x) = \|x\|^2$ geprobeerd (zoals in Voorbeeld 3.6) of algemener $V(x) = x^T Q x$ (met Q zodanig dat V positief definiet is). Voor mechanische systemen is de totale energie vaak een goede kandidaat-Lyapunovfunctie. Een methode die bekend staat als *Zubov's methode* kiest eerst $\mathcal{L}_f V(x)$. Als $\mathcal{L}_f V(x)$ bekend is kan men namelijk altijd uitsluitsel over stabiliteit verkrijgen. Als van $\mathcal{L}_f V(x)$ wordt uitgegaan, dan moet V worden bepaald uit de partiële differentiaalvergelijking $\langle \nabla V, f \rangle = \mathcal{L}_f V$. Dit komt echter neer op het oplossen van de DV (67), zodat deze methode niet erg praktisch is en slechts theoretisch nut heeft. Een andere methode die in veel gevallen een oplossing biedt is de *variabele gradiënt methode*. Hierbij probeert men een vectorfunctie $g(x)$ te bepalen zodat $\langle g(x), f(x) \rangle$ negatief definiet is, terwijl $g(x)$ de gradiënt van een functie V is: $g(x) = \nabla V(x)$. De laatste eis legt de volgende voorwaarde op:

$$\frac{\partial g_1}{\partial x_2} = \frac{\partial g_2}{\partial x_1} \tag{74}$$

Als aan deze voorwaarde is voldaan, kan V eenvoudig uit g worden bepaald. Vaak wordt $g(x)$ van de vorm $g(x) = A(x)x$ gekozen en wordt in eerste instantie geprobeerd de (2,2)-matrix $A(x)$ constant te houden.

Voorbeeld 3.6 De laatste keer Beschouw weer het stelsel (55) en neem voor het gemak $k = -1$. We zullen de variabele gradiënt methode demonstreren. Stel dat

$$g(x) = \begin{pmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{pmatrix}$$

Uit (74) volgt dat a, b, c, d moeten voldoen aan

$$\frac{\partial a}{\partial x_2}x_1 + \frac{\partial b}{\partial x_2}x_2 + b = \frac{\partial c}{\partial x_1}x_1 + c + \frac{\partial d}{\partial x_1}x_2 \tag{75}$$

Verder geldt

$$\langle f(x), g(x)\rangle = -cx_1^2 + (a-d)x_1x_2 + bx_2^2 - cx_1x_2^3 - dx_2^4 \tag{76}$$

We zullen eerst laten zien dat a, b, c, d niet allemaal constant kunnen zijn. In dat geval levert (75) namelijk dat $b = c$ en dus

$$\langle f(x), g(x)\rangle = -bx_1^2 + (a-d)x_1x_2 + bx_2^2 - bx_1x_2^3 - dx_2^4$$

Kies dan $\alpha \in \mathbb{R}$ willekeurig en

$$\bar{x} = \begin{cases} (\alpha, 0) & \text{als } b < 0 \\ (0, \alpha) & \text{als } b > 0 \\ (\alpha, 0) & \text{als } b = 0 \end{cases}$$

We vinden dan

$$\langle f(\bar{x}), g(\bar{x})\rangle \begin{cases} > 0 & \text{als } b < 0 \text{ of } b > 0 \\ = 0 & \text{als } b = 0 \end{cases}$$

hetgeen betekent dat $\langle f(x), g(x)\rangle$ niet negatief definiet is. We zullen nu proberen a, b, c, d te bepalen zodat $\langle f, g\rangle$ de volgende vorm heeft:

$$\langle f(x), g(x)\rangle = -(x_1^2 + \beta x_2^2)^2 - \gamma x_2^4 + r(x) \tag{77}$$

waar $\gamma > 0$ en $r(x)$ sneller naar nul gaat dan $-(x_1^2 + \beta x_2^2)^2 - \gamma x_2^4$ voor $x \to 0$. In dit geval is $\langle f, g\rangle$ namelijk negatief definiet (ga na). Inspectie van (76) leert dat dit mogelijk is met

$$\begin{aligned} b &= 0 \\ c &= x_1^2 \\ d &= \text{constant} \\ a - d &= \delta x_1 x_2,\ \delta \in \mathbb{R} \end{aligned} \tag{78}$$

Invullen van (78) in (75) levert

$$\delta x_1^2 = 2x_1^2 + x_1^2 = 3x_1^2 \Rightarrow \delta = 3 \tag{79}$$

Dus:

$$\begin{aligned}\langle f(x), g(x)\rangle = \quad & -x_1^4 + 3x_1^2x_2^2 - x_1^3x_2^3 - dx_2^4 = \\ & -(x_1^4 - 3x_1^2x_2^2 + dx_2^4) - x_1^3x_2^3 = \\ & -(x_1^2 - \tfrac{3}{2}x_2^2)^2 - (d - \tfrac{9}{4})x_2^4 - x_1^3x_2^3\end{aligned} \tag{80}$$

Als we in (80) $d > \frac{9}{4}$ nemen, krijgen we de vorm (77). Neem $d = 3$. We bepalen nu een V zodanig dat $\nabla V = g$. V moet voldoen aan:

$$\frac{\partial V}{\partial x_1} = 3x_1^2x_2 + 3x_1 \tag{81}$$

$$\frac{\partial V}{\partial x_2} = x_1^3 + 3x_2 \tag{82}$$

(82) levert

$$V(x_1, x_2) = x_1^3 + \frac{3}{2}x_2^2 + \phi(x_1) \tag{83}$$

waar we de functie ϕ nog nader moeten bepalen. Dit doen we door (83) in te vullen in (81):

$$3x_1^2x_2 + \phi'(x_1) = 3x_1^2 + 3x_1 \Rightarrow \phi'(x_1) = 3x_1 \Rightarrow \phi(x_1) = \frac{3}{2}x_1^2$$

We vinden dus de Lyapunovfunctie

$$V(x_1, x_2) = \frac{3}{2}(x_1^2 + x_2^2) + x_1^3x_2$$

■

4 Meerdimensionale stelsels

In de vorige sectie hebben we tweedimensionale stelsels DV's behandeld. We zullen in deze sectie deze resultaten generaliseren naar meerdimensionale stelsels, dat wil zeggen stelsels van de vorm

$$\dot{x} = f(x) \tag{84}$$

waar $x \in \mathbb{R}^n$, met $n \geq 2$, en waar f weer continu differentieerbaar is. We beperken ons tot de stabiliteitstheorie gebaseerd op linearisatie. De generalisatie van stabiliteitstheorie via Lyapunovfuncties is in feite direct.

4.1 Lineaire stelsels

4.1.1 Oplossingen van meerdimensionale lineaire stelsels

Beschouw het lineaire stelsel

$$\dot{x} = Ax \tag{85}$$

waar $x \in \mathbb{R}^n$ en A een (n,n)-matrix is. Net als in Paragraaf 3.1.1 definiëren we het *karakteristieke polynoom* $p(\lambda)$ van A:

$$p(\lambda) = \det(\lambda I - A) \tag{86}$$

waar I de (n,n)-eenheidsmatrix is. Het kan worden aangetoond dat $p(\lambda)$ de volgende vorm heeft:

$$p(\lambda) = \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0 \tag{87}$$

Uit de Hoofdstelling van de Algebra volgt dat er $N, M \in \mathbb{N}$, $n_1, \cdots, n_N$, $m_1, \cdots, m_M$ met $\sum_{i=1}^{N} n_i + 2\sum_{i=1}^{M} m_i = n$, en $\lambda_1, \cdots, \lambda_N \in \mathbb{R}$, $\mu_1, \cdots$, $\mu_M \in \mathbb{C}\backslash\mathbb{R}$ bestaan zodanig dat we kunnen schrijven

$$p(\lambda) = \left(\prod_{i=1}^{N}(\lambda - \lambda_i)^{n_i}\right)\left(\prod_{i=1}^{M}(\lambda - \mu_i)^{m_i}(\lambda - \overline{\mu}_i)^{m_i}\right) \tag{88}$$

waar $\overline{\mu}_i$ de *complex toegevoegde* van μ_i is. Het getal n_i (m_i) wordt de *algebraïsche multipliciteit* van λ_i (μ_i) genoemd.

We keren nu even terug naar tweedimensionale stelsels, dus beschouw (85) met $n = 2$. Geef de eigenwaarden van A aan met λ_1, λ_2. Onderscheid nu

- $\lambda_1, \lambda_2 \in \mathbb{R}$ en als $\lambda_1 = \lambda_2 = \mu$, dan $(\mu I - A) = 0$.

  Laat v_1 en v_2 eigenwaarden bij λ_1 en λ_2 zijn en vorm weer de matrix $V = (v_1 \quad v_2)$. Uit Sectie 3 volgt dat de oplossingen van (85) van de volgende vorm zijn:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = V \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{89}$$

  met $c_1, c_2 \in \mathbb{R}$. Verder uitschrijven van (89) levert

$$x(t) = c_1 v_1 e^{\lambda_1 t} + c_2 v_2 e^{\lambda_2 t} \tag{90}$$

- $\lambda_1 = \lambda_2 = \mu$ en $(\mu I - A) \neq 0$.

  Laat v een eigenvector bij μ zijn en w een vector die voldoet aan $Aw = \mu w + v$. Uit Sectie 3 volgt dan dat de oplossingen van (85) van de volgende vorm zijn:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = V \begin{pmatrix} e^{\mu t} & te^{\mu t} \\ 0 & e^{\mu t} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{91}$$

  met $c_1, c_2 \in \mathbb{R}$. Verder uitschrijven levert

$$x(t) = \tilde{c}_1 \tilde{v} e^{\mu t} + \tilde{c}_2 \tilde{w} t e^{\mu t} \tag{92}$$

  waar $\tilde{c}_1, \tilde{c}_2 \in \mathbb{R}$, $\tilde{w} = v$ en $\tilde{v}$ voldoet aan

$$A\tilde{v} = \mu\tilde{v} + \tilde{w}$$

- $\lambda_1 = \alpha + i\beta$, $\lambda_2 = \alpha - i\beta$, $\beta \neq 0$.

  Laat $r, s \in \mathbb{R}^2$ zodanig zijn dat $v_1 = r + is$ een eigenvector bij λ_1 is. Vorm $V = (r \;\; s)$. Uit Sectie 3 volgt dan dat de oplossingen van (85) van de volgende vorm zijn:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = V \begin{pmatrix} r_0 e^{\alpha t} \cos(\phi_0 + \beta t) \\ r_0 e^{\alpha t} \sin(\phi_0 + \beta t) \end{pmatrix} \tag{93}$$

  Verder uitschrijven levert:

$$x(t) = r_0 r e^{\alpha t} \cos(\phi_0 + \beta t) + r_0 s e^{\alpha t} \sin(\phi_0 + \beta t) \tag{94}$$

We zullen bovenstaande resultaten nu generaliseren naar het geval $n \geq 2$. Het gegeven resultaat is gebaseerd op [8].

Stelling 4.1 *Beschouw het lineaire stelsel (85) en laat* N, M, n_i, λ_i $(i = 1, \cdots, N)$, m_i, μ_i $(i = 1, \cdots, M)$ *gedefinieerd zijn als hierboven. Laat* $\mu_i = \alpha_i + i\beta_i$ $(i = 1, \cdots, M)$.

(i) Dan worden de oplossingen van (85) gegeven door

$$\begin{aligned} x(t) = & \sum_{i=1}^{N} \sum_{j=0}^{n_i - 1} c_{ij} B_{ij} t^j e^{\lambda_i t} + \\ & \sum_{i=1}^{M} \sum_{j=0}^{m_i - 1} r_{ij} e^{\alpha_i t} [R_{ij} \cos(\phi_{ij} + \beta_i t) + \\ & \qquad S_{ij} \sin(\phi_{ij} + \beta_i t)] \end{aligned} \tag{95}$$

waar $c_{ij} \in \mathbb{R}$, $r_{ij} \in \mathbb{R}^+$, $\phi_{ij} \in [0, 2\pi]$ *willekeurig en* B_{ij}, R_{ij}, S_{ij} *voldoen aan*

$$\begin{aligned} AB_{in_i-1} &= \lambda_i B_{in_i-1} \\ AB_{ij} &= \lambda_i B_{ij} + (j+1)B_{ij+1} \end{aligned} \tag{96}$$

voor $j = n_i - 2, \cdots, 0$ *en*

$$\begin{aligned} AR_{in_i-1} &= \alpha_i R_{in_i-1} - \beta_i S_{in_i-1} \\ AS_{in_i-1} &= \beta_i R_{in_i-1} + \alpha_i S_{in_i-1} \\ AR_{ij} &= \alpha_i R_{ij} - \beta_i S_{ij} + (j+1)R_{ij+1} \\ AS_{ij} &= \beta_i R_{ij} + \alpha_i S_{ij} + (j+1)S_{ij+1} \end{aligned} \tag{97}$$

voor $j = n_i - 2, \cdots, 0$.

(ii) Door een geschikte keuze van de constanten $c_{ij}, r_{ij}, \phi_{ij}$ *in (95) kunnen we n lineair onafhankelijke oplossingen van (85) vinden.*

■

4.1.2 De variatie-van-constanten-formule

Volgens Stelling 4.3 kunnen we n lineair onafhankelijke oplossingen $x^1(t), \cdots, x^n(t)$ van (85) vinden. Met behulp van deze oplossingen kunnen we de algemene oplossing van (85) ook op een andere manier representeren. Vorm de matrix $X(t)$ met kolommen $x^1(t), \cdots, x^n(t)$:

$$X(t) = (x^1(t) \ \cdots \ x^n(t)) \tag{98}$$

Definieer vervolgens de matrix

$$\mathrm{e}^{At} := X(t)X(0)^{-1} \tag{99}$$

Merk op dat $X(t)$ voldoet aan

$$\frac{d}{dt}X(t) = AX(t) \tag{100}$$

Met de kettingregel vinden we dan

$$\frac{d}{dt}\mathrm{e}^{At} = A\mathrm{e}^{At} \tag{101}$$

Beschouw nu (85) met $x(0) = x_0$. De oplossing wordt dan gegeven door

$$x(t) = \mathrm{e}^{At} \tag{102}$$

Namelijk, uit (101) en (102) volgt met de kettingregel

$$\frac{d}{dt}(\mathrm{e}^{At}x_0) = \left(\frac{d}{dt}\mathrm{e}^{At}\right)x_0 = A\mathrm{e}^{At}x_0 = Ax(t) \tag{103}$$

en uit (99) volgt

$$x(0) = X(0)X(0)^{-1}x_0 = x_0 \tag{104}$$

Opmerking 4.2 e^{At} in (99) wordt een *matrixexponentiaal* genoemd. Een andere manier om e^{At} te definiëren is via

$$\mathrm{e}^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!}A^k \tag{105}$$

waar $A^0 = I$, $A^1 = A$, $A^k = A(A^{k-1})$ $(k = 2, 3, \cdots)$. De "suggestieve" notatie wordt geheel gerechtvaardigd als we de eigenschappen van e^{At} vergelijken met de eigenschappen van de "normale" exponentiële functie e^{at} met $a \in \mathbb{R}$:

- $\mathrm{e}^{at}x_0$, met $x_0 \in \mathbb{R}$, is de oplossing van het BWP

$$\dot{x}(t) = ax(t), \quad x(0) = x_0$$

- De Taylorontwikkeling van e^{at} wordt gegeven door

$$\mathrm{e}^{at} = \sum_{k=0}^{\infty} \frac{t^k}{k!}a^k$$

In de praktijk komen we vaak lineaire differentiaalvergelijkingen tegen van de vorm

$$\dot{x}(t) = Ax(t) + b(t) \tag{106}$$

De term $b(t)$ wordt de *bronterm* genoemd en kan worden geïnterpreteerd als een aandrijf- of besturingsterm. We zullen nu een uitdrukking geven voor de oplossing van het BWP

$$\begin{cases} \dot{x}(t) = Ax(t) + b(t) \\ x(0) = x_0 \end{cases} \tag{107}$$

Uit het voorgaande weten we dat de oplossing van de *homogene vergelijking* (dit is vergelijking (106) met $b(t) \equiv 0$) wordt gegeven door

$$x(t) = \mathrm{e}^{At}x_0 \tag{108}$$

We proberen nu met *variatie van constanten* een *particuliere oplossing* $y(t)$ te vinden van de vorm

$$y(t) = \mathrm{e}^{At}\xi(t) \tag{109}$$

met $\xi(0) = 0$. Met de kettingregel volgt uit (109)

$$\dot{y}(t) = A\mathrm{e}^{At}\xi(t) + \mathrm{e}^{At}\dot{\xi}(t) = Ay(t) + \mathrm{e}^{At}\dot{\xi}(t) \tag{110}$$

Uit (106) volgt dat $y(t)$ moet voldoen aan

$$\dot{y}(t) = Ay(t) + b(t) \tag{111}$$

en dus volgt uit (110) en (111)

$$\mathrm{e}^{At}\dot{\xi}(t) = b(t) \tag{112}$$

ofwel

$$\dot{\xi}(t) = (\mathrm{e}^{At})^{-1}b(t) \tag{113}$$

Met behulp van bijvoorbeeld (105) kan worden bewezen dat $(\mathrm{e}^{At})^{-1} = \mathrm{e}^{-At}$. Dan volgt uit (113)

$$\xi(t) = \int_0^t \mathrm{e}^{-A\tau}b(\tau)d\tau \tag{114}$$

waar de integraal wordt berekend door ieder van de termen van de vector $\mathrm{e}^{-A\tau}b(\tau)$ apart te integreren. (109) en (114) leveren

$$y(t) = \int_0^t \mathrm{e}^{A(t-\tau)}b(\tau)d\tau \tag{115}$$

De oplossing van het BWP (107) wordt nu verkregen door de oplossing van de homogene vergelijking en de particuliere oplossing bij elkaar op te tellen:

$$x(t) = \mathrm{e}^{At}x_0 + \int_0^t \mathrm{e}^{A(t-\tau)}b(\tau)d\tau \tag{116}$$

Deze formule wordt de *variatie-van-constanten-formule* genoemd.

4.1.3 Stabiliteit van meerdimensionale lineaire stelsels

De vorm van (95) suggereert dat ook hier, net als in het tweedimensionale geval, de eigenwaarden van A bepalend zijn voor de stabiliteit van (85). Dit is inderdaad het geval. Voor we het volledige resultaat kunnen formuleren, moeten we eerst nog een begrip introduceren. Beschouw een eigenwaarde λ_i van A met algebraïsche multipliciteit n_i. We weten dat het aantal eigenvectoren bij λ_i overaftelbaar is (als v_i een eigenvector is, dan is αv_i ($\alpha \in \mathbb{R}\backslash\{0\}$ ook een eigenvector). Het verhaal wordt echter anders als we op zoek gaan naar eigenvectoren $v_1, \cdots, v_d$ bij λ_i die lineair onafhankelijk zijn. Dan kan worden aangetoond dat er een $g_i \in \mathbb{N}$ bestaat die voldoet aan $1 \leq g_i \leq n_i$ en de volgende eigenschappen heeft:

- Er bestaan eigenvectoren $v_1, \cdots, v_{g_i}$ bij λ_i die lineair onafhankelijk zijn.
- Laat $v_1, \cdots, v_d$ eigenvectoren bij λ_i zijn met $d > g_i$. Dan zijn $v_1, \cdots, v_d$ lineair afhankelijk.

Het getal g_i wordt de *geometrische multipliciteit* van λ_i genoemd.

We hebben nu

Stelling 4.3 *Beschouw het lineaire stelsel (85) en laat $M, N, n_i, m_i, \lambda_i, \mu_i$ gedefinieerd zijn als hierboven. Laat $\mu_i = \alpha_i + i\beta_i$. Dan is de oorsprong*

(i) een asymptotisch stabiel evenwichtspunt voor (85) dan en slechts dan als

$$\begin{aligned} \lambda_i &< 0 \qquad (i = 1, \cdots, N) \\ \alpha_i &< 0 \qquad (i = 1, \cdots, M) \end{aligned}$$

(ii) een stabiel evenwichtspunt voor (85) dan en slechts dan als

$$\begin{aligned} \lambda_i &\leq 0 \qquad (i = 1, \cdots, N) \\ \alpha_i &\leq 0 \qquad (i = 1, \cdots, M) \end{aligned}$$

én voor alle λ_i met $\lambda_i = 0$ en alle μ_i met $\alpha_i = 0$ geldt dat

$$n_i = g_i$$

■

Uit bovenstaande stelling volgt dat om (asymptotische) stabiliteit van (85) na te gaan we het karakakteristieke polynoom $p(\lambda)$ van A moeten bepalen en daarna de nulpunten van $p(\lambda)$. Omdat $p(\lambda)$ een n-de graads polynoom is, kan dit voor grote n (≥ 5) zeer lastig zijn. Als we slechts geïnteresseerd zijn in asymptotische stabiliteit, kan de volgende stelling uitkomst bieden.

Stelling 4.4 (Routh-Hurwitz)
Beschouw het stelsel (85). Laat $p(\lambda)$ het karakteristieke polynoom van A zijn, gegeven door

$$p(\lambda) = \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0$$

Vorm de (n,n)-matrix

$$D = \begin{pmatrix} p_{n-1} & p_{n-3} & p_{n-5} & \cdots & \cdots & 0 \\ 1 & p_{n-2} & p_{n-4} & \cdots & \cdots & 0 \\ 0 & p_{n-1} & p_{n-3} & \cdots & \cdots & 0 \\ 0 & 1 & p_{n-2} & \cdots & \cdots & 0 \\ 0 & 0 & p_{n-1} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & p_0 \end{pmatrix}$$

Laat $D_1, \cdots, D_n$ de hoofdminoren van D zijn (d.w.z. D_k is de determinant van de matrix bestaande uit de eerste k kolommen en rijen van A). Dan is de oorsprong een asymptotisch stabiel evenwichtspunt voor (85) dan en slechts dan als

$$D_k > 0 \quad (k = 1, \cdots, n)$$

■

4.2 Stabiliteit via linearisatie

Beschouw het n-dimensionale stelsel

$$\dot{x} = f(x) \tag{117}$$

waar $x \in \mathbb{R}^n$ en $f(0) = 0$. Neem weer aan dat f continu differentieerbaar is. Dan kunnen we (117) schrijven als

$$\dot{x} = Ax + g(x) \tag{118}$$

waar

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

en waar $g(x)$ voldoet aan

- $g(x)$ is continu.
- $g(0) = 0$.
- $\lim_{\|x\| \to 0} \frac{\|g(x)\|}{\|x\|} = 0$.

Net als in Sectie 3 beschouwen we het gelineariseerde stelsel

$$\dot{x} = Ax \tag{119}$$

Dan hebben we de volgende generalisatie van Stelling 3.5.

Stelling 4.5 *Beschouw het n-dimensionale stelsel (117) en het gelineariseerde stelsel (119). Laat $\lambda_1, \cdots, \lambda_n$ de eigenwaarden van A zijn.*

(i) Als $Re(\lambda_i) < 0$ $(i = 1, 2, \cdots, n)$ dan is de oorsprong een asymptotisch stabiel evenwichtspunt voor (117).

(ii) Als ten minste één van de eigenwaarden van A een reëel deel groter dan nul heeft, dan is de oorsprong een instabiel evenwichtspunt voor (117). ■

Net als Stelling 3.5 doet Stelling 4.5 geen uitspraak voor het geval dat ten minste één van A op de imaginaire as ligt. Om in dit geval stabiliteitsuitspraken te kunnen doen, moeten we de toevlucht nemen tot andere methoden, bijvoorbeeld de methode van Lyapunovfuncties. Deze methode voor meerdimensionale stelsels is een directe generalisatie van dezelfde methode voor tweedimensionale stelsels die we in Sectie 3 hebben behandeld. Daarom laten we de behandeling van Lyapunovfuncties voor meerdimensionale stelsels achterwege.

Referenties

Onderstaande referenties geven allemaal een goede inleiding in de theorie van gewone differentiaalvergelijkingen en in de stabiliteitstheorie. [1] en [2] zijn klassiekers op het gebied van gewone differentiaalvergelijkingen, alhoewel [1] soms wat te bondig is geschreven en [2] wat ontoegankelijk is. [3] en [10] (trouwens ook in een engelse versie verkrijgbaar) zijn klassiekers op het gebied van stabiliteitstheorie. [3] is echter niet meer in de handel. [4],[6],[11] geven een moderne inleiding op (vooral) het gebied van de lineaire gewone differentiaalvergelijkingen. [4],[6] en [5] zijn vooral interessant omdat in deze boeken voornamelijk systeemtheoretische vraagstukken worden behandeld. Uitstekende boeken (en goedkoop) in het nederlands zijn [7] en [9].

Referenties

[1] Bellman, R., *Stability theory of differential equations*, Dover Publications, New York, 1953.

[2] Coddington, E.A., en N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, New York, 1955.

[3] Hahn, W., *Stability of motion*, Springer Verlag, Berlijn, 1967.

[4] Kailath, T., *Linear systems*, Prentice Hall, Englewood Cliffs, New Jersey, 1980.

[5] Khalil, H.K., *Nonlinear systems*, Macmillan Publishing Company, New York, 1992.

[6] Luenberger, D.G., *Introduction to dynamic systems*, John Wiley & Sons, New York, 1979.

[7] Mattheij, R.M.M., en J. Molenaar, *Beginwaardeproblemen in theorie en praktijk*, Epsilon Uitgaven, Utrecht, 1991.

[8] Polderman, J.W., en J.C. Willems, *Inleiding wiskundige systeemtheorie*, Collegedictaat, Universiteit Twente, 1991.

[9] Verhulst, F., *Nietlineaire differentiaalvergelijkingen en dynamische systemen*, Epsilon Uitgaven, Utrecht, 1985.

[10] Willems, J.L., *Stabilität dynamischer Systeme*, R. Oldenbourg Verlag, München, 1973.

[11] Wilson, H.K., *Ordinary differential equations*, Addison-Wesley, Reading, Mass., 1971.

Stochastiek

J.Th.M. Wijnen

§1. STOCHASTISCHE GROOTHEDEN

1.1. In de kansrekening houden we ons bezig met de studie van *modellen* van situaties (experimenten, verschijnselen), waarbij het 'toeval' een rol speelt. Bij deze modellen blijft doorgaans het 'toeval' onbesproken, even-als het mechanisme achter de verschijnselen dat tot het model leidt. Het model geeft meestal aan welke gebeurtenissen kunnen optreden en wat de kansen op deze gebeurtenissen zijn. Wanneer we een (toevals-) experiment verrichten zullen daarbij één of meer grootheden $x, y, \ldots$ een rol spelen, op de uitkomst waarvan het toeval een zekere invloed heeft. Dergelijke grootheden noemen we *stochastische grootheden* of *stochasten*. Dat een grootheid x opgevat wordt als een stochast brengen we in de notatie tot uitdrukking door deze grootheid vet te drukken, dat wil zeggen we noteren $\mathbf{x}$.

1.2. VOORBEELDEN.
1) We werpen met een munt en beschouwen de stochast

$$\mathbf{x} = \begin{cases} 1 & \text{als er munt wordt geworpen,} \\ -1 & \text{als er kruis wordt geworpen.} \end{cases}$$

2) We kopen een TV-toestel en beschouwen de stochast

$\mathbf{x}$ = levensduur van het gekochte TV-toestel.

1.3. Laat een stochast $\mathbf{x}$ gegeven zijn. Het kan zijn dat we een eindig aantal getallen $u_1, u_2, \ldots, u_n$ of een rij getallen $u_1, u_2, \ldots$ kunnen aangeven zó dat de stochast $\mathbf{x}$ altijd één van de waarden $u_1, u_2, \ldots$ aanneemt. In dat geval zeggen we dat de stochast $\mathbf{x}$ *discreet verdeeld* is.
Wanneer de stochast $\mathbf{x}$ niet discreet verdeeld is, zeggen we dat de stochast $\mathbf{x}$ *continu verdeeld* is.

1.4. VOORBEELDEN.
1) In voorbeeld 1) van 1.2 is de stochast $\mathbf{x}$ discreet verdeeld.
2) In voorbeeld 2) van 1.2 is de stochast $\mathbf{x}$ continu verdeeld.

1.5. DEFINITIE. Laat $\mathbf{x}$ een discreet verdeelde stochast zijn met mogelijke uitkomsten $u_1, u_2, \ldots$
Een *kansverdeling* P voor de stochast $\mathbf{x}$ is een rij getallen $P(u_1), P(u_2), \ldots$ zo dat

1) $P(u_i) \geq 0$ voor $i = 1, 2, \ldots$

2) $P(u_1) + P(u_2) + \ldots = \sum_i P(u_i) = 1.$

1.6. VOORBEELD. We gooien met een dobbelsteen en beschouwen de stochast

$\mathbf{x} =$ aantal gegooide ogen.

Omdat de stochast $\mathbf{x}$ slechts de waarden $1, 2, \ldots, 6$ kan aannemen is $\mathbf{x}$ discreet verdeeld. Wanneer we aannemen met een 'zuivere' dobbelsteen te maken te hebben, kunnen we dat in de keuze van de kansverdeling P voor de stochast $\mathbf{x}$ tot uitdrukking brengen, namelijk

$$P(1) = P(2) = \ldots = P(6) = 1/6.$$

Omdat iedere mogelijke uitkomst van de stochast $\mathbf{x}$ dezelfde kans heeft, noemen we zo'n kansverdeling P *symmetrisch.*
Wanneer we van een stochast $\mathbf{x}$ de kansverdeling kennen, kunnen we kansen op zekere *gebeurtenissen* uitrekenen, bijvoorbeeld: Laat A de gebeurtenis zijn: $\mathbf{x}$ is even. Dan geldt

$$P(A) = P(\{2, 4, 6\}) = P(2) + P(4) + P(6) = \frac{1}{2}.$$

1.7. Laat $\mathbf{x}$ een continu verdeelde stochast zijn. Om nu kansen op gebeurtenissen met betrekking tot de stochast $\mathbf{x}$ te kunnen uitrekenen, kunnen we niet dezelfde methode gebruiken als die voor discreet verdeelde stochasten (zie 1.5). In dit geval zullen wij werken met een kansdichtheid.

DEFINITIE. Een *kansdichtheid* $f(x)$ is een functie $f : \mathbb{R} \to \mathbb{R}$ zo dat
1) $f(x) \geq 0$ voor alle $x \in \mathbb{R}$,

2) $\int_{-\infty}^{\infty} f(x)dx = 1.$ (Vergelijk 1.5.)

Wanneer $\mathbf{x}$ een continu verdeelde stochast is met een kansdichtheid $f_{\mathbf{x}}(x)$ (N.B. de index $\mathbf{x}$ in $f_{\mathbf{x}}(x)$ geeft aan dat de kansdichtheid $f_{\mathbf{x}}(x)$ bij de stochast $\mathbf{x}$ behoort) kunnen we hiermee allerlei kansen uitrekenen, bijvoorbeeld

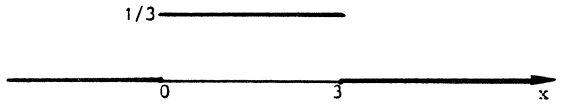
$$P(a \leq \mathbf{x} \leq b) = \int_a^b f_{\mathbf{x}}(x)dx.$$

1.8. VOORBEELD. We beschouwen een staaf van 3 m. lengte. Ten gevolge van een zeker proces zal deze staaf ergens breken. We beschouwen de stochast

$\mathbf{x} = x -$ coördinaat $(0 \leq x \leq 3)$ van het breekpunt.

Het is duidelijk dat de stochast $\mathbf{x}$ continu verdeeld is. Om tot uitdrukking te brengen dat ieder punt langs de staaf dezelfde kans heeft als ieder ander punt om als breekpunt op te treden, kiezen we als kansdichtheid $f_{\mathbf{x}}(x)$

$$f_{\mathbf{x}}(x) = \begin{cases} 1/3 & \text{als } 0 \leq x \leq 3, \\ 0 & \text{anders.} \end{cases}$$

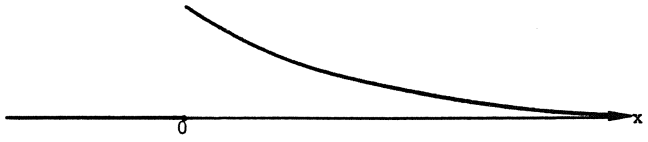


(Ga na dat aan de twee voorwaarden in 1.7 is voldaan.)
We zeggen dat de stochast $\mathbf{x}$ *homogeen verdeeld* is op [0,3].
Laat A de gebeurtenis zijn dat het breekpunt niet ligt tussen 1 en 2 meter langs de staaf. Dan geldt

$$P(A) = P(0 \leq \mathbf{x} \leq 1 \text{ of } 2 \leq \mathbf{x} \leq 3) = \int_0^1 f_{\mathbf{x}}(x)dx + \int_2^3 f_{\mathbf{x}}(x)dx = \frac{2}{3}.$$

1.9. VOORBEELD. Laat $\mathbf{x}$ de stochast zijn uit voorbeeld 2) in 1.2. We veronderstellen dat de kansdichtheid $f_{\mathbf{x}}(x)$ van $\mathbf{x}$ gegeven wordt door

$$f_{\mathbf{x}}(x) = \begin{cases} 0.1e^{-0.1x} & \text{als } x \geq 0, \\ 0 & \text{anders.} \end{cases}$$



Ga na dat $f_{\mathbf{x}}(x)$ een kansdichtheid is. Laat A de gebeurtenis zijn dat de levensduur van de gekochte TV groter is dan 10 jaar of kleiner dan 5 jaar. Er geldt

$$P(A) = P(\mathbf{x} > 10 \text{ of } 0 \leq \mathbf{x} < 5) = \int_{10}^{\infty} f_{\mathbf{x}}(x)dx + \int_0^5 f_{\mathbf{x}}(x)dx =$$

$$= 1 - e^{-\frac{1}{2}} + e^{-1} = 0.762.$$

1.10. DEFINITIE. We zeggen dat een stochast $\mathbf{x}$ *homogeen verdeeld* is op het interval $[a, b]$, indien voor de kansdichtheid $f_{\mathbf{x}}(x)$ van $\mathbf{x}$ geldt

$$f_{\mathbf{x}}(x) = \begin{cases} \frac{1}{b-a} & \text{als } a \leq x \leq b, \\ 0 & \text{anders.} \end{cases}$$

DEFINITIE. We zeggen dat een stochast $\mathbf{x}$ *exponentieel verdeeld* is met parameter $\lambda > 0$, indien voor de kansdichtheid $f_{\mathbf{x}}(x)$ van $\mathbf{x}$ geldt

$$f_{\mathbf{x}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{als } x \geq 0, \\ 0 & \text{anders.} \end{cases}$$

1.11. We gaan uit van een stochast $\mathbf{x}$. Wanneer we onafhankelijk van elkaar n maal het experiment uitvoeren waarop de stochast $\mathbf{x}$ betrekking heeft, dan zullen we n *realiseringen* $x_1, x_2, \ldots, x_n$ van de stochast $\mathbf{x}$ verkrijgen. Voor grote waarden van n zullen we **bij benadering** voor het gemiddelde

$$\bar{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

steeds eenzelfde getal vinden, dat gelijk is aan het getal dat we de verwachting van de stochast $\mathbf{x}$ (notatie $E\mathbf{x}$) zullen noemen en hierna gaan definiëren.
De intuïtieve betekenis van de verwachting $E\mathbf{x}$ van een stochast $\mathbf{x}$ is dus de waarde die deze stochast $\mathbf{x}$ **gemiddeld genomen** zal aannemen.

1.12. DEFINITIE. Laat $\mathbf{x}$ een discreet verdeelde stochast zijn met mogelijke uitkomsten $u_1, u_2, \ldots$ en kansverdeling $P(u_1), P(u_2), \ldots$. Dan definiëren we de *verwachting* $E\mathbf{x}$ van $\mathbf{x}$ door

$$E\mathbf{x} = u_1 P(u_1) + u_2 P(u_2) + \ldots = \sum_i u_i P(u_i).$$

OPMERKING. Het is belangrijk te zien dat de verwachting $E\mathbf{x}$ het met de bijbehorende kansen *gewogen gemiddelde* is van alle mogelijke uitkomsten van de stochast $\mathbf{x}$.

1.13. DEFINITIE. Laat $\mathbf{x}$ een continu verdeelde stochast zijn met kansdichtheid $f_{\mathbf{x}}(x)$. Dan definiëren we de verwachting $E\mathbf{x}$ van $\mathbf{x}$ door

$$E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x) dx \qquad \text{(mits deze integraal bestaat).}$$

OPMERKING. Deze definitie is het continue analogon van de definitie in 1.12, wanneer we de onderstaande **interpretatie van de kansdichtheid** voor ogen houden:

$$P(x \leq \mathbf{x} \leq x + \Delta x) = f_{\mathbf{x}}(x)\Delta x, \text{ voor alle } x \in \mathbb{R} \text{ en 'kleine' } \Delta x.$$

Dus ook in dit geval moet de verwachting $E\mathbf{x}$ van $\mathbf{x}$ gezien worden als een met de bijbehorende kansen *gewogen gemiddelde* van alle mogelijke uitkomsten van de stochast $\mathbf{x}$.

1.14. VOORBEELDEN.
1) Beschouw de stochast $\mathbf{x}$ uit voorbeeld 1.6. Deze stochast $\mathbf{x}$ is discreet verdeeld met mogelijke uitkomsten $1, 2, ..., 6$ met kansverdeling

$$P(1) = P(2) = \ldots = P(6) = \frac{1}{6}.$$

Voor de verwachting $E\mathbf{x}$ geldt

$$E\mathbf{x} = 1.P(1) + \ldots + 6P(6) = \frac{1}{6}(1 + 2 + \ldots + 6) = \frac{21}{6}.$$

Dus $E\mathbf{x} = 3.5$.
2) Laat $\mathbf{x}$ de continu verdeelde stochast zijn uit voorbeeld 2) in 1.2 met kansdichtheid

$$f_{\mathbf{x}}(x) = \begin{cases} 0.1e^{-0.1x} & \text{als } x \geq 0, \\ 0 & \text{anders} \quad (\text{zie} \quad 1.9). \end{cases}$$

Voor de verwachting $E\mathbf{x}$ geldt

$$E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x)dx = 0.1 \int_{0}^{\infty} xe^{-0.1x}dx.$$

Na partiële integratie vinden we $E\mathbf{x} = 10$. (Ga dit na!)
3) Laat $\mathbf{x}$ een op het interval $[a, b]$ homogeen verdeelde stochast zijn. Dat wil zeggen dat voor de kansdichtheid $f_{\mathbf{x}}(x)$ van $\mathbf{x}$ geldt (zie 1.10)

$$f_{\mathbf{x}}(x) = \begin{cases} \frac{1}{b-a} & \text{als } a \leq x \leq b, \\ 0 & \text{anders.} \end{cases}$$

Voor de verwachting $E\mathbf{x}$ geldt

$$E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x)dx = \frac{1}{b-a} \int_{a}^{b} xdx = \frac{a+b}{2}.$$

4) Laat $\mathbf{x}$ een exponentieel verdeelde stochast zijn met parameter $\lambda > 0$. Dat wil zeggen dat voor de kansdichtheid $f_{\mathbf{x}}(x)$ van $\mathbf{x}$ geldt (zie 1.10)

$$f_{\mathbf{x}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{als } x \geq 0, \\ 0 & \text{anders.} \end{cases}$$

Voor de verwachting $E\mathbf{x}$ geldt

$$E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x)dx = \lambda \int_{0}^{\infty} x\, e^{-\lambda x}\, dx.$$

Na partiële integratie vinden we $E\mathbf{x} = 1/\lambda$. (Ga dit na.)

1.15. We gaan uit van een gegeven stochast $\mathbf{x}$. In de praktijk zal het vaak voorkomen dat men nieuwe stochasten gaat creëren, die functies zijn van de stochast $\mathbf{x}$, dat wil zeggen, laat $g(x)$ een functie zijn $g : \mathbb{R} \to \mathbb{R}$ en definieer de stochast $\mathbf{y}$ door

$$\mathbf{y} = g(\mathbf{x}).$$

1.16. VOORBEELD. Laat $\mathbf{x}$ een stochast zijn en neem als functie g in 1.15

$$g(x) = x^2 \quad \text{voor alle } x \in \mathbb{R}.$$

De in 2.1.15 bedoelde stochast $\mathbf{y}$ wordt nu

$$\mathbf{y} = g(\mathbf{x}) = \mathbf{x}^2.$$

1.17. Wanneer we verwachtingen moeten uitrekenen van stochasten die functies zijn van een gegeven stochast $\mathbf{x}$ dan zijn de drie volgende stellingen van belang. We zullen geen bewijs van deze stellingen geven.

1.18. STELLING. Laat $\mathbf{x}$ een discreet verdeelde stochast zijn met mogelijke uitkomsten $u_1, u_2, \ldots$ en kansverdeling $P(u_1), P(u_2) \ldots$
Verder is een functie $g : \mathbb{R} \to \mathbb{R}$ gegeven. Beschouw de stochast

$$\mathbf{y} = g(\mathbf{x}).$$

Er geldt

$$E\mathbf{y} = Eg(\mathbf{x}) = g(u_1)P(u_1) + g(u_2)P(u_2) + \ldots = \sum_i g(u_i)P(u_i).$$

OPMERKING. Ook hier geldt weer dat $E\mathbf{y}$ het met de bijbehorende kansen gewogen gemiddelde is van alle mogelijke uitkomsten $g(u_1), g(u_2), \ldots$ van de stochast $\mathbf{y}$.

1.19. STELLING. Laat $\mathbf{x}$ een continu verdeelde stochast zijn met kansdichtheid $f_{\mathbf{x}}(x)$. Verder is een functie $g : \mathbb{R} \to \mathbb{R}$ gegeven en beschouw de stochast

$$\mathbf{y} = g(\mathbf{x}).$$

Er geldt

$$E\mathbf{y} = Eg(\mathbf{x}) = \int_{-\infty}^{\infty} g(x) f_{\mathbf{x}}(x) dx.$$

OPMERKING. Ook hier geldt weer dat $E\mathbf{y}$ het met de bijbehorende kansen gewogen gemiddelde is van alle mogelijke uitkomsten van de stochast $\mathbf{y}$.

1.20. STELLING.
1) Voor willekeurige functies g and h en iedere stochast $\mathbf{x}$ geldt

$$E(g(\mathbf{x}) + h(\mathbf{x})) = Eg(\mathbf{x}) + Eh(\mathbf{x}).$$

2) Voor $a, b \in \mathbb{R}$ geldt

$$E(a\mathbf{x} + b) = aE\mathbf{x} + b.$$

1.21. VOORBEELD. Boven op een paal is een appel bevestigd. Met een geweer proberen we de appel te raken. We beschouwen de stochast

$\mathbf{x} =$ aantal schoten nodig om de appel te raken.

De stochast $\mathbf{x}$ is discreet verdeeld met mogelijke uitkomsten 1,2,...
Uit het verleden weten we gemiddeld 10 schoten nodig te hebben om de appel te raken. Dit brengt ons ertoe de volgende kansverdeling P voor de stochast $\mathbf{x}$ te definiëren:

$$P(1) = \frac{1}{10},\ P(2) = \frac{9}{10} \cdot \frac{1}{10},\ P(3) = (\frac{9}{10})^2 \frac{1}{10}, \cdots .$$

In het algemeen:

$$P(n) = (\frac{9}{10})^{n-1} \frac{1}{10}, \qquad (n = 1, 2, 3, \ldots).$$

We gaan eerst na of aan de twee eisen uit 1.5 is voldaan. Eis 1), namelijk $P(u_n) \geq 0$, is triviaal. Verder geldt

$$\sum_i P(u_i) = \sum_{n=1}^{\infty} P(u_n) = \frac{1}{10} \sum_{n=1}^{\infty} (\frac{9}{10})^{n-1} = 1,$$

omdat voor $|x| < 1$ geldt

$$\sum_{n=1}^{\infty} x^{n-1} = \frac{1}{1-x}.$$

Voor de verwachting $E\mathbf{x}$ geldt

$$E\mathbf{x} = \sum_i u_i P(u_i) = \sum_{n=1}^{\infty} nP(n) = \frac{1}{10} \sum_{n=1}^{\infty} n(\frac{9}{10})^{n-1}.$$

Omdat voor $|x| < 1$ geldt $\sum_{n=1}^{\infty} x^n = \frac{x}{1-x}$, vinden we door differentiatie

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2} \quad \text{als } |x| < 1.$$

Dus

$$E\mathbf{x} = \frac{1}{10} \cdot 10^2 = 10.$$

Wanneer we de appel na n schoten raken wordt een bedrag 100 - $9n$ uitgekeerd. We beschouwen nu de stochast

$\mathbf{y}$ = de te ontvangen uitkering.

Er geldt

$$\mathbf{y} = 100 - 9\mathbf{x}.$$

Voor de te verwachten uitkering $E\mathbf{y}$ volgt nu uit stelling 1.20 ,

$$E\mathbf{y} = E(100 - 9\mathbf{x}) = 100 - 9E\mathbf{x} = 10.$$

1.22. VOORBEELD. Gegeven is een exponentieel verdeelde stochast $\mathbf{x}$ met parameter $\lambda > 0$. We beschouwen de stochast

$$\mathbf{y} = \mathbf{x}^2.$$

Uit stelling 1.19 volgt nu

$$E\mathbf{y} = E\mathbf{x}^2 = \int_{-\infty}^{\infty} x^2 f_{\mathbf{x}}(x)dx = \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx.$$

Na tweemaal partieel integreren vinden we

$$E\mathbf{x}^2 = \frac{2}{\lambda^2} \qquad \text{(Ga dit na!)}.$$

1.23. Wanneer we een stochast $\mathbf{x}$ bestuderen zullen we in het algemeen niet alleen geïnteresseerd zijn in de verwachting $E\mathbf{x}$ van $\mathbf{x}$ maar ook in de mate van spreiding van de stochast $\mathbf{x}$ rond zijn verwachting $E\mathbf{x}$. De grootheid die een maat is voor de spreiding noemen we de variantie van $\mathbf{x}$.

DEFINITIE. De *variantie* var $\mathbf{x}$ van een stochast $\mathbf{x}$ wordt gedefinieerd door

$$\text{var}\,\mathbf{x} = E(\mathbf{x} - E\mathbf{x})^2.$$

1.24. VOORBEELD. We beschouwen een stochast $\mathbf{x}$, die homogeen verdeeld is op het interval $[a, b]$. In voorbeeld 3) van 1.14 hebben we uitgerekend

$$E\mathbf{x} = \frac{a+b}{2}.$$

Op grond van stelling 1.19 geldt nu

$$\operatorname{var} \mathbf{x} = E(\mathbf{x} - E\mathbf{x})^2 = \int_{-\infty}^{\infty} (x - \frac{a+b}{2})^2 f_{\mathbf{x}}(x)dx =$$

$$= \frac{1}{b-a} \int_a^b (x - \frac{a+b}{2})^2 dx = \frac{(a-b)^2}{12}.$$

1.25. Voor het uitrekenen van varianties zijn de volgende twee stellingen van belang:
STELLING.

$$\operatorname{var} \mathbf{x} = E\mathbf{x}^2 - (E\mathbf{x})^2.$$

BEWIJS.

$$\operatorname{var} \mathbf{x} = E(\mathbf{x} - E\mathbf{x})^2 = E(\mathbf{x}^2 - 2\mathbf{x}\, E\mathbf{x} + (E\mathbf{x})^2) =$$

$$= E\mathbf{x}^2 - 2(E\mathbf{x})^2 + (E\mathbf{x})^2 \text{ (zie 1.20)} = E\mathbf{x}^2 - (E\mathbf{x})^2.$$

1.26. STELLING. Voor iedere $a, b \in \mathbb{R}$ geldt

$$\operatorname{var}(a\mathbf{x} + b) = a^2 \operatorname{var} \mathbf{x}.$$

BEWIJS.

$$\begin{aligned} \operatorname{var}(a\mathbf{x} + b) &= E(a\mathbf{x} + b)^2 - (E(a\mathbf{x} + b))^2 = \\ &= E(a^2\mathbf{x}^2 + 2ab\mathbf{x} + b^2) - (aE\mathbf{x} + b)^2 = \\ &= a^2 E\mathbf{x}^2 + 2abE\mathbf{x} + b^2 - a^2(E\mathbf{x})^2 - 2abE\mathbf{x} - b^2 = \\ &= a^2(E\mathbf{x}^2 - (E\mathbf{x})^2) = a^2 \operatorname{var} \mathbf{x}. \end{aligned}$$

1.27 VOORBEELD. Laat $\mathbf{x}$ een exponentieel verdeelde stochast zijn met parameter $\lambda > 0$. In 1.14 voorbeeld 4) is aangetoond dat $E\mathbf{x} = 1/\lambda$.
In 1.22 is aangetoond dat $E\mathbf{x}^2 = 2/\lambda^2$. Dus

$$\operatorname{var} \mathbf{x} = E\mathbf{x}^2 - (E\mathbf{x})^2 = \frac{1}{\lambda^2}.$$

§2. SIMULTANE VERDELINGEN

2.1. In zeer veel gevallen spelen bij het uitvoeren van een experiment twee of meer stochasten tegelijkertijd een rol. In deze paragraaf bekijken we het geval dat twee stochasten $\mathbf{x}$ en $\mathbf{y}$ simultaan worden bestudeerd. We spreken in dit geval van een *paar stochasten* $(\mathbf{x}, \mathbf{y})$.

2.2. Wanneer we het experiment, waarin het paar stochasten $(\mathbf{x}, \mathbf{y})$ een rol speelt, hebben uitgevoerd beschikken we over een *realisering* (x, y) van het paar stochasten $(\mathbf{x}, \mathbf{y})$. Voor deze realisering (x, y) geldt $(x, y) \in \mathbb{R}^2$. Om kansen op zekere gebeurtenissen met betrekking tot een paar stochasten $(\mathbf{x}, \mathbf{y})$ uit te rekenen maken we vaak gebruik van een simultane kansdichtheid.

DEFINITIE. Een functie $f : \mathbb{R}^2 \to \mathbb{R}$ heet een *simultane kansdichtheid* indien

i) $f(x,y) \geq 0$ voor alle $(x,y) \in \mathbb{R}^2$.

ii) $\int\int_{\mathbb{R}^2} f(x,y)dx\,dy = 1.$ (Vergelijk 1.7.)

Om aan te geven dat een simultane kansdichtheid $f(x,y)$ bij het paar stochasten $(\mathbf{x},\mathbf{y})$ behoort noteren we deze kansdichtheid met $f_{\mathbf{x},\mathbf{y}}(x,y)$.

2.3. VOORBEELD. We kopen tegelijkertijd een radiotoestel en een TV en beschouwen het paar stochasten $(\mathbf{x},\mathbf{y})$:

$\mathbf{x}$ = levensduur van de radio,

$\mathbf{y}$ = levensduur van de TV.

We veronderstellen dat het paar stochasten $(\mathbf{x},\mathbf{y})$ een simultane kansdichtheid heeft

$$f_{\mathbf{x},\mathbf{y}}(x,y) = \begin{cases} 0.02e^{-0.1x}\,e^{-0.2y} & \text{als } x \geq 0 \text{ èn } y \geq 0, \\ 0 & \text{anders.} \end{cases}$$

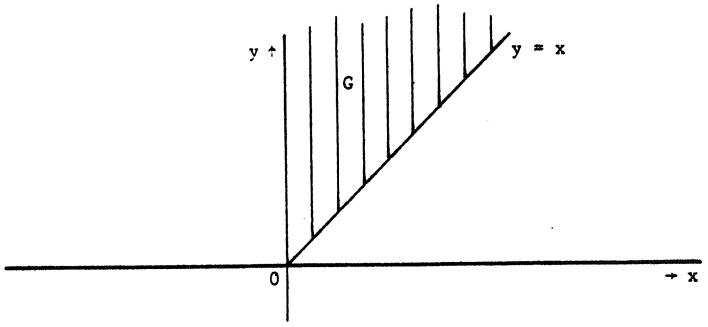
Ga na dat aan de eisen i) en ii) uit 2.2 is voldaan.
Laat A de gebeurtenis zijn dat het TV-toestel een langere levensduur heeft dan het radiotoestel. Wanneer we de kans op de gebeurtenis A, dat wil zeggen $P(A)$, willen uitrekenen kunnen we dat doen door de simultane kansdichtheid $f_{\mathbf{x},\mathbf{y}}(x,y)$ te **integreren** over het gebied

$$G = \{(x,y) \in \mathbb{R}^2 \mid y > x\}.$$

Omdat $f_{\mathbf{x},\mathbf{y}}(x,y) = 0$ buiten het eerste kwadrant, vinden we

$$P(A) = \int_G\int 0.02\,e^{-0.1x}\,e^{-0.2y}\,dx\,dy,$$

met G als in de figuur.



Dus

$$P(A) = 0.02 \int_0^\infty e^{-0.1y} \left(\int_0^y e^{-0.1x} dx \right) dy = \\ = 0.2 \int_0^\infty (e^{-0.2y} - e^{-0.3y}) dy = \frac{1}{3}.$$

2.4. Het kan ook voorkomen dat er een eindig aantal of een rij punten $(x_1, y_1), (x_2, y_2), \ldots$ in $\mathbb{R}^2$ bestaat, zo dat iedere realisering van het paar stochasten $(\mathbf{x}, \mathbf{y})$ een punt uit bovengenoemde rij is. In dit geval wordt de simultane kansverdeling P van het paar stochasten $(\mathbf{x}, \mathbf{y})$ gegeven door aan ieder punt (x_i, y_i) van de rij een kans $P(x_i, y_i)$ toe te kennen zo dat

i) $P(x_i, y_i) \geq 0$ voor alle $i = 1, 2, \ldots,$

ii) $P(x_1, y_1) + P(x_2, y_2) + \ldots = \sum_i P(x_i, y_i) = 1.$ (Vergelijk 1.5.)

2.5. VOORBEELD. We werpen met een dobbelstaan en beschouwen het paar stochasten $(\mathbf{x}, \mathbf{y})$

$$\mathbf{x} = \begin{cases} 1 \text{ als een 1 of 5} & \text{wordt gegooid,} \\ 2 \text{ als een 2 of 6} & \text{wordt gegooid,} \\ 3 \text{ als een 3} & \text{wordt gegooid,} \\ 4 \text{ als een 4} & \text{wordt gegooid.} \end{cases}$$

$$\mathbf{y} = \begin{cases} 0 \text{ als een even getal wordt gegooid,} \\ 1 \text{ als een oneven getal wordt gegooid.} \end{cases}$$

De simultane kansverdeling van het paar $(\mathbf{x}, \mathbf{y})$ kan nu overzichtelijk in een matrix weergegeven worden. We veronderstellen dat het paar $(\mathbf{x}, \mathbf{y})$ de in de volgende matrix aangegeven simultane kansverdeling bezit.

1	$\frac{1}{3}$	0	$\frac{1}{6}$	0	
0	0	$\frac{1}{3}$	0	$\frac{1}{6}$	
$\underline{y}$ / $\underline{x}$	1	2	3	4	

Laat A de gebeurtenis zijn dat $\mathbf{x} + \mathbf{y} = 4$. Wanneer we de kans op de gebeurtenis A, dat wil zeggen $P(A)$, willen uitrekenen kunnen we

dat doen door de simultane kansverdeling te **sommeren** over de punten (x_1, y_i) waarvoor geldt

$$x_i + y_i = 4.$$

Dus

$$P(A) = P(3,1) + P(4,0) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

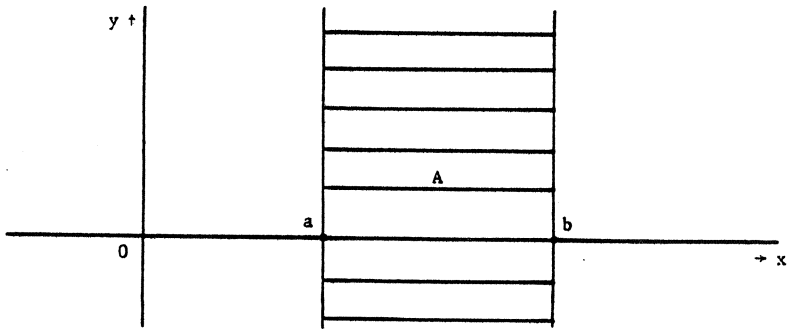
2.6. Wanneer men de beschikking heeft over de simultane kansverdeling van het paar stochasten $(\mathbf{x}, \mathbf{y})$, dan kan men ook de kansverdelingen van de stochasten $\mathbf{x}$ en $\mathbf{y}$ afzonderlijk berekenen. We noemen deze kansverdelingen van $\mathbf{x}$ en $\mathbf{y}$ afzonderlijk de *marginale kansverdelingen* van $\mathbf{x}$ respectievelijk $\mathbf{y}$.
Een manier waarop men uit de simultane kansverdeling van $(\mathbf{x}, \mathbf{y})$ de marginale kansverdelingen van $\mathbf{x}$ respectievelijk $\mathbf{y}$ kan berekenen, wordt toegelicht in de volgende stelling.

2.7. STELLING. Gegeven is een paar stochasten $(\mathbf{x}, \mathbf{y})$ met simultane kansdichtheid $f_{\mathbf{x},\mathbf{y}}(x, y)$. De stochasten $\mathbf{x}$ en $\mathbf{y}$ afzonderlijk hebben beide een *marginale kansdichtheid* $f_{\mathbf{x}}(x)$ respectievelijk $f_{\mathbf{y}}(y)$ die wordt gegeven door

$$f_{\mathbf{x}}(x) = \int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(x, y)dy \text{ èn } f_{\mathbf{y}}(y) = \int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(x, y)dx.$$

AFLEIDING. Laat $[a, b] \subset \mathbb{R}$ een interval zijn. Dan geldt

$$P(\mathbf{x} \in [a, b]) = \int_A \int f_{\mathbf{x},\mathbf{y}}(x, y)dx\,dy \quad \text{met } A \subset \mathbb{R}^2$$



$$= \int_a^b \left(\int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(x,y)dy \right) dx.$$

Dus

$$f_{\mathbf{x}}(x) = \int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(x,y)dy.$$

2.8. VOORBEELD. Beschouw het paar $(\mathbf{x},\mathbf{y})$ uit 2.3 en laat zien dat

$$f_{\mathbf{x}}(x) = \begin{cases} 0.1\, e^{-0.1x} & \text{als } x \geq 0, \\ 0 & \text{als } x < 0; \end{cases}$$

$$f_{\mathbf{y}}(y) = \begin{cases} 0.2\, e^{-0.2y} & \text{als } y \geq 0, \\ 0 & \text{als } y < 0. \end{cases}$$

2.9. VOORBEELD. Beschouw het paar stochasten $(\mathbf{x},\mathbf{y})$ uit 2.5. De marginale kansverdelingen van $\mathbf{x}$ en $\mathbf{y}$ kunnen in dit geval eenvoudig berekend worden:

$$P(\mathbf{x}=1) = P(\mathbf{x}=2) = \frac{1}{3} \text{ èn } P(\mathbf{x}=3) = P(\mathbf{x}=4) = \frac{1}{6},$$

$$P(\mathbf{y}=0) = P(\mathbf{y}=1) = \frac{1}{2}.$$

Deze marginale kansverdelingen zijn ook in de matrix uit 2.5 aan te geven. In onderstaande matrix zijn dus de **simultane kansverdeling** van $(\mathbf{x},\mathbf{y})$ en de **marginale kansverdelingen** van $\mathbf{x}$ en $\mathbf{y}$ overzichtelijk aangegeven.

	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	
1	$\frac{1}{3}$	0	$\frac{1}{6}$	0	$\frac{1}{2}$
0	0	$\frac{1}{3}$	0	$\frac{1}{6}$	$\frac{1}{2}$
$\underline{y}$ / $\underline{x}$	1	2	3	4	

2.10. Van de nu volgende stelling geven we geen afleiding.
STELLING. Laat $(\mathbf{x}, \mathbf{y})$ een paar stochasten zijn en g een functie van twee variabelen. Voor de verwachting van de stochast $\mathbf{z} = g(\mathbf{x}, \mathbf{y})$ geldt:

$$E\mathbf{z} = Eg(\mathbf{x}, \mathbf{y}) = \sum_i g(x_i, y_i) P(x_i, y_i) \quad \text{in het discrete geval,}$$

en

$$E\mathbf{z} = Eg(\mathbf{x}, \mathbf{y}) = \int\int_{\mathbb{R}^2} g(x, y) f_{\mathbf{x},\mathbf{y}}(x, y) dx\, dy$$

in het continue geval (aangenomen dat deze verwachting bestaat).

2.11. OPMERKING. Ook in stelling 2.10 komt weer tot uitdrukking dat de verwachting van een stochast niets anders is dan het met de bijbehorende kansen gewogen gemiddelde van alle mogelijke waarden van de stochast. De **interpretatie** van de simultane kansdichtheid $f_{\mathbf{x},\mathbf{y}}(x, y)$ is dan ook

$$P(x < \mathbf{x} < x + \Delta x \text{ èn } y < \mathbf{y} < y + \Delta y) = f_{\mathbf{x},\mathbf{y}}(x, y) \Delta x\, \Delta y.$$

2.12. STELLING. Voor ieder paar stochasten $(\mathbf{x}, \mathbf{y})$ geldt

$$E(\mathbf{x} + \mathbf{y}) = E\mathbf{x} + E\mathbf{y}.$$

AFLEIDING. We veronderstellen dat het paar $(\mathbf{x}, \mathbf{y})$ een simultane kansdichtheid $f_{\mathbf{x},\mathbf{y}}(x, y)$ heeft en geven voor dit geval een afleiding.

$$\begin{aligned}
E(\mathbf{x}+\mathbf{y}) &= \int\int_{\mathbb{R}^2} (x+y) f_{\mathbf{x},\mathbf{y}} dx dy = \\
&= \int\int_{\mathbb{R}^2} x f_{\mathbf{x},\mathbf{y}}(x, y) dx dy + \int\int_{\mathbb{R}^2} y f_{\mathbf{x},\mathbf{y}}(x, y) dx dy = \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(x, y) dy \right) dx + \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(x, y) dx \right) dy = \\
&= \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x) dx + \int_{-\infty}^{\infty} y f_{\mathbf{y}}(y) dy = E\mathbf{x} + E\mathbf{y}.
\end{aligned}$$

2.13. OPMERKING. Voor andere bewerkingen dan de som geldt een dergelijke eigenschap doorgaans niet. Zo is bijvoorbeeld in het algemeen $E(\mathbf{xy}) \neq E\mathbf{x}E\mathbf{y}$.

2.14. VOORBEELD. Beschouw het paar $(\mathbf{x}, \mathbf{y})$ uit 2.5. Er geldt

$$\begin{aligned}
E\mathbf{x} &= 1.P(\mathbf{x}=1) + 2P(\mathbf{x}=2) + 3P(\mathbf{x}=3) + 4P(\mathbf{x}=4) = \\
&= \frac{1}{3} + \frac{2}{3} + \frac{3}{6} + \frac{4}{6} = 2\frac{1}{6}. \\
E\mathbf{y} &= 0.P(\mathbf{y}=0) + 1.P(\mathbf{y}=1) = \frac{1}{2} \\
E(\mathbf{xy}) &= 1.P(\mathbf{x}=1, \mathbf{y}=1) + 2P(\mathbf{x}=2, \mathbf{y}=1) + 3P(\mathbf{x}=3, \mathbf{y}=1) + \\
&\quad + 4P(\mathbf{x}=4, \mathbf{y}=1) = \frac{1}{3} + 0 + \frac{3}{6} + 0 = \frac{5}{6}.
\end{aligned}$$

Hier geldt dus $E\mathbf{xy} \neq E\mathbf{x}\, E\mathbf{y}$.

2.15. Laat $(\mathbf{x}, \mathbf{y})$ een paar stochasten zijn en $a, b \in \mathbb{R}$. We beschouwen nu drie gebeurtenissen:

$$A = (\mathbf{x} \leq a),\ B = (\mathbf{y} \leq b) \text{ èn } C = A \cap B = (\mathbf{x} \leq a \text{ èn } \mathbf{y} \leq b).$$

Voor de hierbij behorende kansen geldt in het algemeen

$$P(C) = P(A \cap B) \neq P(A)P(B).$$

DEFINITIE. Laat $(\mathbf{x}, \mathbf{y})$ een paar stochasten zijn. Dan noemen we de stochasten $\mathbf{x}$ en $\mathbf{y}$ (*stochastisch*) *onafhankelijk* indien voor alle $a, b \in \mathbb{R}$ geldt

$$(\mathbf{x} \leq a \text{ èn } y \leq b) = P(\mathbf{x} \leq a)P(\mathbf{y} \leq b).$$

Wanneer $\mathbf{x}$ en $\mathbf{y}$ niet onafhankelijk zijn, heten ze *afhankelijk*.

2.16. Wanneer het paar $(\mathbf{x}, \mathbf{y})$ een simultane kansdichtheid heeft of wanneer $(\mathbf{x}, \mathbf{y})$ een discrete kansverdeling heeft, kunnen we aan de simultane kansdichtheid zien of de stochasten $\mathbf{x}$ and $\mathbf{y}$ al dan niet onafhankelijk zijn. Dit wordt geformuleerd in de volgende stelling, die we niet bewjzen.

STELLING. Twee stochasten $\mathbf{x}$ en $\mathbf{y}$ zijn onafhankelijk dan en slechts dan indien

$$P(\mathbf{x} = x_i \text{ èn } \mathbf{y} = y_i) = P(\mathbf{x} = x_i)P(\mathbf{y} = y_i)$$

voor alle i in het discrete geval, en

$$f_{\mathbf{x},\mathbf{y}}(x, y) = f_{\mathbf{x}}(x) f_{\mathbf{y}}(y)$$

voor alle x en y in het continue geval.

2.17 STELLING. Laat $\mathbf{x}$ en $\mathbf{y}$ onafhankelijke stochasten zijn. Dan geldt voor ieder tweetal verzamelingen $A \subset \mathbb{R}$ en $B \subset \mathbb{R}$

$$P(\mathbf{x} \in A \text{ èn } \mathbf{y} \in B) = P(\mathbf{x} \in A)P(\mathbf{y} \in B).$$

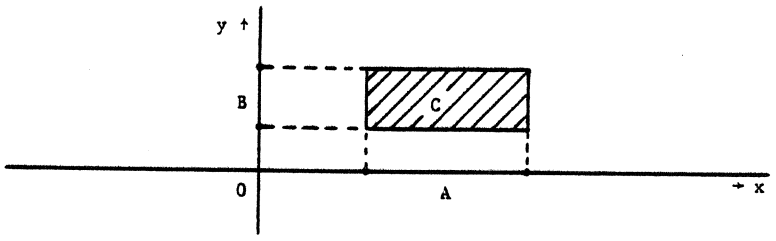
AFLEIDING. We geven een afleiding voor het geval dat het paar $(\mathbf{x}, \mathbf{y})$ een simultane kansdichtheid $f_{\mathbf{x},\mathbf{y}}(x, y)$ heeft. Uit 2.16 volgt

$$f_{\mathbf{x},\mathbf{y}}(x, y) = f_{\mathbf{x}}(x) f_{\mathbf{y}}(y).$$

Er geldt

$$P(\mathbf{x} \in A \text{ èn } \mathbf{y} \in B) = \int_C \int f_{\mathbf{x},\mathbf{y}}(x, y)\, dx\, dy$$

met C als in de figuur.



$$= \int_C \int f_{\mathbf{x}}(x) f_{\mathbf{y}}(y)\, dx\, dy = \int_A f_{\mathbf{x}}(x) \left(\int_B f_{\mathbf{y}}(y) dy \right) dx =$$

$$= \int_A f_{\mathbf{x}}(x) dx \cdot \int_B f_{\mathbf{y}}(y) dy = P(\mathbf{x} \in A) \cdot P(\mathbf{y} \in B).$$

2.18 VOORBEELD. Beschouw het paar stochasten $(\mathbf{x}, \mathbf{y})$ uit 2.3. Uit 2.8 volgt

$$f_{\mathbf{x},\mathbf{y}}(x, y) = f_{\mathbf{x}}(x) f_{\mathbf{y}}(y).$$

Op grond van 2.16 kunnen we concluderen dat de stochasten $\mathbf{x}$ en $\mathbf{y}$ onafhankelijk zijn.

2.19. VOORBEELD. Beschouw het paar stochasten $(\mathbf{x}, \mathbf{y})$ uit 2.5. Er geldt

$$P(\mathbf{x} = 2 \text{ èn } \mathbf{y} = 1) = 0$$

en

$$P(\mathbf{x} = 2) P(\mathbf{y} = 1) = \frac{1}{6}.$$

Op grond van 2.16 kunnen we concluderen dat de stochasten $\mathbf{x}$ en $\mathbf{y}$ afhankelijk zijn.

2.20. STELLING. Als $\mathbf{x}$ en $\mathbf{y}$ onafhankelijke stochasten zijn, dan geldt

i) $E(\mathbf{xy}) = E\mathbf{x}\, E\mathbf{y}$.

ii) $\operatorname{var}(\mathbf{x} + \mathbf{y}) = \operatorname{var} \mathbf{x} + \operatorname{var} \mathbf{y}$.

AFLEIDING. i) We geven een afleiding voor het geval $(\mathbf{x}, \mathbf{y})$ een simultane kansdichtheid heeft.

$$E(\mathbf{xy}) = \int \int_{\mathbb{R}^2} xy f_{\mathbf{x},\mathbf{y}}(x, y) dx dy = \int \int_{\mathbb{R}^2} xy f_{\mathbf{x}}(x) f_{\mathbf{y}}(y) dx dy =$$

$$= \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x)dx \int_{-\infty}^{\infty} y f_{\mathbf{y}}(y)dy = E\mathbf{x}E\mathbf{y}.$$

ii) $\begin{aligned} \operatorname{var}(\mathbf{x}+\mathbf{y}) &= E(\mathbf{x}+\mathbf{y})^2 - (E(\mathbf{x}+\mathbf{y}))^2 = \\ &= E\mathbf{x}^2 + 2E(\mathbf{xy}) + E\mathbf{y}^2 - (E\mathbf{x})^2 - 2E\mathbf{x}E\mathbf{y} - (E\mathbf{y})^2 = \\ &= (E\mathbf{x}^2 - (E\mathbf{x})^2) + (E\mathbf{y}^2 - (E\mathbf{y})^2) = \operatorname{var}\mathbf{x} + \operatorname{var}\mathbf{y}. \end{aligned}$

2.21. STELLING. Laat $\mathbf{x}$ en $\mathbf{y}$ twee onafhankelijke stochasten zijn en g en h twee willekeurige functies. Dan zijn de stochasten $\mathbf{u} = g(\mathbf{x})$ en $\mathbf{v} = h(\mathbf{y})$ onafhankelijk.

AFLEIDING. Zij $a, b \in \mathbb{R}$ en beschouw de gebeurtenissen

$$A = (\mathbf{u} \leq a) = (g(\mathbf{x}) \leq a) \text{ en } B = (\mathbf{v} \leq b) = (h(\mathbf{y}) \leq b).$$

Definieer de verzamelingen

$$\tilde{A} = \{x \in \mathbb{R} | g(x) \leq a\} \text{ en } \tilde{B} = \{y \in \mathbb{R} | h(y) \leq b\},$$

dan geldt

$$A = (\mathbf{x} \in \tilde{A}) \text{ èn } B = (\mathbf{y} \in \tilde{B}\}.$$

Uit 2.2.17 volgt

$$\begin{aligned} P(\mathbf{u} \leq a \text{ èn } \mathbf{v} \leq b) &= P(\mathbf{x} \in \tilde{A} \text{ èn } \mathbf{y} \in \tilde{B}) = P(\mathbf{x} \in \tilde{A})P(\mathbf{y} \in \tilde{B}) = \\ &= P(g(\mathbf{x}) \leq a)P(h(\mathbf{y}) \leq b) = P(\mathbf{u} \leq a)P(\mathbf{v} \leq b). \end{aligned}$$

2.22. VOORBEELD. Gegeven zijn twee onafhenkelijke stochasten $\mathbf{x}$ en $\mathbf{y}$ met

$$E\mathbf{x} = 1 \text{ en } \operatorname{var}\mathbf{x} = 2, \quad E\mathbf{y} = 3 \text{ en } \operatorname{var}\mathbf{y} = 3.$$

Er geldt

$$E(\mathbf{x}^2\mathbf{y}) = E\mathbf{x}^2 E\mathbf{y} = 3(\operatorname{var}\mathbf{x} + (E\mathbf{x})^2) = 9$$

en

$$\operatorname{var}(3\mathbf{x} - \mathbf{y}) = \operatorname{var}(3\mathbf{x}) + \operatorname{var}(-\mathbf{y}) = 9 \operatorname{var}\mathbf{x} + \operatorname{var}\mathbf{y} = 21.$$

2.23. Wanneer twee stochasten $\mathbf{x}$ en $\mathbf{y}$ onafhankelijk zijn dan geldt $E\mathbf{xy} = E\mathbf{x}E\mathbf{y}$ (zie 2.20). Het omgekeerde is echter **niet waar**, d.w.z. als voor twee stochasten $\mathbf{x}$ en $\mathbf{y}$ geldt $E\mathbf{xy} = E\mathbf{x}E\mathbf{y}$, dan zijn $\mathbf{x}$ en $\mathbf{y}$ niet noodzakelijk onafhankelijk.
DEFINITIE. We noemen twee stochasten $\mathbf{x}$ en $\mathbf{y}$ *ongecorreleerd* indien

$$E\mathbf{xy} = E\mathbf{x}\, E\mathbf{y}.$$

Dat twee ongecorreleerde stochasten niet noodzakelijk onafhankelijk zijn wordt geïllustreerd in het volgende voorbeeld.

2.24. We beschouwen het paar stochasten $(\mathbf{x}, \mathbf{y})$ dat homogeen verdeeld is op de cirkel met straal 1 en middelpunt (0,0). Dat wil zeggen:

$$f_{\mathbf{x},\mathbf{y}}(x,y) = \begin{cases} \frac{1}{\pi} & \text{als } x^2 + y^2 \leq 1 \\ 0 & \text{anders} \end{cases}$$

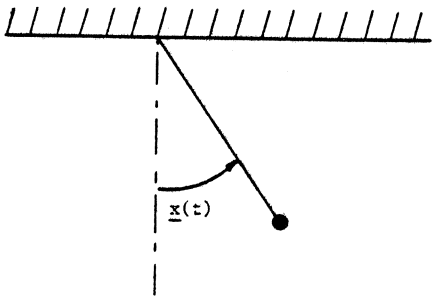
Ga na dat de stochasten $\mathbf{x}$ en $\mathbf{y}$ dezelfde marginale kansdichtheid hebben en dat er geldt:

$$f_{\mathbf{x}}(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{als } |x| \leq 1 \\ 0 & \text{anders.} \end{cases}$$

Omdat $f_{\mathbf{x},\mathbf{y}}(x,y) \neq f_{\mathbf{x}}(x) f_{\mathbf{y}}(y)$ concluderen we dat de stochasten $\mathbf{x}$ en $\mathbf{y}$ **afhankelijk** zijn (zie 2.16).
Ga na dat op grond van symmetrie overwegingen geldt $E\mathbf{xy} = E\mathbf{x} = E\mathbf{y} = 0$. Omdat $E\mathbf{xy} = E\mathbf{x}E\mathbf{y}$ zijn $\mathbf{x}$ en $\mathbf{y}$ **ongecorreleerd**.

§3. Stochastische processen

3.1. We beschouwen de slinger in onderstaande figuur en veronderstellen dat voor de hoek $\mathbf{x}(t)$ als functie van de tijd t geldt
(1) $\mathbf{x}(t) = \cos(2\pi t + \boldsymbol{\theta})$.



Omdat de beginfase $\boldsymbol{\theta}$ van het toeval afhangt, dat wil zeggen een stochast is, is ook de hoek $\mathbf{x}(t)$ voor ieder tijdstip $t \in \mathbb{R}$ een stochast. Indien voor ieder tijdstip $t \in \mathbb{R}$ op een of andere manier een stochast $\mathbf{x}(t)$ is vastgelegd spreken we van een *stochastisch proces* $\mathbf{x}(t)$.

Wanneer we in werkelijkheid de beweging van de slinger registreren zullen we één van de vele *realiseringen* $x(t)$ van het stochastisch proces $\mathbf{x}(t)$ waarnemen, bijvoorbeeld

$$x(t) = \cos(2\pi t + 1).$$

De collectie van alle realiseringen van een stochastisch proces $\mathbf{x}(t)$ noemen we het *ensemble* van het proces $\mathbf{x}(t)$. Het ensemble van het proces $\mathbf{x}(t)$ uit (1) bestaat dus uit de collectie signalen

$$x(t) = \cos(2\pi t + \theta) \quad \text{met } \theta \in \mathbb{R},$$

dat wil zeggen alle harmonische signalen met amplitude 1 en frequentie 1.

We veronderstellen dat de beginfase $\boldsymbol{\theta}$ in het proces $\mathbf{x}(t)$ uit (1) een op het interval $[0, \frac{1}{2}]$ homogeen verdeelde stochast is, dat wil zeggen voor de kansdichtheid $f_{\boldsymbol{\theta}}(\theta)$ van $\boldsymbol{\theta}$ geldt:

$$(2) \qquad f_{\boldsymbol{\theta}}(\theta) = \begin{cases} 2 & \text{als } 0 \leq \theta \leq \frac{1}{2} \\ 0 & \text{anders.} \end{cases}$$

Laat $t \in \mathbb{R}$ een vast tijdstip zijn en beschouw de *verwachtingsfunctie* $\mu_{\mathbf{x}}(t)$ van het proces $\mathbf{x}(t)$ gedefinieerd door

$$\mu_{\mathbf{x}}(t) = E\mathbf{x}(t).$$

De verwachting $E\mathbf{x}(t)$ wordt uitgerekend door voor iedere realisering in het ensemble van het proces $\mathbf{x}(t)$ zijn waarde op tijdstip t te bepalen en deze waarde te vermenigvuldigen (te wegen) met de kans dat de beschouwde realisering wordt waargenomen. Vervolgens dient men te sommeren (integreren) over alle realiseringen in het ensemble van het proces $\mathbf{x}(t)$.
Voor het proces $\mathbf{x}(t)$ uit (1) waarbij $\boldsymbol{\theta}$ de kansdichtheid uit (2) bezit, vinden we

$$\mu_{\mathbf{x}}(t) = E\mathbf{x}(t) = \int_{-\infty}^{\infty} \cos(2\pi t + \theta) f_{\boldsymbol{\theta}}(\theta) d\theta =$$

$$= 2 \int_0^{\frac{1}{2}} \cos(2\pi t + \theta) d\theta = 2(\sin(2\pi t + \frac{1}{2}) - \sin 2\pi t) =$$

$$= 4 \sin(\frac{1}{4}) \cos(2\pi t + \frac{1}{4}).$$

De verwachtingsfunctie $\mu_{\mathbf{x}}(t)$ is een voorbeeld van een *ensemble gemiddelde*, dat wil zeggen men berekent voor een vaste tijd t een (gewogen) gemiddelde over alle realiseringen in het ensemble van een proces $\mathbf{x}(t)$.
De *variantiefunctie* $v_{\mathbf{x}}(t)$ van een proces $\mathbf{x}(t)$ wordt gedefinieerd door

$$v_{\mathbf{x}}(t) = \operatorname{var} \mathbf{x}(t) = E(\mathbf{x}(t) - \mu_{\mathbf{x}}(t))^2.$$

Voor het proces $\mathbf{x}(t)$ uit (1) waarbij $\boldsymbol{\theta}$ de kansdichtheid uit (2) bezit vinden we

$$v_{\mathbf{x}}(t) = E(\mathbf{x}(t))^2 - 16 \sin^2(\tfrac{1}{4}) \cos^2(2\pi t + \tfrac{1}{4}).$$

Het ensemble gemiddelde $E(\mathbf{x}(t))^2$ wordt nu dus als volgt uitgerekend

$$E(\mathbf{x}(t))^2 = \int_{-\infty}^{\infty} \cos^2(2\pi t + \theta) f_{\boldsymbol{\theta}}(\theta) d\theta = 2 \int_0^{\frac{1}{2}} \cos^2(2\pi t + \theta) d\theta =$$
$$= \frac{1}{2} + \sin(\tfrac{1}{2}) \cos(4\pi t + \tfrac{1}{2}).$$

Dus

$$v_{\mathbf{x}}(t) = \frac{1}{2} + \sin(\tfrac{1}{2}) \cos(4\pi t + \tfrac{1}{2}) - 16 \sin^2(\tfrac{1}{4}) \cos^2(2\pi t + \tfrac{1}{4}).$$

Laten $t_1 \in \mathbb{R}$ en $t_2 \in \mathbb{R}$ twee vaste tijdstippen zijn en beschouw de *stochastische autocorrelatiefunctie* $R_{\mathbf{xx}}(t_1, t_2)$ van het proces $\mathbf{x}(t)$ gedefinieerd door

$$R_{\mathbf{xx}}(t_1, t_2) = E\mathbf{x}(t_1)\mathbf{x}(t_2).$$

De stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(t_1, t_2)$ is dus weer een ensemble gemiddelde en wordt voor het proces $\mathbf{x}(t)$ uit (1), waarbij $\boldsymbol{\theta}$ de kansdichtheid uit (2) bezit, als volgt uitgerekend:

$$\begin{aligned} R_{\mathbf{xx}}(t_1, t_2) &= E\mathbf{x}(t_1)\mathbf{x}(t_2) = \\ &= \int_{-\infty}^{\infty} \cos(2\pi t_1 + \theta) \cos(2\pi t_2 + \theta) f_{\boldsymbol{\theta}}(\theta) d\theta = \\ &= 2 \int_0^{\frac{1}{2}} \cos(2\pi t_1 + \theta) \cos(2\pi t_2 + \theta) d\theta = \\ &= \sin(\tfrac{1}{2}) \cos(2\pi(t_1 + t_2) + \tfrac{1}{2}) + \frac{1}{2} \cos(2\pi(t_1 - t_2)). \end{aligned}$$

3.2. De belangrijkste begrippen uit 3.1 vatten we hier nog eens samen.

DEFINITIE. We spreken van een *stochastisch proces* $\mathbf{x}(t)$ indien voor ieder tijdstip $t \in \mathbb{R}$ een stochast $\mathbf{x}(t)$ is vastgelegd.

DEFINITIE. Wanneer we een stochastisch proces $\mathbf{x}(t)$ waarnemen (registreren) nemen we één *realisering* $x(t)$ van het proces $\mathbf{x}(t)$ waar. De collectie van alle mogelijke realiseringen (*ensemble signalen, random signalen*) noemen we het *ensemble* van het proces $\mathbf{x}(t)$.

Enkele belanrijke *ensemble gemiddelden*:

DEFINITIE. Zij $\mathbf{x}(t)$ een stochastisch proces. Voor ieder tijdstip $t \in \mathbb{R}$ definiëren we de *verwachtingsfunctie* $\mu_{\mathbf{X}}(t)$ door

$$\mu_{\mathbf{X}}(t) = E\mathbf{x}(t),$$

en de *variantiefunctie* $v_{\mathbf{X}}(t)$ door

$$v_{\mathbf{X}}(t) = \operatorname{var} \mathbf{x}(t) = E(\mathbf{x}(t) - \mu_{\mathbf{X}}(t))^2 = E(\mathbf{x}(t))^2 - (\mu_{\mathbf{X}}(t))^2.$$

Voor ieder tweetal tijstippen $t_1 \in \mathbb{R}$ en $t_2 \in \mathbb{R}$ definiëren we de *stochastische autocorrelatiefunctie* $R_{\mathbf{XX}}(t_1, t_2)$ door

$$R_{\mathbf{XX}}(t_1.t_2) = E\mathbf{x}(t_1)\mathbf{x}(t_2).$$

3.3. DEFINITIE. Zij $\mathbf{x}(t)$ een stochastisch proces met verwachtingsfunctie $\mu_{\mathbf{X}}(t)$ en stochastische autocorrelatiefunctie $R_{\mathbf{XX}}(t_1, t_2)$.
Het proces $\mathbf{x}(t)$ heet *stationair* indien:
i) De verwachtingsfunctie $\mu_{\mathbf{X}}(t)$ niet van de tijd t afhangt, dat wil zeggen dat er een getal $\mu_{\mathbf{X}}(t)$ bestaat zo dat

$$\mu_{\mathbf{X}}(t) = \mu_{\mathbf{X}} \quad \text{voor alle } t \in \mathbb{R}.$$

ii) De stochastische autocorrelatiefunctie $R_{\mathbf{XX}}(t_1, t_2)$ slechts afhangt van het verschil $t_2 - t_1$, dat wil zeggen dat er een functie $R_{\mathbf{XX}}(\tau)$ van één variabele τ bestaat zo dat

$$R_{\mathbf{XX}}(t_1, t_2) = R_{\mathbf{XX}}(t_1 - t_2) \quad \text{voor alle } t_1 \in \mathbb{R} \text{ en } t_2 \in \mathbb{R}.$$

3.4. OPMERKINGEN. 1) In het geval van een stationair proces $\mathbf{x}(t)$ noemen we het getal $\mu_{\mathbf{X}}$ de *verwachting* van het proces $\mathbf{x}(t)$.
2) In het geval van een stationair proces $\mathbf{x}(t)$ noemen we de functie $R_{\mathbf{XX}}(\tau)$ van één variabele de *stochatische autocorrelatiefunctie* van het proces $\mathbf{x}(t)$. Voor een stationair proces $\mathbf{x}(t)$ geldt dus

$$R_{\mathbf{XX}}(t_1, t_2) = R_{\mathbf{XX}}(0, t_2 - t_1) = R_{\mathbf{XX}}(t_2 - t_1) \quad \text{voor alle } t_1, t_2 \in \mathbb{R}.$$

3.5. STELLING. Bij een stationair proces $\mathbf{x}(t)$ hangt de variantiefunctie niet van de tijd t af. De constante waarde van de variantiefunctie duiden we aan met $v_{\mathbf{X}}$ en wordt de *variantie* van het proces $\mathbf{x}(t)$ genoemd.

BEWIJS.

$$v_{\mathbf{X}}(t) = E\mathbf{x}(t)\mathbf{x}(t) - \mu_{\mathbf{X}}^2 = R_{\mathbf{XX}}(0) - \mu_{\mathbf{X}}^2.$$

3.6. VOORBEELD. Ga na dat het proces $\mathbf{x}(t)$ gegeven door (1) in 3.1, waarbij de kansdichtheid van $\boldsymbol{\theta}$ gegeven wordt door (2) in 3.1, niet stationair is.

3.7. VOORBEELD. We beschouwen het stochastische proces

$$\mathbf{x}(t) = \mathbf{a}\cos(2\pi t + \boldsymbol{\theta}),$$

waarbij we veronderstellen dat de amplitude $\mathbf{a}$ en de beginfase $\boldsymbol{\theta}$ onafhankelijke stochasten zijn. De amplitude $\mathbf{a}$ is exponentieel verdeeld met parameter $\lambda = 1$ en de beginfase $\boldsymbol{\theta}$ is homogeen verdeeld op $[0, 2\pi]$, dat wil zeggen, voor de kansdichtheden $f_{\mathbf{a}}(a)$ en $f_{\boldsymbol{\theta}}(\theta)$ geldt:

$$f_{\mathbf{a}}(a) = \begin{cases} e^{-a} & \text{als } a \geq 0, \\ 0 & \text{anders} \end{cases}$$

en

$$f_{\boldsymbol{\theta}}(\theta) = \begin{cases} \dfrac{1}{2\pi} & \text{als } 0 \leq \theta \leq 2\pi, \\ 0 & \text{anders.} \end{cases}$$

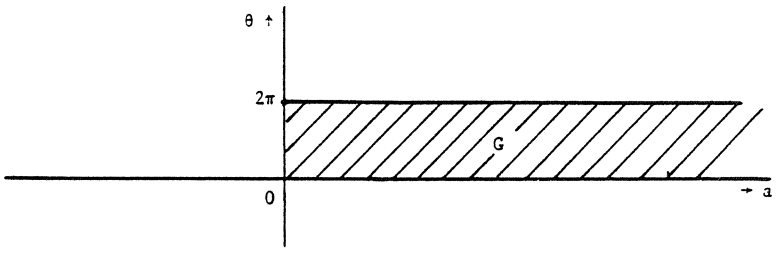
Het ensemble van dit proces $\mathbf{x}(t)$ is de collectie van alle harmonische signalen met frequentie $f = 1$ Hz.
Om te onderzoeken of dit proces $\mathbf{x}(t)$ stationair is bepalen we de verwachtingsfuctie $\mu_{\mathbf{x}}(t)$ en de stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(t_1, t_2)$ van het proces $\mathbf{x}(t)$. Er geldt

$$\mu_{\mathbf{x}}(t) = E\mathbf{x}(t) = \int\int_{\mathbb{R}^2} a\cos(2\pi t + \theta) f_{\mathbf{a},\boldsymbol{\theta}}(a, \theta)\, da\, d\theta.$$

Omdat de stochasten $\mathbf{a}$ en $\boldsymbol{\theta}$ onafhankelijk zijn geldt voor de simultane kansdichtheid $f_{\mathbf{a},\boldsymbol{\theta}}(a, \theta)$ van het paar stochasten $(\mathbf{a}, \boldsymbol{\theta})$:

$$f_{\mathbf{a},\boldsymbol{\theta}}(a, \theta) = f_{\mathbf{a}}(a) f_{\boldsymbol{\theta}}(\theta) = \begin{cases} \dfrac{1}{2\pi} e^{-a} & \text{als } a \geq 0 \text{ èn } 0 \leq \theta \leq 2\pi \\ 0 & \text{anders.} \end{cases}$$

In onderstaande figuur is een schets gemaakt van het gebied G waarbuiten de simultane kansdichtheid $f_{\mathbf{a},\boldsymbol{\theta}}(a, \theta)$ nul is.



Dus

$$\mu_{\mathbf{X}}(t) = \frac{1}{2\pi}\int_G\int a\,\cos(2\pi t+\theta)e^{-a}\,da\,d\theta =$$

$$= \frac{1}{2\pi}\int_0^{\infty} ae^{-a}\left(\int_0^{2\pi}\cos(2\pi t+\theta)d\theta\right)da = 0.$$

We zien dus dat $\mu_{\mathbf{X}}(t)$ **niet van de tijd** t **afhangt.**
Voor ieder tweetal tijdstippen $t_1 \in \mathbb{R}$ en $t_2 \in \mathbb{R}$ geldt:

$$R_{\mathbf{XX}}(t_1,t_2) = E\mathbf{x}(t_1)\mathbf{x}(t_2) =$$

$$= \int\int_{\mathbb{R}^2} a^2\,\cos(2\pi t_1+\theta)\cos(2\pi t_2+\theta)f_{\mathbf{a},\boldsymbol{\theta}}(a,\theta)da\,d\theta =$$

$$= \frac{1}{2\pi}\int_0^{\infty} a^2e^{-a}\left(\int_0^{2\pi}\cos(2\pi t_1+\theta)\cos(2\pi t_2+\theta)d\theta\right)da =$$

$$= \frac{1}{4\pi}\int_0^{\infty} a^2e^{-a}\left(\int_0^{2\pi}[\cos(2\pi t_1+2\pi t_2+2\theta)+\cos(2\pi t_2-2\pi t_2)]d\theta\right)da =$$

$$= \frac{1}{2}\cos[2\pi(t_2-t_1)]\int_0^{\infty} a^2e^{-a}da = \cos[2\pi(t_2-t_2)].$$

We zien dus dat $R_{\mathbf{XX}}(t_1,t_2)$ slechts **afhangt van het verschil** t_2-t_1.
Op grond van 3.3 kunnen we concluderen dat dit proces $\mathbf{x}(t)$ stationair is. Voor de verwachting $\mu_{\mathbf{X}}$ en de stochastische autocorrelatie $R_{\mathbf{XX}}(\tau)$ van dit stationaire proces $\mathbf{x}(t)$ geldt

$$\mu_{\mathbf{X}} = 0 \text{ èn } R_{\mathbf{XX}}(\tau) = \cos\,2\pi\tau.$$

Ga na dat voor de variante $v_{\mathbf{X}}$ van het proces $\mathbf{x}(t)$ geldt

$$v_{\mathbf{X}} = 1.$$

3.8. Voor de feitelijke meting van de stochastische autocorrelatiefunctie $R_{\mathbf{XX}}(\tau)$ van een stationair proces $\mathbf{x}(t)$ zou men in principe als volgt te werk kunnen gaan: we nemen n identieke en onafhankelijk werkende apparaten die elk een realisering (random signaal) uit het ensemble van het proces $\mathbf{x}(t)$ leveren. Daarmee krijgen we de beschikking over n ensemble signalen $x_1(t),\ldots,x_n(t)$ uit het ensemble van het proces $\mathbf{x}(t)$.
Fixeer nu een tijdstip τ. Dan zal voor grote n de waarde van $R_{\mathbf{XX}}(\tau)$ goed benaderd worden door

$$\frac{1}{n}(x_1(0)x_1(\tau)+\ldots+x_n(0)x_n(\tau)).$$

Merk op dat ook hier weer duidelijk naar voren komt dat $R_{\mathbf{XX}}(\tau)$ een ensemble gemiddelde is.
Doe dit voor vele waarden van τ, dan is ook $R_{\mathbf{XX}}$ als functie van τ bekend.
Het behoeft geen betoog, dat de hier geschetste methode zeer bewerkelijk is.
Voor vele stationaire processen bestaat er echter een eenvoudiger en snellere methode.

3.9. DEFINITIE. Laat $\mathbf{x}(t)$ een *stationair* proces zijn. Dan definiëren we voor iedere realisering (random signaal) $x(t)$ in het ensemble van het stationaire proces $\mathbf{x}(t)$

$$\bar{x} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt,$$

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau)dt.$$

OPMERKINGEN. 1) Naar aanleiding van de hierboven gegeven definities rijst natuurlijk de vraag of de voorkomende limieten bestaan. We volstaan met te vermelden dat G.D. Birkhoff in 1931 bewezen heeft dat voor alle in de praktijk voorkomende realiseringen $x(t)$ uit het ensemble van een stationair proces $\mathbf{x}(t)$ de bovenstaande limieten inderdaad bestaan.
2) Het is belangrijk op te merken dat in de definities van $\bar{x}$ en $R_{xx}(\tau)$ we met **één vaste realisering x(t)** werken en over de **tijd** middelen. De grootheden $\bar{x}$ en $R_{xx}(\tau)$ noemen we dan ook *tijdsgemiddelden* behorende bij de realisering $x(t)$ van het stationaire proces $\mathbf{x}(t)$. Dit in tegenstelling tot de eerder besproken ensemble gemiddelden waar de tijd gefixeerd is en gemiddeld wordt over de realiseringen uit het ensemble van het proces.
3) We noemen $\bar{x}$ het *gemiddelde* en $R_{xx}(\tau)$ de *autocorrelatie* van het random signaal $x(t)$.
4) In het algemeen zullen $\bar{x}$ en $R_{xx}(\tau)$ afhangen van de beschouwde realisering $x(t)$ uit het ensemble van het stationaire proces $\mathbf{x}(t)$.

3.10. DEFINITIE. Een stationair proces $\mathbf{x}(t)$ met verwachting $\mu_{\mathbf{X}}$ en stochastische autocorrelatiefunctie $R_{\mathbf{XX}}(\tau)$ heet *ergodisch* wanneer voor iedere realisering $x(t)$ uit het esemble van het stationaire proces $\mathbf{x}(t)$ geldt

i) $\bar{x} = \mu_{\mathbf{X}}$

ii) $R_{xx}(\tau) = R_{\mathbf{XX}}(\tau)$ voor alle $\tau \in \mathbb{R}$.

3.11. OPMERKINGEN. 1) Een ergodisch proces wordt aldus gekenmerkt door de eis dat de **tijdsgemiddelden** $\bar{x}$ en $R_{xx}(\tau)$ gelijk moeten zijn aan de

corresponderende **ensemble gemiddelden** $\mu_{\mathbf{X}}$ en $R_{\mathbf{XX}}(\tau)$. Voor ergodische processen kunnen bijgevolg $\mu_{\mathbf{X}}$ en $R_{\mathbf{XX}}(\tau)$ gemeten worden door aan **één enkele realisering** $x(t)$ de grootheden $\bar{x}$ en $R_{xx}(\tau)$ te meten.
2) Van vele stochastische processen is bekend dat ze stationair en zelfs ergodisch zijn. Van vele stochastische processen, die men in de praktijk ontmoet, kan men op min of meer fysische gronden vermoeden dat ze ergodisch zijn. Als voorbeelden hiervan geven we: de krachten ten gevolge van golfslag op de onderbouw van een off-shore boorinstallatie; de kracht op een schokdemper van een auto die over een oneffen weg rijdt; de temperatuur op een bepaald punt van het zonsoppervlak.
3) In de ingenieurspraktijk komt het nogal eens voor, dat stochastische processen bestudeerd moeten worden, waarover men a priori nagenoeg geen informatie heeft. Men begint dan vaak te werken vanuit de veronderstelling, dat het proces ergodisch is; in feite werkt men dan dus met een aangenomen model van een zeer eenvoudige structuur.Zolang de metingen geen aanleiding geven om het model te verwerpen, is dit een vruchtbare wijze van werken.

3.12. Voor de feitelijke meting van de verwachting $\mu_{\mathbf{X}}$ en de stochastische autocorrelatiefunctie $R_{\mathbf{XX}}(\tau)$ van een ergodisch proces $\mathbf{x}(t)$ kan men nu als volgt te werk gaan: Men bepaalt voor **één realisering** $x(t)$ van het proces $\mathbf{x}(t)$ voor een (grote) waarde van T het **tijdsgemiddelde**

$$\bar{x}_T = \frac{1}{2T}\int_{-T}^{T} x(t)dt.$$

De gevonden waarde $\bar{x}_T$ is in feite een *schatting* van het gemiddelde $\bar{x}$ van het random signaal $x(t)$ omdat de limietovergang in de definitie van $\bar{x}$ in de meetpraktijk niet gemaakt kan worden. Omdat het proces $\mathbf{x}(t)$ ergodisch is, is de gevonden waarde $\bar{x}_T$ dus ook een schatting van de verwachting $\mu_{\mathbf{X}}$ van het ergodische proces $\mathbf{x}(t)$.
Vervolgens bepaalt men voor een zekere waarde van τ en een (grote) waarde van T voor **één realisering** $x(t)$ van het proces $\mathbf{x}(t)$ het **tijdsgemiddelde**

$$R_{xx}^{T}(\tau) = \frac{1}{2T}\int_{-T}^{T} x(t)x(t+\tau)dt.$$

Op grond van dezelfde overwegingen als hierboven is de gevonden waarde $R_{xx}^{T}(\tau)$ een **schatting** van de waarde $R_{\mathbf{XX}}(\tau)$ van de stochastische autocorrelatieffunctie van het ergodische proces $\mathbf{x}(t)$.

3.13. Vaak beschouwt men twee stochastische processen $\mathbf{x}(t)$ en $\mathbf{y}(t)$ tegelijkertijd. Men kan hierbij bijvoorbeeld denken aan een linear tijdinvariant systeem met impulsrespons $h(t)$ dat geëxciteerd wordt door (de realiseringen van) een stochastisch proces $\mathbf{x}(t)$ en het proces $\mathbf{y}(t)$ is de responsie hierop. Het verband tussen de processen $\mathbf{x}(t)$ en $\mathbf{y}(t)$ wordt in dit geval gegeven door

$$\mathbf{y}(t) = (h * \mathbf{x})(t) = \int_{-\infty}^{\infty} h(t-\sigma)\mathbf{x}(\sigma)d\sigma.$$

We spreken in het algemeen van een *paar stochastische processen* $(\mathbf{x}(t),\mathbf{y}(t))$.

3.14. DEFINITIE. Laat $(\mathbf{x}(t), \mathbf{y}(t))$ een paar stochastische processen zijn, dan wordt de *stochastische kruiscorrelatiefunctie* $R_{\mathbf{xy}}(t_1, t_2)$ gedefinieerd door

$$R_{\mathbf{xy}}(t_1, t_2) = E\mathbf{x}(t_1)\mathbf{y}(t_2) \quad \text{voor alle } t_1, t_2 \in \mathbb{R}.$$

3.15. DEFINITIE. Laat $(\mathbf{x}(t), \mathbf{y}(t))$ een paar stochastische processen zijn met stochastische kruiscorrelatiefunctie $R_{\mathbf{xy}}(t_1, t_2)$. Het paar $(\mathbf{x}(t), \mathbf{y}(t))$ heet *stationair* indien
i) $\mathbf{x}(t)$ en $\mathbf{y}(t)$ stationaire processen zijn;
ii) de stochastische kruiscorrelatie $R_{\mathbf{xy}}(t_1, t_2)$ slechts afhangt van het verschil $t_2 - t_1$, dat wil zeggen er bestaat een functie $R_{\mathbf{xy}}(\tau)$ van één variabele τ zo dat

$$R_{\mathbf{xy}}(t_1, t_2) = R_{\mathbf{xy}}(0, t_2 - t_1) = R_{\mathbf{xy}}(t_2 - t_1) \quad \text{voor alle } t_1, t_2 \in \mathbb{R}.$$

3.16. In het geval van een een *stationair paar* processen $(\mathbf{x}(t), \mathbf{y}(t))$ onderscheiden we de volgende *ensemble-gemiddelden*

$u_{\mathbf{x}}$	: verwachting van $\mathbf{x}(t)$
$R_{\mathbf{xx}}(\tau)$	: stochastische autocorrelatiefunctie van $\mathbf{x}(t)$.
$\mu_{\mathbf{y}}$	: verwachting van $\mathbf{y}(t)$.
$R_{\mathbf{yy}}(\tau)$	: stochastische autocorrelatiefunctie van $\mathbf{y}(t)$.
$R_{\mathbf{xy}}(\tau)$	: stochastische korrelcorrelatiefunctie van $(\mathbf{x}(t), \mathbf{y}(t))$.

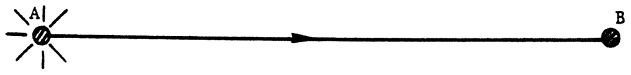
3.17. VOORBEELD. We beschouwen een stochastisch proces $\mathbf{x}(t)$. In het punt A (zie tekening hieronder) zenden we de realiseringen uit het ensemble van het proces $\mathbf{x}(t)$ uit. Kortheidshalve zeggen we dat we het proces $\mathbf{x}(t)$ in A uitzenden. Wanneer we in A een random signaal $x(t)$ uitzenden wordt ten gevolge hiervan in het punt B een signaal $y(t)$ ontvangen. We nemen aan dat er een constante vetragingstijd $\tau_0 > 0$ bij het ontvangen van een uitgezonden signaal optreedt en een constante versterkingsfactor $\alpha > 0$, dat wil zeggen

$$y(t) = \alpha x(t - \tau_0) \quad \text{voor alle } t \in \mathbb{R}.$$

Wanneer we dus in het punt A een proces $\mathbf{x}(t)$ uitzenden dan ontvangen we in B een proces $\mathbf{y}(t)$ en er geldt

$$\mathbf{y}(t) = \alpha \mathbf{x}(t - \tau_0).$$

In deze situatie is er dus sprake van een paar stochastische processen $(\mathbf{x}(t), \mathbf{y}(t))$.



We karakteriseren het uitgezonden proces $\mathbf{x}(t)$ door de verwachtingsfunctie $\mu_{\mathbf{x}}(t)$ en de stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(t_1, t_2)$ als bekend te veronderstellen.
We gaan het paar stochastische processen $(\mathbf{x}(t), \mathbf{y}(t))$ karakteriseren door de ensemble-gemiddelden $\mu_{\mathbf{y}}(t), R_{\mathbf{yy}}(t_1, t_2)$ en $R_{\mathbf{xy}}(t_1, t_2)$ uit te drukken in de bekend veronderstelde ensemble-gemiddelden $\mu_{\mathbf{x}}(t)$ en $R_{\mathbf{xx}}(t_1, t_2)$. Er geldt
(1) $\mu_{\mathbf{y}}(t) = E\mathbf{y}(t) = E(\alpha\mathbf{x}(t-\tau_0)) = \alpha\mu_{\mathbf{x}}(t-\tau_0)$.
(2) $R_{\mathbf{yy}}(t_1, t_2) = E\mathbf{y}(t_1)\mathbf{y}(t_2) = E\alpha^2\mathbf{x}(t_1-\tau_0)\mathbf{x}(t_2-\tau_0) = \alpha^2 R_{\mathbf{xx}}(t_1-\tau_0, t_2-\tau_0)$.
(3) $R_{\mathbf{xy}}(t_1, t_2) = E\mathbf{x}(t_1)\mathbf{y}(t_2) = E\alpha\mathbf{x}(t_1)\mathbf{x}(t_2-\tau_0) = \alpha R_{\mathbf{xx}}(t_1, t_2-\tau_0)$.
Laten we nu veronderstellen dat het uitgezonden proces $\mathbf{x}(t)$ stationair is. We gaan nu aantonen dat het paar $(\mathbf{x}(t), \mathbf{y}(t))$ stationair is. Omdat $\mathbf{x}(t)$ stationair is wordt het dus gekaraktiseerd door zijn verwachting $\mu_{\mathbf{x}}$ en stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(\tau)$ bekend te veronderstellen. Uit (1) en (2) volgt

$$\mu_{\mathbf{y}}(t) = \alpha\mu_{\mathbf{x}}(t-\tau_0) = \alpha\mu_{\mathbf{x}}.$$

$$R_{\mathbf{yy}}(t_1, t_2) = \alpha^2 R_{\mathbf{xx}}(t_1, -\tau_0, t_2-\tau_0) = \alpha^2 R_{\mathbf{xx}}(t_2-t_1).$$

We concluderen dat het proces $\mathbf{y}(t)$ stationair is en er geldt
(4) $\mu_{\mathbf{y}} = \alpha\mu_{\mathbf{x}}$ en $R_{\mathbf{xx}}(\tau) = \alpha^2 R_{\mathbf{xx}}(\tau)$.
Verder volgt uit (3)

$$R_{\mathbf{xy}}(t_1, t_2) = \alpha R_{\mathbf{xx}}(t_1, t_2-\tau_0) = \alpha R_{\mathbf{xx}}(t_2-t_1-\tau_0).$$

Omdat de stochastische kruiscorrelatiefunctie $R_{\mathbf{xy}}(t_1, t_2)$ slechts van $t_2 - t_1$ afhangt, is het paar $(\mathbf{x}(t), \mathbf{y}(t))$ stationair. Er geldt
(5) $R_{\mathbf{xy}}(\tau) = \alpha R_{\mathbf{xx}}(\tau-\tau_0)$.

3.18. DEFINITIE. Laat $(\mathbf{x}(t), \mathbf{y}(t))$ een **stationair paar** stochastische processen zijn en $(x(t), y(t))$ een paar realiseringen van $(\mathbf{x}(t), \mathbf{y}(t))$. Dan definiëren we

$$R_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)y(t+\tau)dt \quad (t \in \mathbb{R}).$$

OPMERKINGEN. 1) De functie $R_{xy}(\tau)$ is de **kruiscorrelatie** van random signalen $x(t)$ en $y(t)$.
2) De limiet in de definitie van $R_{xy}(\tau)$ bestaat voor alle in de praktijk voorkomende paren realiseringen $(x(t), y(t))$ van een **stationair paar** $(\mathbf{x}(t), \mathbf{y}(t))$ (zie opmerking 1) in 3.9).

3) In het algemeen zal de kruiscorrelatie $R_{xy}(\tau)$ afhangen van het paar realiseringen $(x(t), y(t))$ van het stationaire paar $(\mathbf{x}(t), \mathbf{y}(t))$.

3.19. DEFINITIE. Laat $(\mathbf{x}(t), \mathbf{y}(t))$ een **stationair paar** stochastische processen zijn. Dan heet $(\mathbf{x}(t), \mathbf{y}(t))$ een *ergodisch paar* wanneer geldt:
i) $\mathbf{x}(t)$ en $\mathbf{y}(t)$ zijn ergodisch;
ii) voor alle paren realiseringen $(x(t), y(t))$ van het paar $(\mathbf{x}(t), \mathbf{y}(t))$ geldt

$$R_{xy}(\tau) = R_{\mathbf{xy}}(\tau) \quad \text{voor alle } \tau \in \mathbb{R}.$$

3.20. OPMERKING. Ook hier is de essentie weer: **tijdsgemiddelden** zijn gelijk aan de corresponderende **ensemble-gemiddelden**.

3.21. Laat $\mathbf{x}(t)$ een stationair proces zijn. Dan hangen $\mu_{\mathbf{x}} = E\mathbf{x}(t)$ en $R_{\mathbf{xx}}(\tau) = E\mathbf{x}(t)\mathbf{x}(t+\tau)$ niet van de tijd t af. Dit maakt het aannemelijk, dat voor realiseringen $x(t)$ van $\mathbf{x}(t)$ zal gelden dat ermee samenhangende tijdsgemiddelden niet echt afhangen van de ligging van het tijdsinterval waarover men middelt, althans voor grote intervallengten. Iets dergelijks kan men verwachten bij realiseringen $(x(t), y(t))$ van een stationair paar stochastische processen $(\mathbf{x}(t), \mathbf{y}(t))$. We formulieren hierover twee stellingen, die we niet van een afleiding zullen voorzien.

3.22. STELLING. Zij $\mathbf{x}(t)$ een stationair proces en laat $I(T)$ voor iedere $T > 0$ een interval ter lengte $2T$ zijn. Dan geldt voor alle realiseringen $x(t)$ van $\mathbf{x}(t)$:

i) $\bar{x} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt = \lim_{T\to\infty} \frac{1}{2T} \int_{I(T)} x(t)dt.$

ii) $R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau)dt = \lim_{T\to\infty} \frac{1}{2T} \int_{I(T)} x(t)x(t+\tau)dt.$

3.23. STELLING. Zij $(\mathbf{x}(t), \mathbf{y}(t))$ een stationair paar stochastische processen en laat $I(T)$ voor iedere $T > 0$ een interval ter lengte $2T$ zijn. Dan geldt voor alle realiseringen $(x(t), y(t))$ van $(\mathbf{x}(t), \mathbf{y}(t))$:

$$R_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)y(t+\tau)dt = \lim_{T\to\infty} \frac{1}{2T} \int_{I(T)} x(t)y(t+\tau)dt.$$

3.24 VOORBEELD. We beschouwen het paar stochastische processen $(\mathbf{x}(t), \mathbf{y}(t))$ uit 3.17 waarbij

$$\mathbf{y}(t) = \alpha\mathbf{x}(t - \tau_0).$$

We veronderstellen dat het uitgezonden proces $\mathbf{x}(t)$ ergodisch is met verwachting $\mu_{\mathbf{x}}$ en stochastische autocorrellatiefunctie $R_{\mathbf{xx}}(\tau)$. We gaan aantonen dat $(\mathbf{x}(t), \mathbf{y}(t))$ een ergodisch paar vormt. In 3.17 is reeds aangetoond dat $(\mathbf{x}(t), \mathbf{y}(t))$ een stationair paar is. Er geldt:

$\mu_{\mathbf{y}} = \alpha\mu_{\mathbf{x}}$ (zie formule (4) in 3.17),
$R_{\mathbf{yy}}(\tau) = \alpha^2 R_{\mathbf{xx}}(\tau)$ (zie formule (4) in 3.17),
$R_{\mathbf{xy}}(\tau) = \alpha R_{\mathbf{xx}}(\tau - \tau_0)$ (zie formule (5) in 3.17).

Laat $x(t)$ een realisering (random signaal) van het uitgezonden proces $\mathbf{x}(t)$ zijn. Voor het corresponderende ontvangen signaal $y(t)$ geldt

$$y(t) = \alpha x(t - \tau_0),$$

en $(x(t), y(t))$ is een realisering van $(\mathbf{x}(t), \mathbf{y}(t))$. Er geldt

$$\bar{y} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} y(t)dt = \alpha \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t - \tau_0)dt =$$

$$= \alpha \lim_{T\to\infty} \frac{1}{2T} \int_{-T-\tau_0}^{T-\tau_0} x(w)dw = \alpha\bar{x} \text{ (zie 3.22)} = \alpha\mu_{\mathbf{x}} = \mu_{\mathbf{y}}$$

Dus $\bar{y} = \mu_{\mathbf{y}}$. Verder geldt

$$R_{yy}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} y(t)y(t + \tau)dt =$$

$$= \alpha^2 \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t - \tau_0)x(t + \tau - \tau_0)dt =$$

$$= \alpha^2 \lim_{T\to\infty} \frac{1}{2T} \int_{-T-\tau_0}^{T-\tau_0} x(w)x(w + \tau)dw = \alpha^2 R_{xx}(\tau) \text{ (zie 3.22)} =$$

$$= \alpha^2 R_{\mathbf{xx}}(\tau) = R_{\mathbf{yy}}(\tau).$$

Dus $R_{yy}(\tau) = R_{\mathbf{yy}}(\tau)$ en hiermee is aangetoond dat $\mathbf{y}(t)$ een ergodisch proces is.

Tenslotte:

$$R_{xy}(\tau) = \lim_{T\to} \frac{1}{2T} \int_{-T}^{T} x(t)y(t + \tau)dt =$$

$$= \alpha \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t + \tau - \tau_0)dt =$$

$$= \alpha R_{xx}(\tau - \tau_0) = \alpha R_{\mathbf{xx}}(\tau - \tau_0) = R_{\mathbf{xy}}(\tau).$$

Dus $R_{xy}(\tau) = R_{\mathbf{xy}}(\tau)$ en we kunnen nu concluderen dat $(\mathbf{x}(t), \mathbf{y}(t))$ een ergodisch paar stochastische processen is.

3.25. Wanneer men berekeningen moet uitvoeren met betrekking tot een zeker stochastisch proces $\mathbf{x}(t)$ zal men vaak voldoende informatie hebben

als men de beschikking heeft over de **verwachtingsfunctie** $\mu_{\mathbf{X}}(t)$ en de **stochastische autocorrelatiefunctie** $R_{\mathbf{XX}}(t_1, t_2)$. Deze ensemblegemiddelden hebben we in 3.1 en voorbeeld 3.7 kunnen berekenen omdat we ons een goed beeld konden vormen van het ensemble van het beschouwde proces $\mathbf{x}(t)$ en van de kansen dat zekere realiseringen ook inderdaad optreden. In feit konden we toen het ensemble **parametriseren** (door middel van één of meer stochastische parameters met bekende kansverdeling). In de praktijk is het echter vaak zo dat men geen informatie heeft over de structuur van het ensemble en dus de functies $\mu_{\mathbf{X}}(t)$ en $R_{\mathbf{XX}}(t_1, t_2)$ niet kan berekenen zoals in 3.1 en 3.7. In sommige situaties beschikt men echter wel over de **simultane kansverdeling** van de stochasten $\mathbf{x}(t_1)$ en $\mathbf{x}(t_2)$ voor iedere tweetal tijdstippen t_1 en t_2. In deze situatie kan men de ensemble-gemiddelden $\mu_{\mathbf{X}}(t)$ en $R_{\mathbf{XX}}(t_1, t_2)$ berekenen zoals geïllustreerd in het volgende voorbeeld.

3.26. VOORBEELD. We beschouwen een stochastisch proces $\mathbf{x}(t)$ en veronderstellen dat voor ieder tijdstip t de stochast $\mathbf{x}(t)$ exponentieel verdeeld is met parameter $\lambda = 2$ en dat voor ieder tweetal tijdstippen t_1 en t_2 met $t_1 \neq t_2$ de stochasten $\mathbf{x}(t_1)$ en $\mathbf{x}(t_2)$ ongecorreleerd zijn. Nu geldt voor ieder tijdstip $t \in \mathbb{R}$:

$$\mu_{\mathbf{X}}(t) = E\mathbf{x}(t) = \int_{-\infty}^{\infty} x f_{\mathbf{X}(t)}(x)dx = 2\int_{0}^{\infty} xe^{-2x}dx = 0.5.$$

Voor ieder tweetal tijdstippen t_1 en t_2 met $t_1 \neq t_2$ geldt

$$R_{\mathbf{XX}}(t_1, t_2) = E\mathbf{x}(t_1)\mathbf{x}(t_2) = E\mathbf{x}(t_1)E\mathbf{x}(t_2) = 0.25,$$

omdat voor $t_1 \neq t_2$ de stochasten $\mathbf{x}(t_1)$ en $\mathbf{x}(t_2)$ ongecorreleerd zijn. Indien $t_1 = t_2$ dan geldt

$$R_{\mathbf{XX}}(t_1, t_2) = E\mathbf{x}^2(t_1) = \int_{-\infty}^{\infty} x^2 f_{\mathbf{X}(t_1)}(x)dx = 2\int_{0}^{\infty} x^2 e^{-2x}dx = 0.5$$

We hebben gevonden

$$\mu_{\mathbf{X}}(t) = 0.5 \quad \text{en} \quad R_{\mathbf{XX}}(t_1, t_2) = \begin{cases} 0.25 & \text{als } t_1 \neq t_2 \\ 0.5 & \text{als } t_1 = t_2\,. \end{cases}$$

We zien dus dat dit proces $\mathbf{x}(t)$ stationair is met

$$\mu_{\mathbf{X}} = 0.5 \quad \text{en} \quad R_{\mathbf{XX}}(\tau) = \begin{cases} 0.25 & \text{als } \tau \neq 0 \\ 0.5 & \text{als } \tau = 0\,. \end{cases}$$

Omdat we te weinig informatie over het ensemble van $\mathbf{x}(t)$ hebben kunnen we de vraag of dit proces $\mathbf{x}(t)$ ergodisch is niet door berekening beantwoorden. Metingen zullen moeten uitwijzen of de ergodiciteitshypothese een acceptabele aaname is.

3.27. Tot nu toe hebben we een beschrijving gegeven van stationaire stochastische processen in **het tijdsdomein** met behulp van twee belangrijke ensemble-gemiddelden:
i) de verwachting $\mu_{\mathbf{x}} = E\mathbf{x}(t)$,
ii) de stochastische autocorrelatie $R_{\mathbf{xx}}(\tau) = E\mathbf{x}(t)\mathbf{x}(t+\tau)$.
Men kan stationaire processen ook bestuderen in het *frequentiedomein* en hierbij speelt het *vermogenspectrum* een belangrijke rol.

3.28. DEFINITIE. Laat $\mathbf{x}(t)$ een **stationair** proces zijn met stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(\tau)$. Het *vermogenspectrum* $G_{\mathbf{xx}}(f)$ van het stationaire proces $\mathbf{x}(t)$ wordt gedefinieerd door

$$G_{\mathbf{xx}}(f) = F(R_{\mathbf{xx}}(\tau)) \quad (f \in \mathbb{R}).$$

3.29. STELLING. Laat $\mathbf{x}(t)$ een **ergodisch** proces zijn. Dan geldt voor iedere realisering $x(t)$ van het proces $\mathbf{x}(t)$ dat het vermogensspectrum $G_{\mathbf{xx}}(f)$ van het proces $\mathbf{x}(t)$ gelijk is aan de spectrale vermogensdichtheid $G_{\mathbf{xx}}(f)$ van het random signaal $x(t)$.

AFLEIDING. Laat $x(t)$ een realisering zijn van het ergodische proces $\mathbf{x}(t)$, dan geldt

$$G_{\mathbf{xx}}(f) = F(R_{\mathbf{xx}}(\tau)) = F(R_{xx}(\tau)) = G_{xx}(f).$$

3.30. DEFINITIE. Laat $(\mathbf{x}(t), \mathbf{y}(t))$ een **station paar** stochastische processen zijn met stochastische kruiscorrelatiefunctie $R_{\mathbf{xy}}(\tau)$. Het *kruis-vermogensspectrum* $G_{\mathbf{xy}}(f)$ van het stationaire paar $(\mathbf{x}(t), \mathbf{y}(t))$ wordt gedefinieerd door

$$G_{\mathbf{xy}}(f) = F(R_{\mathbf{xy}}(\tau) \quad (f \in \mathbb{R}).$$

3.31. STELLING. Laat $(\mathbf{x}(t), \mathbf{y}(t))$ een **ergodisch paar** stochastische processen zijn. Dan geldt voor iedere realisering $(x(t), y(t))$ van het paar processen $(\mathbf{x}(t), \mathbf{y}(t))$ dan het kruis-vermogensspectrum $G_{\mathbf{xy}}(f)$ van het paar processen $(\mathbf{x}(t), \mathbf{y}(t))$ gelijk is aan de spectrale-kruisvermogensdichtheid $G_{xy}(f)$ van het paar random signalen $(x(t), y(t))$.

AFLEIDING. Laat $(x(t), y(t))$ een realisering zijn van het ergodische paar stochastische processen $(\mathbf{x}(t), \mathbf{y}(t))$, dan geldt

$$G_{\mathbf{xy}}(f) = F(R_{\mathbf{xy}}(\tau)) = F(R_{xy}(\tau)) = G_{xy}(f).$$

3.32. OPMERKING. In de stellingen 3.29 en 3.31 komt weer naar voren dat de vermogensspectra van **ergodische processen** bepaald kunnen worden door meting **aan één realisering**.

3.33. STELLING. Laat $\mathbf{x}(t)$ een stationair proces zijn met stochastisch autocorrelatiefunctie $R_{\mathbf{xx}}(\tau)$ en vermogensspectrum $G_{\mathbf{xx}}(f)$. Dan geldt

i) De functies $R_{\mathbf{xx}}(\tau)$ en $G_{\mathbf{xx}}(f)$ zijn even.

ii) Voor alle $f \in \mathbb{R}$ geldt $G_{\mathbf{xx}}(f) \in \mathbb{R}$ en $G_{\mathbf{xx}}(f) \geq 0$.

iii) $F(G_{\mathbf{xx}}(f)) = R_{\mathbf{xx}}(\tau)$.

AFLEIDING.
i) $R_{\mathbf{xx}}(\tau) = E\mathbf{x}(t)\mathbf{x}(t+\tau) = E\mathbf{x}(t-\tau)\mathbf{x}(t) = R_{\mathbf{xx}}(-\tau)$.

Omdat $G_{\mathbf{xx}}(f) = F(R_{\mathbf{xx}}(\tau))$ en $R_{\mathbf{xx}}(\tau)$ even is, geldt dat $G_{\mathbf{xx}}(f)$ even is.

ii) $$G^*_{\mathbf{xx}}(f) = \left(\text{“}\lim_{T\to\infty}\text{”} \int_{-T}^{T} R_{\mathbf{xx}}(\tau) e^{-2\pi j f \tau} d\tau \right)^* =$$

$$= \text{“}\lim_{T\to\infty}\text{”} \int_{-T}^{T} R_{\mathbf{xx}}(\tau) e^{2\pi j f \tau} d\tau = G_{\mathbf{xx}}(-f) = G_{\mathbf{xx}}(f).$$

Omdat $G^*_{\mathbf{xx}}(f) = G_{\mathbf{xx}}(f)$ geldt $G_{\mathbf{xx}}(f) \in \mathbb{R}$
We geven geen afleiding van het feit dat $G_{\mathbf{xx}}(f) \geq 0$.
iii) Er geldt

$$R_{\mathbf{xx}}(\tau) = \text{“}\lim_{T\to\infty}\text{”} \int_{-T}^{T} G_{\mathbf{xx}}(f) e^{2\pi j f \tau} df =$$

$$= \left(\text{“}\lim_{T\to\infty}\text{”} \int_{-T}^{T} G_{\mathbf{xx}}(f) e^{2\pi j f \tau} df \right)^* =$$

$$= \text{“}\lim_{T\to\infty}\text{”} \int_{-T}^{T} G_{\mathbf{xx}}(f) e^{-2\pi j f \tau} df.$$

Dus $R_{\mathbf{xx}}(\tau) = F(G_{\mathbf{xx}}(f))$.

3.34. Laat $x(t)$ een random signaal zijn uit het ensemble van een stationair proces $\mathbf{x}(t)$. Fixeer een tijdstip t_0 en beschouw voor 'kleine' Δt de energieinhoud $E_x[t_0, t_0 + \Delta t]$ van het random signaal $x(t)$ over het tijdsinterval $[t_0, t_0 + \Delta t]$. Dan geldt

$$E_x[t_0, t_0 + \Delta t] = \int_{t_0}^{t_0+\Delta t} x^2(t)dt \approx x^2(t_0)\Delta t.$$

Naaraanleiding hiervan definiëren we het *vermogen* $P_x(t)$ van het signaal $x(t)$ op tijdstip t door

$$P_x(t) = x^2(t).$$

DEFINITIE. Laat $\mathbf{x}(t)$ een **stationair** proces zijn, dan definiëren we het *vermogen* $P_{\mathbf{x}}$ van het stationaire proces $\mathbf{x}(t)$ door

$$P_{\mathbf{x}}(t) = E\mathbf{x}^2(t).$$

OPMERKING. Omdat voor een stationair proces $\mathbf{x}(t)$ met stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(\tau)$ geldt $E\mathbf{x}^2(t) = R_{\mathbf{xx}}(0)$, hangt $E\mathbf{x}^2(t)$ niet van de tijd t af.

3.35. STELLING. Laat $\mathbf{x}(t)$ een stationair proces zijn met vermogen $P_{\mathbf{x}}$ en vermogensspectrum $G_{\mathbf{xx}}(f)$, dan geldt

$$P_{\mathbf{x}} = \int_{-\infty}^{\infty} G_{\mathbf{xx}}(f)df = R_{\mathbf{xx}}(0).$$

AFLEIDING. Omdat $F(G_{\mathbf{xx}}(f)) = R_{\mathbf{xx}}(\tau)$ (zie 3.33 iii) geldt

$$R_{\mathbf{xx}}(0) = \int_{-\infty}^{\infty} G_{\mathbf{xx}}(f)df.$$

Dus

$$P_{\mathbf{x}} = E\mathbf{x}^2(t) = R_{\mathbf{xx}}(0) = \int_{-\infty}^{\infty} G_{\mathbf{xx}}(f)df.$$

OPMERKING. Omdat $G_{\mathbf{xx}}(f) \geq 0$ (zie 3.33 ii) geldt dat, wanneer $\int_{-\infty}^{\infty} G_{\mathbf{xx}}(f)df$ niet convergent is, we kunnen concluderen $P_{\mathbf{x}} = +\infty$. Een stationair proces $\mathbf{x}(t)$ met een oneindig groot vermogen kan in werkelijkheid natuurlijk niet bestaan.

3.36. VOORBEELD. We beschouwen een reëel lineair tijdinvariant systeem S met impulsrespons $h(t)$ en systeemfunctie $H(f)$. We veronderstellen dat de impulsrespons $h(t)$ absoluut integreerbaar is, dat wil zeggen

$$\int_{-\infty}^{\infty} |h(t)|dt \quad \text{is convergent .}$$

Een dergelijke lineair tijdinvariant systeem is *stabiel.* Op dit begrip zullen we echter niet verder ingaan. We veronderstellen dat het systeem S geëxciteerd wordt door (een realisering van) een **stationair** proces $\mathbf{x}(t)$ met verwachting $\mu_{\mathbf{x}}$, stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(\tau)$ en vermogensspectrum $G_{\mathbf{xx}}(f)$. Nu geldt voor het stochastisch proces $\mathbf{y}(t)$ aan de uitgang van het systeem S

$$\mathbf{y}(t) = \int_{-\infty}^{\infty} \mathbf{x}(\tau)h(t-\tau)d\tau.$$

Het paar stochastische processen $(\mathbf{x}(t), \mathbf{y}(t))$ is stationair. We geven hiervan geen afleiding. Verder geldt

i) $\mu_{\mathbf{y}} = \mu_{\mathbf{x}} \int_{-\infty}^{\infty} h(t)dt = \mu_{\mathbf{x}} H(0).$

ii) $R_{\mathbf{xy}}(\tau) = (R_{\mathbf{xx}} * h)(\tau)$.

iii) $R_{\mathbf{yy}}(\tau) = (R_{\mathbf{xy}} * \tilde{h})(\tau)$, waarbij $\tilde{h}(t) = h(-t)$.

iv) $G_{\mathbf{xy}}(f) = G_{\mathbf{xx}}(f)H(f)$.

v) $G_{\mathbf{yy}}(f) = G_{\mathbf{xx}}(f)|H(f)|^2$.

AFLEIDING.

$$\text{i)} \quad \mu_{\mathbf{y}} = E\mathbf{y}(t) = E\int_{-\infty}^{\infty} \mathbf{x}(\tau)h(t-\tau)d\tau = \int_{-\infty}^{\infty} E\mathbf{x}(\tau)h(t-\tau)d\tau =$$

$$= \mu_{\mathbf{x}} \int_{-\infty}^{\infty} h(t-\tau)d\tau = \mu_{\mathbf{x}} \int_{-\infty}^{\infty} h(w)dw = \mu_{\mathbf{x}} H(0).$$

$$\text{ii)} \quad R_{\mathbf{xy}}(\tau) = E\mathbf{x}(t)\mathbf{y}(t+\tau) = E\int_{-\infty}^{\infty} \mathbf{x}(t)\mathbf{x}(u)h(t+\tau-u)du =$$

$$= \int_{-\infty}^{\infty} E\mathbf{x}(t)\mathbf{x}(u)h(t+\tau-u)du =$$

$$= \int_{-\infty}^{\infty} R_{\mathbf{xx}}(u-t)h(t+\tau-u)du =$$

$$= \int_{-\infty}^{\infty} R_{\mathbf{xx}}(w)h(\tau-w)dw = (R_{\mathbf{xx}} * h)(\tau).$$

$$\text{iii)} \quad R_{\mathbf{yy}}(\tau) = E\mathbf{y}(t)\mathbf{y}(t+\tau) = E\int_{-\infty}^{\infty} \mathbf{y}(t+\tau)\mathbf{x}(u)h(t-u)du =$$

$$= \int_{-\infty}^{\infty} E\mathbf{x}(u)\mathbf{y}(t+\tau)h(t-u)du = \int_{-\infty}^{\infty} R_{\mathbf{xy}}(t+\tau-u)h(t-u)du =$$

$$= \int_{-\infty}^{\infty} R_{\mathbf{xy}}(\tau-w)h(-w)dw = \int_{-\infty}^{\infty} R_{\mathbf{xy}}(\tau-w)\tilde{h}(w)dw = (R_{\mathbf{xy}} * \tilde{h})(\tau).$$

iv) volgt uit ii) door Fourier transformatie.
v) Uit iii) volgt

$$G_{\mathbf{yy}}(f) = G_{\mathbf{xy}}(f)H(-f) = G_{\mathbf{xx}}(f)H(f)H(-f).$$

Omdat het systeem S reëel is, geldt $H(-f) = H*(f)$ (zie 1.6.17). Dus

$$G_{\mathbf{yy}}(f) = G_{\mathbf{xx}}(f)|H(f)|^2.$$

Relatie v) geeft de mogelijkheid om de **interpretatie** van het vermogensspectrum $G_{\mathbf{xx}}(f)$ van een startionair proces $\mathbf{x}(t)$ wat nader uit te werken. Fixeer voor een gegeven frequentie f_0 de frequentieband $[f_0, f_0 + \Delta f]$ met 'kleine' Δf en veronderstel dat we beschikken over een systeem F (*filter*) met systeemfunctie

$$H(f) = \begin{cases} 1 & \text{als } f_0 \leq f \leq f_0 + \Delta f \\ 0 & \text{anders.} \end{cases}$$

Het filter F laat dus de frequentiecomponenten van het ingangssignaal $x(t)$, waarvan de frequentie binnen de band $[f_0, f_0 + \Delta f]$ ligt, door en de andere frequentiecomponenten worden niet doorgelaten, dat wil zeggen het uitgangssignaal $y(t)$ bestaat precies uit de frequentiecomponenten van het ingangssignaal $x(t)$, die binnen de band $[f_0, f_0 + \Delta f]$ liggen. Wanneer we dus op de ingang (random signalen van) een stationair proces $\mathbf{x}(t)$ zetten, zal op de uitgang het stationaire proces $\mathbf{y}(t)$ staan, dat we kunnen opvatten als het met betrekking tot de frequentieband $[f_0, f_0 + \Delta f]$ gefilterde proces $\mathbf{x}(t)$. Nu geldt voor het vermogen $P_{\mathbf{y}}$ van het stationaire proces $\mathbf{y}(t)$ op grond van relatie v) en 3.35

$$P_{\mathbf{y}} = \int_{-\infty}^{\infty} G_{\mathbf{yy}}(f)df = \int_{-\infty}^{\infty} G_{\mathbf{xx}}(f)|H^2(f)|df =$$

$$= \int_{-0}^{f_0+\Delta f} G_{\mathbf{xx}}(f)df \approx G_{\mathbf{xx}}(f_0)\Delta f\,.$$

Dus

$$P_{\mathbf{y}} \approx G_{\mathbf{xx}}(f_0)\Delta f.$$

Wanneer we het vermogen van de frequentiecomponenten van een stationair proces $\mathbf{x}(t)$ binnen de frequentieband $[f, f + \Delta f]$ aanduiden met $P_{\mathbf{x}}[f, f + \Delta f]$, dan geldt voor 'kleine' Δf

$$P_{\mathbf{x}}[f, f + \Delta f] \approx G_{\mathbf{xx}}(f)\Delta f\,.$$

(Vergelijk deze relatie met een soortgelijke relatie voor kansdichtheden in 1.13 en 2.11).

3.37. *Random signalen* uit het ensemble van **ergodische** processen spelen bij het meten van systeemkarakteristieken van lineair tijdinvariante systemen een belangrijke rol. We lichten dit toe in twee situaties.

3.38. STELLING. Laat $x(t)$ een random signaal uit het ensemble van een ergodisch proces $\mathbf{x}(t)$ zijn, dat als ingangssignaal fungeert van een lineaire tijdinvariant systeem. Het corresponderende uitgangssignaal is $y(t)$. De amplitudeversterking $|H(f)|$ kan bepaald worden door meting van de spectrale vermogensdichtheden $G_{xx}(f)$ en $G_{yy}(f)$ van in- en uitgang. Er geldt

$$|H(f)| = \sqrt{\frac{G_{yy}(f)}{G_{xx}(f)}}, \quad \text{indien } G_{xx}(f) \neq 0.$$

AFLEIDING. Op grond van relatie v) uit 3.36 geldt

$$|H(f)| = \sqrt{\frac{G_{\mathbf{yy}}(f)}{G_{\mathbf{xx}}(f)}} = \sqrt{\frac{F(R_{\mathbf{yy}}(\tau))}{F(R_{\mathbf{xx}}(\tau))}} = \sqrt{\frac{F(R_{yy}(\tau))}{F(R_{xx}(\tau))}},$$

als $G_{\mathbf{xx}}(f) \neq 0$. Er volgt nu

$$|H(f)| = \sqrt{\frac{G_{yy}(f)}{G_{xx}(f)}}.$$

3.39. STELLING. Laat $x(t)$ een random signaal uit het ensemble van een ergodisch proces $\mathbf{x}(t)$ zijn, dat als ingangssignaal fungeert van een lineair tijdinvariant systeem. Het corresponderende uitgangssignaal is $y(t)$. De systeemfunctie $H(f)$ kan bepaald worden door meting van de spectrale vermogensdichtheid $G_{xx}(f)$ van de ingang en de spectrale-kruisvermogensdichtheid $G_{xy}(f)$ van de in- en uitgang. Er geldt

$$H(f) = \frac{G_{xy}(f)}{G_{xx}(f)} \quad \text{indien } G_{xx}(f) \neq 0.$$

AFLEIDING. Op grond van relatie iv) uit 3.36 geldt

$$H(f) = \frac{G_{\mathbf{xy}}(f)}{G_{\mathbf{xx}}(f)} = \frac{F(R_{\mathbf{xy}}(\tau))}{F(R_{\mathbf{xx}}(\tau))} = \frac{F(R_{xy}(\tau))}{F(R_{xx}(\tau))} = \frac{G_{xy}(f)}{G_{xx}(f)},$$

indien $G_{xx}(f) \neq 0$.

3.40. OPMERKING. Bij de afleiding van de stellingen 3.38 en 3.39 hebben we gebruik gemaakt van het feit dat wanneer aan de ingang van een lineair tijdinvariant systeem een ergodisch proces $\mathbf{x}(t)$ staat, het uitgangsproces $\mathbf{y}(t)$ ook ergodisch is en het paar processen $(\mathbf{x}(t), \mathbf{y}(t))$ ook ergodisch is. We geven hiervan geen afleiding.

§4. RUIS

4.1. Wij kennen allen intuïtief het begrip 'ruis': dat wat men uit een luidspreker hoort komen wanneer de versterker vol aanstaat zonder dat aan de versterker een ingangssignaal wordt aangeboden. Ruis treedt vaak op als een (ongewenste) storing van een te onderzoeken signaal.
Neemt men diverse ruissignalen en past men daarop een frequentie-analyse toe, dan blijkt er ruis van verschillende soorten te zijn: het magnitudespectrum kan smal of breed zijn, bij lage of hoge frequenties gecentreerd, of nagenoeg constant tot op hoge frequenties.
Ruis heeft altijd iets toevalligs in zich: het is (realisering van) een stochastisch proces. Daar verschuiving in de tijd het karakter van ruis niet verandert, is dit proces ook stationair. Meestal (afhankelijk van de aard van de bron van de ruis) is ruis zelfs ergodisch.

4.2. Een voor de theorie belangrijke vorm van ruis is de zogenaamde 'witte ruis'. Wij zullen witte ruis niet definiëren, maar geven in plaats daarvan een aantal eigenschappen.
We noemen een **ergodisch** proces $\mathbf{n}(t)$ *witte ruis* indien
i) $\mu_{\mathbf{n}} = 0$.
ii) Er een positief reël getal a bestaat zo dat $G_{\mathbf{nn}}(f) = a$.

OPMERKINGEN.
1) De naam witte ruis vindt zijn oorsprong in eigenschap ii); immers het vermogensspectrum $G_{\mathbf{nn}}(f)$ heeft voor alle frequenties (kleuren) dezelfde waarde.
2)Voor het vermogen $P_{\mathbf{n}}$ van witte ruis geldt

$$P_{\mathbf{n}} = \int_{-\infty}^{\infty} G_{\mathbf{nn}}(f)df = +\infty \quad \text{(zie 3.35)} .$$

Hieruit volgt dat witte ruis een wiskundige idealisatie is van de praktische situatie, waarbij het vermogensspectrum over een **grote relevante frequentieband** een constante waarde heeft en daarbuiten naar nul gaat (zie ook de opmerking in 3.35).
3) Het positivieve getal a noemen we de *intensiteit* van de witte ruis.

4.3. STELLING. Laat $\mathbf{n}(t)$ een witte ruis proces zijn met intensiteit $a > 0$. Dan geldt voor de stochastische autocorrelatiefunctie

$$R_{\mathbf{nn}}(\tau) = a\delta(\tau).$$

AFLEIDING. Uit 3.33 iii) volgt

$$R_{\mathbf{nn}}(\tau) = F(G_{\mathbf{nn}}(f)) = a\delta(\tau).$$

4.4. STELLING. Laat $\mathbf{n}(t)$ een witte ruis proces zijn, dan geldt voor $t_1 \neq t_2$ dat de stochasten $\mathbf{n}(t_1)$ en $\mathbf{n}(t_2)$ **ongecorreleerd** zijn.

AFLEIDING. Voor $t_1 \neq t_2$ geldt

$$E\mathbf{n}(t_1)\mathbf{n}(t_2) = R_{\mathbf{nn}}(t_2 - t_1) = a\delta(t_2 - t_1) = 0.$$

Verder geldt $E\mathbf{n}(t_1)E\mathbf{n}(t_2) = \mu_{\mathbf{n}}^2 = 0$. Dus

$$E\mathbf{n}(t_1)\mathbf{n}(t_2) = E\mathbf{n}(t_1)E\mathbf{n}(t_2).$$

OPMERKING. Vaak veronderstelt men dat bij witte ruis de stochasten $\mathbf{n}(t_1)$ en $\mathbf{n}(t_2)(t_1 \neq t_2)$ behalve ongecorreleerd ook **onafhankelijk** zijn.

4.5. VOORBEELD. Zij $\mathbf{x}(t)$ een stationair proces met verwachting $\mu_{\mathbf{x}}$ en stochastische autocorrelatiefunctie $R_{\mathbf{xx}}(\tau)$. Wanneer we dit stationaire proces $\mathbf{x}(t)$ willen waarnemen zal het vaak zo zijn dat het proces $\mathbf{n}(t)$ op een of andere manier gestoord wordt door bijvoorbeeld witte ruis. Het stationaire proces $\mathbf{x}(t)$ dat in feite wordt waargenomen is het door witte ruis $\mathbf{n}(t)$ gestoorde proces $\mathbf{x}(t)$ en er geldt

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{n}(t).$$

We zullen in deze situatie steeds aannemen dat voor alle tijdstippen t_1 en t_2 de stochasten $\mathbf{x}(t_1)$ en $\mathbf{n}(t_2)$ ongecorreleerd zijn.
We gaan nu aantonen dat $(\mathbf{x}(t), \mathbf{y}(t))$ een paar stationaire processen vormt met
i) $\mu_{\mathbf{y}} = \mu_{\mathbf{x}}$.

ii) $R_{\mathbf{xy}}(\tau) = R_{\mathbf{xx}}(\tau)$.

iii) $R_{\mathbf{yy}}(\tau) = R_{\mathbf{xx}}(\tau) + a\delta(\tau)$, waarin a de intensiteit van het witte ruis proces $\mathbf{n}(t)$ is.

Voor ieder tijdstip t geldt

$$\mu_{\mathbf{y}}(t) = E\mathbf{y}(t) = E[\mathbf{x}(t) + \mathbf{n}(t)] = E\mathbf{x}(t) + E\mathbf{n}(t) = \mu_{\mathbf{x}}.$$

Dus $\mu_{\mathbf{y}}(t)$ hangt niet van de tijd t af.
Voor ieder tweetal tijdstippen t_1 en t_2 geldt

$$\begin{aligned} R_{\mathbf{yy}}(t_1, t_2) &= E\mathbf{y}(t_1)\mathbf{y}(t_2) = E[\mathbf{x}(t_1)\mathbf{x}(t_2) + \mathbf{x}(t_1)\mathbf{n}(t_2) + \mathbf{n}(t_1)\mathbf{x}(t_2) + \\ &+ \mathbf{n}(t_1)\mathbf{n}(t_2)] = R_{\mathbf{xx}}(t_2 - t_1) + E\mathbf{x}(t_1)E\mathbf{n}(t_1) + \\ &+ E\mathbf{n}(t_1)E\mathbf{x}(t_2) + R_{\mathbf{nn}}(t_2 - t_1) = \\ &= R_{\mathbf{xx}}(t_2 - t_1) + a\delta(t_2 - t_1). \end{aligned}$$

Dus $R_{\mathbf{yy}}(t_1, t_2)$ hangt slechts van $t_2 - t_1$ af en we kunnen concluderen dat $\mathbf{y}(t)$ een stationair proces is. Tevens zijn i) en iii) bewezen.
Verder geldt

$$\begin{aligned} R_{\mathbf{xy}}(t_1, t_2) &= E\mathbf{x}(t_1)\mathbf{y}(t_2) = E[\mathbf{x}(t_1)\mathbf{x}(t_2) + \mathbf{x}(t_1)\mathbf{n}(t_2)] = \\ &= R_{\mathbf{xx}}(t_2 - t_1) + E\mathbf{x}(t_1)E\mathbf{n}(t_2) = R_{\mathbf{xx}}(t_2 - t_1). \end{aligned}$$

Dus $R_{\mathbf{xy}}(t_1, t_2)$ hangt slechts van $t_2 - t_1$ af en we kunnen concluderen dat het paar processen $(\mathbf{x}(t), \mathbf{y}(t))$ stationair is. Tevens is ii) bewezen.
Wanneer men in de praktijk uit het gestoorde proces $\mathbf{y}(t)$ het oorspronkelijke proces $\mathbf{x}(t)$ wil reconstrueren is relatie iii) van belang; immers de witte ruis manifisteert zich in de stochastische autocorrelatiefunctie 'slechts' als een delta-puls.

Bronvermelding
Deze tekst is afkomstig uit:
Dr. D.A. Overdijk, Fouriertheorie, Stochastiek en Signaalverwerking, collegedictaat nr. 2360, Technische Universiteit Eindhoven.

STUREN EN WAARNEMEN

J.W. van der Woude
Tecnische Universiteit Delft

INLEIDING

In deze notitie zullen we een aantal begrippen introduceren die van fundamenteel belang zijn voor de lineaire systeemtheorie. Daartoe gaan we uit van een systeem beschreven door de volgende lineaire vergelijkingen.

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t). \end{aligned} \tag{1}$$

Hierbij staat de vector x voor de toestand van het systeem, de vector u voor de ingang of besturing, en de vector y voor de uitgang of meting. De vectoren x, u en y zijn reëelwaardige vectorfuncties van de tijd t en hebben afmetingen n, m en p. A, B en C zijn reële constante matrices met afmetingen $n \times n$, $n \times m$ en $p \times n$. We nemen hier aan dat de componenten van de vector u continue functies zijn met op slechts eindig veel plaatsen een discontinuïteit. We geven dit aan door te schrijven : $u \in \Omega$. Onder deze aanname is de klasse van besturingen groot genoeg om realistische problemen aan te kunnen pakken terwijl tevens de vergelijkingen in (1) nog opgelost kunnen worden.

BESTUURBAARHEID EN WAARNEEMBAARHEID

Het eerste begrip dat we invoeren geeft de mate aan waarin we een systeem kunnen beinvloeden via de ingang u. We geven de oplossing x van (1) bij de beginwaarde $x(0) = x_0$ aan met $x(t, x_0, u)$. Meer expliciet volgt met de variatie van constanten formule

$$x(t, x_0, u) = e^{tA}x_0 + \int_0^t e^{(t-\tau)A}Bu(\tau)d\tau. \tag{2}$$

Definitie. *Het systeem (1) heet (volledig) bestuurbaar als er voor elk tweetal vectoren x_0 en x_1 een $t_1 > 0$ en een $u \in \Omega$ bestaan zo dat $x(t_1, x_0, u) = x_1$.*

We noemen dus een systeem (1) bestuurbaar als we vanuit een willekeurige toestand x_0 elke willekeurig andere toestand x_1 kunnen bereiken door toepassing van een geschikte besturing u.

Het tweede begrip dat we invoeren geeft aan in hoe verre men uit de ingang en de uitgang van een systeem de toestand kan bepalen. We geven de uitgang die het systeem geeft bij de begintoestand x_0 en ingang u aan met $y(t, x_0, u)$. Dus

$$y(t,x_0,u) = Cx(t,x_0,u) = Ce^{tA}x_0 + \int_0^t Ce^{(t-\tau)A}Bu(\tau)d\tau. \tag{3}$$

Definitie. *Het systeem (1) heet (volledig) waarneembaar als er een $t_1 > 0$ bestaat zodat voor een willekeurige $u \in \Omega$ uit $y(t,x_0,u) = y(t,x_1,u)$ voor $0 \leq t \leq t_1$ volgt dat $x_0 = x_1$.*

We zeggen dus dat een systeem waarneembaar is als we uit de kennis van de ingang u en de uitgang y over een voldoend lange tijd de begintoestand x_0 eenduidig kunnen bepalen. We merken op dat vanwege de lineariteit geldt dat systeem (1) waarneembaar is dan en slechts dan als er een $t_1 > 0$ bestaat zodat uit $y(t,x_0,0) = Ce^{tA}x_0 = 0$ voor $0 \leq t \leq t_1$ volgt dat $x_0 = 0$.

De eigenschappen bestuurbaarheid en waarneembaarheid van (1) worden volkomen bepaald door het matrixtripel (C,A,B). We zullen daarom (C,A,B) bestuurbaar respectievelijk waarneembaar noemen als het systeem (1) deze eigenschappen heeft.

In het volgende zullen we expliciete voorwaarden behandelen voor bestuurbaarheid en waarneembaarheid van (1). Belangrijk bij de afleiding van deze voorwaarden is een gevolg van de stelling van Cayley-Hamilton. Namelijk ; iedere positieve gehele macht van de $n \times n$ matrix A kan geschreven worden als een lineaire combinatie met scalaire coëfficienten van de matrices I, A,..., A^{n-1}.

BESTUURBAARHEID

De volgende bestuurbaarheidsvoorwaarde kan worden afgeleid.

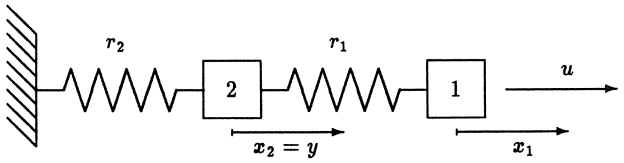
Stelling. *Het systeem (1) is bestuurbaar dan en slechts dan indien*

$$\text{rang}\left(B, AB, A^2B, \ldots, A^{n-1}B\right) = n. \tag{4}$$

Opmerking. In (4) staat $(B, AB, A^2B, \ldots, A^{n-1}B)$ voor de $n \times nm$ matrix verkregen door de matrices $B, AB, A^2B, \ldots, A^{n-1}B$ achter elkaar te plaatsen. Om de stelling enigszins aannemelijk te maken veronderstel dat rang conditie (4) niet geldt. Dan zijn de rijen van de matrix $(B, AB, A^2B, \ldots, A^{n-1}B)$ onderling afhankelijk. Er bestaat dus een vector $\eta \in R^n$ zodat $\eta'(B, AB, A^2B, \ldots, A^{n-1}B) = 0$, ofwel $\eta' A^i B = 0$ voor alle $i = 0, 1, \ldots, n-1$ (hierbij staat $'$ voor de getransponeerde). Vanwege de stelling van Cayley-Hamilton volgt hieruit dat $\eta' A^i B = 0$ voor alle $i \geq 0$. Op zijn beurt betekent dit dat $\eta' e^{tA} B = 0$ voor alle t. Kies nu $x_0 = 0$ en x_1 zodat $\eta' x_1 \neq 0$. Uit (2) volgt nu $\eta' x(t_1, x_0, u) = 0$ voor alle $t_1 > 0$ en alle $u \in \Omega$. Dus $x_1 \neq x(t_1, x_0, u)$ voor alle $t_1 > 0$ en alle $u \in \Omega$. Volgens de definitie is het systeem niet bestuurbaar. We hebben hiermee de noodzaak van de rang conditie (4) bewezen.

Voorbeeld. (i) In de onderstaande figuur zien we een massa-veer systeem bestaande uit twee eenheidsmassa's verbonden met veren met veerconstante r_1 en r_2. Op de rechter massa oefenen we een kracht uit (de ingang) en de verplaatsing van de linker massa meten

we (de uitgang). De verplaatsing van de massa's geven we aan met x_1 en x_2.



De bijbehorende vergelijkingen zijn (de afhankelijkheid van de tijd t is weggelaten)

$$\begin{aligned} \ddot{x}_1 &= u - r_1(x_1 - x_2), \\ \ddot{x}_2 &= r_1(x_1 - x_2) - r_2 x_2, \\ y &= x_2. \end{aligned}$$

De vergelijkingen vormen nog geen systeem als (1) omdat er de tweede orde afgeleiden in voorkomen. We kunnen dit verhelpen door de invoering van nieuwe variabelen $x_3 = \dot{x}_1$ en $x_4 = \dot{x}_2$. We krijgen dan de volgende eerste orde vergelijkingen

$$\begin{aligned} \dot{x}_1 &= x_3, \\ \dot{x}_2 &= x_4, \\ \dot{x}_3 &= u - r_1(x_1 - x_2), \\ \dot{x}_4 &= r_1(x_1 - x_2) - r_2 x_2. \end{aligned}$$

Uit deze vergelijkingen volgt een systeem van de vorm (1) met

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -r_1 & r_1 & 0 & 0 \\ r_1 & -r_1 - r_2 & 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, C = (0\ 1\ 0\ 0).$$

Er geldt nu dat

$$\left(B, AB, A^2B, A^3B\right) = \begin{pmatrix} 0 & 1 & 0 & -r_1 \\ 0 & 0 & 0 & r_1 \\ 1 & 0 & -r_1 & 0 \\ 0 & 0 & r_1 & 0 \end{pmatrix}.$$

Deze matrix is niet singulier dan en slechts dan als $r_1 \neq 0$. Vanwege de voorgaande stelling volgt dat het massa-veer systeem bestuurbaar is dan en slechts dan als $r_1 \neq 0$.
Het is duidelijk dat indien $r_1 = 0$ de beide massa's niet gekoppeld zijn. Een kracht (een besturing) uitgeoefend op de rechter massa is dan niet van invloed op de beweging van de linker massa. Deze kan dus niet naar een andere plaats gebracht of een andere snelheid gegeven worden door het toepassen van welke besturing dan ook.

(*ii*) Beschouw het systeem (zonder de uitgangsvergelijking)

$$\begin{aligned} \dot{x}_1 &= -2x_1 - 6x_2 - 3u, \\ \dot{x}_2 &= 2x_1 + 5x_2 + 2u. \end{aligned}$$

Geschreven als een systeem van de vorm (1) wil dit zeggen dat

$$A = \begin{pmatrix} -2 & -6 \\ 2 & 5 \end{pmatrix}, B = \begin{pmatrix} -3 \\ 2 \end{pmatrix}.$$

Dan is

$$\text{rang}\,(B, AB) = \text{rang} \begin{pmatrix} -3 & -6 \\ 2 & 4 \end{pmatrix} = 1$$

zodat het systeem niet bestuurbaar is. Inderdaad, als $z = 2x_1 + 3x_2$ dan volgt $\dot{z} = z$ en dus $z(t) = e^t z_0$. Als nu bijvoorbeeld $z_0 = 0$ dan geldt $z(t) = 0$ voor alle $t > 0$ en zijn toestanden $x = (x_1, x_2)'$ waarvoor $2x_1 + 3x_2 = 0$ niet te verplaatsen.

De bestuurbaarheid van (1) of het matrixtripel (C, A, B) hangt niet af van C. Daarom spreken we ook wel van de bestuurbaarheid van het matrixpaar (A, B).

WAARNEEMBAARHEID

De volgende waarneembaarheidsvoorwaarde kan worden afgeleid.

Stelling. *Het systeem (1) is waarneembaar dan en slechts dan indien*

$$\text{rang} \begin{pmatrix} C \\ CA \\ . \\ . \\ CA^{n-1} \end{pmatrix} = n. \tag{5}$$

Opmerking. De matrix in (5) heeft afmetingen $np \times n$ en is verkregen door de matrices $C, CA, CA^2, \ldots, CA^{n-1}$ onder elkaar te plaatsen. Stel eens dat de rang conditie (5) niet geldt. Dan zijn de kolommen in de matrix in de conditie lineair afhankelijk. Dit betekent dat er een vector $x_0 \neq 0$ bestaat zodat $CA^i x_0 = 0$ voor $i = 0, 1, \ldots, n-1$. Vanwege de stelling van Cayley-Hamilton volgt dat $CA^i x_0 = 0$ voor alle $i \geq 0$ waaruit weer volgt dat $Ce^{tA} x_0 = 0$. Merk op dat $y(t, x_0, 0) = Ce^{tA} x_0$. Omdat $x_0 \neq 0$ volgt nu direct met de definitie dat het systeem niet waarneembaar kan zijn. Hiermee is de noodzaak van de rang conditie (5) bewezen.

Voorbeeld. (*i*) Beschouw het massa-veer systeem uit het vorige voorbeeld. Er volgt dat

$$\begin{pmatrix} C \\ CA \\ CA^2 \\ CA^3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ r_1 & -r_1 - r_2 & 0 & 0 \\ 0 & 0 & r_1 & -r_1 - r_2 \end{pmatrix}.$$

Deze matrix heeft rang 4 dan en slechts dan als $r_1 \neq 0$. Het systeem is dus waarneembaar dan en slechts dan als $r_1 \neq 0$.

Als $r_1 = 0$ dan zijn de twee massa's niet gekoppeld. Dit betekent dat de bewegingen van de rechter massa niet van invloed zijn op de bewegingen van de linker massa. Dus de bewegingen van de rechter massa kunnen niet uit metingen aan de linker massa worden waargenomen.

(*ii*) Beschouw het systeem

$$\begin{array}{rcrcrcl} \dot{x}_1 & = & -x_1 & + & 3x_2 & + & u, \\ \dot{x}_2 & = & -2x_1 & + & 4x_2 & + & u, \\ y & = & x_1 & - & x_2. & & \end{array}$$

Dan is

$$A = \begin{pmatrix} -1 & 3 \\ -2 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, C = (1 \ -1).$$

Er geldt

$$\begin{pmatrix} C \\ CA \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$$

zodat het systeem niet waarneembaar is. Het blijkt dat als $x_0 = (1 \ 1)'$ dan $y(t, x_0, 0) = 0$ voor alle $t \geq 0$ zodat de uitgang geen informatie levert over het feit dat $x_0 \neq 0$.

De waarneembaarheid van (1) blijkt alleen af te hangen van A en C en niet van B. Daarom zeggen we in plaats van (1) of (C, A, B) is waarneembaar ook wel : (C, A) is waarneembaar.

Er bestaat een analogie tussen bestuurbaarheids- en waarneembaarheidseigenschappen. Zo volgt uit de voorgaande stellingen dat (A, B) bestuurbaar is dan en slechts dan als (B', A') waarneembaar is en dat (C, A) waarneembaar is dan en slechts dan als (A', C') bestuurbaar is. Dit betekent dat we bij elke stelling over de bestuurbaarheid een stelling over de waarneembaarheid kunnen afleiden door de formules te transponeren. Men noemt dit het dualiteitsprincipe. Dit principe stelt ons in staat sommige resultaten alleen voor bestuurbaarheid te bewijzen en de overeenkomstige resultaten voor de waarneembaarheid zonder bewijs op te schrijven.

BASISTRANSFORMATIES IN DE TOESTANDSRUIMTE

Als we in de toestandsruimte van het systeem (1) een nieuwe basis kiezen met overgangsmatrix S dan komt dit overeen met de transformatie

$$x(t) = S\bar{x}(t). \tag{6}$$

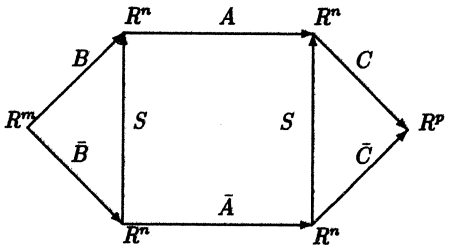
Substitutie van deze transformatie in (1) geeft het systeem

$$\begin{array}{rcrcl} \dot{\bar{x}}(t) & = & \bar{A}\bar{x}(t) & + & \bar{B}u(t), \\ y(t) & = & \bar{C}\bar{x}(t), & & \end{array} \tag{7}$$

waarbij

$$\bar{A} = S^{-1}AS, \bar{B} = S^{-1}B, \bar{C} = CS. \tag{8}$$

De twee systemen (1) en (7) noemen we isomorf als er een inverteerbare matrix S bestaat zodat de tripels (C, A, B) en $(\bar{C}, \bar{A}, \bar{B})$ voldoen aan (8). We kunnen de relaties (8) als volgt aangeven in een diagram.



We zeggen dat het diagram commuteert als de samengestelde afbeeldingen van afbeeldingen in het diagram alleen afhangen van het begin- en eindpunt in het diagram en niet van de weg waarmee het begin- met het eindpunt verbonden is. Zo kan men uit het diagram de formules (8) gemakkelijk aflezen. Ook andere gelijkheden kan men gemakkelijk aflezen. Bijvoorbeeld $CA^iB = \bar{C}\bar{A}^i\bar{B}$ voor alle $i \geq 0$.

Uit de interpretatie van S als matrix voor de overgang naar een andere basis volgt dat bestuurbaarheid, waarneembaarheid en het ingang-uitgangsgedrag (zie (3)) invariant blijven bij een basistransformatie in de toestandsruimte.

Stelling. *Als (C, A, B) en $(\bar{C}, \bar{A}, \bar{B})$ isomorf zijn dan geldt het volgende.*
i) (A, B) is bestuurbaar dan en slechts dan als $(\bar{A}, \bar{B})$ bestuurbaar is.
ii) (C, A) is waarneembaar dan en slechts dan als $(\bar{C}, \bar{A})$ waarneembaar is.
iii) $Ce^{tA}B = \bar{C}e^{t\bar{A}}\bar{B}$ voor alle t.

Men kan proberen een matrix S te zoeken waarmee het tripel $(\bar{C}, \bar{A}, \bar{B})$ een eenvoudige gedaante krijgt. Bijvoorbeeld in het geval dat het systeem (1) mogelijk niet bestuurbaar is kan men het volgende bewijzen.

Stelling. *Er bestaat een inverteerbare matrix S zodat het tripel $(\bar{C}, \bar{A}, \bar{B})$ gegeven door (8) de volgende blokdecompositie heeft*

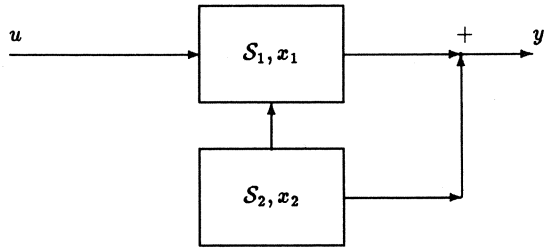
$$\bar{A} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \bar{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}, \bar{C} = (C_1\ C_2) \tag{9}$$

waarbij (A_{11}, B_1) volledig bestuurbaar is.

Bovenstaande betekent dat $\bar{x}$ in (7) gesplitst kan worden als $\bar{x} = (x_1 \; x_2)'$ en dat het systeem (7) in meer detail geschreven kan worden als

$$\begin{array}{rcl} \dot{x}_1 &=& A_{11}x_1 + A_{12}x_2 + B_1u, \\ \dot{x}_2 &=& A_{22}x_2, \\ y &=& C_1x_1 + C_2x_2. \end{array}$$

Schematisch kan het systeem als volgt weergegeven worden.



Het systeem kan dus worden ontbonden in twee stukken : een volledig bestuurbaar stuk $\mathcal{S}_1$ met toestandsvariabele x_1 en een volledig onbestuurbaar stuk $\mathcal{S}_2$ met toestandsvariabele x_2. Er gaat wel invloed uit van $\mathcal{S}_2$ op $\mathcal{S}_1$ maar niet omgekeerd. Ook geldt dat het ingang-uitgangsgedrag van (1) gelijk is aan het ingang-uitgangsgedrag van het bestuurbare deel $\mathcal{S}_1$. Dit wordt uitgedrukt door : $Ce^{tA}B = C_{11}e^{tA_{11}}B_{11}$ voor alle t.

De voorgaande resultaten laten zich direct dualiseren. We krijgen dan het volgende.

Stelling. *Er bestaat een inverteerbare matrix S zodat het tripel $(\bar{C}, \bar{A}, \bar{B})$ gegeven door (8) de volgende blokdecompositie heeft*

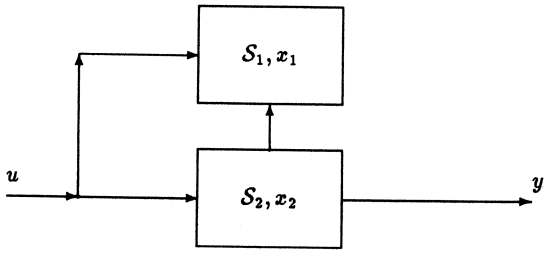
$$\bar{A} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \bar{B} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \bar{C} = (0 \; C_2) \tag{10}$$

waarbij (C_2, A_{22}) volledig waarneembaar is.

In getransformeerde variabelen luiden de vergelijkingen

$$\begin{array}{rcl} \dot{x}_1 &=& A_{11}x_1 + A_{12}x_2 + B_1u, \\ \dot{x}_2 &=& A_{22}x_2 + B_2u, \\ y &=& C_2x_2. \end{array}$$

Hier kunnen we het systeem ontbinden in een waarneembaar deelsysteem $\mathcal{S}_2$ en een onwaarneembaar deelsysteem $\mathcal{S}_1$.



Het ingang-uitgangsgedrag van het systeem (1) en van het waarneembare deelsysteem $\mathcal{S}_2$ zijn aan elkaar gelijk: $Ce^{tA}B = C_{22}e^{tA_{22}}B_{22}$ voor alle t.

Tenslotte kunnen we bij een eventueel niet-bestuurbaar en niet-waarneembaar systeem een basis kiezen die beide facetten naar voren brengt.

Stelling. *Gegeven systeem (1) bestaat er een inverteerbare matrix S zodat het tripel $(\bar{C}, \bar{A}, \bar{B})$ gegeven door (8) de volgende gedaante heeft*

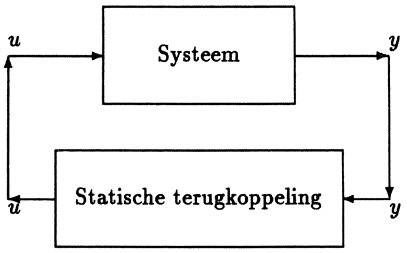
$$\bar{A} = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & 0 & 0 & A_{44} \end{pmatrix}, \bar{B} = \begin{pmatrix} B_1 \\ 0 \\ B_3 \\ 0 \end{pmatrix}, \bar{C} = (0\ 0\ C_3\ C_4) \tag{11}$$

waarbij $\left(\begin{pmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{pmatrix}, \begin{pmatrix} B_1 \\ B_3 \end{pmatrix}\right)$ *volledig bestuurbaar is en* $\left((C_3\ C_4), \begin{pmatrix} A_{33} & A_{34} \\ 0 & A_{44} \end{pmatrix}\right)$ *volledig waarneembaar is.*

De vergelijkingen in de getransformeerde variabelen volgen eenvoudig uit (11). Ook een schematische weergave kan verkregen worden uit (11). Het schema is een combinatie van de twee eerder gegeven schema's. Bovenstaande splitsing van een systeem in delen die (on)bestuurbaar en (on)waarneembaar zijn wordt de Kalman decompositie genoemd.

STABILISATIE

Een besturing waarbij de ingang u wordt bepaald op grond van de uitgang y of de toestand x heet terugkoppeling. Het eenvoudigste type terrugkoppeling is de statische terugkoppeling of proportionele regelaar $u = Fy$ waar F een matrix is van geschikte afmetingen.



Omdat de uitgang in tegenstelling tot de toestand in het algemeen niet voldoende informatie geeft over de toekomstige responsie van het systeem zijn de mogelijkheden van uitgangsterugkoppeling beperkt. Beter hanteerbaar is de toestandsterugkoppeling waarin de besturing een functie is van de toestand

$$u(t) = Fx(t). \tag{12}$$

Samenstelling met systeem (1) levert dan het volgende autonome systeem (= systeem zonder besturingen)

$$\dot{x}(t) = (A + BF)x(t). \tag{13}$$

We merken op dat de oplossingen van het autonome systeem $\dot{x}(t) = Ax(t)$ voldoen aan $x(t) \to 0$ $(t \to \infty)$ dan en slechts dan als A een stabiliteitsmatrix is (alle eigenwaarden van A hebben een negatief reëel deel). Indien A een stabiliteitsmatrix is wordt het systeem $\dot{x}(t) = Ax(t)$ daarom ook wel (asymptotisch) stabiel genoemd.

Definitie. *Het systeem (1) heet stabiliseerbaar indien er een terugkoppelmatrix F bestaat zodat $A + BF$ een stabiliteitsmatrix is, of ook, zodat (13) stabiel is.*

Bij de afleiding van voorwaarden waaronder stabiliseerbaarheid mogelijk is wordt een belangrijke rol gespeeld door de volgende zeer fundamentele stelling, de *poolplaatsingsstelling.*

Stelling. *Het systeem (1) is bestuurbaar dan en slechts dan als er bij elk polynoom $p(z) = z^n + p_1 z^{n-1} + \ldots + p_{n-1} z + p_n$ een matrix F bestaat zodat $\det(zI - (A + BF)) = p(z)$.*

Opmerking. We bewijzen de bovenstaande stelling hier niet. Wel brengen we in herinnering dat $\det(zI - (A + BF))$ het karakteristieke polynoom is van de matrix $A + BF$. In geval van volledige bestuurbaarheid zijn de reële coefficienten van het polynoom $p(z)$ willekeurig te kiezen en daarmee ook de wortels van $p(z)$ (mits symmetrisch om de reële as). De eigenwaarden van $A + BF$ (ook wel de polen van het systeem genoemd) kunnen

dus op willekeurige plaatsen in het complexe vlak gelegd worden (mits symmetrisch om de reële as). Er volgt nu : als het systeem (1) bestuurbaar is, dan is het ook stabiliseerbaar.

Voorbeeld. Beschouw het massa-veer systeem uit het vorige voorbeeld. Het systeem is bestuurbaar mits $r_1 \neq 0$. Veronderstel dat $r_1 = 1$. Een eenvoudige berekening laat zien dat de eigenwaarden van de matrix A op de imaginaire as liggen : op $\pm i\frac{\sqrt{5}\pm 1}{2}$. Dit correspondeert met ongedempte eigentrillingen. Veronderstel nu dat het gewenst is door een stabiliserende terugkoppeling $u = Fx$ de eigenwaarden van $A + BF$ op -1,-2 en -1+i en -1-i te leggen. Kies dan $p(z) = (z+1)\,(z+2)\,(z+i-\mathrm{i})\,(z+i+\mathrm{i}) = z^4 + 5z^3 + 10z^2 + 10z + 4$ en bereken F zodat $\det(zI - (A + BF)) = p(z)$. Dit levert $F = (-7, 11, -5, 0)$.

Beschouw nu het systeem (1). Eerder hebben we gezien dat door een geschikte basiskeuze het systeem gesplitst kan worden in een volledig bestuurbaar deel en een volledig onbestuurbaar deel. Met behulp van de poolplaatsingsstelling kan nu het volgende bewezen worden.

Stelling. *Het systeem (1) is stabiliseerbaar dan en slechts dan als het volledig onbestuurbare deel van (1) stabiel is.*

Voorbeeld. Zij

$$A = \begin{pmatrix} -3 & -4 & -6 \\ 2 & 3 & 3 \\ 1 & 2 & 2 \end{pmatrix}, B = \begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix}.$$

Dan geldt

$$\left(B, AB, A^2B\right) = \begin{pmatrix} 2 & 0 & -4 \\ 0 & 1 & 3 \\ -1 & 0 & 2 \end{pmatrix}.$$

Deze matrix is singulier en het systeem is dus niet bestuurbaar. Merk op dat de kolomruimte van de matrix (B, AB, A^2B) opgespannen wordt door de vectoren $(2, 0, -1)'$ en $(0, 1, 0)'$. Samen met de vector $(0, 0, 1)'$ vormen de drie vectoren een basis voor de toestandsruimte R^3. Definieer nu de overgangsmatrix

$$S = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Er geldt dan

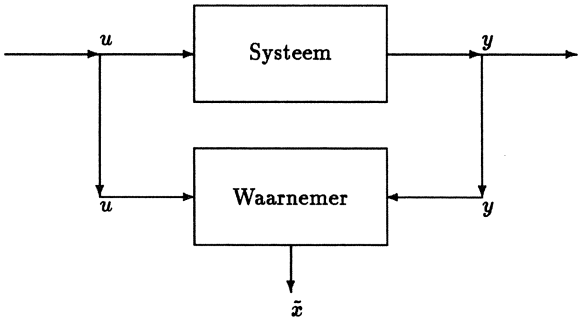
$$\bar{A} = S^{-1}AS = \begin{pmatrix} 0 & -2 & -3 \\ 1 & 3 & 3 \\ 0 & 0 & -1 \end{pmatrix}, \bar{B} = S^{-1}B = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Schrijf de matrices $\bar{A}$ en $\bar{B}$ nu in de vorm (9). Er volgt bijvoorbeeld dat $A_{22} = -1$. Het onbestuurbare gedeelte van het systeem is dus stabiel. Immers de eigenwaarden van A_{22} hebben negatief reëel deel. Het bestuurbare gedeelte van het systeem vastgelegd door de

matrices A_{11} en B_1 kan stabiel gemaakt worden. Vanwege de poolplaatsingsstelling is er een matrix F_1 zodat $A_{11}+B_1F_1$ een stabiliteitsmatrix is. Er bestaat dus een samengestelde matrix $\bar{F} = \begin{pmatrix} \bar{F}_1 & \bar{F}_2 \end{pmatrix}$ met $\bar{F}_2$ willekeurig zodat $\bar{A} + \bar{B}\bar{F}$ een stabiliteitsmatrix is. Definieer $F = \bar{F}S^{-1}$ en er volgt dat $A + BF$ een stabiliteitsmatrix is. Het systeem is dus niet bestuurbaar is maar wel stabiliseerbaar.

WAARNEMERS

Gewoonlijk is de veronderstelling dat de toestand rechtstreeks meetbaar is onrealistisch. Men kan echter wel vaak een toestandswaarnemer bouwen. Dit is een systeem dat met behulp van de gegeven ingang en uitgang van (1) een schatting geven van de toestand. Zo'n waarnemer heeft dus u en y als ingang en als uitgang een grootheid $\tilde{x}$ die een benadering is voor de toestand x.



We verlangen van een waarnemer dat het verschil $x(t) - \tilde{x}(t)$ op den duur klein is voor elk paar beginwaarden x_0 en $\tilde{x}_0$. Tevens willen we dat als $\tilde{x}$ eenmaal de waarde van x heeft dat dit ook zo blijft : als $x(t_0) = \tilde{x}(t_0)$ voor een zeker tijdstip t_0 dan moet $x(t) = \tilde{x}(t)$ voor alle $t \geq t_0$. Aangezien $\tilde{x}$ een benadering van x moet zijn is het verder aannemelijk dat $\tilde{x}$ samen met de besturing u bij benadering voldoet aan de vergelijkingen van (1). We merken op dat een eventueel verschil tussen x en $\tilde{x}$ alleen waar te nemen is door een verschil tussen de bijbehorende metingen $y = Cx$ en $\tilde{y} = C\tilde{x}$.

Het blijkt nu mogelijk te veronderstellen dat de waarnemer de volgende vorm heeft.

$$\begin{aligned} \dot{\tilde{x}} &= A\tilde{x} + Bu + R(y - \tilde{y}), \\ \tilde{y} &= C\tilde{x}, \end{aligned} \tag{14}$$

waarbij R een nog geschikt te kiezen matrix is. We kunnen de waarnemer dus als een duplicaat van het oorspronkelijke systeem beschouwen met een extra ingang om de afwijking van y en $\tilde{y}$ te corrigeren.

We zullen nu verifiëren dat de waarnemer (14) aan de gestelde eisen voldoet. Definieer daartoe $e(t) = x(t) - \tilde{x}(t)$. Er volgt eenvoudig dat

$$\dot{e} = (A - RC)e.$$

Stel nu dat $x(t_0) = \tilde{x}(t_0)$ voor een zekere t_0. Dus $e(t_0) = 0$. Er volgt direct dat $e(t) = 0$ voor alle $t \geq t_0$. Verder volgt dat $e(t) \to 0$ $(t \to \infty)$ dan en slechts dan als $A - RC$ een stabiliteitsmatrix is. We kunnen dus een waarnemer construeren die aan de bovengestelde eisen voldoet als we een matrix R kunnen vinden zodat alle eigenwaarden van $A - RC$ een negatief reëel deel hebben.

Definitie. *Het systeem (1) heet detecteerbaar indien er een matrix R bestaat zodat alle eigenwaarden van $A - RC$ een negatief reëel deel hebben.*

Het is duidelijk dat het paar (C, A) detecteerbaar is dan en slechts dan als het paar (A', C') stabiliseerbaar is. Gebruikmakend van dualiteit volgt tevens.

Stelling. *Het systeem (1) is detecteerbaar dan en slechts dan als het volledig onwaarneembaar deel van (1) stabiel is.*

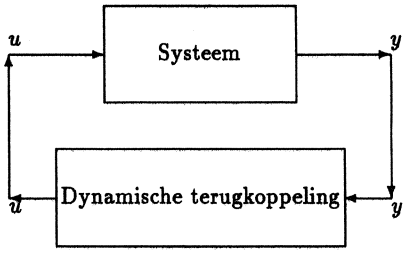
Er geldt : als systeem (1) waarneembaar is, dan is het ook detecteerbaar. Verder volgt uit het voorgaande.

Stelling. *Er bestaat een toestandswaarnemer van de vorm (14) voor het systeem (1) dan en slechts dan als (1) detecteerbaar is.*

DYNAMISCHE TERUGKOPPELING

Als we nu het systeem (1) willen stabiliseren maar we hebben de toestand niet tot onze beschikking dan kunnen we dit trachten te doen door een dynamische terugkoppeling bestaande uit de combinatie van een toestandswaarnemer en een stabiliserende toestandsterugkoppeling. (Neem (14), vul in $\tilde{y} = C\tilde{x}$ en $u = F\tilde{x}$.)

$$\begin{aligned} \dot{\tilde{x}} &= (A + BF - RC)\tilde{x} + Ry, \\ u &= F\tilde{x}. \end{aligned} \tag{15}$$



We laten nu zien dat we op deze manier een stabiel systeem krijgen. Neem daarom aan dat F en R zo gekozen zijn dat $A + BF$ en $A - RC$ stabiliteitsmatrices zijn. Beschouw nu

het systeem gedefinieerd door samenstelling van (1) en (15). Dus

$$\begin{aligned} \dot{x} &= Ax + BF\tilde{x}, \\ \dot{\tilde{x}} &= RCx + (A + BF - RC)\tilde{x}. \end{aligned}$$

Met de variabele $e = x - \tilde{x}$ volgt

$$\begin{aligned} \dot{x} &= (A + BF)x - BFe, \\ \dot{e} &= (A - RC)e. \end{aligned}$$

De coefficientenmatrix van dit systeem is

$$\tilde{M} = \begin{pmatrix} A + BF & -BF \\ 0 & A - RC \end{pmatrix}.$$

Omdat $A + BF$ en $A - RC$ stabiliteitsmatrices zijn is ook $\tilde{M}$ een stabiliteitsmatrices. We hebben nu het volgende resultaat.

Stelling. *Er bestaat een stabiliserende dynamische terugkoppeling (15) voor het systeem (1) dan en slechts dan als (1) stabiliseerbaar en detecteerbaar is.*

Conclusie. Nodig en voldoende voor de stabilisatie door terugkoppeling is dat het systeem (1) stabiliseerbaar en detecteerbaar is. Hieraan is zeker voldaan indien het systeem (1) bestuurbaar en waarneembaar is.

Wat betreft de feitelijk terugkoppeling kunnen we twee gevallen onderscheiden.
$C = I$. In dit geval is de waarneembaarheid, en dientengevolge de detecteerbaarheid, vanzelfsprekend. Dus alleen stabiliseerbaarheid dient onderzocht te worden. Omdat de gehele toestand gemeten wordt kan voor de stabilisatie met een statische (= proportionele) toestandsterugkoppeling zoals in (12) volstaan worden.
$C \neq I$. In dit geval dienen zowel stabiliseerbaarheid als detecteerbaarheid onderzocht te worden. Voor de stabilisatie zal nu in het algemeen een dynamische (= proportionele + integrerende) terugkoppeling zoals in (15) toegepast moeten worden.

LITERATUUR.
Er bestaan veel boeken die het onderwerp van deze notitie uitgebreid behandelen. In het onderstaand overzicht zijn slechts enkele hiervan vermeld. Alle vermelde boeken hebben een inleidend karakter en geven een goed beeld van de systeemtheorie.

D.G.Luenberger, *Introduction to Dynamic Systems*, Wiley, New York, 1979.
T. Chen, *Introduction to Linear System Theory*, Holt, Rinehart, Winston, New York, 1984.
T. Kailath, *Linear Systems*, Prentice Hall, 1980.
H.W. Knobloch, H. Kwakernaak, *Lineare Kontrolltheorie*, Springer, 1985.
D.M. Wiberg, *State Space and Linear Systems*, Schaum's Outline Series, Mc Graw-Hill, New York, 1971.

TOT SLOT.

Bij het samenstellen van deze notitie is dankbaar gebruik gemaakt van collegedictaten van de systeemtheoriegroepen van diverse nederlandse universiteiten. Met name zijn delen overgenomen uit het dictaat 'Lineaire Multivariabele Systemen' geschreven door prof.dr.ir. M.L.J. Hautus (bewerkt door dr. F. Eising) van de Technische Universiteit Eindhoven. Hiervoor hartelijke dank.

Tijdoptimale besturing van lineaire systemen

M.L.J. Hautus

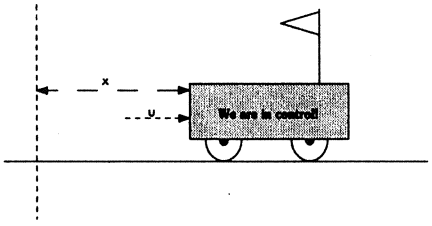
1. Probleemformulering

Vele systemen kunnen worden gestuurd. De mogelijkheid tot sturen kan bijvoorbeeld voorkomen bij mechanische, chemische of economische systemen. Wiskundig kan zo'n situatie vaak beschreven worden als een (gewone) differentiaalvergelijking met een tijdsafhankelijke parameter erin, die we de besturing noemen. Door middel van die parameter kan invloed worden uitgeoefend op het verloop van het proces dat door de differentiaalvergelijking wordt beschreven. Het ligt voor de hand dat men deze mogelijkheid tot sturen wil benutten om een gegeven doel te bereiken. Zo zal men bijvoorbeeld proberen een raket naar de maan te sturen, of, in een chemisch proces, materiaal te produceren van een gewenste kwaliteit.

Gewoonlijk zullen er meerdere besturingen zijn die het beoogde bewerkstelligen. Het ligt dan voor de hand dat men onder deze besturingen een optimale kiest. Om te definiëren wat dit betekent, moet men een optimaliteitscriterium aangeven. In het geval van de raket die naar de maan vliegt, kan dat bijvoorbeeld minimaal brandstofgebruik zijn, voor een chemisch proces een maximale hoeveelheid geproduceerde stof.

De algemene behandeling van optimale-besturings-problemen is nogal ingewikkeld. We zullen ons hier beperken tot een heel eenvoudige situatie: *De tijdoptimale besturing van een lineair systeem*. We beginnen met enkele voorbeelden.

1.1 Voorbeeld Veronderstel dat we een wagentje hebben dat op het tijdstip $t = 0$ een bepaalde positie en snelheid heeft. We kunnen het wagentje met een beperkte kracht vooruit duwen of tegenhouden. Onze taak is het wagentje in een zo kort



mogelijke tijd op een voorgeschreven positie tot stilstand te krijgen. Een eenvoudig wiskundig model van het probleem ziet er alsvolgt uit:

De bewegingsvergelijking van de wagen luidt:

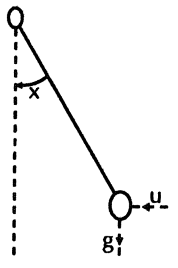
(1.1) $\ddot{x} = u,$

waar x de positie van de wagen t.o.v. de gewenste positie voorstelt, en u de uitgeoefende kracht en dus de besturing. Voor het gemak stellen we alle constanten gelijk aan 1 (We kunnen dit altijd bereiken door middel van een geschikte schaling). Verder geven we de beperking op de besturing aan door te eisen dat $|u(t)| \leq 1$ geldt voor alle t. We kunnen het optimaliseringsprobleem dus alsvolgt weergeven:

Probleem *Gegeven getallen x_0 en v_0, bepaal een functie u: $[0, \infty) \to [-1, 1]$ zo dat voor de oplossing van de differentiaalvergelijking* (1.1) *met beginwaarden $x(0) = x_0$, $\dot{x}(0) = v_0$ de gewenste eindtoestand $x(T) = \dot{x}(T) = 0$ bereikt wordt met minimale $T > 0$.*

Bij dit probleem is de oplossing niet zo moeilijk te raden. Bij het volgende probleem is dat niet zo eenvoudig.

1.2 Voorbeeld Laat nu een slinger gegeven zijn met een gegeven begin-positie en -snelheid. Weer kunnen we een beperkte kracht u uitoefenen als aangegeven in de tekening en ons doel is de slinger in de evenwichtsstand tot rust te brengen. Als we de bewegingsvergelijking opstellen volgens de wetten van Newton, krijgen we een niet-lineaire differentiaalvergelijking. Eenvoudigheidshalve lineariseren we deze. Bovendien schalen we de variabelen zo dat alle constanten gelijk aan 1 worden. Op soortgelijke manier als in het vorige voorbeeld krijgen we dan het volgende probleem:



Probleem *Gegeven getallen x_0 en v_0, bepaal een functie u:$[0, \infty) \to [-1, 1]$ zo dat de oplossing van de differentiaalvergelijking $\ddot{x} + x = u$ met beginwaarden $x(0) = x_0$, $\dot{x}(0) = v_0$ voldoet aan $x(T) = \dot{x}(T) = 0$ met minimale T.*

We zien dat de problemen afgezien van de differentiaalvergelijking het zelfde zijn. Het is gemakkelijk in te zien hoe men het probleem kan

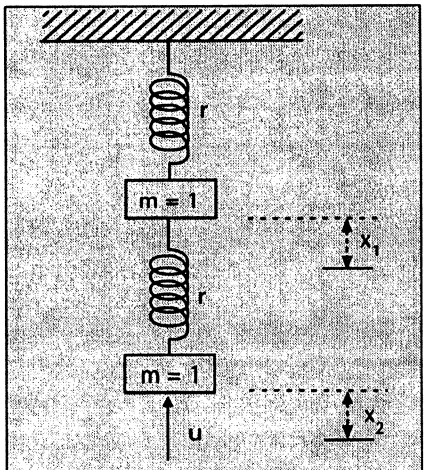
generaliseren door een algemenere vergelijking van de tweede orde. Voor een algemene formulering van het optimaliseringsprobleem dat we gaan behandelen, maken we gebruik van vector-matrix-notatie:

1.3 Tijdoptimaliseringsprobleem
Gegeven een n×n-matrix A, een n-vector b, en een n-vector x_0, bepaal een functie u: [0, ∞) → [-1, 1] *zo dat de oplossing van het stelsel differentiaalvergelijkingen*

(1.2) $\dot{x} = Ax + bu,\ x(0) = x_0$

voldoet aan $x(T) = 0$ *voor minimale T.*

De problemen uit bovenstaande voorbeelden kunnen op de standaardmanier tot dit probleem herleid worden.



Men kan problemen van hogere orde formuleren als men bijv. ingewikkeldere mechanische systemen beschouwt. Een voorbeeld hiervan is het systeem weergegeven in de hierboven aangegeven tekening. Het systeem voldoet aan de vergelijking:

$$\ddot{x}_1 = r(x_2 - 2x_1),$$
$$\ddot{x}_2 = r(x_1 - x_2) + u,$$

waarin r de veerconstante van beide veren is. Door de substituties $x_3 = \dot{x}_1$, $x_4 = \dot{x}_2$, krijgen we het stelsel (1.2), waar

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2r & r & 0 & 0 \\ r & -r & 0 & 0 \end{bmatrix},\ b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

We werken dit voorbeeld hier niet verder uit.

We hebben nog niet gespecificeerd tot welke functieklasse u moet behoren. Men zou in eerste instantie kunnen denken aan de klasse der continue functies. Het zal echter blijken dat het probleem gewoonlijk geen oplossing heeft in deze klasse. We zullen daarom algemenere besturingen toelaten en slechts eisen dat u stuksgewijs continu is. Dit betekent dat we wel toestaan dat u discontinuïteiten vertoont, maar dat er hiervan maar eindig veel mogen voorkomen, terwijl in elk discontinuïteitspunt de linker- en rechterlimiet bestaan. Als we zulke besturingen toelaten, moeten we ook een algemenere interpretatie geven aan het begrip oplossing van de differentiaalvergelijking (1.2). We zullen van de oplossing eisen dat ze voldoet aan de differentiaalvergelijking in de punten waar u continu is, en dat ze continu is

in de discontinuïteitspunten van u. Uit deze interpretatie volgt dat de waarde van u in de discontinuïteitspunten geen invloed heeft op de oplossing.

Vragen die naar aanleiding van dit optimaliseringsprobleem naar voren komen zijn: Bestaat er een oplossing, is deze uniek en hoe vindt men een oplossing? Het existentieprobleem valt uiteen in twee onderdelen: Is het mogelijk de oorsprong vanuit het punt x_0 te bereiken, en zo ja, is er een besturing u die dit in minimale tijd doet? Deze vragen zullen we in de § 3 bespreken. Wil er voor elke begintoestand een besturing bestaan die de toestand naar nul toe brengt, dan moet (A, b) kennelijk bestuurbaar zijn, d.w.z., de vectoren $b, Ab, ..., A^{n-1}b$ moeten onafhankelijk zijn. We zullen verder steeds aannemen dat dit het geval is.

1.4 Veronderstelling *Het paar (A, b) is bestuurbaar.*

Vanwege de beperking op u geeft de bestuurbaarheid van (A, b) nog geen garantie dat de oorsprong vanuit elk punt bereikt kan worden.

In de volgende paragraaf behandelen we eerst het richtingoptimaliseringprobleem, dat veel gemakkelijker is dan het tijdoptimaliseringsprobleem.

2. Richtingoptimalisering

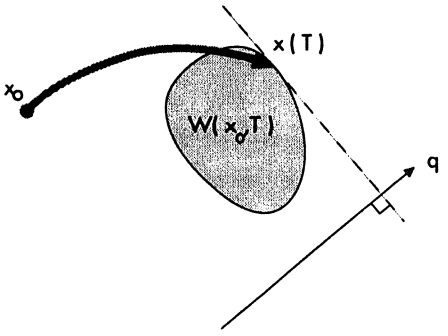
In deze paragraaf bestuderen we het volgende optimaliseringsprobleem:

2.1 Richtingoptimaliseringprobleem *Gegeven een $n \times n$-matrix A, een n-vector b, een n-vector x_0, een n-vector $q \neq 0$ en een getal $T > 0$, bepaal een functie u: $[0, T] \to [-1, 1]$ zo dat voor de oplossing van het stelsel differentiaalvergelijkingen*

(2.1) $\dot{x} = Ax + bu,\ x(0) = x_0$

$(q, x(T))$ wordt gemaximaliseerd.

Hier stelt $(q, x(T))$ het inwendige product van q en $x(T)$ voor. Om dit probleem een meetkundige interpretatie te geven, introduceren we de vanuit het punt x_0 op het tijdstip T **bereikbare verzameling** van (2.1). Deze wordt met $W(x_0,T)$ aangegeven en gedefinieerd als de verzameling



van alle waarden die de toestand op het tijdstip T kan hebben wanneer we alle mogelijke toegelaten besturingen kiezen (d.w.z., stuksgewijs continue functies met waarden in het interval [-1, 1]). Voor elk punt x in $W(x_0,T)$ bestaat er een besturing u zo dat de resulterende baan op het tijdstip T in x terecht komt. Het probleem houdt dus in dat we in $W(x_0,T)$ naar het punt zoeken dat zo ver mogelijk in de richting q ligt. Een optimale besturing, dat wil zeggen een oplossing van Probleem 2.1, zullen we een **q-optimale** besturing noemen.

Dit probleem is enerzijds vrij gemakkelijk en vormt anderzijds een belangrijk hulpmiddel bij het in de vorige paragraaf ingevoerde probleem. Om probleem 2.1 op te lossen introduceren we de aan (2.1) **geadjungeerde vergelijking**:

(2.2) $\dot{p} = -A'p,\ p(T) = q,$

waar A' de getransponeerde matrix van A voorstelt. Als we een oplossing $p(t)$ van (2.2) en een oplossing $x(t)$ van (2.1) hebben kunnen we een belangrijke eigenschap van het inwendig product (p,x) geven. We maken hierbij gebruik van de bekende relatie $(A'x,y) = (x,Ay)$. Er geldt

$$\frac{d}{dt}(p,x) = (\dot{p},x) + (p,\dot{x}) = (-A'p,x) + (p,Ax+bu) = (p,b)u$$

Op grond hiervan kunnen we een eenvoudige formule voor de grootheid vinden die we willen maximaliseren. Immers:

$$(q,x(T)) = (p(T),x(T)) = (p(0),x(0)) + \int_0^T \frac{d}{dt}(p(t),x(t))dt =$$

(2.3)

$$= (p(0),x_0) + \int_0^T (p(t),b)u(t)dt$$

Aangezien $(p(0), x_0)$ onafhankelijk is van de besturing u, zien we dat het maximaliseren van $(q, x(T))$ equivalent is met het maximaliseren van de integraal in het rechterlid. Uit de vorm van deze integraal ziet men gemakkelijk dat we hiervoor de besturing u zo moeten kiezen dat voor elke waarde van $t \in [0, T]$ de uitdrukking $(p(t), b)u(t)$ maximaal is. De waarde van u wordt kennelijk bepaald door de grootheid

(2.4) $\rho(t) := (p(t), b).$

We vinden dat u q-optimaal is dan en slechts dan als

(2.5a) $u(t) = 1$ voor die t waarvoor $\rho(t) > 0$,

(2.5b) $u(t) = -1$ voor die t waarvoor $\rho(t) < 0$.

De waarden van u in de punten waar $\rho(t) = 0$, hebben geen invloed op $(q, x(T))$. We kunnen de formules (2.5) samenvatten met behulp van de **tekenfunctie sgn**, gedefinieerd door

$$\operatorname{sgn}\alpha := \begin{cases} -1 & (\alpha < 0) \\ 0 & (\alpha = 0) \\ 1 & (\alpha > 0) \end{cases}$$

Met de tekenfunctie kunnen we u alsvolgt weergeven:

(2.5) $u(t) = \operatorname{sgn}\rho(t)$, voor $\rho(t) \neq 0$.

Het Maximumprincipe

De karakterisering die leidt tot (2.4), is een van de eenvoudigste voorbeelden van het bekende algemene maximumprincipe van Pontryagin. Dit maximumprincipe geldt voor algemene, niet noodzakelijk lineaire systemen, d.w.z., systemen van de vorm $\dot{x} = f(x,u)$. Het principe wordt geformuleerd m.b.v. de **Hamiltoniaan**,

$$H(p,x,u) := (p, f(x,y)),$$

dus de functie die wordt verkregen door de vectorvariabele p inwendig met het rechterlid van de differentiaalvergelijking te vermenigvuldigen. Voor het geval van vergelijking (2.1) wordt dit $H(p,x,u) := (p, Ax + bu)$. Verder wordt in het algemeen de **geadjungeerde differentiaalvergelijking** alsvolgt ingevoerd:

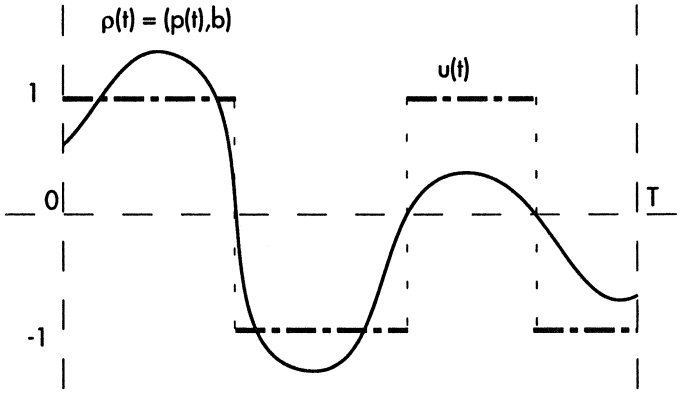
$$\dot{p} = H_x(x,u)p,$$

waar H_x de partiële afgeleide van H naar x voorstelt. Het is duidelijk dat dit overeenstemt met (2.2).

Het maximumprincipe zegt nu dat een optimale besturing $u(t)$ de uitdrukking $H(p(t),x(t),u(t))$ maximaliseert, waar $p(t)$ een geschikte oplossing van de geadjungeerde vergelijking is en $x(t)$ de corresponderende oplossing van de oorspronkelijke differentiaalvergelijking. Merk op dat het maximumprincipe slechts een noodzakelijke voorwaarde voor optimaliteit geeft.

Op deze manier hebben we een expliciete methode om de optimale besturingen te berekenen:

- Bereken de oplossing van de lineaire differentiaalvergelijking $\dot{p} = -A'p$ met eindwaarde $p(T) = q$ en ρ uit (2.4).
- Bereken vervolgens u uit formule (2.5).



De oplossing p van de geadjungeerde vergelijking is uniek bepaald door de beginwaarde. Uit bovenstaande volgt dat ook u overal uniek is, behalve in de punten waar $\rho(t) = 0$. Hiervan kunnen er echter slechts eindig veel zijn. Als $\rho(t)$ immers oneindig veel nulpunten heeft op het interval $[\,0, T]$, is $\rho(t)$ identiek gelijk aan nul, want de functie is analytisch. Dan is ook $\dot{\rho}(t) = \ddot{\rho}(t) = \cdots = 0$. Er geldt dan

$$\dot{\rho}(t) = (\dot{p}(t), b) = -(A'p(t), b) = -(p(t), Ab).$$

Met volledige inductie vinden we zo dat

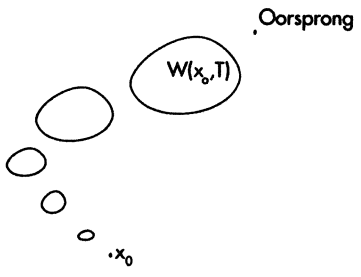
$$\rho^{(k)}(t) = (-1)^{(k)}(p(t), A^k b) = 0$$

voor $k = 0,1,\ldots$ en $0 \leqslant t \leqslant T$. In het bijzonder is $(q, A^k b) = 0$ voor $k = 0,1,\ldots$ De vector q staat dus loodrecht op de vectoren $b, Ab, \ldots, A^{n-1}b$. Aangezien $q \neq 0$, houdt dit in dat de vectoren $b, Ab, \ldots, A^{n-1}b$ afhankelijk zijn, hetgeen in strijd is met de bestuurbaarheid van het paar (A,b). We vinden zo dat u *essentieel* uniek bepaald is, d.w.z. op een eindig aantal punten na, welke, zoals in § 1 opgemerkt, geen invloed hebben op de oplossing van de differentiaalvergelijking. De besturing die zo onstaat wordt in de Engelse literatuur een **Bang-Bang**-besturing genoemd. De tijdstippen waarop de

besturing schakelt van +1 naar -1 of omgekeerd, heten **schakeltijden**. We kunnen constateren dat de optimale besturing volledig bepaald is door de beginwaarde (+1 of - 1) en de schakeltijden.

3. Tijdoptimalisering

Het begrip bereikbare verzameling, ingevoerd in §2, geeft ons de mogelijkheid een meetkundige interpretatie te geven aan het tijdoptimaliseringsprobleem. M.b.v. deze interpretatie kunnen we het tijdoptimaliseringsprobleem herleiden tot een richtingoptimaliseringsprobleem. De minimale tijd waarin vanuit het beginpunt x_0 de oorsprong bereikt kan worden, is de kleinste waarde van T waarvoor $0 \in W(x_0,T)$ geldt. De bereikbare verzameling heeft een aantal algemene eigenschappen die maken dat we deze observatie kunnen gebruiken. Deze worden weergegeven in de volgende stelling:



3.1 Stelling *Er geldt*:

- $W(x_0,T)$ *is een begrensde, gesloten en convexe verzameling voor alle* T *en* x_0.
- *De functie* $T \mapsto W(x_0,T)$ *is continu.*

De tweede eigenschap moet nog nader worden uitgelegd. Daartoe moeten we een afstand definiëren tussen twee begrensde, gesloten en convexe verzamelingen, d.w.z., we moeten aangeven wat we bedoelen als we zeggen dat zulke verzamelingen dicht bij elkaar liggen. Daarna kunnen we de functie $T \mapsto W(x_0,T)$ continu noemen als $W(x_0,T)$ dicht bij $W(x_0,S)$ ligt zodra T dicht bij S ligt. We definiëren eerst de afstand van een punt tot een verzameling als

$$\delta(x, V) := \min \{ |x - y| \mid y \in V \}.$$

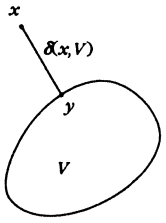
De afstand tussen twee verzamelingen V en W wordt alsvolgt gedefinieerd:

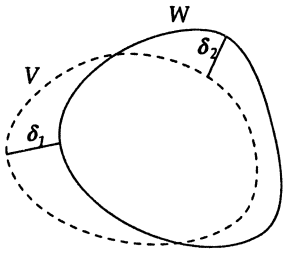
$$d(V, W) := \max \{ d_1(V, W), d_1(W, V) \},$$

waar

$$d_1(V, W) := \max \{ \delta(x, W) \mid x \in V \}.$$

De volgende tekeningen illustreren deze begrippen:



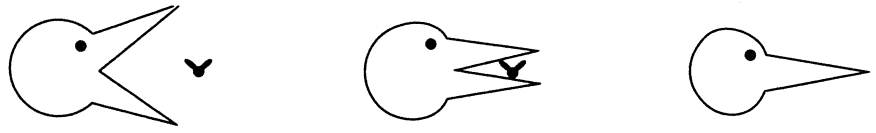


$\delta_1 = d_1(V, W), \delta_2 = d_1(W, V)$,
$d(V,W) = \max\{\delta_1, \delta_2\}$

Stelling 3.1 heeft een belangrijk gevolg:

3.2 Stelling *Voor de kleinste waarde* T^* *van* T *waarvoor* $0 \in W(x_0,T)$ geldt $0 \in \partial W(x_0,T^*)$.

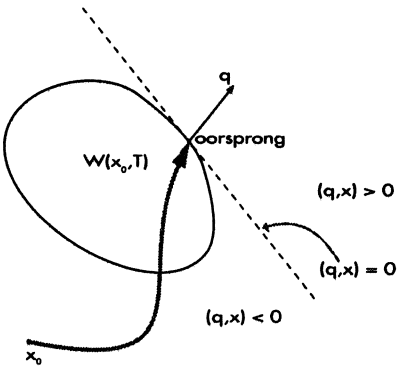
Hierbij stelt $\partial W(x_0,T)$ de rand van $W(x_0,T)$ voor. Deze eigenschap lijkt misschien vanzelfsprekender dan ze is. De convexiteit $W(x_0,T)$ is daarbij bijv. wezenlijk. De volgende tekeningen geven aan dat zonder convexiteit de uitspraak van Stelling 3.2 niet waar is.



Jammer voor het vliegje dat de vogel niet convex is!

Nu we weten dat $0 \in \partial W(x_0,T^*)$, kunnen we een stutvlak vinden aan $W(x_0,T^*)$ in 0. Dit betekent dat er een $q \neq 0$ bestaat zo dat voor alle $x \in W(x_0,T^*)$ geldt $(q,x) \leq (q,0) = 0$. We zien dus dat een besturing die de toestand op het tijdstip T^* in de oorsprong brengt (dat is een tijdoptimale besturing), ook q-optimaal is op het tijdsinterval $[0, T^*]$.

3.3 Stelling *Een besturing u^* is tijdoptimaal met eindtijd T^* dan en slechts dan als er een $q \neq 0$ bestaat zo dat u^* q-optimaal is op $[0, T^*]$, terwijl voor de bijbehorende toestand $x^*(T^*) = 0$ geldt. Verder is de tijdoptimale besturing (essentieel) uniek.*



Bewijs Het "slechts dan" gedeelte volgt uit het voorafgaande. Om te bewijzen dat de voorwaarde ook voldoende is veronderstellen we dat u een willekeurige besturing is en u^* een q-optimale besturing die voldoet aan $x^*(T^*) = 0$. We introduceren de functies $\omega^*(t) := (p(t), x^*(t))$ en $\omega(t) := (p(t), x(t))$. We weten uit de vorige paragraaf dat $\dot{\omega}^*(t) = (p(t), b)u^*(t) = |(p(t), b)|$ en $\dot{\omega}(t) = (p(t), b)u(t) \leqslant \dot{\omega}^*(t)$. Hieruit volgt in de allereerste plaats dat $\omega^*(t)$ strikt stijgend is, want $\dot{\omega}^*(t)$ is ten hoogste in eindig veel punten nul en elders positief. Verder is $\omega^*(T^*) = 0$. Kennelijk is $\omega^*(t) < 0$ voor $t < T^*$. Tenslotte is $\omega^*(0) = \omega(0) = (p(t), x_0)$, en omdat $\dot{\omega}(t) \leqslant \dot{\omega}^*(t)$ vinden we $\omega(t) \leqslant \omega^*(t) < 0$ voor $t < T^*$. Maar dit betekent dat $x(t)$ niet gelijk aan nul kan zijn voor $t < T^*$. De baan die correspondeert met de besturing u kan de oorsprong dus niet bereiken op een tijdstip $T < T^*$. De besturing u^* is tijdoptimaal. We moeten nog laten zien dat de tijdoptimale besturing uniek is. Het is duidelijk dat $\omega(T^*) < 0$ is tenzij $\dot{\omega}(t) = \dot{\omega}^*(t)$ geldt voor $0 \leqslant t \leqslant T$. In dat geval moet ook u q-optimaal zijn. We vinden op grond van de vorige paragraaf dat $u(t) = u^*(t)$ in alle continuïteitspunten van u. □

Bovenstaand resultaat houdt in dat we de resultaten van de vorige paragraaf kunnen toepassen. Dit levert ons echter nog niet direct een expliciete procedure om een tijdoptimale besturing te vinden, omdat q niet gegeven is. Ook moeten we bedenken dat T^* niet van te voren bekend is. De vector q en het getal T^* moeten bepaald worden uit de voorwaarde dat de toestand op het tijdstip T^* de waarde 0 aanneemt. Deze voorwaarde geeft ons in principe voldoende informatie, hoewel we $n+1$ onbekenden en n vergelijkingen hebben. Uit het voorafgaande volgt immers gemakkelijk dat de lengte van q onbelangrijk is. We mogen, als dat handig is, bijv. de lengte van q gelijk aan 1 veronderstellen.

Tenslotte gaan we in op vraag naar de existentie van optimale besturingen voor elke beginwaarde x_0. Zoals reeds in §1 opgemerkt, valt deze vraag

uiteen in twee delen:

1. Kan de oorsprong vanuit elk punt worden bereikt?
2. Als de oorsprong kan worden bereikt, is er dan een besturing u die dat in minimale tijd doet?

In verband met de eerste vraag hebben we al geëist dat (A,b) bestuurbaar is, maar zoals al opgemerkt, zal vanwege de beperking op u dit niet hoeven te impliceren dat het antwoord altijd 'ja' is. Het volgende is hiervan een voorbeeld:

3.4 Voorbeeld Het ééndimensionaal systeem $\dot{x} = x + u$ is bestuurbaar. Als we echter $x_0 > 1$ kiezen, dan zal voor elke besturing u die aan $|u| \leqslant 1$ voldoet, $\dot{x} > 0$ moeten gelden. Het is dan onmogelijk om de oorsprong te bereiken.

Het is duidelijk dat de instabiliteit van het systeem de oorzaak van het probleem is. Ook zien we dat de oorsprong wel bereikt kan worden als de absolute waarde van de begintoestand maar klein genoeg is. We kunnen in verband hiermee het begrip lokale bestuurbaarheid invoeren. We noemen het systeem (1.2) **lokaal nulbestuurbaar** als er een $r > 0$ bestaat zo dat voor iedere x_0 die voldoet aan $|x_0| < r$, er een besturing u en een tijdstip T bestaat zo dat de resulterende toestand op het tijdstip T de waarde 0 aanneemt. Men kan dan bewijzen dat (1.2) lokaal nulbestuurbaar is dan en slechts dan als (A, b) bestuurbaar is. Het is daarom niet moeilijk aan te tonen dat voor een asymptotisch stabiel systeem de bestuurbaarheid van (A,b) voldoende is voor een positief antwoord op de eerste vraag. In feite is een zwakkere eis voldoende:

3.5 Stelling *Neem aan dat (A,b) bestuurbaar is en dat elke eigenwaarde λ van A voldoet aan* Re $\lambda \leqslant 0$. *Dan is er voor elke x_0 een besturing u en een getal $T > 0$ zo dat de oplossing van (1.2) op het tijdstip T de waarde 0 heeft.*

Het bewijs laten we achterwege.

De tweede vraag over de existentie wordt altijd positief beantwoord vanwege Stelling 3.1.

De locatie van de eigenwaarden is ook van belang voor het aantal schakel-

punten van een optimale besturing. Als de matrix A niet-reële eigenwaarden heeft, zal de functie $\rho(t)$ een (co)sinus bevatten, en daarom veel nulpunten kunnen hebben. Als A echter alleen reële nulpunten heeft, kan men bewijzen dat $\rho(t)$ ten hoogste n - 1 nulpunten kan hebben. In dat geval zijn er voor de optimale besturing ook ten hoogste n - 1 schakelpunten. Als we deze punten met $t_1, \ldots, t_{n-1}$ aangeven en T met t_n dan moeten we, om het optimaliseringsprobleem op te lossen, de n variabelen $t_1, \ldots, t_n$ bepalen uit de n vergelijkingen $x(T) = 0$.

4 Voorbeelden

We passen de theorie van de vorige paragraaf toe op Voorbeeld 1.1. De bewegingsvergelijking kan alsvolgt geschreven worden:

$$\dot{x} = y,$$
$$\dot{y} = u.$$

De theorie uit §3 zegt dat er een niet-triviale oplossing p van de geadjungeerde vergelijking moet zijn zo dat de optimale besturing, zo die bestaat, gegeven wordt door $u = \operatorname{sgn} \rho$, waar $\rho = (p, b)$. De geadjungeerde vergelijking $\dot{p} = -A'p$ kan hier geschreven worden als

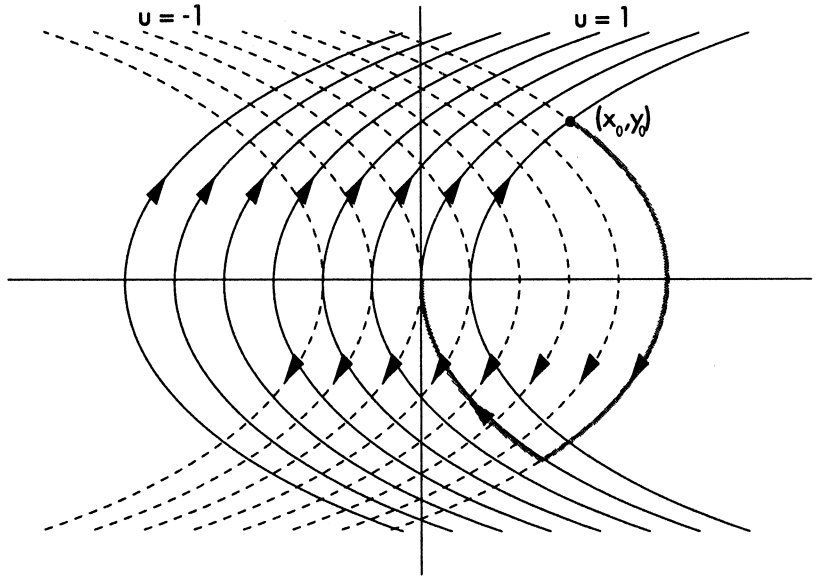
$$\dot{p}_1 = 0,$$
$$\dot{p}_2 = p_1,$$

waar p_1 en p_2 de componenten van de vector p zijn. Hieruit volgt dat p_1 constant en p_2 een lineaire functie is, zeg, $p_2 = \alpha t + \beta$. Verder is $\rho = p_2$. We concluderen dat er voor een optimale besturing u getallen α en β bestaan, die niet allebei nul zijn, zo dat $u = \operatorname{sgn}(\alpha t + \beta)$. (Als α en β beide gelijk nul waren dan hadden we de nuloplossing van de geadjungeerde vergelijking en dit zou inhouden dat $q = 0$, hetgeen we uitgesloten hebben.) Dit is equivalent met de uitspraak dat u alleen de waarden -1 en +1 aanneemt, en dat er ten hoogste één schakelpunt is. Deze informatie is alles wat we van de theorie krijgen, en het is genoeg om de optimale besturing te vinden voor elke beginpositie $(x_0, \dot{x}_0)$. We schetsen daartoe de banen (d.w.z., oplossingen van de differentiaalvergelijking voor x en y). Eerst nemen we de stukken van de oplossing die corresponderen met $u = 1$. De vergelijking voor (x,y) luidt daarvoor:

$$\dot{x} = y,\ \dot{y} = 1.$$

Door de twee vergelijkingen op elkaar te delen krijgen we $dy/dx = 1/y$, van welke de oplossingen voldoen aan $y^2 = x + C$. De oplossingskrommen in het fasevlak zijn dus parabolen met de x-as als as en top naar links. De parabool die correspondeert met $C = 0$, is de enige die door de oorsprong gaat. Op soortgelijke manier vinden we voor de oplossingen die corresponderen met $u = -1$, de familie $y^2 = -x - C$, die bestaat uit parabolen die ook de x-as als as hebben maar die de top naar rechts hebben.

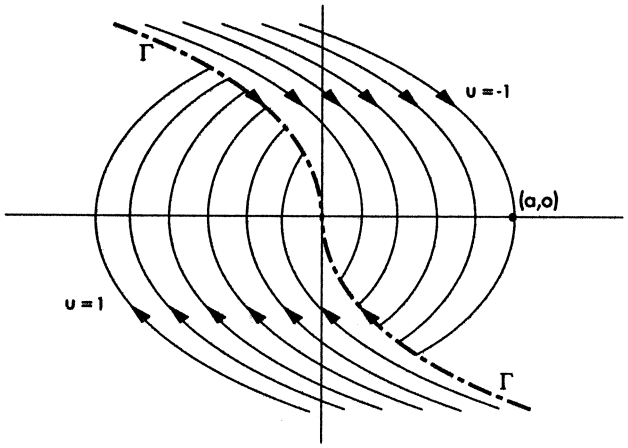
De optimale baan in het fasevlak bestaat uit twee stukken parabool, één corresponderend met $u = 1$, en één met $u = -1$. Het beginpunt van deze gecombineerde baan ligt in het gegeven punt (x_0, y_0), het eindpunt ligt in de oorsprong. Het is gemakkelijk in te zien dat dit maar op een manier kan (zie de grijze lijn in de tekening).



Op deze manier hebben we een expliciete representatie van de oplossing van het optimaliseringsprobleem. De parabolen die door de oorsprong gaan, spelen hierbij een bijzondere rol: Het laatste stuk van een optimale baan ligt altijd op één van deze twee parabolen, in feite op de helft ervan die in het tweede of vierde kwadrant ligt. De halve parabolen vormen tezamen een kromme Γ, die we de **schakelkromme** noemen. Deze heeft

de eigenschap dat in de punten erboven de optimale besturing de waarde -1 en eronder de waarde +1 heeft. De volgende tekening geeft de oplossing grafisch weer.

De oplossing is gemakkelijk te interpreteren. Neem bijv. aan dat het wagentje aanvankelijk in het punt $x = a > 0$ stil staat. Dan moeten we beginnen de wagen zo hard mogelijk naar links te duwen tot het moment dat we door zo hard mogelijk tegen te houden nog net kunnen bereiken dat de wagen op de positie $x = 0$ tot stilstand komt.



We bekijken nu voorbeeld 1.2, de slinger.

De bewegingsvergelijking ziet er nu alsvolgt uit:

$$\dot{x} = y,$$
$$\dot{y} = -x + u.$$

We zien dat nu

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \text{ en } b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

De theorie uit §3 zegt dat er een niet-triviale oplossing p van de geadjungeerde vergelijking moet zijn zo dat de optimale besturing gegeven wordt door $u = \text{sgn } \rho$, waar $\rho = (p, b)$. De geadjungeerde vergelijking $\dot{p} = -A'p$ kan hier geschreven worden als

$$\dot{p}_1 = -p_2,$$
$$\dot{p}_2 = p_1,$$

waar p_1 en p_2 de componenten van de vector p zijn. Omdat $\rho = p_2$, zijn we vooral in deze variabele geïnteresseerd. Deze voldoet aan de differentiaalvergelijking $\ddot{p}_2 + p_2 = 0$. De algemene oplossing hiervan is $p_2 = r \cos(t+\vartheta)$, waar r en ϑ constanten zijn. We willen de tekenwisselingen, en dus de nulpunten van p_2 weten. We zien dat er tussen twee opeenvolgende waarden steeds een afstand π is, terwijl het eerste en laatste nulpunt op een afstand niet groter dan π van de rand liggen. We tekenen weer de banen behorende bij de differentiaalvergelijkingen met $u = 1$ en $u = -1$. Het stelsel differentiaalvergelijkingen voor $u = 1$ luidt:

$$\dot{x} = y, \quad \dot{y} = -x + 1. \tag{4.1}$$

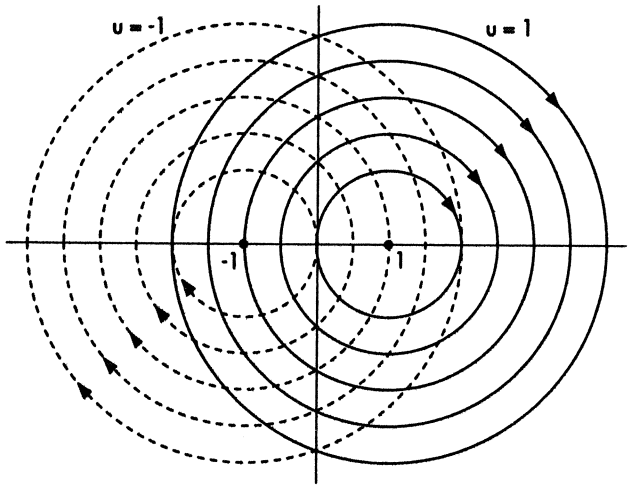
Als we deze vergelijkingen op elkaar delen vinden we

$$ydy/dx = -x + 1,$$

welke we kunnen oplossen d.m.v. separatie van variabelen. We vinden dat de vergelijking $y^2 + (x - 1)^2 = C$ geldt voor de oplossingen van (4.1). Op soortgelijke manier vinden we $y^2 + (x + 1)^2 = C$ voor de oplossingen die corresponderen met $u = -1$. De oplossingskrommen in het fasevlak zijn hier dus cirkels met middelpunt (1, 0) resp. (-1, 0). De cirkels die corresponderen met $C = 1$, zijn de enige die door de oorsprong gaan. (Zie de figuur op de volgende pagina.)

De optimale baan bestaat uit afwisselend een stuk van de ene en de andere familie van cirkels. Tussen twee keer schakelen ligt steeds een tijdsinterval π. Het is niet moeilijk in te zien dat in zo'n interval precies een halve cirkel doorlopen wordt. Van de vergelijking (4.1) bijvoorbeeld is de constante functie $x = 1$, $y = 0$ een oplossing. Verder wordt de algemene oplossing van de corresponderende homogene vergelijking gegeven door $x = r \cos(t + \vartheta)$, $y = r \sin(t + \vartheta)$. De algemene oplossing van (4.1) is dus

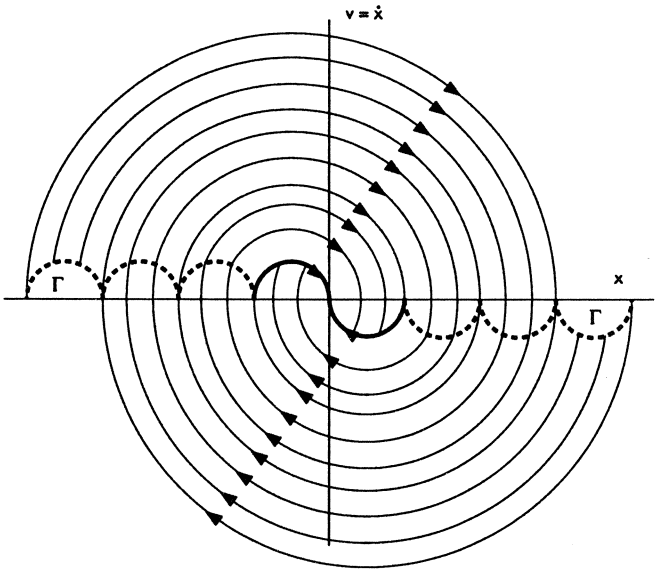
$x = r \cos(t + \vartheta) + 1,$
$y = r \sin(t + \vartheta).$



De oplossingen zijn periodiek met periode 2π, in welk tijdsinterval een cirkel doorlopen wordt. Een halve cirkel heeft dus een tijdsduur π nodig.

De stukken waaruit een optimale baan bestaat zijn dus halve cirkels met uitzondering van het eerste en het laatste stuk, die minder kunnen zijn. We zien dat we de oorsprong zonder sprong kunnen bereiken vanuit de halve cirkels die eindigen in de oorsprong. De verzameling van die punten noemen we Σ_1. De verzameling van de punten van waaruit men met een cirkelboog die niet groter is dan een halve cirkel op Σ_1 terecht kan komen geven we aan met Σ_2. Vanuit deze punten wordt de oorsprong met één sprong bereikt. Zo kan men successievelijk het hele platte vlak opvullen met verzamelingen Σ_i, van waaruit men de oorsprong met i schakelingen kan bereiken. Een baan die start in Σ_n doorloopt achtereen volgens de verzamelingen Σ_{n-1}, Σ_{n-2}, . . . , Σ_1. Schakelingen vinden plaats op de schakelkromme Γ, die in dit geval bestaat uit een vereniging van halve cirkels. Weer zien we dat de optimale besturing boven Γ de waarde -1 en onder Γ de waarde +1 aanneemt.

Onderstaande figuur geeft hiervan een overzicht.



Literatuur

G. Boltjanski — *Mathematische Methoden der optimalen Steuerung*, Academisch Verlagsgeselschaft, Leipzig, 1971
Een uitgebreide behandeling van de problemen die hier besproken zijn, met allerlei voorbeelden, wordt gegeven in "Kapitel 2"

H.Hermes & J.P. LaSalle — *Functional Analysis and Time-Optimal Control*,
Academic Press, New York, 1966
Hier wordt in "Part II" vooral de theorie en generalisaties van de hier behandelde stof besproken.

E.B. Lee & L. Markus — *Foundations of Optimal Control Theory*,
Wiley, New York, 1968
In §1.1 wordt de stof behandeld op een manier die lijkt op wat er in deze syllabus gebeurt. Een uitgebreidere behandeling wordt gegeven in "Chapter II"

Dit document is geproduceerd op een Archimedes m.b.v. van de programma's Impression II, Draw + en BestForm

Kalman Filtering

A.W. Heemink

1 Inleiding

Van tal van verschijnselen is het van belang nauwkeurige voorspellingen te kunnen berekenen. Denk hierbij bijvoorbeeld aan het weer of de waterstanden langs de Nederlandse kust of de baan van een sateliet. Hierbij wordt veelvuldig gebruik gemaakt van mathematisch fysische modellen. Deze modellen bestaan uit differentiaalvergelijkingen die bepaalde wetten beschrijven waar de verschijnselen bij benadering aan voldoen. Bij dergelijke modellen wordt de beschikbare fysische kennis van het verschijnsel gebruikt om het zo goed mogelijk te voorspellen.

Een andere methode om de voorspellingen te berekenen kan bereikt worden met behulp van technieken uit de tijdreeksanalyse. Hierbij wordt getracht de systematische verbanden die mogelijk binnen één meetreeks of tussen meerdere waargenomen reeksen bestaan, op te sporen en te modelleren. Dit resulteert in een 'black-box' model, dat wil zeggen een model waarbij de relaties tussen de verschillende reeksen op statistische wijze zijn bepaald uit de beschikbare meetinformatie.

De eenvoudige black-box modellen hebben ten opzichte van mathematisch fysische modellen een aantal voordelen. Zo kan bijvoorbeeld 'on-line' meetinformatie onmiddelijk worden gebruikt om de modelvoorstellingen en eventueel ook het black-box model zelf, voortdurend aan te passen aan de zich wijzigende omstandigheden. Dit maakt dergelijke modellen bij uitstek geschikt voor het berekenen van 'on-line' voorspellingen. Bij gebruikmaking van de meeste bestaande mathematisch fysische modellen voor het berekenen van voorspellingen kan deze extra meetinformatie niet eenvoudig, of slechts ten dele, worden benut.

Black-box modellen hebben echter niet alleen voordelen. Het grote nadeel van het toepassen van deze modellen is dat er geen gebruik kan worden gemaakt van de differentiaalvergelijkingen die de verschijnselen beschrijven. Dit heeft tot gevolg dat voor de meeste toepassingen mathematisch fysische modellen nauwkeuriger zijn en een veel gedetaileerder beeld geven van het verschijnsel.

Het bovenstaande geeft aan dat er behoefte is aan een methode die de voordelen van mathematisch fysische modellen en black-box modellen combineert. Een dergelijke aanpak

is realiseerbaar met behulp van Kalman filtering.

De theorie van Kalman filtering werd voor het eerst gepubliceerd door R.E. Kalman in 1960. Hoewel het artikel zeer theoretisch was en derhalve moeilijk leesbaar voor praktische ingenieurs, duurde het niet lang voordat duidelijk was dat Kalman filtering van groot praktisch belang zou worden. De publikatie van deze techniek was een doorbraak in de zin dat het vanaf dat moment mogelijk werd om wiskundige modellen, bestaande uit differentiaalvergelijkingen op wiskundig zeer elegante wijze te kunnen combineren met beschikbare meetinformatie. Op deze wijze is het mogelijk tot een zo goed mogelijke beschrijving van het verschijnsel te komen. Met behulp van Kalman filtering is het mogelijk om, gebruik makend van de metingen, de modelresultaten, en eventueel ook onzekere parameters in het model zelf, voortdurend te verbeteren en aan te passen aan de zich wijzigende omstandigheden.

De eerste belangrijke toepassing liet niet lang op zich wachten. Al in de begin jaren zestig werden Kalman filters ontwikkeld voor de navigatie van ruimteschepen. Bij het vaststellen van de positie van de ruimteschip kwamen namelijk waarnemingen beschikbaar uit verschillende bronnen zoals radar, gyroscopen en visuele waarnemingen, ieder met een bepaalde nauwkeurigheid. Naast meetinformatie was er ook kennis beschikbaar omtrent het dynamische gedrag van het ruimteschip, meestal in de vorm van een stelsel differentiaalvergelijkingen. Kalman filtering bleek bij uitstek geschikt om alle beschikbare waarnemingen op de beste manier te combineren met de informatie die het oplossen van de bewegingsvergelijkingen omtrent de positie van ruimteschip gaf. Het gebruik van Kalman filters ten behoeve van navigatie resulteerde in zeer nauwkeurige positiebepalingssystemen. Het gebruik van deze systemen is vooral van belang geweest bij het navigeren van de Apollo-ruimteschepen bij de terugkeer in de dampkring, dat zeer zorgvuldig diende te geschieden.

Na de bovenstaande, succesvolle toepassing van Kalman filtering bleek dat deze techniek ook bij heel andere problemen nuttig gebruikt kon worden, zoals bijvoorbeeld het voorspellen van de verspreiding van luchtvervuiling of bij het voorspellen van waterstanden. Deze toepassingen werden mede mogelijk gemaakt door de steeds groter wordende rekencapaciteit van computers in combinatie met de ontwikkeling van steeds efficiëntere algoritmen voor het oplossen van de Kalman filtervergelijkingen.

Om het principe van Kalman filtering te illustreren is in figuur 1.1, zeer schematisch de

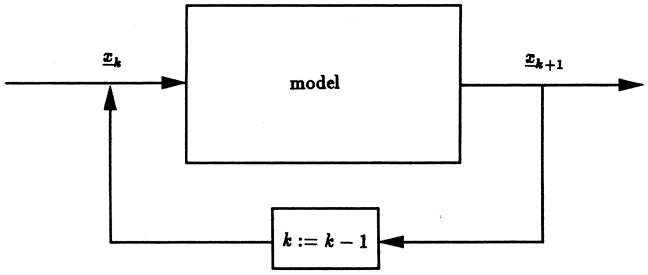


Figure 1.1: Mathematisch fysisch model.

werking van een mathematisch fysisch model weergegeven. Uitgaande van een toestand $\underline{x}_k$, die het systeem op tijdstip k volledig beschrijft, en de invoer gegevens kan het verschijnsel worden gesimuleerd tot, bijvoorbeeld, tijdstip $k+1$ waarna alle resultaten $\underline{x}_{k+1}$ worden uitgevoerd en weggeschreven. Een volgende periode kan op analoge wijze worden gesimuleerd door deze resultaten als begincondities weer in te voeren.

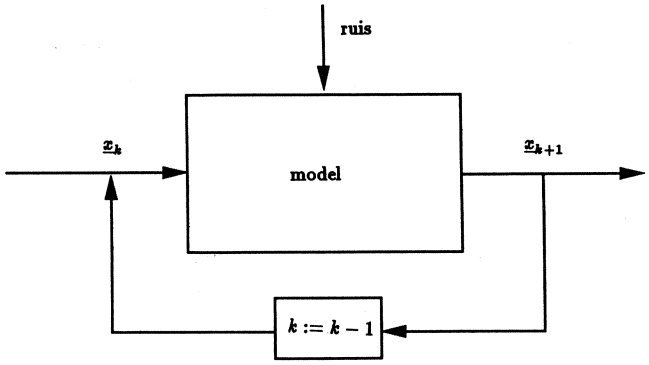


Figure 1.2: Stochastisch mathematisch fysisch model.

Het combineren van het model met een Kalman filter geschiedt in twee stappen. Eerst worden de nauwkeurigheid van de invoer en van het model beschreven door er stochastische stoortermen of ruisprocessen aan toe te voegen. Dit is schematisch weergegeven in figuur 1.2. Na het op deze wijze 'inbedden' van het mathematisch fysisch model in een 'stochastische omgeving', kan in figuur 1.3 het Kalman filter worden geïntroduceerd.

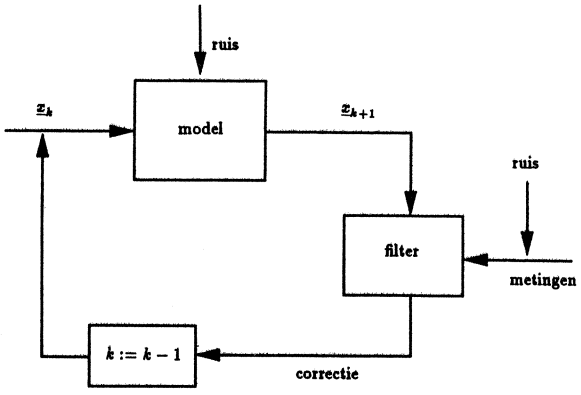


Figure 1.3: Kalman filter.

Zoals in dit figuur is weergegeven bepaalt dit filter, uitgaande van de modelresultaten en de beschikbare meetinformatie, een zo goed mogelijke correctie van de modelresultaten, de inputgegevens en eventueel enkele onzekere parameters in het model. Dit laatste is van belang omdat bij het toepassen van mathematisch fysische modellen een aantal modelparameters meestal niet nauwkeurig bekend zijn. Hierdoor moet ieder model worden afgeregeld: de onzekere coefficienten worden zo ingesteld dat de modelresultaten zo goed mogelijk overeenkomen met de beschikbare meetgegevens. Een nadeel van het door middel van 'trial en error' met de hand afregelen van een model is dat een dergelijke aanpak meestal niet leidt tot het 'best mogelijke' resultaat en bovendien subjectief is: verschillende modelbouwers komen tot verschillende keuzen voor de modelparameters. Een nog groter nadeel echter is het tijdrovende karakter van deze aanpak. Modelresultaten moeten na iedere simulatie uitvoering worden geanalyseerd om te onderzoeken of, en zo ja hoe, de waarden van de onzekere modelparameters verbeterd kunnen worden. Dit vereist veel kennis van de specifieke eigenschappen van het af te regelen model. Dit nadeel wordt nog versterkt door het feit dat de modelparameters vaak niet constant zijn en regelmatig met

behulp van recente meetgegevens moeten worden aangepast. Het Kalman filter kan nu ook worden gebruikt om deze parameters voortdurend aan te passen aan de zich wijzigende omstandigheden.

Voorbeeld 1.1

Stel we beschikken over model informatie x_m van een scalaire grootheid x in de vorm:

$$x_m = x + w \tag{1.1}$$

met w als een onbekende modelfout met statistiek:

$$\begin{array}{rcl} E\{w\} & = & 0 \\ E\{w^2\} & = & P \end{array} \tag{1.2}$$

Voorts is er een meting z verricht:

$$z = x + v \tag{1.3}$$

met v als een onbekende meetfout met statistiek

$$\begin{array}{rcl} E\{v\} & = & 0 \\ E\{v^2\} & = & R \end{array} \tag{1.4}$$

We willen nu de toestand x zo goed mogelijk reconstrueren met behulp van de modelinformatie en de meetinformatie. We kiezen als schatting $\hat{x}$ voor x een lineaire combinatie van modelresultaat en meting

$$\hat{x} = (1-k)x_m + kz \tag{1.5}$$

waarbij $0 \leq k \leq 1$ een weegfactor is.
Er geldt:

$$\begin{array}{rcl} E\{\hat{x}\} & = & (1-k)E\{x_m\} + kE\{z\} = (1-k)E\{x\} + kE\{x\} = E\{x\} \\ var\{\hat{x}\} & = & E\{(\hat{x} - E\{x\})^2\} = (1-k)^2P + k^2R \end{array} \tag{1.6}$$

We zien dat voor $k = 0$ er geen gewicht aan de meting wordt toegekend en dat de variantie van de schatting $\hat{x}$ reduceert tot de variantie van x_m. Bij $k = 1$ geldt het omgekeerde en wordt juist de modelinformatie niet meegenomen en reduceert de variantie van $\hat{x}$ tot de variantie van de meting z. In het algemeen zal het voordelen bieden beide bronnen van informatie te gebruiken en $k > 0$ en $k < 1$ te kiezen. De variantie van $\hat{x}$ zal dan meestal kleiner zijn dan bij $k = 0$ of $k = 1$ (zie figuur 1.4). een optimaal resultaat kan verkregen worden door k zo te kiezen dat de variantie van $\hat{x}$ minimaal is:

$$\frac{d}{dk} var\{\hat{x}\} = -2(1-k)P + 2kR = 0 \tag{1.7}$$

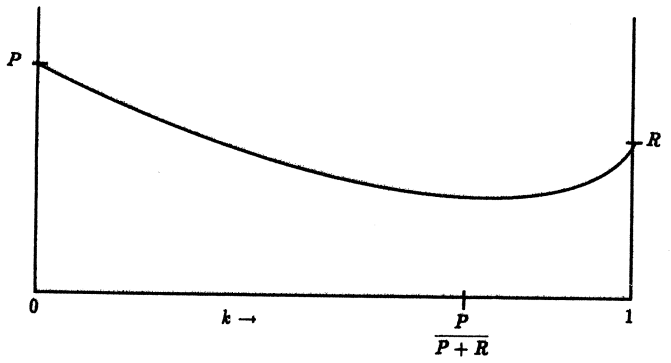


Figure 1.4: Variantie van de schatting.

zodat

$$k = \frac{P}{P+R} \tag{1.8}$$

en

$$var\{\hat{x}\} = \frac{PR}{P+R} \tag{1.9}$$

De schatting $\hat{x}$ wordt in dit geval minimum variantie schatting van x genoemd.

□

De filtertheorie houdt zich in het algemeen bezig met het onderdrukken of tegenhouden van toevaligheden in een systeembeschrijving. Meetinformatie wordt benut om modelfouten op te sporen en te elimineren, het model zelf wordt gebruikt om toevallige uitschieters in de metingen te onderdrukken. Het bekendste resultaat is het Kalman filter voor lineaire systemen. Het begrip filteren kan men onderscheiden in

1. Filteren.
   Hierbij willen we, gegeven alle beschikbare metingen tot en met tijdstip k de gehele toestand op datzelfde tijdstip k reconstrueren.

2. Voorspellen.
   Hierbij willen we, gebaseerd op de metingen die tot en met tijdstip k beschikbaar zijn, de toestand op een later tijdstip $l > k$ voorspellen.

3. Vereffenen ("smoothing").
   Hierbij willen we uitgaande van een serie metingen die tot en met tijdstip k verricht zijn, een toestand op een tijdstip $l < k$ reconstrueren.

In dit dictaat ligt de nadruk op het filteren en voorspellen omdat deze begrippen voor de meeste toepassingen belangrijker zijn dan verefferen. Voorts beschouwen we hier discrete tijd systemen. Continue tijd systemen zijn veel lastiger te analyseren en bovendien komen dit soort systemen in de praktijk nauwelijks voor. Immers het uiteindelijk filter zal meestal worden geimplementeerd op een digitale computer en zal daarom vroeg of laat gediscretiseerd moeten worden.

2 Het discrete tijd model

De klasse van lineaire stochastische discrete tijdsystemen die we in dit dictaat bestuderen heeft de vorm:

$$\underline{x}_{k+1} = F_k \underline{x}_k + G_k \underline{w}_k \tag{2.1}$$

$$\underline{z}_k = M_k \underline{x}_k + \underline{v}_k \tag{2.2}$$

met:

k	=	0,1,2, ...
$\underline{x}_k$	=	toestandsvektor op tijdstip k
$\underline{z}_k$	=	de meetvektor op tijdstip k
$\underline{w}_k$	=	de systeemruis op tijdstip k
$\underline{v}_k$	=	de meetruis op tijdstip k
F_k	=	systeemmatrix
G_k	=	ruis input matrix
M_k	=	meetmatrix

De processen $\underline{w}_k$ en $\underline{v}_k$ zijn Gaussisch en onderling onafhankelijk met statistiek:

$$\begin{aligned} E\{\underline{w}_k\} &= 0 \\ E\{\underline{w}_k \underline{w}_l^T\} &= Q_k, \quad k = l \\ &= 0, \quad k \neq l \end{aligned} \tag{2.3}$$

$$\begin{aligned} E\{\underline{v}_k\} &= 0 \\ E\{\underline{v}_k \underline{v}_l^T\} &= R_k, \quad k = l \\ &= 0, \quad k \neq l \end{aligned} \tag{2.4}$$

Hierbij zijn Q_k en R_k respectievelijk positief semidefiniet en positief definiet.

De beginvektor $\underline{x}_0$ is ook Gaussisch en onafhankelijk van de ruisprocessen $\underline{w}_k$ en $\underline{v}_k$. De statistiek van $\underline{x}_0$ is

$$\begin{aligned} E\{\underline{x}_0\} &= \hat{\underline{x}}_0 \\ E\{[\underline{x}_0 - \hat{\underline{x}}_0][x_0 - \hat{x}_0]^T\} &= P_0 \end{aligned} \tag{2.5}$$

waarbij P_0 positief definiet is.

De tijdstap tussen twee opeenvolgende metingen hoeft niet constant te zijn.

Het bovenbeschreven model is niet het meest algemene model dat bestudeerd kan worden. Het is mogelijk dat zowel P_0 als R_k positief semidefiniet is. Voorts mogen de systeem- en meetruis gecorreleerd zijn of een gemiddelde ongelijk 0 hebben. In dit dictaat worden deze

generalisaties echter buiten beschouwing gelaten.

Herschrijving van het model (2.1) levert:

$$\begin{aligned}
\underline{x}_1 &= F_0\underline{x}_0 + G_0\underline{w}_0 \\
\underline{x}_2 &= F_1x_1 + G_1\underline{w}_1 \\
&= F_1F_0\underline{x}_0 + F_1G_0\underline{w}_0 + G_1\underline{w}_1 \\
\underline{x}_3 &= F_2\underline{x}_2 + G_2\underline{w}_2 \\
&= F_2F_1F_0\underline{x}_0 + F_2F_1G_0\underline{w}_0 + F_2G_1\underline{w}_1
\end{aligned}$$

of in meer algemene vorm

$$\underline{x}_{k+1} = \Phi_{k+1,0}\underline{x}_0 + \sum_{l=0}^{k} \Phi_{k,l+1}G_l\underline{w}_l \tag{2.6}$$

waarbij $\Phi_{k,l}$ de toestandsovergangsmatrix is:

$$\begin{aligned}
\Phi_{k,l} &= F_{k-1}F_{k-2}\cdots F_l \\
\Phi_{k,k} &= I
\end{aligned} \tag{2.7}$$

Uit (2.6) volgt onmiddelijk dat $\underline{x}_k$ een lineaire combinatie is van de Gaussische vektoren $\underline{x}_0$ en $\underline{w}_l, l = 0, 1, \cdots, k$, en daarmee zelf ook een Gaussisch proces is . Hetzelfde geldt voor $\underline{z}_k$ als lineaire combinatie van de Gaussische vektoren $\underline{x}_k$ en $\underline{v}_k$. Dit is een zeer belangrijke eigenschap van het model (2.1) - (2.2) omdat de kansdichtheid van een Gaussisch proces zich volledig laat beschrijven door de verwachtingswaarde en de kovariantiematrix van dit proces.

Voorbeeld 2.1

Het model (2.1) is zeer algemeen. Zo is bijvoorbeeld een AR(1)-model:

$$x_{k+1} = ax_k + w_k \tag{2.8}$$

van de vorm (2.1). Het AR(2) model:

$$x_{k+1} = ax_k + bx_{k-1} + w_k \tag{2.9}$$

schijnbaar niet. Als we echter de nieuwe toestand $\underline{y}_k$ definiëren als:

$$\underline{y}_k = [x_k \;\; x_{k-1}]^T \tag{2.10}$$

dan kunnen we het AR(2) model als volgt herschrijven:

$$\underline{y}_{k+1} = \begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix} \underline{y}_k + \begin{bmatrix} 1 \\ 0 \end{bmatrix} w_k \tag{2.11}$$

Deze vergelijking is toch weer van de vorm (2.1).

3 Het discrete Kalman filter

We willen nu alle beschikbare metingen $\underline{z}_k$, gemodelleerd via vergelijking (2.2), combineren met de informatie die via het model (2.1) beschikbaar is, om een optimale schatting van de toestand $\underline{x}_k$ te kunnen berekenen. Bij het oplossen van dit filterprobleem zijn we daarom geinteresseerd in de kansdichtheid:

$$P(\underline{x}_k|\underline{z}_1,\cdots,\underline{z}_k) \tag{3.1}$$

Als deze kansdichtheid expliciet beschreven is, dan kunnen we een optimale schatting $\hat{\underline{x}}_k$ van de toestand $\underline{x}_k$ definiëren. Bijvoorbeeld de maximum a posteriori schatting:

$$P(\hat{\underline{x}}_k|\underline{z}_1,\cdots,\underline{z}_k) \text{ is maximaal} \tag{3.2}$$

of de voorwaardelijke gemiddelde schatting:

$$\hat{\underline{x}}_k = E\{\underline{x}_k\} \tag{3.3}$$

of de minimum variantie of kleinste kwadraten schatting:

$$E\{\|\underline{x}_k - \hat{\underline{x}}_k\|^2\} \text{ is minimaal} \tag{3.4}$$

Als de kansdichtheid $P(\underline{x}_k|\underline{z}_1,\cdots,\underline{z}_k)$ Gaussisch is, zijn alle bovenbeschreven schattingen identiek. In het algemeen kan bewezen worden dat de minimum variantie schatting gelijk is aan de voorwaardelijke gemiddelde schatting.

Voorbeeld 3.1

Stel we beschikken, analoog aan voorbeeld 1.1, over model informatie x_m van een scalaire grootheid x in de vorm:

$$x_m = x + w \tag{3.5}$$

met w als een onbekende modelfout met kansdichtheid

$$w \sim N(0, P) \tag{3.6}$$

zodat:

$$x \sim N(x_m, P) \,. \tag{3.7}$$

Voorts is er een meting z verricht:

$$z = x + v \tag{3.8}$$

met v als de meetruis met statistiek

$$v \sim N(0, R) \tag{3.9}$$

We willen nu net als bijvoorbeeld 1.1 de model kennis over x combineren met de meetinformatie z om de toestand x zo goed mogelijk te kunnen reconstrueren. Hiertoe gaan we eerst de kansdichtheid van x gegeven z berekenen:

$$\begin{aligned} P_{x|z}(x|z) &= \frac{P_{z|x}(z|x)P_x(x)}{P_z(z)} \\ &= \frac{P_{z|x}(z|x)P_x(x)}{\int_{-\infty}^{\infty} P_{z|x}(z|x)P_x(x)dx} \\ &= \frac{P_{x+v|x}(x+v|x)P_x(x)}{\int_{-\infty}^{\infty} P_{x+v|x}(x+v|x)P_x(x)dx} \\ &= \frac{P_v(v)P_x(x)}{\int_{-\infty}^{\infty} P_v(v)P_x(x)dx} \end{aligned} \tag{3.10}$$

De kansdichtheden $P_v(v)$ en $P_x(x)$ zijn gegeven door respectievelijk de vergelijkingen (3.9) en (3.7). Deze kansdichtheden invullen in de uitdrukking (3.10) levert na enig rekenwerk:

$$P_{x|z}(x|z) \sim N(\frac{R}{P+R}x_m + \frac{P}{P+R}z, \frac{PR}{P+R}) \tag{3.11}$$

zodat vrijwel voor iedere zinvolle schatter $\hat{x}$ geldt:

$$\hat{x} = \frac{R}{P+R}x_m + \frac{P}{P+R}z \tag{3.12}$$

□

Vergelijken we voorbeeld 1.1 met voorbeeld 3.1 dan zien we dat het resultaat identiek is. Bij het voorbeeld 1.1 werden alleen het gemiddelde en de variantie van w en v gegeven en werd vervolgens de optimale (minimum variantie) lineaire schatting geconstrueerd. Bij voorbeeld 3.1 daarentegen werden de kansdichtheden van w en v helemaal gegeven en bleek de optimale (minimum variantie, maximum a posteriori) schatting lineair te zijn.

Definieer $\underline{\hat{x}}_{k|l}$ als een minimum variantie schatting van $\underline{x}_k$ gegeven de metingen $\underline{z}_1, \underline{z}_2, \cdots, \underline{z}_l$ en $P_{k|l}$ als de kovariantiematrix van deze schatting. Deze grootheden kunnen nu met behulp van de volgende recursieve vergelijkingen worden berekend:

Het Kalman filter

Begin condities:

$$\hat{x}_{0|0} = \hat{x}_0 \tag{3.13}$$

$$P_{0|0} = P_0 \tag{3.14}$$

Tijd propagatie:

$$\underline{\hat{x}}_{k|k-1} = F_{k-1}\underline{\hat{x}}_{k-1|k-1} \tag{3.15}$$

$$P_{k|k-1} = F_{k-1}P_{k-1|k-1}F_{k-1}^T + G_{k-1}Q_{k-1}G_{k-1}^T \tag{3.16}$$

Meet aanpassing:

$$\underline{\hat{x}}_{k|k} = [I - K_kM_k]\underline{\hat{x}}_{k|k-1} + K_z z_k \tag{3.17}$$

$$P_{k|k} = [I - K_kM_k]P_{k|k-1}[I - K_kM_k]^T + K_kR_kK_k^T \tag{3.18}$$

met als Kalman versterkingsmatrix

$$K_k = P_{k|k-1}M_k^T[M_kP_{k|k-1}M_k^T + R_k]^{-1} \tag{3.19}$$

De vergelijking (3.18) geldt voor willekeurige K_k. Substitutie van de optimale K_k gegeven door vergelijking (3.19) in de vergelijking (3.18) levert het volgende eenvoudige alternatief voor vergelijking (3.18):

$$P_{k|k} = [I - K_kM_k]P_{k|k-1} \tag{3.20}$$

We kunnen analoog aan voorbeeld 3.1 het algemene Kalman filter afleiden door alle stochastische processen Gaussisch te veronderstellen. Het optimale filter blijkt dan lineair te zijn. Ook kunnen we net als bij voorbeeld 1.1 laten zien dat als alleen de eerste twee momenten van de stochastische processen bekend zijn en alleen de klasse van lineaire filters beschouwd wordt, het Kalman filter optimaal is in minimum variantie zin.

Het Kalman filter heeft een predictie-correctie structuur. Gebaseerd op alle beschikbare informatie, is op tijdstip $k-1$ een optimale schatting gemaakt van de toestand. Nu wordt

eerst met behulp van de vergelijkingen (3.15) en (3.16) een voorspelling van de toestand berekend op tijdstip k. Is deze voorspelling bekend, dan is het mogelijk met behulp van vergelijking (2.2) de volgende meting te voorspellen. Als deze meting beschikbaar is dan wordt het verschil tussen deze meting en de voorspelde waarde gebruikt om de voorspelling van de toestand te corrigeren met behulp van de vergelijkingen (3.17) - (3.19). In figuur 3.1 is een schema van het Kalman filter gepresenteerd. Merk op dat de metingen zelf niet van belang zijn voor het berekenen van de Kalman gain K_k. Deze kan daardoor eventueel van te voren worden berekend.

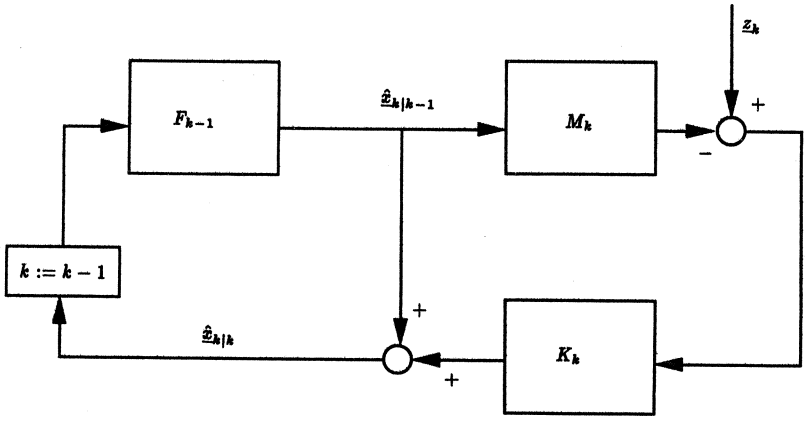


Figure 3.1: Schema van het Kalman filter.

Het filter produceert, naast een optimale schatting, ook de kovariantie van deze schatting. Dit is een belangrijke maat voor de nauwkeurigheid van de schatting. Bovendien kan deze kovariantiematrix gebruikt worden om inzicht te krijgen in het effect van een gewijzigde meetinspanning op de nauwkeurigheid van de schattingen.

Voorbeeld 3.2

Stel we willen een constante x schatten met behulp van metingen die behept zijn met ruis. Dat kan als volgt met het Kalman filter gebeuren:

$$x_{k+1} = x_k \tag{3.21}$$
$$z_k = x_k + v_k \tag{3.22}$$

met R als de variantie van de meetruis.

De beginschatting $\hat{x}_0$ heeft variantie P_0. Voor de metingen beschikbaar zijn kunnen we reeds de variantie van de schattingen berekenen.

tijdstip $k = 1$

tijd propagate:

$$P_{1|0} = P_0 \tag{3.23}$$

meet aanpassing:

$$P_{1|1} = \frac{P_0}{1 + P_0/R} \tag{3.24}$$

tijdstip $k = 2$

tijd propagatie:

$$P_{2|1} = \frac{P_0}{1 + P_0/R} \tag{3.25}$$

meet aanpassing

$$P_{2|2} = \frac{P_0}{1 + 2P_0/R} \tag{3.26}$$

tijdstip $k = k$

tijd propagatie:

$$P_{k|k-1} = \frac{P_0}{1 + (k-1)P_0/R} \tag{3.27}$$

meet aanpassing:

$$P_{k|k} = \frac{P_0}{1 + kP_0/R} \tag{3.28}$$

Merk op dat voor $k \to \infty$ $P_{k|k-1} \to 0$ en $P_{k|k} \to 0$.

We gaan nu de beschikbare metingen $z_1, z_2, \cdots$ verwerken. Omdat

$$K_k = \frac{P_{k|k-1}}{R + P_{k|k-1}} \tag{3.29}$$

geldt:

$$\hat{x}_{k|k-1} = \hat{x}_{k-1|k-1} \tag{3.30}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - \hat{x}_{k|k-1}) \tag{3.31}$$

Merk op dat voor $k \to \infty$ ook $K_k \to 0$. Dit houdt in dat metingen een steeds minder zwaar gewicht krijgen. Door gebruik te maken van het Kalman filter kan de constante x op recursieve wijze geschat worden. Het is eenvoudig aan te tonen dat het resultaat precies hetzelfde is als er gebruikt wordt gemaakt van de bekende, niet-recursieve, kleinste kwadraten methode:

Beschouw hiertoe het geval dat er geen informatie over de beginschatting is: $P_0 \to \infty$. Nu volgt uit vergelijking (3.27) dat:

$$P_{k|k-1} = \lim_{P_0 \to \infty} \frac{P_0}{1 + (k-1)P_0/R} = \frac{R}{k-1} \tag{3.32}$$

en dus dat

$$K_k = \frac{1}{k} \tag{3.33}$$

Het invullen van K_k in vergelijking (3.31) levert samen met vergelijking (3.30):

$$\begin{aligned} \hat{\underline{x}}_{k|k} &= \hat{\underline{x}}_{k-1|k-1} + \frac{1}{k}(z_k - \hat{x}_{k-1|k-1}) \\ &= \frac{k-1}{k}\hat{\underline{x}}_{k-1|k-1} + \frac{1}{k}z_k \end{aligned} \tag{3.34}$$

Het herhaaldelijk toepassen van de relatie (3.34) levert:

$$\begin{aligned} \hat{\underline{x}}_{k|k} &= \frac{k-1}{k}\left\{\frac{k-2}{k-1}\hat{\underline{x}}_{k-2|k-2} + \frac{1}{k-1}z_{k-1}\right\} + \frac{1}{k}z_k \\ &= \frac{k-2}{k}\hat{\underline{x}}_{k-2|k-2} + \frac{1}{k}z_{k-1} + \frac{1}{k}z_k \\ &= \frac{k-i}{k}\hat{\underline{x}}_{k-i|k-i} + \frac{1}{k}\sum_{j=1}^{i} z_{k-j+1} \end{aligned}$$

$$= \frac{1}{k}\sum_{j=1}^{k} z_{k-j+1}$$

$$= \frac{1}{k}\sum_{l=1}^{k} z_l \tag{3.35}$$

Voorbeeld 3.3

Beschouw nu het scalaire systeem:

$$x_{k+1} = x_k + w_k \tag{3.36}$$

$$z_k = x_k + v_k \tag{3.37}$$

Ten opzichte van het systeem (3.21)-(3.22) is er een systeem ruis geintroduceerd met variantie Q. De parameter x_k mag nu in de tijd in stochastische zin fluctueren. De niet-recursieve, kleinste kwadraten methode is nu niet toepasbaar. Het gebruik van het Kalman filter levert:

tijdstip $k = 1$

tijd propagatie:

$$P_{1|0} = P_0 + Q \tag{3.38}$$

meet aanpassing

$$P_{1|1} = \frac{P_0 + Q}{1 + (P_0 + Q)/R} \tag{3.39}$$

tijdstip $k = k$

$$P_{k|k-1} = P_{k|k-1} + Q \tag{3.40}$$

$$P_{k|k} = \frac{P_{k-1|k-1} + Q}{1 + (P_{k-1|k-1} + Q)/R} = \frac{R}{1 + \frac{R}{P_{k-1|k-1}+Q}} \tag{3.41}$$

Dit is een iteratieproces van de vorm:

$$x_k = f(x_{k-1}) \tag{3.42}$$

met

$$f(x) = \frac{x + Q}{1 + (x + Q)/R} \tag{3.43}$$

Omdat voor $x > 0$ geldt dat:

$$\frac{df}{dx} < 1$$

bestaat $\lim_{k\to\infty} x_k$ en dus bestaat $\lim_{k\to\infty} P_{k|k} = P_\infty$. Voor P_∞ geldt nu:

$$P_\infty = \frac{R}{1 + \frac{R}{P_\infty + Q}} \tag{3.44}$$

of

$$P_\infty = -\frac{1}{2}Q + \frac{1}{2}\sqrt{Q^2 + 4RQ} \tag{3.45}$$

Uit vergelijking (3.29) volgt voor

$$\lim_{k\to\infty} K_k = \lim_{k\to\infty} \frac{P_{k-1|k}}{R} = \frac{P_{k-1|k-1} + Q}{R} = \frac{P_\infty + Q}{R} > 0 \tag{3.46}$$

In tegenstelling tot het resultaat bij voorbeeld 3.2 is K_k nu altijd groter dan 0 en dus blijven nieuwe metingen invloed hebben op de schattingen.

□

Voorbeeld 3.4

Beschouw het AR(2)-model

$$x_{k+1} = ax_k + bx_{k-1} + w_k \tag{3.47}$$

en stel dat we x_k foutloos kunnen meten:

$$z_k = x_k \tag{3.48}$$

Substitutie van (3.38) in (3.37) levert de volgende meetvergelijking:

$$z_{k+1} = az_k + bz_{k-1} + w_k \tag{3.49}$$

Stel nu dat we de coefficient a en b niet nauwkeurig weten. Definieer in dit geval:

$$\underline{y}_k = [a \;\; b]^T \tag{3.50}$$

Omdat de coefficient a en b constant zijn geldt als systeemvergelijking voor $\underline{y}_k$:

$$\underline{y}_{k+1} = \underline{y}_k \tag{3.51}$$

De meetvergelijking (3.49) kan nu herschreven worden als:

$$z_{k+1} = [z_k \;\; z_{k-1}]\underline{y}_{k+1} + w_k \tag{3.52}$$

Het model (3.51) - (3.52) is nu van de vorm (2.1) - (2.2) zodat we het Kalman filter kunnen gebruiken om de parameters van het AR(2)-model te schatten. Door aan de systeemvergelijking (3.51) systeemruis toe te voegen kunnen we net als bij voorbeeld 3.3 modelleren dat de parameters in de tijd in stochastische zin kunnen fluctueren.

Recent Developments in Mathematical System Theory

G.J. Olsder
Delft University of Technology

Abstract

Many disciplines within mathematical system theory exist in which exiting new developments are going on. In the talk we will briefly mention some of them and then concentrate on one new development, viz. the one of discrete event dynamic systems. This paper is exclusively devoted to an introduction to discrete events.

About Difference Equations, Algebras and Discrete Events

Geert Jan Olsder

Abstract

An introduction to the theory of discrete event dynamic systems is given. Discrete event dynamic systems (DEDS) are nonlinear in the conventional algebra, but are linear in the max-plus algebra. Of many concepts and results within the conventional linear algebra and linear systems theory duplicates exist in the max-plus algebra and the theory of DEDS. The motivation to study DEDS comes from the description of flows in networks. Such networks are for instance related to computer systems, traffic systems and flexible manufacturing in production planning.

1 Difference Equations

1.1 Introduction

A well known equation in the theory of difference equations is the linear equation

$$x(t+1) = Ax(t), \quad t = 0, 1, 2, \ldots \tag{1}$$

The vector $x \in R^n$ represents the 'state' of an underlying model and this state evolves in time according to this equation; $x(t)$ denotes the state at time instant t. The symbol A represents a given $n \times n$ matrix. If an initial condition

$$x(0) = x_0 \tag{2}$$

is given, then the whole future evolution of (1) is determined.

Implicit in the text above is that (1) is a vector equation. Written out in scalar equations it becomes

$$x_i(t+1) = \sum_{j=1}^{n} a_{ij} x_j(t), \quad i = 1, \ldots, n; \quad t = 0, 1, \ldots \tag{3}$$

The symbol x_i denotes the i-th component of the vector x; the elements a_{ij} are the entries of the square matrix A. If $a_{ij}, i, j = 1, \ldots, n$ and $x_j(t), j = 1, \ldots, n$ are given, then $x_j(t+1), j = 1, \ldots, n$, can be calculated according to (1) or (3).

As an example take $n = 2$, such that A is a 2×2 matrix. Take

$$A = \begin{pmatrix} 3 & 7 \\ 2 & 4 \end{pmatrix} \tag{4}$$

and as initial condition

$$x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} . \tag{5}$$

The time evolution of (1) becomes for this example

$$x(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \; x(1) = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \; x(2) = \begin{pmatrix} 23 \\ 14 \end{pmatrix}, \; x(3) = \begin{pmatrix} 167 \\ 102 \end{pmatrix}, \ldots \tag{6}$$

1.2 Solution by means of Eigenvectors

Assume that the initial vector (2) equals an eigenvector of A; the corresponding eigenvalue is denoted by λ. The solution of (1) can be written as

$$x(t) = \lambda^t x_0, \quad t = 0, 1, \ldots \tag{7}$$

More generally, if the initial vector can be written as a linear combination of the set of linearly independent eigenvectors;

$$x_0 = \sum_j c_j v_j \; , \tag{8}$$

where v_j is the j-th eigenvector with corresponding eigenvalue λ_j, the c_j are coefficients, then

$$x(t) = \sum_j c_j \lambda_j^t v_j \; .$$

If the matrix A is diagonizable, then the set of linearly independent eigenvectors spans R^n, and any initial condition x_0 can be expressed as in (8). If A is not diagonizable, then one must work with generalized eigenvectors and the formula which expresses $x(t)$ in terms of eigenvalues and x_0 is slightly more complicated. This complication will not occur in the current context and therefore will not be dealt with explicitly.

2 Changing the Algebra

2.1 The Max-Plus Algebra

The only operations used in (1) or (3) are multiplication ($a_{ij} \times x_j(t)$) and addition (the $\sum$ symbol). Most of this paper can be considered as a study of formulas of the form (1), in which the operations are changed. Suppose that the two operations in (3) are changed in the following way; addition becomes maximization and multiplication becomes addition. Then (3) becomes

$$\begin{aligned} x_i(k+1) &= \max(a_{i1} + x_1(k), a_{i2} + x_2(k), \ldots, a_{in} + x_n(k)) \\ &= \max_j(a_{ij} + x_j(k)), \quad i = 1, \ldots, n \ . \end{aligned} \tag{9}$$

If the initial condition (2) also holds for (9), then the time evolution of (9) is completely determined again. Of course the time evolutions of (3) and (9) will be different in general. Equation (9), as it stands, is a nonlinear difference equation. As an example take A from (4) and $x(0)$ from (5). Then the time evolution of (9) becomes

$$x(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ x(1) = \begin{pmatrix} 7 \\ 4 \end{pmatrix}, \ x(2) = \begin{pmatrix} 11 \\ 9 \end{pmatrix}, \ x(3) = \begin{pmatrix} 16 \\ 13 \end{pmatrix}, \ldots \tag{10}$$

2.2 Motivation

We are used to thinking of the argument t in $x(t)$ as a time instant; at time instant t the state is $x(t)$. With respect to (9) we will introduce a different meaning for this argument. In order to emphasize this different meaning, the argument t has already been replaced by k. For a practical motivation we need to think of a network, which consists of a number of nodes and some arcs connecting these nodes. The network corresponding to (9) has n nodes; one for each component x_i. Entry a_{ij} corresponds to the arc from node j to node i. In terms of graph theory such a network is called a directed graph ('directed' because the individual arcs between the nodes are one way

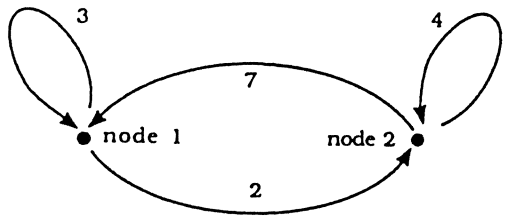


Figure 1: Network corresponding to Equation (4)

arrows). Therefore the arcs corresponding to a_{ij} and a_{ji}, if both exist, are considered to be different.

The nodes in the network can perform certain activities; each node has its own kind of activity. Such activities take a finite time, called holding time, to be performed. These holding times may be different for different nodes. It is assumed that an activity at a certain node can only start when all preceding ('directly upstream') nodes have finished their activities and sent the results of these activities along the arcs to the current node. Thus the arc corresponding to a_{ij} can be interpreted as an output channel for node j and simultaneously as an input channel for node i. Suppose that this node i starts its activity as soon as all preceding nodes have sent their results (the rather neutral word 'results' is used, it could equally have been messages, ingredients or products,...) to node i, then (9) describes when the activities take place. The interpretation of the quantities used is:

- $x_i(k)$: is the earliest time instant at which node i becomes active for the k-th time;
- a_{ij} : is the sum of the holding time (i.e. time duration of the activity) at node j and the travelling time (the rather neutral 'travelling time' is used rather than for instance 'transportation time' or 'communication time') from node j to node i.

The fact that we write a_{ij} rather than a_{ji} for a quantity connected to the arc from node j to node i has to do with matrix equations which will be written in the classical way with column vectors, as will be seen later on. (This is in contrast with queuing theory, where it is customary to work with row vectors.) For the example given above, the network has two nodes and four arcs, as given in Figure 1. The interpretation of the number 3 in this figure is that if node 1 has started an activity, the next activity cannot start within

the next 3 time units. Similarly, the time between two subsequent activities of node 2 is at least 4 time units. Node 1 sends its results to node 2 and once an activity starts in node 1, it takes 2 time units before the result of this activity reaches node 2. Similarly it takes 7 time units after the initiation of an activity of node 2 for the result of that activity to reach node 1. Suppose that an activity refers to some production. The production time of node 1 could for instance be 1 unit of time; after that, node 1 needs 2 time units for recovery (lubrication say) and the travelling time of the result (the final product) from node 1 to node 2 is 1 unit of time. Thus the number $a_{11} = 3$ is made up of a production time 1 and a recovery time 2 and the number $a_{21} = 2$ is made up of the same production time 1 and a travelling time 1. Similarly, if the production time at node 2 is 4, the this node does not need any time for recovery (because $a_{22} = 4$), and the travelling time from node 2 to node 1 is 3 (because $a_{12} = 7 = 4 + 3$).

If we now look at the sequence (10) again, the interpretation of the vectors $x(k)$ is different from the initial one. The argument k is not a time instant anymore, but a counter which states how many times the various nodes have been active. At time 14 node 1 has been active twice (more precisely, node 1 has started two activities, respectively at times 7 and 11). At the same time 14, node 2 has been active three times (it started activities at times 4, 9 and 13). The counting of the activities is such that it coincides with the argument of the x vector. The initial condition is henceforth considered to be the 0-th activity.

In Figure 1 there was an arc from any node to any other node. In many networks referring to more practical situations, this will not be the case. If there is no arc from node j to node i then node i does not need any result from node j. Therefore node j does not have a direct influence on the behavior of node i. In such a situation it is useful to consider the element a_{ij} to be equal to $-\infty$. In (9) a term $-\infty + x_j(k)$ does not influence $x_i(k+1)$ as long as $x_j(k)$ is finite. The number $-\infty$ will occur frequently in the sequel and it will be indicated by ε.

2.3 Some Notation and Some Calculus

For reasons which will become clear later on, (9) will be written as

$$x_i(k+1) = \bigoplus_j a_{ij} \otimes x_j(k), \quad i = 1, \ldots, n \ ,$$

or in vector notation,

$$x(k+1) = A \otimes x(k) \ . \tag{11}$$

The symbol $\bigoplus_j c(j)$ refers to the maximum of the elements $c(j)$ with respect to all appropriate j, and $\otimes$ refers to addition. Later on the symbol $\oplus$ will also be used; $a \oplus b$ refers to the maximum of the scalars a and b. If the initial condition for (11) is $x(0) = x_0$, then

$$x(1) = A \otimes x_0 \ ,$$

$$x(2) = A \otimes x(1) = A \otimes (A \otimes x_0) = (A \otimes A) \otimes x_0 = A^2 \otimes x_0 \ .$$

It can be shown that indeed $A \otimes (A \otimes x_0) = (A \otimes A) \otimes x_0$. For the example given above it is easy to check this by hand. Instead of $A \otimes A$ we simply write A^2. We get

$$x(3) = A \otimes x(2) = A \otimes (A^2 \otimes x_0) = (A \otimes A^2) \otimes x_0 = A^3 \otimes x_0 \ ,$$

and in general

$$x(k) = (\underbrace{A \otimes A \otimes \cdots \otimes A}_{\text{k times}}) \otimes x_0 = A^k \otimes x_0 \ .$$

The matrices A^2, A^3,..., can be calculated directly. Let us consider the A-matrix of (4) again, then

$$A^2 = \begin{pmatrix} \max(3+3, 7+2) & \max(3+7, 7+4) \\ \max(2+3, 4+2) & \max(2+7, 4+4) \end{pmatrix} = \begin{pmatrix} 9 & 11 \\ 6 & 9 \end{pmatrix} \ .$$

In general

$$(A^2)_{ij} = \bigoplus_l a_{il} \otimes a_{lj} = \max_l (a_{il} + a_{lj}) \ . \tag{12}$$

The quantity $(A^2)_{ij}$ can be interpreted as the maximum (with respect to l) of all connections from node j via node l to node i. One speaks of paths of length two between the nodes j and i. More generally, $(A^k)_{ij}$ denotes the maximum of all paths of length k, starting at node j and ending at node i.

The multiplication of two matrices in the max-plus algebra follows the standard pattern as shown by the example

$$\begin{pmatrix} 1 & 2 \\ \varepsilon & 0 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 5 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 6 & 5 \\ 1 & 3 \\ 3 & 4 \end{pmatrix} .$$

2.4 Axiomatics

The operations $\oplus$ and $\otimes$ defined on the set R can also be defined with respect to a more general set of elements $\mathcal{D}$. One then speaks of a *dioid.*

Definition .1 (Dioid) *A* dioid *is a set $\mathcal{D}$ endowed with two operations denoted $\oplus$ and $\otimes$ (called 'sum' or 'addition', and 'product' or 'multiplication') obeying the following axioms:*

Axiom .2 (Associativity of addition)

$$\forall a, b, c \in \mathcal{D}, (a \oplus b) \oplus c = a \oplus (b \oplus c) \ .$$

Axiom .3 (Commutativity of addition)

$$\forall a, b \in \mathcal{D}, a \oplus b = b \oplus a \ .$$

Axiom .4 (Associativity of multiplication)

$$\forall a, b, c \in \mathcal{D}, (a \otimes b) \otimes c = a \otimes (b \otimes c) \ .$$

Axiom .5 (Distributivity)

$$\begin{aligned} \forall a, b, c \in \mathcal{D}, \quad & (a \oplus b) \otimes c = (a \otimes c) \oplus (b \otimes c) \ , \\ & c \otimes (a \oplus b) = c \otimes a \oplus c \otimes b \ . \end{aligned}$$

This is right, *respectively* left, *distributivity of multiplication with respect to addition. One statement does not follow from the other since multiplication is not assumed to be commutative.*

Axiom .6 (Existence of a zero element)

$$\exists \varepsilon \in \mathcal{D} : \forall a \in \mathcal{D}, a \oplus \varepsilon = a \ .$$

Axiom .7 (Absorbing zero element)

$$\forall a \in \mathcal{D}, a \otimes \varepsilon = \varepsilon \otimes a = \varepsilon \ .$$

Axiom .8 (Existence of an identity element)

$$\exists e \in \mathcal{D} \ : \forall a \in \mathcal{D}, a \otimes e = e \otimes a = a \ .$$

Axiom .9 (Idempotency of addition)

$$\forall a \in \mathcal{D}, a \oplus a = a \ .$$

Definition .10 (Commutative dioid) *A dioid is* commutative *if multiplication is commutative.*

With the noticeable exception of Axiom .9, most the axioms of dioids are required for rings too. Indeed, Axiom .9 is the most distinguishing feature of dioids. Because of this axiom, addition cannot be cancellative, that is, $a \oplus b = a \oplus c$ does not imply $b = c$ in general. Multiplication is not necessarily cancellative either (of course, because of Axiom .7, cancellation would anyway only apply to elements different from ε). For an example in which multiplication is not cancellative take $\mathcal{D} = R \cup \{-\infty\} \cup \{+\infty\}$ and define $\oplus$ as max and $\otimes$ as min.

It is easily shown that in dioids the distributivity with respect to matrices also holds, i.e. $A \otimes (B \otimes C) = (A \otimes B) \otimes C$, where these multiplications only make sense if the matrices have appropriate dimensions.

2.5 Systems with Inputs and Outputs

An extension of (11) is

$$\left.\begin{array}{rcl} x(k+1) & = & (A \otimes x(k)) \oplus (B \otimes u(k)) \ , \\ y(k) & = & C \otimes x(k) \ . \end{array}\right\} \tag{13}$$

The symbol $\oplus$ in this formula refers to componentwise maximization. The m-vector u is called the input to the system; the p-vector y is the output of the system.The components of u refer to nodes which have no predecessors. Similarly, the components of y refer to nodes with no successors. The components of x now refer to internal nodes, i.e. to nodes with have both successors and predecessors. The matrices $B = \{b_{ij}\}$ and $C = \{c_{ij}\}$ have sizes $n \times m$ and $p \times n$ respectively. The traditional way of writing (13) would be

$$\begin{array}{rcl} x_i(k+1) & = & \max(a_{i1} + x_1(k), \ldots, a_{in} + x_n(k), \\ & & \qquad b_{i1} + u_1(k), \ldots, b_{im} + u_m(k)), \quad i = 1, \ldots, n \ ; \\ y_i(k) & = & \max(c_{i1} + x_1(k), \ldots, c_{in} + x_n(k)), \quad i = 1, \ldots, p \ . \end{array}$$

Usually (13) is written as

$$\left.\begin{array}{rcl} x(k+1) & = & Ax(k) \oplus Bu(k) \ , \\ y(k) & = & Cx(k) \ . \end{array}\right\}, \tag{14}$$

where it is understood that multiplication has priority over addition. If it is clear where the '$\otimes$'-symbols are used, they are sometimes omitted, as shown

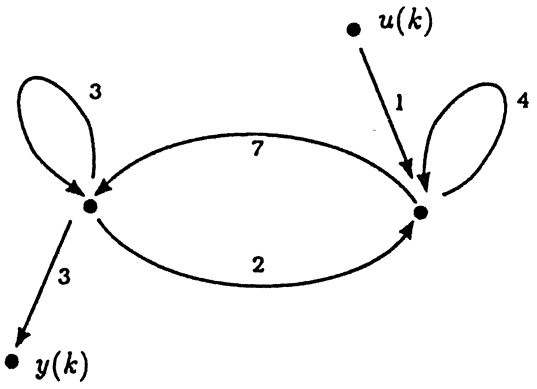


Figure 2: Network with input and output

in (14). This practice is exactly the same one as with respect to the more common multiplication ' × ' or ' . ' symbol in conventional algebra. In the same vein, in conventional algebra $1 \times x$ is the same as $1x$, which is usually written as x. Within the context of the $\otimes$ and $\oplus$ symbols, $0 \otimes x$ is exactly the same as x. The symbol ε is the neutral element with respect to maximization; its numerical value equals $-\infty$. Similarly, the symbol e will denote the neutral element with respect to addition; it assumes the numerical value 0. Note that $1 \otimes x$ is different from x.

If one wants to think in terms of a network again, then $u(k)$ is a vector indicating when certain resources become available for the k-th time. Subsequently it takes b_{ij} time units before the j-th resource reaches node i of the network. The vector $y(k)$ refers to the time instant at which the final products of the network are delivered to the outside world.

Take for example

$$\begin{aligned} x(k+1) &= \begin{pmatrix} 3 & 7 \\ 2 & 4 \end{pmatrix} x(k) \oplus \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix} u(k) \ , \\ y(k) &= \begin{pmatrix} 3 & \varepsilon \end{pmatrix} x(k) \ . \end{aligned} \tag{15}$$

The corresponding network is shown in Figure 2. Because $b_{11} = \varepsilon(= -\infty)$, the input $u(k)$ only goes to node 2. If one would replace B by $(\ 2,\ 1\ )'$ for instance, then each input would 'spread' itself over the two nodes. In this example with $B = (\ 2,\ 1\ )'$, from time instant $u(k)$ on, it takes 2 time units for the input to reach node 1 and 1 time unit to reach node 2. In many

practical situations an input will enter the network through one node. That is why in (15) only one b_i-component is different from ε. Similar remarks can be made with respect to the output. Suppose that we have (5) as an initial condition and that

$$u(0) = 1, u(1) = 7, u(2) = 13, u(3) = 19, \ldots,$$

then it easily follows that

$$x(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, x(1) = \begin{pmatrix} 7 \\ 4 \end{pmatrix}, x(2) = \begin{pmatrix} 11 \\ 9 \end{pmatrix}, x(3) = \begin{pmatrix} 16 \\ 14 \end{pmatrix}, \ldots$$

$$y(0) = 4, y(1) = 10, y(2) = 14, y(3) = 19, \ldots$$

2.6 Higher Order Difference Equations

We started this section with the difference equation (1), which is a first order linear vector difference equation. It is well known that a higher order linear scalar difference equation

$$z(k+1) = a_1 z(k) + a_2 z(k-1) + \cdots + a_n z(k-n+1) \tag{16}$$

can be written in the form of equation (1). If we introduce the vector $(z(k), z(k-1), \ldots, z(k-n+1))'$, then (16) can be written as

$$\begin{pmatrix} z(k+1) \\ z(k) \\ \vdots \\ \vdots \\ z(k-n+2) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \ldots & \ldots & a_n \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & & & & \\ \vdots & & & & \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} z(k) \\ z(k-1) \\ \vdots \\ \vdots \\ z(k-n+1) \end{pmatrix} . \tag{17}$$

This equation has exactly the form of (1). If we change the operations in (16) in the standard way; addition becomes maximization and multiplication becomes addition, then the numerical evaluation of (16) becomes

$$z(k+1) = \max(a_1 + z(k), a_2 + z(k-1), \ldots, a_n + z(k-n+1)) . \tag{18}$$

This equation can also be written as a first order linear vector difference equation. In fact this equation is almost Equation (17), which must now be evaluated with the operations maximization and addition. The only difference is that the 1's and 0's in the matrix in (17) must be replaced by e's and ε's respectively.

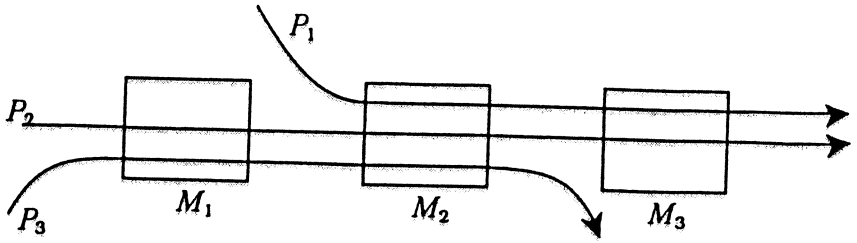


Figure 3: Routing of parts along machines

Table 1: Processing times

	P_1	P_2	P_3
M_1		1	5
M_2	3	2	3
M_3	4	3	

3 Example on Production

Consider a manufacturing system consisting of three machines. It is supposed to produce three kinds of parts according to a certain product mix. The routes to be followed by each part and each machine are depicted in Figure 3; in which $M_i, i = 1,2,3$, are the machines and $P_i, i = 1,2,3$, are the parts. Processing times are given in Table 1. Note that this manufacturing system has a flow-shop structure, i.e. all parts follow the same sequence on the machines (although they may skip some) and every machine is visited at most once by each part. This manufacturing system is automated and there are no set-up times on machines when they switch from one part type to another. Parts are carried on a limited number of pallets (or, equivalently, product carriers) by means of fixtures. For reasons of simplicity it is assumed that

1. only one pallet is available for each part type;
2. there are no set-up times or travelling times;
3. the sequencing of part types on the machines is known and it is (P_2, P_3) on M_1, (P_1, P_2, P_3) on M_2 and (P_1, P_2) on M_3.

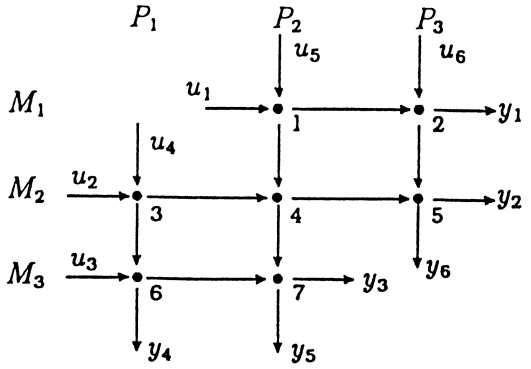


Figure 4: The ordering of activities in the flexible manufacturing system

The last point mentioned is not for reasons of simplicity. If any machine would start working on the part which would arrive first instead of waiting for the appropriate part, the modelling cannot be done within the context of the max-plus algebra. In order to accomodate the rule 'first in, first served', one would need a third operation, viz. *min* and that would make the general modeling more difficult and less transparant.

We can draw a graph in which each node corresponds to a combination of a machine and a part. Since M_1 works on 2 parts, M_2 on 3 and M_3 on 2, this graph has seven nodes. The arcs between the nodes express the precedence constraints between operations due to the sequencing of operations on the machines. To each node i in Figure 4 corresponds a number x_i; it denotes the earliest time instant at which the node can start its activity. In order to be able to calculate these quantities, the time instants at which the machines and parts (together called the resources) are available must be given. This is done by means of a six-dimensional input vector u (six since there are six resources: three machines and three parts). There is an output vector also; the elements of the six-dimensional vector y denote the time instants at which the parts are ready and the machines have finished their jobs (for one cycle). The model becomes

$$x = Ax \oplus Bu \ ; \tag{19}$$

$$y = Cx \ , \tag{20}$$

in which the matrices are

$$A = \begin{pmatrix} \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1 & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1 & \varepsilon & 3 & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & 5 & \varepsilon & 2 & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & 3 & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & 2 & \varepsilon & 4 & \varepsilon \end{pmatrix}; \quad B = \begin{pmatrix} e & \varepsilon & \varepsilon & \varepsilon & e & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & e \\ \varepsilon & e & \varepsilon & e & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & e & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \end{pmatrix}; \tag{21}$$

$$C = \begin{pmatrix} \varepsilon & 5 & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & 3 & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & 3 \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & 4 & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & 3 \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & 3 & \varepsilon & \varepsilon \end{pmatrix}.$$

Equation (19) is an implicit equation in x. Let us see what we get by repeated substitution of the complete right-hand side of (19) for x of this same right-hand side. After one substitution:

$$\begin{aligned} x &= A^2x \oplus ABu \oplus Bu \\ &= A^2x \oplus (A \oplus e)Bu \ , \end{aligned}$$

and after k substitutions:

$$x = A^kx \oplus (A^{k-1} \oplus A^{k-2} \oplus \cdots \oplus A \oplus e)Bu \ .$$

In the formulas above e refers to the identity matrix; zeros on the diagonal and ε's elsewhere. The symbol e will be used as the identity element for all spaces that will be encountered in this paper. Similarly, ε will be used to denote the zero element of any space to be encountered.

For this example it is easily shown that $A^n = \varepsilon = -\infty$ for $n \geq 3$. In the next session a graph theoretic explanation for these equalities will be given (the precedence graph of A is acyclic). Therefore the solution x in the current example becomes

$$x = (A^2 \oplus A \oplus e)Bu \ ,$$

for which we can write

$$x = A^*Bu \ ,$$

where A^* is defined as

$$A^* \stackrel{\text{def}}{=} e \oplus A \oplus \cdots \oplus A^n \oplus A^{n+1} \oplus \cdots . \tag{22}$$

In our example on production, $A^k, k > 2$, does not contribute to the sum in (22). For later reference, we also introduce the notation

$$A^+ \stackrel{\text{def}}{=} A \oplus \cdots \oplus A^n \oplus A^{n+1} \oplus \cdots . \tag{23}$$

Remark .11 With the conventional matrix calculus in mind one might be tempted to write for (22):

$$(e \oplus A \oplus A^2 \oplus \cdots) = (e \ominus A)^{-1} \ . \tag{24}$$

Of course, we have not defined the inverse of a matrix within the current setting and (24) is an empty statement. It is also strange to have a 'minus' sign $\ominus$ in (24) and it is not known how to interpret this sign in the context of the max-operation at the left-hand side of the equation. It should be the reverse operation of $\oplus$. If we dare to continue along these shaky lines, one could calculate the solution of (19) as

$$(e \ominus A)x = Bu \Rightarrow x = (e \ominus A)^{-1}Bu \ ,$$

which equals $x = A^*Bu$ if we believe (24) to make sense. Quite often one can guide one's intuition by considering formal expressions of the kind (24). One tries to find formal analogies in the notation with the conventional analysis. It can be shown [1] that an inverse as in (24) does not exist in general and therefore we get 'stuck' with the series expansion. ■

Now we add feedback arcs to Figure 4 as illustrated in Figure 5. In this graph the feedback arcs are indicated by dotted lines. The meaning of these feedback arcs is the following. After a machine has finished a sequence of products, it starts with the next sequence. If the pallet on which product P_i was mounted is at the end, the finished product is removed and the empty pallet immediately goes back to the starting point to pick up a new part P_i. If it is assumed that the feedback arcs have zero time duration, then $u(k) = y(k-1)$, where $u(k)$ is the k-th input cycle and $y(k)$ the k-th output. Thus we can write

$$\begin{aligned} y(k) \ = \ Cx(k) \ &= \ CA^*Bu(k) \\ &= \ CA^*By(k-1) \ . \end{aligned} \tag{25}$$

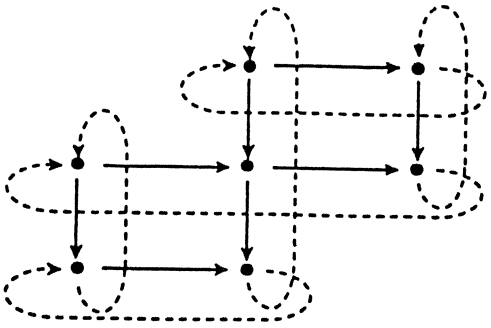


Figure 5: Production system with feedback arcs

The transition matrix from $y(k-1)$ to $y(k)$ can be calculated (a simple computer program does the job, but it can be done by hand);

$$M \stackrel{\text{def}}{=} CA^*B = \begin{pmatrix} 6 & \varepsilon & \varepsilon & \varepsilon & 6 & 5 \\ 9 & 8 & \varepsilon & 8 & 9 & 8 \\ 6 & 10 & 7 & 10 & 6 & \varepsilon \\ \varepsilon & 7 & 4 & 7 & \varepsilon & \varepsilon \\ 6 & 10 & 7 & 10 & 6 & \varepsilon \\ 9 & 8 & \varepsilon & 8 & 9 & 8 \end{pmatrix} . \tag{26}$$

This matrix M determines the speed with which the manufacturing system can work. We will come back to this issue in the next section.

4 Some Graph Theory and the Spectral Theory of Matrices

4.1 Graph Theory

Informally, we already encountered several items directly related to graphs. We will formalize some of these concepts here. A *directed graph* $\mathcal{G}$ is defined as a pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of elements called *nodes* and where $\mathcal{E}$ is a set of which the elements are ordered (not necessarily different) pairs of nodes, called *arcs*. The possibility of several arcs between two nodes exists (one then speaks about a multigraph); in this paper, however, we exclusively deal with directed graphs in which there is at most one (i.e. zero or one) arc between any two nodes. Instead of directed graph one often uses the shorter

word 'digraph', or even 'graph' if it is clear from the context that digraph is meant.

Denote the number of nodes by n, and number the individual nodes $1, 2, \ldots, n$. If $(i,j) \in \mathcal{E}$, then i is called the initial node or the origin of the arc (i,j), and j the final node or the destination of the arc (i,j). Graphically, the nodes are represented by points, and the arc (i,j) is represented by an 'arrow' from i to j.

We now give a list of some concepts of graph theory which will be used later on.

Predecessor, successor. If in a graph $(i,j) \in \mathcal{E}$ then i is called a predecessor of j and j is called a successor of i. The set of all predecessors of j is indicated by $\pi(j)$ and the set of all successors of i is indicated by $\sigma(i)$. A predecessor is also called an *upstream node* and a successor is also called a *downstream node*.

Source, sink. If $\pi(i) = \emptyset$ then node i is called a source; if $\sigma(i) = \emptyset$ then i is called a sink. Depending on the application, a source, respectively sink, is also called an *input(-node)*, respectively an *output(-node)* of the graph.

Path, circuit, loop. A path ρ is a sequence of nodes $i_1, i_2, \ldots, i_p$, $p > 1$, such that $i_j \in \pi(i_{j+1}), j = 1, \ldots, p-1$. Node i_1 is the initial node and i_p is the final one of this path. The *length* of the path is equal to the sum of the lengths of the arcs of which it is composed, the lengths of the arcs being 1. The length of path ρ is denoted $|\rho|_l$ (equal to $p-1$ in the above example). The subscript 'l' here refers to the word 'length' (later on another subscript 'w' will appear for a different concept). Equivalently, one also says that a path is a sequence of arcs which connects a sequence of nodes. An *elementary path* is a path in which no node appears more than once. A circuit is a path where the initial and the final node coincide. An *elementary circuit* $i_1, i_2, \ldots, i_p = i_1$ is a circuit in which the path $i_1, i_2, \ldots, i_{p-1}$ is elementary. A loop is a circuit involving a single node. A digraph is said to be *acyclic* if it contains no circuits.

Descendant, ascendant. The set of descendants $\sigma^+(i)$ of node i consists of all nodes j such that a path exists from i to j. Similarly the set of ascendants $\pi^+(i)$ of node i is the set of all nodes j such that a path exists from j to i. One has, e.g., $\pi^+(i) = \pi(i) \cup \pi(\pi(i)) \cup \ldots$. The

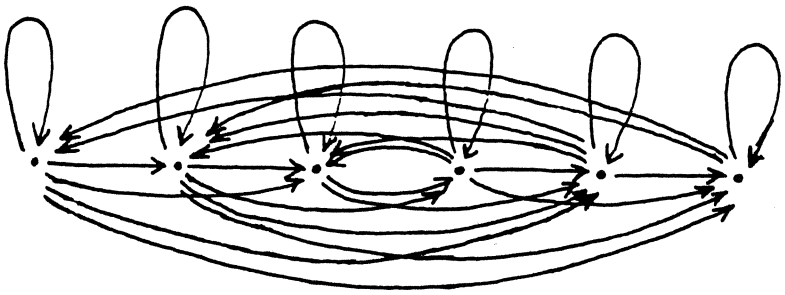


Figure 6: Precedence graph of the matrix M

mapping $i \mapsto \pi^*(i) = \{i\} \cup \pi^+(i)$ is the transitive closure of π; the mapping $i \mapsto \sigma^*(i) = \{i\} \cup \sigma^+(i)$ is the transitive closure of σ.

Strongly connected graph. A graph is called strongly connected if for any two different nodes i and j there exists a path from i to j. Equivalently, $i \in \sigma^*(j)$ for all $i, j \in \mathcal{V}$, with $i \neq j$. Note that according to this definition an isolated node, with or without a loop, is a strongly connected graph.

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. If we associate a real number a_{ij} to each arc $(j,i) \in \mathcal{E}$, then $\mathcal{G}$ is called a *weighted graph.* The quantity a_{ij} is called the *weight* of arc (j,i). Note that the second subscript of a_{ij} refers to the initial (and not the final) node. The reason is that in the algebraic context we work with column vectors (and not with row vectors).

In the following definition the starting point is a square matrix, the entries of which may again assume the 'value' ε.

Definition .12 (Precedence graph) *The precedence graph of a square matrix A, of size $n \times n$, is a weighted digraph with n nodes and an arc (j,i) if $a_{ij} \neq \varepsilon$, in which case the weight of this arc receives the numerical value of a_{ij}. The precedence graph is denoted $\mathcal{G}(A)$.*

It is not difficult to see that any weighted digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the precedence graph of an appropriately defined square matrix. The weight a_{ij} of the arc from node j to node i is defined as the ij-th entry of a matrix A. If an arc does not exist, the corresponding entry of A becomes ε. The matrix A thus defined has $\mathcal{G}$ as its precedence graph. As an example, the precedence graph of (26) is given in Figure 6; it has many circuits. The precedence graph of A from (21), not shown in a figure here, is acyclic. This latter property yields $A^k = \varepsilon$ for $k \geq n$.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a weighted digraph with n nodes. The weights are combined such as to form the $n \times n$ matrix A. The numerical value of a_{ij}

equals the weight of the arc from node j to node i. If no such arc exists, then $a_{ij} = \varepsilon$. As seen before, the element (i,j) of $A^k = A \otimes \cdots \otimes A$, considered within the max-plus algebra denotes the maximum weight with respect to all paths of length k which go from node j to node i. If no such path exists, then $(A^k)_{ij} = \varepsilon$. Within this algebra, ε gets assigned the numerical value $-\infty$. The weight of a path ρ is denoted $|\rho|_w$.

Definition .13 *The* mean weight of a path *is defined as the sum of the weights of the individual arcs of this path, divided by the length of this path. If the path is denoted by ρ, then the mean weight equals $|\rho|_w/|\rho|_l$. If such a path is a circuit one talks about the mean weight of the circuit, or simply the* cycle mean.

We are interested in the maximum of these cycle means, where the maximum is taken over all circuits in the graph. This number will be called the *maximum cycle mean.*

4.2 Spectral Theory of Matrices

Given a matrix A with entries in the max-plus algebra, we consider the problem of existence of eigenvalues and eigenvectors, that is, the existence of λ and v such that:

$$Av = \lambda v \ . \tag{27}$$

For example,

$$\begin{pmatrix} 3 & 7 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 2.5 \\ e \end{pmatrix} = 4.5 \begin{pmatrix} 2.5 \\ e \end{pmatrix} \ .$$

Thus it is seen that the matrix A of (4) has an eigenvalue 4.5. To exclude degenerate cases, it is assumed that not all elements of v are identical to ε. Equation (7) is also valid in the current setting. If x_0 is an eigenvector of A, with corresponding eigenvalue λ, then the solution of the difference equation (11) can be written as

$$x(k) = \lambda^k x_0 \quad (= \lambda^k \otimes x_0), \quad k = 0, 1, \ldots \tag{28}$$

The numerical evaluation of λ^k in this formula equals $k\lambda$ in conventional analysis. The eigenvalue λ can be interpreted as the cycle time (defined as the inverse of the throughput) of the underlying system; each node of the corresponding network becomes active every λ units of time, as it follows straightforwardly from (28). Also, the relative order in which the nodes

become active for the k-th time, as expressed by the components $x_i(k)$, is exactly the same as the relative order in which the nodes become active for the $(k+1)$-st time. More precisely, equation (28) yields $x_l(k+1) - x_j(k+1) = x_l(k) - x_j(k)$, $j,l = 1,\ldots,n$. Thus the solution (28) exhibits a kind of periodicity. Procedures exist for the calculation of eigenvalues and eigenvectors; an efficient one is the procedure known as Karp's algorithm [5].

The main result on eigenvalues in the max-plus algebra is as follows.

Theorem .14 *We are given a square matrix A. If $\mathcal{G}(A)$ is strongly connected, there exists one and only one eigenvalue and at least one eigenvector. The eigenvalue is equal to the maximum cycle mean of the graph:*

$$\lambda = \max_{\zeta} \frac{|\zeta|_w}{|\zeta|_l} ,$$

where ζ ranges over the set of circuits of $\mathcal{G}(A)$.

Proof

Existence of x and λ. Define the matrix B by subtracting λ, in the conventional way, from each entry of A, where $\lambda = \max_\zeta |\zeta|_w/|\zeta|_l$. Now the maximum circuit weight of $\mathcal{G}(B)$ is e. Hence B^* and $B^+ = BB^*$ exist. The matrix B^+ has some columns with entries e on the diagonal. Indeed, we can pick a node k on a circuit $\xi \in \arg\max_\zeta |\zeta|_w/|\zeta|_l$. The maximum weight of all paths from k to k is e. Therefore we have $e = B^+_{kk}$. Then:

$$B^+_{.k} = B^*_{.k} \Rightarrow BB^*_{.k} = B^+_{.k} = B^*_{.k} \Rightarrow AB^+_{.k} = \lambda B^+_{.k} ,$$

where $B_{.k}$ denotes column k of B. Hence $x = B^+_{.k}$ is an eigenvector. The set of nodes of $\mathcal{G}(A)$ corresponding to indices of the nonzero components of x is called the *support* of x.

If $\mathcal{G}(A)$ is strongly connected, the support of x contains all nodes.
Let us suppose that the support of x does not cover the whole graph. Then there are arcs going from the support of x to other nodes because the graph $\mathcal{G}(A)$ is strongly connected. Then the support of Ax would be larger than the support of x which is contradicted by $Ax = \lambda x$.

Uniqueness of λ. If λ satisfies Equation (27), we have $(Ax)_1 = \lambda x_1$ and there exists a component x_{i_1} such that $A_{1i_1}x_{i_1} = \lambda x_1$. Then $(Ax)_{i_2} =$

λx_{i_1} and there exists a component x_{i_2} such that $A_{i_1 i_2} x_{i_2} = \lambda x_{i_1}$ and so on until we reach a component x_{i_l} that we have already met. In this way we have defined a circuit $\beta = (i_l, i_m, \ldots, i_{l+1}, i_l)$ such that:

$$A_{i_l i_{l+1}} A_{i_{l+1} i_{l+2}} \cdots A_{i_m i_l} x_{i_{l+1}} x_{i_{l+2}} \cdots x_{i_m} x_{i_l} = \lambda^{m-l+1} x_{i_l} x_{i_{l+1}} \cdots x_{i_m} \ .$$

Therefore, because $x_k \neq \varepsilon$ for all k, λ^{m-l+1} is the weight of circuit of length $m - l + 1$. Hence λ is the average weight of circuit β.

Let us now take any circuit $\gamma = (i_1, \ldots, i_p, i_1)$ such that its nodes belong to the support of x (here any node of $\mathcal{G}(A)$). We have:

$$\begin{aligned} A_{i_2 i_1} x_{i_1} &\leq \lambda x_{i_2} \ , \\ &\vdots \\ A_{i_p i_{p-1}} x_{i_{p-1}} &\leq \lambda x_{i_p} \ , \\ A_{i_1 i_p} x_{i_p} &\leq \lambda x_{i_1} \ . \end{aligned}$$

Hence, by $\otimes$-multiplying these inequalities and because $x_k \neq \varepsilon$ for all k, one sees that λ is greater than the average weight of γ. Therefore λ is the maximum cycle mean and it is unique.

■

Remark .15 It is important to understand the role of the support of x in the previous proof. If $\mathcal{G}(A)$ is not strongly connected, the support of x is not necessarily the whole set of nodes and, in general, there is not a unique eigenvalue. ■

Example .16 With the assumption of Theorem .14 , the uniqueness of the eigenvector is not assured as is shown by

$$\begin{pmatrix} 1 & e \\ e & 1 \end{pmatrix} \begin{pmatrix} e \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ e \end{pmatrix} = 1 \begin{pmatrix} e \\ -1 \end{pmatrix},$$

and

$$\begin{pmatrix} 1 & e \\ e & 1 \end{pmatrix} \begin{pmatrix} -1 \\ e \end{pmatrix} = \begin{pmatrix} e \\ 1 \end{pmatrix} = 1 \begin{pmatrix} -1 \\ e \end{pmatrix}.$$

■

Example .17 The following example is a trivial counterexample to the uniqueness of the eigenvalue if the graph is not strongly connected:

$$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 2 \end{pmatrix} \begin{pmatrix} e \\ \varepsilon \end{pmatrix} = 1 \begin{pmatrix} e \\ \varepsilon \end{pmatrix} , \quad \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 2 \end{pmatrix} \begin{pmatrix} \varepsilon \\ e \end{pmatrix} = 2 \begin{pmatrix} \varepsilon \\ e \end{pmatrix} .$$

■

Example .18 The matrix M of (26) has a unique eigenvalue:

$$\begin{pmatrix} 6 & \varepsilon & \varepsilon & \varepsilon & 6 & 5 \\ 9 & 8 & \varepsilon & 8 & 9 & 8 \\ 6 & 10 & 7 & 10 & 6 & \varepsilon \\ \varepsilon & 7 & 4 & 7 & \varepsilon & \varepsilon \\ 6 & 10 & 7 & 10 & 6 & \varepsilon \\ 9 & 8 & \varepsilon & 8 & 9 & 8 \end{pmatrix} \begin{pmatrix} e \\ 3 \\ 3.5 \\ .5 \\ 3.5 \\ 3 \end{pmatrix} = 9.5 \begin{pmatrix} e \\ 3 \\ 3.5 \\ .5 \\ 3.5 \\ 3 \end{pmatrix} .$$

It follows that the eigenvalue equals 9.5, which means in more practical terms that the manufacturing system 'delivers' an item (a product or a machine) at all of its output channels every 9.5 units of time. ■

The eigenvector of this latter example is also unique, apart from adding the same constant to all components. If v is an eigenvector, then cv, where c is a scalar, is also an eigenvector, as it follows directly from the definition of eigenvalue. It is possible that several eigenvectors can be associated with the only eigenvalue of a matrix, i.e. eigenvectors may not be identical up to an additional constant as shown in Example .16.

5 A Stochastic Extension

The evolution equation studied in this section is

$$x(k+1) = A(k) \otimes x(k), \ k = 0, 1, 2, \ldots, \tag{29}$$

with some initial condition $x(0)$. Some (or all) entries of $A(k)$ are stochastic. We assume that

- the underlying distribution functions do not depend on k.

- these stochastic entries can assume only a finite number of different values. It will also be assumed that these values are finite, though the method to be described can be generalized to the case that $-\infty$ is also allowed as a value.
- $A(k)$ and $A(l)$ are independent stochastic matrices for $k \neq l$. Problems where $A(k)$ and $A(k+1)$ are correlated can be treated also, provided there exists a model with the Markov property that describes the evolution of $A(k)$, $k = 0, 1, \ldots$. In such a case the latter model would be added to (29) and the theory to be described should be applied to this augmented model. For reason of simplicity, we will not explicitly deal with such problems.
- no correlation between stochastic entries of $A(k)$ exists, though such correlations can be treated rather routinely.
- $\mathcal{G}(A(k))$ is strongly connected. (If this assumption is true for one k, it automatically is true for all k due to the second assumption above.)

The quantity of central interest in this section is

$$\lim_{k \to \infty} E(x_i(k+1) - x_i(k)), \tag{30}$$

for an arbitrary i, being the average cycle time for component i. This quantity is a kind of 'average cycle time'; it can been proved [1] that this average cycle time is independent of i. The method of calculation of the average cycle time will be shown by means of a simple example. Consider the case that $x \in R^2$ and that for each k the matrix A is one of the following two matrices

$$\begin{pmatrix} 3 & 7 \\ 2 & 4 \end{pmatrix}, \begin{pmatrix} 3 & 5 \\ 2 & 4 \end{pmatrix}.$$

Both matrices occur with probability 1/2 and there is no correlation in time. Starting from an arbitrary $x(0)$-vector, say $x(0) = (0,2)'$, we will set up the reachability tree of all possible states x. This is indicated the following table, being a table of transitions. In order to get a concise notation, the different state vectors are indicated by n_i, $i = 1, \ldots$. The table has been obtained in the following way. The starting point is $n_1 \stackrel{\text{def}}{=} (0,2)'$. From there, two states can be reached in one step: $(9,6)'$ or $(7,6)'$, depending on which A-matrix occurs. The states will be normalized such that the first component equals zero. This results in $(0,-3)'$ and $(0,-1)'$. (Other normalizations are

Table 2: Transitions of stochastic states

initial state	$a_{12} = 7$	$a_{12} = 5$
$n_1 = (0,2)'$	$n_2 + 91$	$n_3 + 71$
$n_2 = (0,-3)'$	$n_4 + 41$	$n_3 + 31$
$n_3 = (0,-1)'$	$n_2 + 61$	$n_3 + 41$
$n_4 = (0,-2)'$	$n_2 + 51$	$n_3 + 31$

possible, and they will lead to the same results.) Both states are new and are therefore added to the list, as n_2 and n_3 respectively. Now we take n_2 as the starting point. Two states can be reached from there: $(4,2)'$ and $(3,2)'$, or, after normalization, $(0,-2)'$ and $(0,-1)'$. Only the first of these states is new and will be added to the list as the next state n_4. In this way we continue: from all states obtained sofar we construct the states which can be reached from there in one step. If a state is found which did not exist sofar, it is added to the list. For the current example it turns out that there exist four different states. The notation $n_i + j1$ in the table refers to the state n_i of which all components are increased by the number j. One directly notices, by viewing the table, that the system never returns to n_1. Hence this node is a transient one. In the stationary situation a Markov chain results with the three states n_2, n_3 and n_4. Let us be slightly more explicit. The elements of this Markov chain, to be denoted by $z(k)$, are, by construction,

$$z(k) = \begin{pmatrix} 0 \\ x_2(k) - x_1(k) \end{pmatrix}.$$

It is easily shown that

$$z(k+1) = \begin{pmatrix} 0 \\ (Ax(k))_2 - (Ax(k))_1 \end{pmatrix} = \begin{pmatrix} 0 \\ (Az(k))_2 - (Az(k))_1 \end{pmatrix},$$

and hence the process $\{z(k)\}$ is indeed Markovian. The transition matrix of the Markov chain is

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix}.$$

The stationary distribution of this chain is easily calculated to be

$$\Pr(n_2) = 1/3,\ \Pr(n_3) = 1/2,\ \Pr(n_4) = 1/6.$$

The average cycle time becomes

$$\begin{aligned}\Pr(n_2)(4\Pr(A_1) + 3\Pr(A_2)) + \Pr(n_3)(6\Pr(A_1) + 4\Pr(A_2)) \\ + \Pr(n_4)(5\Pr(A_1) + 3\Pr(A_2)) = 13/3,\end{aligned}$$

where the coefficients are the appropriate numbers out of the table above. The first term in this expression for instance, $\Pr(n_2)(4\Pr(A_1 + 3\Pr(A_2))$, is obtained as follows. If the state is in n_2, then this happens with (stationary) probability $\Pr(n_2)$. The next step either leads to n_4, with probability $\Pr(A_1)$ and obtained after 4 time units (see Table 2), or it leads to n_3, with probability $\Pr(A_2)$ and obtained after 3 time units. The other terms are obtained similarly. It is the quantity at the right-hand side,13/3, which equals the expression in (30).

This example described a method to calculate the average cycle time. The crucial feature in this method is that the number of different normalized state vectors is finite.

6 Counter versus Dater Description

The new concepts of counter and dater descriptions will be explained by means of the equation

$$x(k+1) = M \otimes x(k), \tag{31}$$

where M equals the matrix introduced in (26). The variable $x_i(k)$ denotes the time instant at which output i delivers its k-th item (being a product or a machine). We now introduce a quantity $\chi_i(t)$ which is related to $x_i(k)$. The argument t of $\chi_i(t)$ refers to the actual clocktime and $\chi_i(t)$ itself refers to the number of times that output i has delivered an item up to (and including) time t. The quantity χ_i can henceforth only assume the values $0, 1, 2, \ldots$. Considering the numerical values of the entries of M, it easily follows that

$$\chi_1(t) = \min\left(\chi_1(t-6)+1, \chi_5(t-6)+1, \chi_6(t-5)+1\right) .$$

For χ_2 one similarly obtains

$$\begin{aligned}\chi_2(t) \;=\; & \min(\chi_1(t-9)+1, \chi_2(t-8)+1, \chi_4(t-8)+1, \\ & \chi_5(t-9)+1, \chi_6(t-8)+1),\end{aligned}$$

etc. One can compactly write

$$\chi(t) = \overline{A}_1 \otimes \chi(t-1) \oplus \overline{A}_2 \otimes \chi(t-2) \oplus \cdots \oplus \overline{A}_l \otimes \chi(t-l) \qquad (32)$$

for some finite l ($l = 10$ in the example with the M-matrix). This latter equation must be read in the so-called min-plus algebra. This min-plus algebra equals the max-plus algebra, except that everywhere where the max-operation appears in the max algebra, one now must read the min-operation. Equation (32) is a higher order difference equation in the min-plus algebra. It can be transformed to a first order differential equation similar to the way as explained in Section 2.6 for scalar higher order equations in the max-plus algebra. Equation (31) (in the max-plus algebra) and Equation (32) (in the min-plus algebra) describe the same system. Equation (32) is referred to as the *counter description* and the other one as the *dater description.* The word 'dater' must be understood as 'timer', but since the word 'time' and its declinations are already used in various ways, the word 'dater' is used. The awareness of these two different descriptions for the same problem has far reaching consequences for the theory of discrete event systems.

The reader should contemplate that the stochastic problem (in which some of the a_{ij} are stochastic) is not very suitable to be given in the counter description, since then the delays in (32) would be stochastic.

7 The z-Transform

Conventional linear systems with inputs and outputs are of the form (13), though (13) itself has the max-plus algebra interpretation. This equation, now considered in the conventional way, is a representation of a linear system in the time domain. Its representation in the z-domain equals

$$Y(z) = C(zI - A)^{-1}BU(z) \ ,$$

where $Y(z), U(z)$ are defined by

$$Y(z) = \sum_{i=0}^{\infty} y(i)z^{-i}, \quad U(z) = \sum_{i=0}^{\infty} u(i)z^{-i} \ ,$$

where it is tacitly assumed that the system was at rest for $t \leq 0$ and where I refers to the unit matrix in the conventional algebra. The matrix $H(z) \stackrel{\text{def}}{=} C(zI - A)^{-1}B$ is called the transfer matrix of the system. The notion of

transfer matrix is especially useful when subsystems are combined to build larger sytems, by means of parallel, series and feedback connections, see [6].

In the max-plus algebra context, the z-transform also exists, but here it is customary to refer to it as the γ-transform where γ operates as z^{-1}. For instance, the γ-transform of u is defined as

$$U(\gamma) = \bigoplus_{i=0}^{\infty} u(i) \otimes \gamma^i \ ,$$

and $Y(\gamma)$ and $X(\gamma)$ are defined likewise. Multiplication of (14) by γ^k yields

$$\left.\begin{array}{rcl} \gamma^{-1}x(k+1)\gamma^{k+1} & = & A \otimes x(k)\gamma^k \oplus B \otimes u(k)\gamma^k, \\ y(k)\gamma^k & = & C \otimes x(k)\gamma^k \ . \end{array}\right\} \tag{33}$$

If these equations are summed with respect to $k = 0, \ldots,$ and if we add $\gamma^{-1}x_0$ to both sides then we obtain

$$\left.\begin{array}{rcl} \gamma^{-1}X(\gamma) & = & A \otimes X(\gamma) \oplus B \otimes U(\gamma) \oplus \gamma^{-1}x_0 \ , \\ Y(\gamma) & = & C \otimes X(\gamma) \ . \end{array}\right\} \tag{34}$$

The first of these equations can be solved by first multiplying (max-plus algebra), equivalently adding (conventional), left- and right-hand side by γ and then repeatedly substituting the right-hand side for $X(\gamma)$ within this right-hand side. This results in

$$X(\gamma) = (\gamma A)^*(\gamma B U(\gamma) \oplus x_0) \ .$$

Thus we obtain $Y(\gamma) = H(\gamma)U(\gamma)$, provided that $x_0 = \varepsilon$, and where the transfer matrix $H(\gamma)$ is defined by

$$H(\gamma) = C \otimes (\gamma A)^* \otimes \gamma \otimes B = \gamma CB \oplus \gamma^2 CAB \oplus \gamma^3 CA^2B \oplus \cdots \tag{35}$$

The expression $Y(\gamma) = H(\gamma)U(\gamma)$ is the max-plus algebra equivalent of $Y(z) = H(z)U(z)$ in the conventional system theory. It can also conveniently be used for building larger systems from subsystems. The transfer matrix is defined by means of an infinite series and the convergence depends on the value of γ. If the series is convergent for $\gamma = \gamma'$, then it is also convergent for all γ's which are smaller than γ'. If the series does not converge, it still has a meaning as a formal series.

Exactly as in conventional system theory, the product of two transfer matrices (of which it is tacitly assumed that the sizes of these matrices is

such that the multiplication is possible), is a new transfer matrix which refers to a system which consists of the original systems put in a series connection. In the same way, the sum of two transfer matrices refers to two systems put in parallel. This section will be concluded by an example of such a parallel connection.

We are given two systems. The first one is given in (15), and is characterized by the 1×1 transfer matrix

$$H_1 = \varepsilon\gamma \oplus 11\gamma^2 \oplus 14\gamma^3 \oplus 20\gamma^4 \oplus 24\gamma^5 \oplus 29\gamma^6 \oplus \cdots$$

It is easily shown that this series converges for $\gamma \leq 4.5$; this bound on γ corresponds to the eigenvalue of A. The second system is given by

$$x(k+1) = \begin{pmatrix} e & \varepsilon & 4 \\ 1 & 1 & \varepsilon \\ \varepsilon & 6 & 3 \end{pmatrix} x(k) \oplus \begin{pmatrix} \varepsilon \\ 2 \\ e \end{pmatrix} u(k) \ ,$$

$$y(k) = \begin{pmatrix} 1 & 1 & 4 \end{pmatrix} x(k) \ ,$$

and its transfer matrix is

$$H_2 = 4\gamma \oplus 12\gamma^2 \oplus 15\gamma^3 \oplus 18\gamma^4 \oplus 23\gamma^5 \oplus 26\gamma^6 \oplus \cdots$$

The transfer matrix of the two systems put in parallel has size 1×1 again (one can talk about a transfer function) and is obtained as

$$H_{\text{par}} = H_1 \oplus H_2 = 4\gamma \oplus 12\gamma^2 \oplus 15\gamma^3 \oplus 20\gamma^4 \oplus 24\gamma^5 \oplus 29\gamma^6 \oplus \cdots \qquad (36)$$

A transfer function can easily be visualized. If $H(\gamma)$ is a scalar function, i.e. the system has one input and one output, then it is a continuous and piecewise linear function. As an example, the transfer function of the parallel connection considered above is pictured in Figure 7.

Above it was shown how to derive the transfer matrix of a system if the representation of the system in the 'time domain' is given. This time domain representation is characterized by the matrices A, B and C. Now one could pose the opposite question; how to obtain a time domain representation, or equivalently, how to find A, B and C if the transfer matrix is given. A partial answer to this question is given in [7]. For the example above, one would like to obtain a time domain representation of the two systems put in parallel starting from (36). This avenue will not be pursued here.

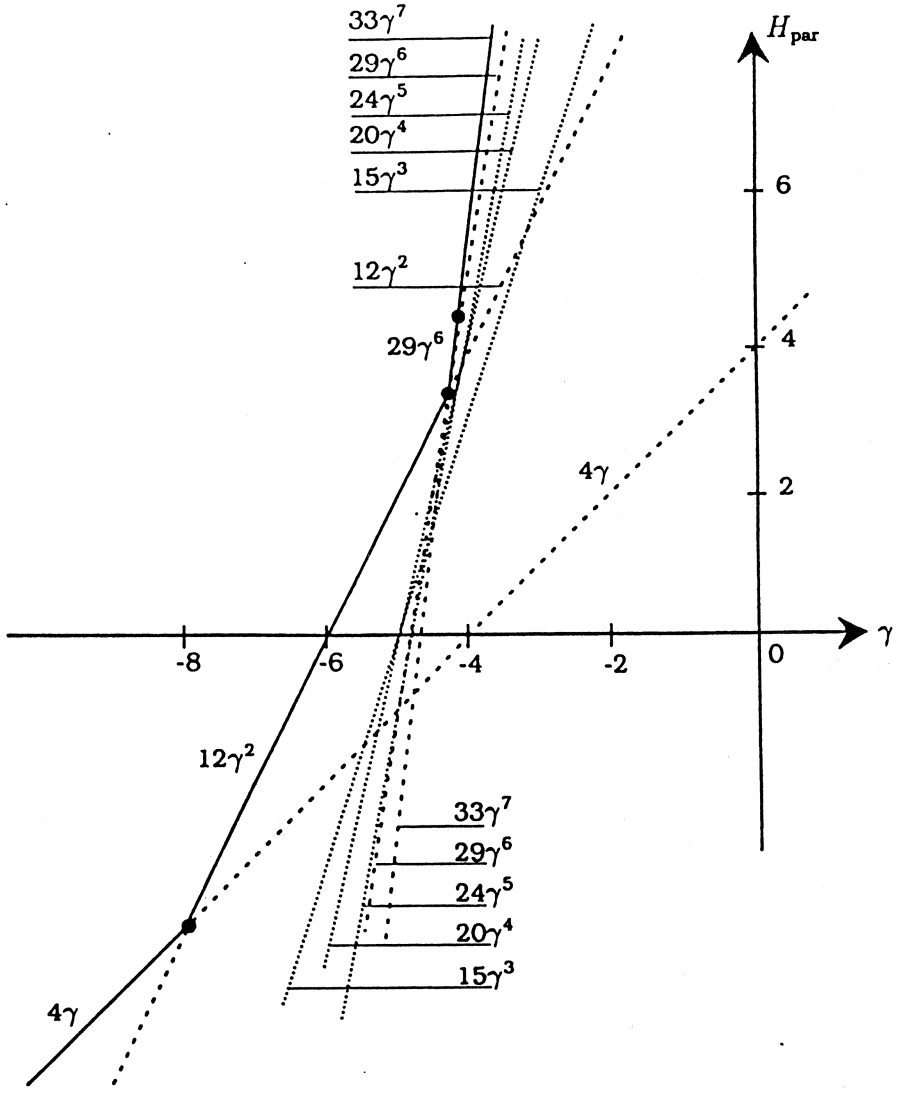


Figure 7: The transfer function H_{par} as a function of γ

8 Conclusions

This paper has shown some recent and exciting developments in the theory of discrete event dynamic systems (DEDS). The theory came into existence around the beginning of the years eighty, though the max-plus algebra was studied earlier, see [3]. The theory of DEDS is based on some other disciplines in (applied) mathematics, specifically on linear systems theory (see [6]) and graph theory (see [4]). The theory of Petri nets is a related topic for which [8] is a good introduction. A comprehensive treatment of DEDS can be found in [1]. The current paper has been based on Chapter 1 of this book for an essential part. In this reference, as in [3], the underlying algebra is not limited to the max-plus algebra, but rather an axiomatic and more abstract point of view is given. It is believed that the development of the theory is only in its childhood and that many more results will follow. The applications of the theory to practical problems is partly parallel to the applications of timed Petri nets. The applications specifically lie in the areas of parallel and distributed processing, where synchronization plays a role. An application to time-table dependent transportation networks is given in [2].

References

[1] F. Baccelli, G. Cohen, G.J. Olsder, and J.P.Quadrat. *Synchronization and Linearity.* John Wiley, 1992.

[2] J.G. Braker. Max-algebra modelling and analysis of time-table dependent networks. In *Proceedings of the first European Control Conference*, pages 1831–1836. Hermes, Paris, 1991.

[3] R.A. Cuninghame Green. *Minimax Algebra.* Lecture Notes in Economics and Mathematical Systems, no 166. Springer Verlag, 1979.

[4] M. Gondran and M. Minoux. *Graphs and Algorithms.* John Wiley, 1986.

[5] Richard M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Mathematics*, 23:309–311, 1978.

[6] Huibert Kwakernaak and Raphael Sivan. *Linear Optimal Control Systems.* Wiley-Interscience, New York, 1972.

[7] G.J. Olsder and R.E. de Vries. On an analogy of minimal realizations in conventional and discrete- event dynamic systems. In P.Varaiya and A.B. Kurzhanski, editors, *Discrete Event Systems: Models and Applications*, volume 103 of *Lecture Notes in Control and Information Sciences*, pages 149–161. Springer Verlag, Berlin, 1988.

[8] James L. Peterson. *Petri net theory and the modeling of systems.* Prentice Hall, Englewood Cliffs, N.J. 07632, 1981.

Sprekers

Prof.Dr. A.W. Grootendorst
(TU Delft) Aardbeistraat 11; 2564 TM Den Haag; 070-3232936.

Prof.Dr.Ir. M.L.J. Hautus
TU Eindhoven, Fac. Wiskunde en Informatica; Postbus 513, 5600 MB Eindhoven; 040-472628.

Prof.Dr.Ir. A.W. Heemink
Rijkswaterstaat, Dienst Getijdewateren; Postbus 20907, 2500 EX Den Haag; TU Delft, Fac. Technische Wiskunde en Informatica; Postbus 356, 2600 AJ Delft; 015-785813.

Dr.Ir. H.J.C. Huijberts
TU Eindhoven, Fac. Wiskunde en Informatica; Postbus 513, 5600 MB Eindhoven; 040-472750.

Dr. A.G.P.M. Nijst
TU Eindhoven, Fact. Wiskunde en Informatica; Postbus 513, 5600 MB Eindhoven, 040-472813.

Prof.Dr. G.J. Olsder
TU Delft, Fac. Technische Wiskunde en Informatica; Postbus 356, 2600 AJ Delft; 015-781912.

Prof.Dr. J.M. Schumacher
Centrum voor Wiskunde en Informatica; Postbus 4079, 1009 AB Amsterdam; 020-5924090.

Dr. A.A. Stoorvogel
TU Eindhoven, Fac. Wiskunde en Informatica; Postbus 513, 5600 MB Eindhoven, 040-472858.

Dr. J.W. van der Woude
TU Delft, Fac. Technische Wiskunde en Informatica; Postbus 356, 2600 AJ Delft; 015-783834.

J.Th.M. Wijnen
TU Eindhoven, Fac. Wiskunde en Informatica; Postbus 513, 5600 MB Eindhoven, 040-472910.

MC SYLLABI

1.1 F. Göbel, J. van de Lune. *Leergang besliskunde, deel 1: wiskundige basiskennis.* 1965.

1.2 J. Hemelrijk, J. Kriens. *Leergang besliskunde, deel 2: kansberekening.* 1965.

1.3 J. Hemelrijk, J. Kriens. *Leergang besliskunde, deel 3: statistiek.* 1966.

1.4 G. de Leve, W. Molenaar. *Leergang besliskunde, deel 4: Markovketens en wachttijden.* 1966.

1.5 J. Kriens, G. de Leve. *Leergang besliskunde, deel 5: inleiding tot de mathematische besliskunde.* 1966.

1.6a B. Dorhout, J. Kriens. *Leergang besliskunde, deel 6a: wiskundige programmering 1.* 1968.

1.6b B. Dorhout, J. Kriens, J.Th. van Lieshout. *Leergang besliskunde, deel 6b: wiskundige programmering 2.* 1977.

1.7a G. de Leve. *Leergang besliskunde, deel 7a: dynamische programmering 1.* 1968.

1.7b G. de Leve, H.C. Tijms. *Leergang besliskunde, deel 7b: dynamische programmering 2.* 1970.

1.7c G. de Leve, H.C. Tijms. *Leergang besliskunde, deel 7c: dynamische programmering 3.* 1971.

1.8 J. Kriens, F. Göbel, W. Molenaar. *Leergang besliskunde, deel 8: minimaxmethode, netwerkplanning, simulatie.* 1968.

2.1 G.J.R. Förch, P.J. van der Houwen, R.P. van de Riet. *Colloquium stabiliteit van differentieschema's, deel 1.* 1967.

2.2 L. Dekker, T.J. Dekker, P.J. van der Houwen, M.N. Spijker. *Colloquium stabiliteit van differentieschema's, deel 2.* 1968.

3.1 H.A. Lauwerier. *Randwaardeproblemen, deel 1.* 1967.

3.2 H.A. Lauwerier. *Randwaardeproblemen, deel 2.* 1968.

3.3 H.A. Lauwerier. *Randwaardeproblemen, deel 3.* 1968.

4 H.A. Lauwerier. *Representaties van groepen.* 1968.

5 J.H. van Lint, J.J. Seidel, P.C. Baayen. *Colloquium discrete wiskunde.* 1968.

6 K.K. Koksma. *Cursus ALGOL 60.* 1969.

7.1 *Colloquium moderne rekenmachines, deel 1.* 1969.

7.2 *Colloquium moderne rekenmachines, deel 2.* 1969.

8 H. Bavinck, J. Grasman. *Relaxatietrillingen.* 1969.

9.1 T.M.T. Coolen, G.J.R. Förch, E.M. de Jager, H.G.J. Pijls. *Colloquium elliptische differentiaalvergelijkingen, deel 1.* 1970.

9.2 W.P. van den Brink, T.M.T. Coolen, B. Dijkhuis, P.P.N. de Groen, P.J. van der Houwen, E.M. de Jager, N.M. Temme, R.J. de Vogelaere. *Colloquium elliptische differentiaalvergelijkingen, deel 2.* 1970.

10 J. Fabius, W.R. van Zwet. *Grondbegrippen van de waarschijnlijkheidsrekening.* 1970.

11 H. Bart, M.A. Kaashoek. H.G.J. Pijls, W.J. de Schipper, J. de Vries. *Colloquium halfalgebra's en positieve operatoren.* 1971.

12 T.J. Dekker. *Numerieke algebra.* 1971.

13 F.E.J. Kruseman Aretz. *Programmeren voor rekenautomaten; de MC ALGOL 60 vertaler voor de EL X8.* 1971.

14 H. Bavinck, W. Gautschi, G.M. Willems. *Colloquium approximatietheorie.* 1971.

15.1 T.J. Dekker, P.W. Hemker, P.J. van der Houwen. *Colloquium stijve differentiaalvergelijkingen, deel 1.* 1972.

15.2 P.A. Beentjes, K. Dekker, H.C. Hemker, S.P.N. van Kampen, G.M. Willems. *Colloquium stijve differentiaalvergelijkingen, deel 2.* 1973.

15.3 P.A. Beentjes, K. Dekker, P.W. Hemker, M. van Veldhuizen. *Colloquium stijve differentiaalvergelijkingen, deel 3.* 1975.

16.1 L. Geurts. *Cursus programmeren, deel 1: de elementen van het programmeren.* 1973.

16.2 L. Geurts. *Cursus programmeren, deel 2: de programmeertaal ALGOL 60.* 1973.

17.1 P.S. Stobbe. *Lineaire algebra, deel 1.* 1973.

17.2 P.S. Stobbe. *Lineaire algebra, deel 2.* 1973.

17.3 N.M. Temme. *Lineaire algebra, deel 3.* 1976.

18 F. van der Blij, H. Freudenthal, J.J. de Iongh, J.J. Seidel, A. van Wijngaarden. *Een kwart eeuw wiskunde 1946-1971, syllabus van de vakantiecursus 1971.* 1973.

19 A. Hordijk, R. Potharst, J.Th. Runnenburg. *Optimaal stoppen van Markovketens.* 1973.

20 T.M.T. Coolen, P.W. Hemker, P.J. van der Houwen, E. Slagt. *ALGOL 60 procedures voor begin- en randwaardeproblemen.* 1976.

21 J.W. de Bakker (red.). *Colloquium programmacorrectheid.* 1975.

22 R. Helmers, J. Oosterhoff, F.H. Ruymgaart, M.C.A. van Zuylen. *Asymptotische methoden in de toetsingstheorie; toepassingen van naburigheid.* 1976.

23.1 J.W. de Roever (red.). *Colloquium onderwerpen uit de biomathematica, deel 1.* 1976.

23.2 J.W. de Roever (red.). *Colloquium onderwerpen uit de biomathematica, deel 2.* 1977.

24.1 P.J. van der Houwen. *Numerieke integratie van differentiaalvergelijkingen, deel 1: eenstapsmethoden.* 1974.

25 *Colloquium structuur van programmeertalen.* 1976.

26.1 N.M. Temme (ed.). *Nonlinear analysis, volume 1.* 1976.

26.2 N.M. Temme (ed.). *Nonlinear analysis, volume 2.* 1976.

27 M. Bakker, P.W. Hemker, P.J. van der Houwen, S.J. Polak, M. van Veldhuizen. *Colloquium discretiseringsmethoden.* 1976.

28 O. Diekmann, N.M. Temme (eds.). *Nonlinear diffusion problems.* 1976.

29.1 J.C.P. Bus (red.). *Colloquium numerieke programmatuur, deel 1A, deel 1B.* 1976.

29.2 H.J.J. te Riele (red.). *Colloquium numerieke programmatuur, deel 2.* 1977.

30 J. Heering, P. Klint (red.). *Colloquium programmeeromgevingen.* 1983.

31 J.H. van Lint (red.). *Inleiding in de coderingstheorie.* 1976.

32 L. Geurts (red.). *Colloquium bedrijfsystemen.* 1976.

33 P.J. van der Houwen. *Berekening van waterstanden in zeeën en rivieren.* 1977.

34 J. Hemelrijk. *Oriënterende cursus mathematische statistiek.* 1977.

35 P.J.W. ten Hagen (red.). *Colloquium computer graphics.* 1978.

36 J.M. Aarts, J. de Vries. *Colloquium topologische dynamische systemen.* 1977.

37 J.C. van Vliet (red.). *Colloquium capita datastructuren.* 1978.

38.1 T.H. Koornwinder (ed.). *Representations of locally compact groups with applications, part I.* 1979.

38.2 T.H. Koornwinder (ed.). *Representations of locally compact groups with applications, part II.* 1979.

39 O.J. Vrieze, G.L. Wanrooy. *Colloquium stochastische spelen.* 1978.

40 J. van Tiel. *Convexe analyse.* 1979.

41 H.J.J. te Riele (ed.). *Colloquium numerical treatment of integral equations.* 1979.

42 J.C. van Vliet (red.). *Colloquium capita implementatie van programmeertalen.* 1980.

43 A.M. Cohen, H.A. Wilbrink. *Eindige groepen (een inleidende cursus).* 1980.

44 J.G. Verwer (ed.). *Colloquium numerical solution of partial differential equations.* 1980.

45 P. Klint (red.). *Colloquium hogere programmeertalen en computerarchitectuur.* 1980.

46.1 P.M.G. Apers (red.). *Colloquium databankorganisatie, deel 1.* 1981.

46.2 P.G.M. Apers (red.). *Colloquium databankorganisatie, deel 2.* 1981.

47.1 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60: general information and indices.* 1981.

47.2 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 1: elementary procedures; vol. 2: algebraic evaluations.* 1981.

47.3 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 3A: linear algebra, part 1.* 1981.

47.4 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 3B: linear algebra, part II.* 1981.

47.5 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 4: analytical evaluations; vol. 5A: analytical problems, part 1.* 1981.

47.6 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 5B: analytical problems, part II.* 1981.

47.7 P.W. Hemker (ed.). *NUMAL, numerical procedures in ALGOL 60, vol. 6: special functions and constants; vol. 7: interpolation and approximation.* 1981.

48.1 P.M.B. Vitányi, J. van Leeuwen, P. van Emde Boas (red.). *Colloquium complexiteit en algoritmen, deel 1.* 1982.

48.2 P.M.B. Vitányi, J. van Leeuwen, P. van Emde Boas (red.). *Colloquium complexiteit en algoritmen, deel 2.* 1982.

49 T.H. Koornwinder (ed.). *The structure of real semisimple Lie groups.* 1982.

50 H. Nijmeijer. *Inleiding systeemtheorie.* 1982.

51 P.J. Hoogendoorn (red.). *Cursus cryptografie.* 1983.