Control theory:

Information in control and regulation

CONTROL THEORY

<u>III. ANALYSIS OF CLOSED LOOP SYSTEMS.</u>

<u>A. Introduction.</u>

1. What is a closed system?

a) Negative feedback.

b) Positive feedback.

2. Schematical representation of a control system.

3. Types of regulators.

a) Continuous regulators.

(1) Proportional regulators (P-regulators).

(2) Differentiating regulators (D-regulators).

(3) Integrating regulators (I-regulators).

(4) Combinations (PI, PD, PID) of these regulators.

b) Discontinuous regulators.

(1) Two or more state regulators.

(2) Regulators with constant velocity.

4. Types of feedback.

a) Proportional feedback.

b) Differentiating feedback.

c) Integrating feedback.

d) Combinations.

<u>B. System analysis.</u>

1. General properties of control systems.

a) Schematic representation.

b) Equation of closed loop systems.

c) Properties.

(1) Gain.

(2) Sensitivity of systems to parameter variation.

(a) Sensitivity of the open system.

(b) Sensitivity of the closed system.

(i) Sensitivity to G.

(ii) Sensitivity to H.

(3) Follower behavior for H = 1.

(4) Approximation.

2. Applications to different systems.

a) Proportional regulators ($\varepsilon_n = k\varepsilon$).

(1) Proportional feedback.

(a) Zero order process.

(i) Equation.

(ii) Properties.

(b) First order process (with disturbance).

(i) Differential equation.

(ii) Properties.

(c) Second order process.

(i) Differential equation.

(ii) Properties.

(2) Differentiating feedback.

Second order process.

(i) Differential equation.

(ii) Properties.

(3) Integrating feedback.

(i) Differential equation.

(ii) Properties.

(4) Combined feedback.

(i) Differential equation.

(ii) Properties.

b) Differentiating regulators ($\varepsilon_n = k \,.\, d\varepsilon/dt$).

c) Integrating regulators ($\varepsilon_n = k \,.\int \varepsilon dt$).

d) All or none regulator ($\varepsilon_n = 0$ or 1).

CONTROL THEORY.

INTRODUCTION.

A. Object of control theory.

Control theory studies the techniques to restrain physical processes to operate in a predetermined manner (according to a certain plan), by means of feedback. E.g. automatic pilot, thermostat. It thus studies how information is used (or can be used) to regulate and steer processes.
A study of the theory in the framework of physiology is very relevant since a large number of processes in biological systems including human organisms are regulated by feedback.
E.g. homeostasis, reflexes, voluntary movement.
The control theory is multidisciplinary. It is applicable to all systems where mutual interactions of some of the variables determine or influence the behavior of system. The theory can thus also be applied to life sciences such as physiology, psychology, economy, sociology, etc. The theory is based on causal relationships. A process is represented as a black box with as input the cause and as output the consequence.

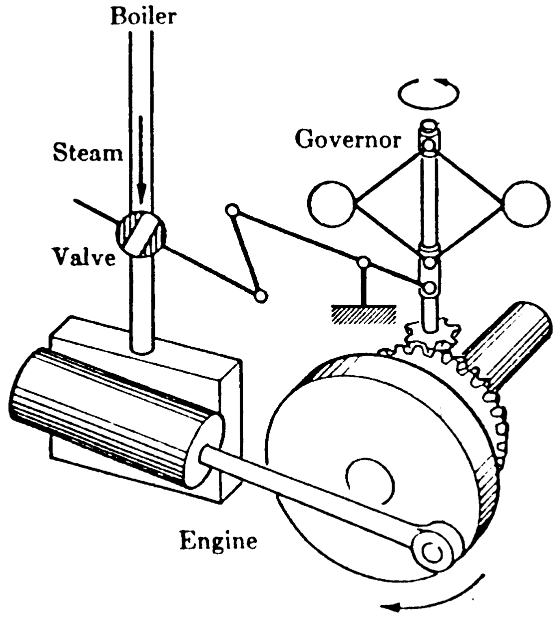
B. Overview of history of control theory.

1. Antiquity.

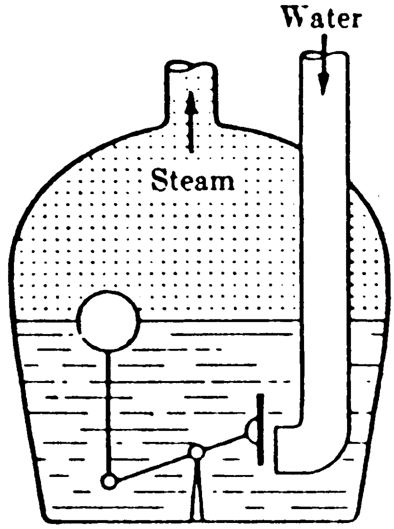
First applications of control mechanisms: float regulator:
water clock of Ktesibios Greece 0-300 BC
oil lamp of Philon of Byzantium 250 BC
"Pneumatica" of Heron of Alexandria 100-0 BC

2. New Age.

Applications:



Watt flyball governor.



Water-level float regulator.

Applications:
temperature regulator of Drebbel (Netherlands 1600)
steam regulator of Papin (France 1700)
flywheel of Watt (England 1769)
float regulator of Polzunow (Russia 1765)

3. Modern times.

a) Mathematical formulation using differential equations.

Maxwell 1868
Vyshnegradskii 1877
Ljapunov 1892
Stodola 1839
Hurwitz 1859

- Graphical methods.
Tolle 1895

4. 20nd century.

a) Before the 2nd world war.

U.S.A.: developments in the frequency domain; starting from the development of the telephone and the electronic amplifier (mainly in the Bell laboratories by Bode, Nyquist, Black)
Russia: analysis mainly in time domain (differential equations); from mathematics and applied mechanics

b) 1940-1950.

Important expansion: automatic control of military equipment (gun, radar, automatic pilot)
mainly in frequency domain
use of Laplace transform and of complex frequency

c) 1950-1960.

More possibilities from use of analog and digital computers:
development of "s-plane" technique
root locus method

d) 1960-1970.

Stimulation from space travel
analysis complex systems
theory of optimal control
mainly time domain (Ljapunov, Minorski, Pontryagin, Bellman)

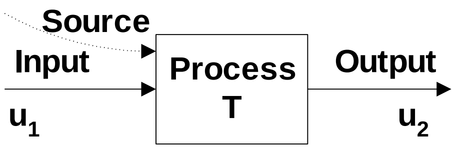
e) 1970 – present time.

Starting from rationalization of industrial processes
optimal control
adaptive processes
"learning systems"

I. DESCRIPTION OF OPEN LOOP SYSTEMS.

A system consists of a combination of a number of physical processes. A *process can be defined as* an e*ffect on a variable of a certain physical process, the quantitative value of which is hereby changed* (e.g. physical process in which the flow of water in a tube is influenced by the position of a valve).

A. Schematic representation of a process.



E.g.: process: change of flow by valve.
source: potential energy of water
output: water flow in tube
input: position of valve

The valve is a regulator: a well-defined relation exists between the flow and the position of the valve.

B. The transfer function of a process.

For each process there is a relation between the output and the input: u_2 is a function of u_1. This function is called the *transfer function* T *of the process.* T *is the operation that is applied by the system on the input*: application of this operation on an input u_1 results in an output u_2.

$\mathbf{u_2 = T\, u_1}$ and thus $T = u_2 / u_1$

N.B.: In mathematically correct form the transfer function is the ratio of the Laplace transforms of output and input, implying that T is an operation. E.g. for an ideal amplifier with gain 5, T is not the number 5, but the operation “x 5” (thus “multiply by 5”)

This transfer function (T) completely determines the dynamic behavior of the process. When the transfer function of a process is given, the output of the process can be calculated for each possible input applied to the process. The transfer function describes the system independent of the input. Therefore the aim of a theoretical or experimental analysis of a process is to determine the transfer function of the process.

C. Influence of a disturbance on a process.

The transfer function of a certain process is not necessarily constant. E.g., wear can change the transfer function. All non-controllable causes of changes in the relation between output an input are called disturbances. It is clear that disturbances have an adverse influence on the outputs variable, since they will cause deviations from the desired output value. It is in the solution of this problem that feedback plays an important role. In feedback control systems, the error of the output as a consequence of (permanent or temporary) changes of T are measured, and the information of this error is used to correct the output of the system by changing the input of the system (adjusting the regulator). This can occur very fast, continuously and automatically, so that already a very small deviation of the output causes a correction of the setting of the regulator, which ensures that the output follows the desired value with high accuracy.

D. Transfer function and differential equation of a process.

The relation between the output and the input of a process can be described by a differential equation of n^{th} order. This equation completely describes the dynamic behavior of the process.

The differential equations contain, in addition to algebraic functions of the input and output variables, also time derivatives (rates, accelerations...) of these variables. Integrals of the variables do not occur in the equations, since they can always be mathematically removed by differentiating the left and right side of the equation.

In this part of the course we limit ourselves to processes with transfer functions that are independent of the amplitude of the input u_1 of the process (within the limit of the working range of the system), and with properties that do not change with the time (a brief introduction to more complex system will be given in a later part). Such linear time-independent (LTI) systems can be described by linear differential equations with constant coefficients; these equations have the form:

$$\mathbf{a_n . \frac{d^n u_2}{dt^n} + a_{n-1} . \frac{d^{n-1} u_2}{d t^{n-1}} + \ldots + a_0 . u_2 = u_1}$$

Replacing the symbols indicating the derivative by the operator symbol D (defined as $D \equiv d/dt$) these differential equations can be rewritten in operator form as

$$(a_n D^n + a_{n-1} D^{n-1} + \ldots. + a_0)\, u_2 = u_1$$

In this operator notation the symbol therefore D represents the operation "the time derivative of". Therefore, $D^2 = D.D.$ represents the 2nd derivative (not the square of the first derivative!). In operator notation the transfer function becomes:

$$\mathbf{T = \frac{u_2}{u_1} = \frac{1}{a_n D^n + a_{n-1} D^{n-1} + \ldots. + a_0}}$$

The transfer function of a process is thus the operator notation of the differential equation of the process. T is therefore itself an operator.

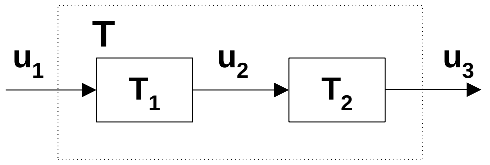
Most handbooks make use of Fourier or Laplace transforms instead of D operators. Within the context of this introductory course however we avoid using these transforms, since they require a more extensive mathematical background, while it is possible to grasp the main conclusions of the control theory without requiring this background.

- In the Fourier notation, the transfer function has the same form as in D operator notation, when D is replaced by an imaginary variable $j\omega$, with $j = \sqrt{-1}$ and ω the frequency in rad/s. A Fourier analysis corresponds to the frequency domain analysis (harmonical analysis) in this course.
- The most common formulation in textbooks makes use of Laplace transforms. This is an extension of the Fourier transform method. In Laplace transform notation, the complex variable $s = \sigma + j\omega$ (with σ a real quantity), takes the place of the imaginary quantity $j\omega$ in Fourier transforms, or of D in the operator notation. The D operator and Fourier notations are special cases of the Laplace operator notation, when the initial values and their derivatives (first and higher order) are all equal to zero.

Analysis using Laplace transforms includes frequency domain analysis as well time domain analysis. In Laplace transformation notation the transfer function becomes an <u>algebraic</u> equation:

$$\mathbf{T = \frac{u_2}{u_1} = \frac{1}{a_n s^n + a_{n-1} s^{n-1} + \ldots. + a_0}}$$

E. <u>Transfer function of 2 processes in series.</u>



$u_2 = T_1 u_1$

$u_3 = T_2 u_2$ Calculate T

$u_3 = T\, u_1$

$u_3 = T u_1 = T_2 u_2 = T_2\,(T_1 u_1) = T_2 T_1 u_1$

$\mathbf{T = T_2\, T_1}$

N.B.: In general operators are non-commutative. $T_2 T_1 \neq T_1 T_2$:

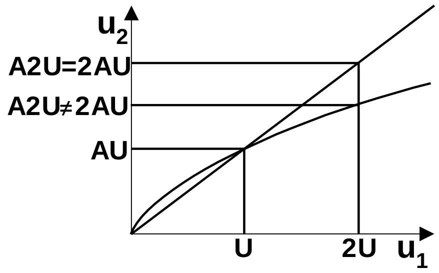
F. Properties of non-ideal processes.

1. Nonlinearity.

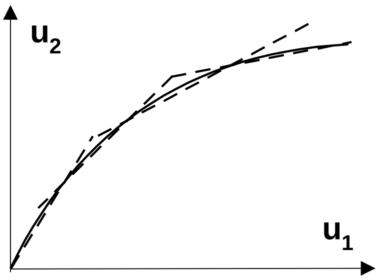
Most physiological systems cannot be described by a linear differential equation.

a) What?

Definition: a system with transfer function T is linear, and can thus be described by a linear differential equation, when it follows the principle of superposition; this is when T(a + b) = Ta + Tb

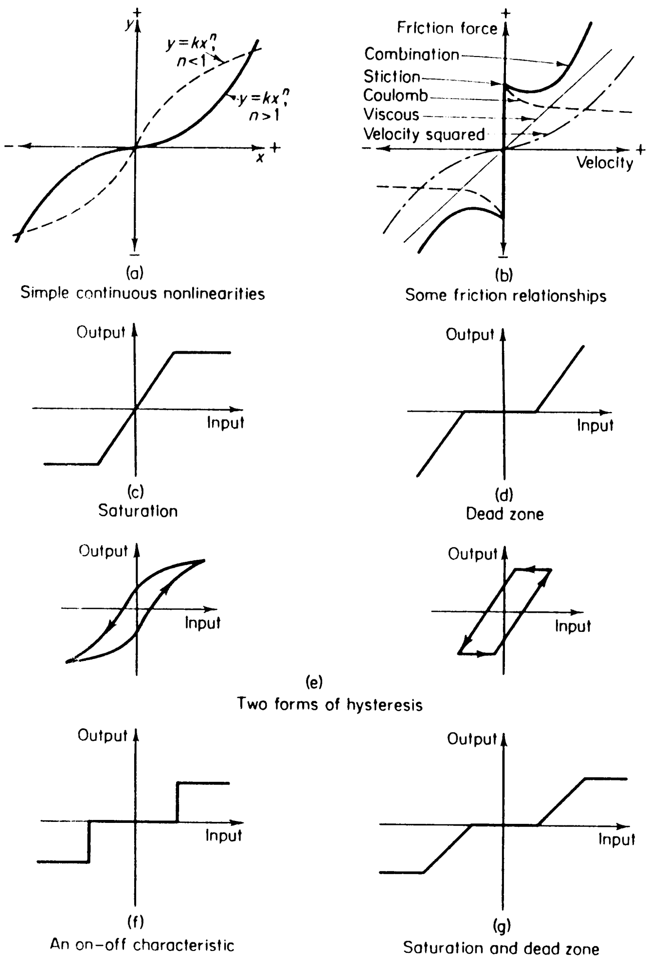


Linear and nonlinear process.

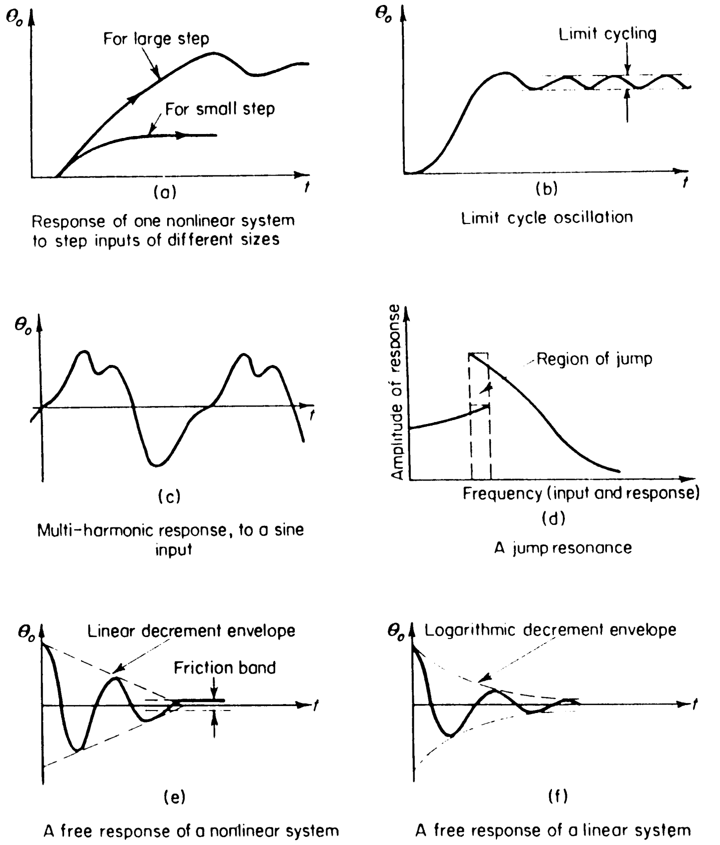


Linear approximation of nonlinear processes.

For nonlinear processes the gain of the transfer function is itself dependent on the amplitude of the input signal. The graph that shows the relation between the amplitude of the output signal and the amplitude of the input signal in steady state is thus nonlinear.



Some common nonlinear relationships.



Some response characteristics of nonlinear systems.

b) Important forms of nonlinearity.

- **Saturation.**

Saturation is a form of nonlinearity where the output cannot exceed a certain value. All real systems have finite saturation levels. The saturation level determines the working range (dynamic range) of the system.

- **Rectification.**

Rectification is a form of nonlinearity where the transfer function is dependent on the sign of the input signal. The curve has another slope for positive values of the input than for negative values.
E.g. one-way valve: the curve has a slope zero for pressure in the wrong direction.

- **Dead zone.**

Dead zone is a form of nonlinearity where the output only changes when the input exceeds a certain value.

- **Hysteresis.**

Hysteresis is a form of nonlinearity where the output corresponding to a particular input can take two values, dependent on the direction (increase or decrease) of the variation of the input.

- **On - Off characteristic.**

On - Off characteristic is a form of nonlinearity where the output can only have 2 possible values: 0 and 1 also called off and on.

- **Combinations.**

All kinds of combinations of these nonlinearities are possible.
E.g. a thermostat is has on - off characteristic with hysteresis.

c) Effect of nonlinearity.

For linear systems, the form of the steady state output of the system has a simple relation to the shape of the input signal, and changing the amplitude of the input signal only causes a proportional change of the amplitude of the output, without otherwise affecting the shape of the output.
In case of nonlinearity, the gain is a function of the amplitude of the input signal. This can cause serious distortions. E.g., upon application of a step function as an input signal, the form of the output signal can be different, dependent on the size of the input step. For sinusoidal inputs signal, the output in steady state is not a simple sinus, but it contains also sinus components of higher frequency than the input signal (harmonical components). Distortion as a consequence of nonlinearity of the system is therefore called harmonic distortion (e.g. in audio amplifiers: output contains some sound components of other frequencies than present in the input signal).

d) Linear approximation of nonlinear systems.

Nonlinearity makes the quantitative (eventually mathematical) analysis of systems much more difficult. In order to be able to use the linear analysis, linear approximation of the system can be used. In this approach the steady state transfer function is approximated by a broken line consisting of a number of straight lines. Each of these parts of the curve can then be (approximately) described by a linear differential equation. The total system is then approximated by a system of linear differential equations each of which are only valid within a certain range of amplitudes of the input signal. The input signals used therefore may not exceed the ranges of the system under study in which particular approximations are valid.

2. Delays.

a) Effect of delay.

In case of a delay in a process, a difference in time between the change of the input and the corresponding change of the output is present.

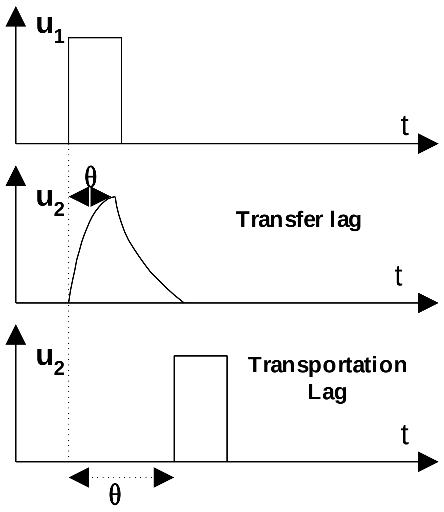
b) Types of delay.

- **Transfer lag.**

This type of delay is due to transformations in the process (by the transfer function).

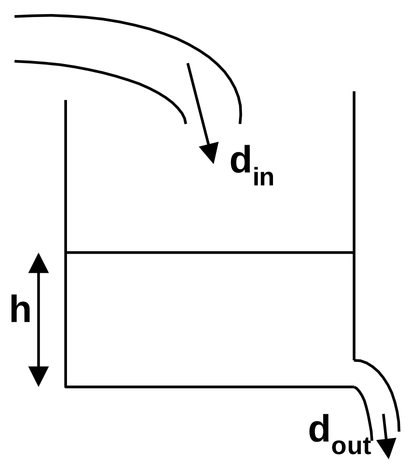
- **Transportation lag.**

This type of delay is caused by the limited rate of propagation of signals between different parts of the system.



G. <u>Derivation of the system transfer function.</u>

Example 1: Filling of a vessel with a hole.



$d_{out} = b.h$

$$\frac{dh}{dt} = a.(d_{in} - d_{out})$$

$$\frac{dh}{dt} = a.(d_{in} - b.h)$$

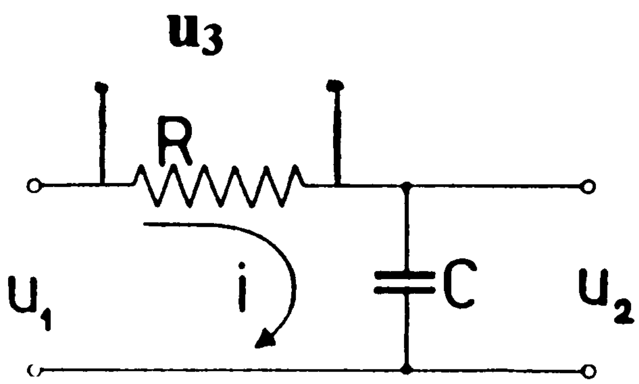
$$\frac{dh}{dt} + a.b.h = a.d_{in}$$

$$\frac{1}{a.b}.\frac{dh}{dt} + h = \frac{1}{b}.\ d_{in}$$

Opl.: steady state (dh/dt =0)

h= d_{in}/b thus since d_{out} = b.h when $d_{out} = d_{in}$

Example 2: Electrical system consisting of resistance and capacitance in series.



$u_1 = u_2 + u_3$

$u_3 = i\ R$

$q = C\ u_2$

$$i = \frac{dq}{dt} = C\,\frac{du_2}{dt}$$

$$u_1 = u_2 + i\ R = u_2 + RC\,\frac{du_2}{dt}$$

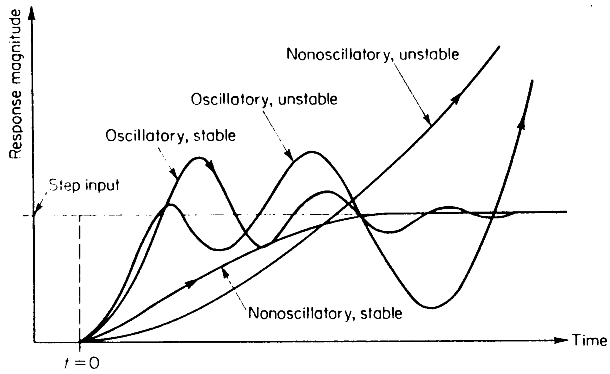
$$RC\,\frac{du_2}{dt} + u_2 = u_1$$

Convention: When writing down the differential equations, we always use the convention of taking the coefficient of u_2 (thus of the zero order term of the output variable) as 1 (this can always be done by dividing left and right side of the equations by the coefficient a_0 of u_2). The reason for using this convention is that this way the parameters of the equations get a much more direct the physical meaning.

II. SYSTEM ANALYSIS.

A. Time domain analysis.

Time domain analysis studies a system by analyzing the response of the system as a function of time to a *change of the input* (most often a stepwise change).
Such a step function can be described as follows: $u_1 = 0$ for $t < 0$, and $u_1 = U$ for $t \geq 0$ (with U a constant).
Upon application of a step function to the input, the output of the system will evolve to a new steady-state or equilibrium. This equilibrium however is not reached instantaneously, but transitory phenomena (transients) will occur. These transients in the output signal will have a different time course dependent on the order of the system and the values of the parameters. The transients are thus dependent on the transfer function of the system, and T can be determined from the study of the transients.



Some possible responses of a system to a step input.

1. First order system.

a) Differential equation.

General form of a 1st order diff. eq.

$$\tau \cdot \frac{du_2}{dt} + u_2 = \alpha \cdot u_1$$

For a step function:

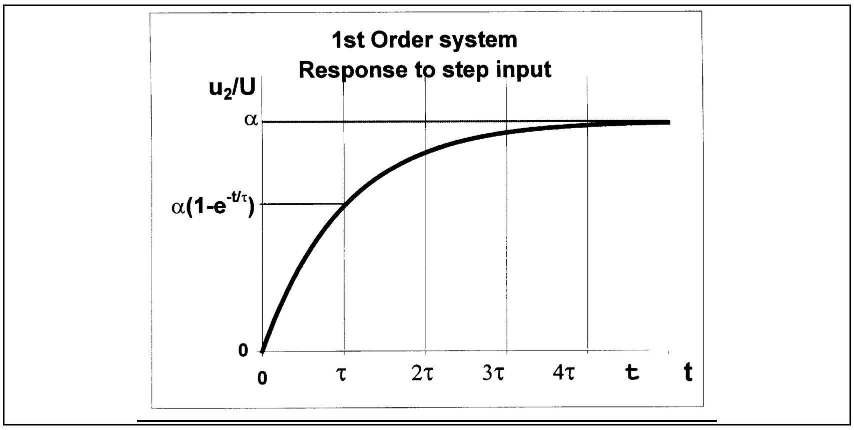
$$\tau \cdot \frac{du_2}{dt} + u_2 = \alpha \cdot U \quad (t \geq 0)$$

E.g.: RC network from previous page: $\alpha = 1$ and $\tau = RC$
Filling of a vessel: $\alpha = 1/b$ and $\tau = 1/ab$

b) Solution.

$u_2 = \alpha U (1 - e^{-t/\tau})$

Graphical representation



c) Meaning of the parameters.

(1) Gain.

$t \to \infty \Rightarrow u_2 \to \alpha U$

The ratio of the steady state output of the system to the input is called the gain A of the system.

$A = u_2(t \to \infty)/u_1 = \alpha U/U = \alpha$

The coefficient α of the term at the right side of the equation is thus the gain of the system, and is dimensionless when u_1 and u_2 are expressed in the same units.

(2) Time constant.

Dimension formula of the diff. eq.:

$$(\dim \tau) \cdot \frac{(\dim u_2)}{t} + \dim u_2 = \dim \alpha \cdot \dim u_1$$

$$\text{thus } \dim \tau \cdot \frac{\dim u_2}{t} = \dim u_2 = \dim \alpha \cdot \dim u_1$$

dim τ is thus time.

When $t = \tau \Rightarrow u_2 = \alpha U(1 - e^{-1})$

$$\frac{u_2(t = \tau)}{u_2(t \to \infty)} = 1 - e^{-1} = 0.632120 \cong 2/3$$

The time constant τ is thus the time required by the system to bring the output to approximately 2/3 of its steady state value. $\tau = t_{1/2} / \ln 2$, and is thus equal to 1.4427 times the half time ($t_{1/2}$).

In order to analyze a first order system, one can thus apply a step function $u_1 = U$ at $t = 0$, and measure the *gain* α *and the time constant* τ of the system. These two numbers *completely determine the dynamic behavior of a first order system*!

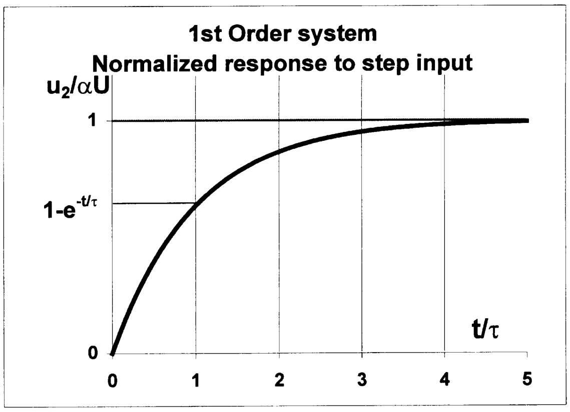
Convention: When plotting the results of a theoretical analysis in a graph, the abscissa and the ordinate are most often normalized. The advantage of normalization is that results that only differ by a scale factor of the independent (t) and / or dependent (u_2) variable are represented by the same graph.

For a first order system:

$$\tau \cdot \frac{du_2}{dt} + u_2 = \alpha \cdot U$$

Substitution $V = u_2/\alpha U$ and $T = t/\tau$:

$$\frac{dV}{dT} + V = 1$$



For a time domain analysis of a first order system most often V is plotted as function of T, thus $u_2/\alpha U$ as function of t/τ. This way one obtains a graph that is valid for all possible values of the parameters. The output is thereby expressed in relative units (%), while the time constant is chosen as the unit of time.

Such type of representation of graphs is so universal that frequently it is not even mentioned that the graph is thus normalized, since most often this is clear from the context. Some of the graphs shown in this course will also use conventional implicit normalization.

E.g.: when in the graph the steady state value is equal to 1, this means most often that the y-axis is normalized, even when as axis variable u_2 is written.

2. Second order system.

a) Differential equation.

General form of 2[nd] order diff. eq.

$$a \cdot \frac{d^2 u_2}{dt^2} + b \cdot \frac{du_2}{dt} + k \cdot u_2 = u_1$$

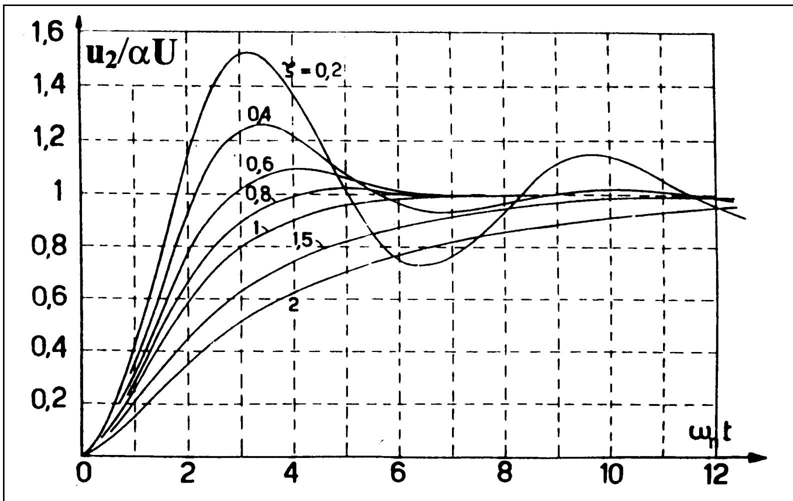
Normalization and writing in the conventional notation (to give the parameters in the equation an explicit physical meaning).

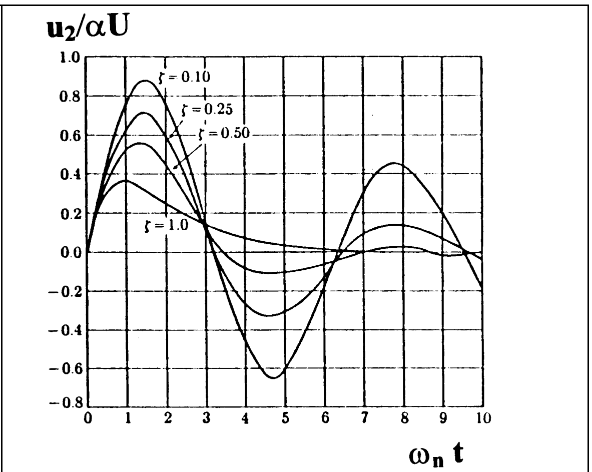
$$\frac{1}{\omega_n^2} \cdot \frac{d^2 u_2}{dt^2} + \frac{2\xi}{\omega_n} \cdot \frac{du_2}{dt} + u_2 = \alpha \cdot u_1$$

For a step function:

$$\frac{1}{\omega_n^2} \cdot \frac{d^2 u_2}{dt^2} + \frac{2\xi}{\omega_n} \cdot \frac{du_2}{dt} + u_2 = \alpha \cdot U \quad (t \geq o)$$

b) Graphical solution.





Response of a second-order system for an impulse function input.

Convention: Often $u_2/\alpha U$ is plotted (sometimes implicitly) as a function of $\omega_n t$.

c) Meaning of the parameters.

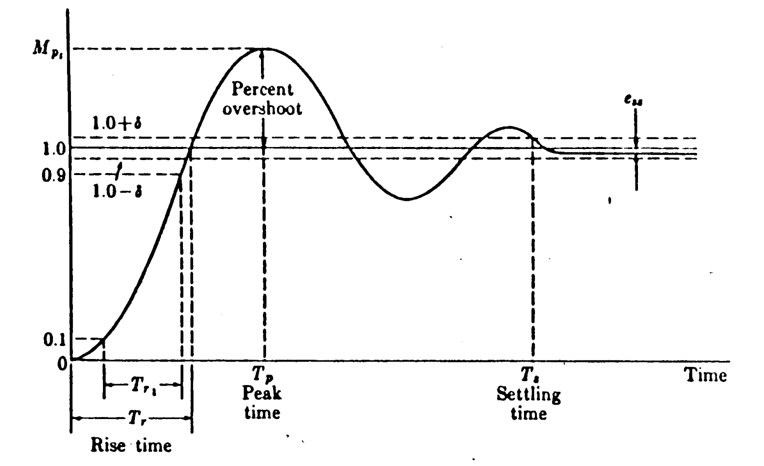
The form of the solution is dependent on the value of ξ:

- for $\xi = 0$ undamped sinusoidal vibration
- for $0 < \xi < 1$ damped sinusoidal vibration
- for $\xi = 1$ critically damped vibration
- for $\xi > 1$ overdamped vibration: u_2 is then a sum of 2 exponential functions

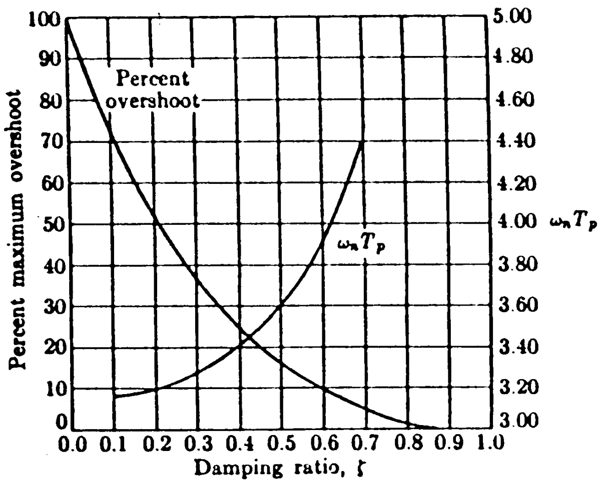
The factor ξ is therefore called the *damping factor of the* system and is a dimensionless quantity. For undamped sinusoidal vibration, $\xi = 0$, and the output of the system is a sinusoidal vibration with a frequency ν_n, which is called the natural frequency (eigen frequency) of the system. $\omega_n = 2\pi\nu_n$ is the *natural pulsation* or natural corner frequency of the system and has as dimension radians per second. For damped vibrations the system goes to a new steady state value αU. Also here, α is the *gain of the* system. *The second order system is thus completely determined by the three parameters* ξ, ω_n and α.

While a time domain analysis enables to quickly make a good estimation of the three parameters of a second order system, it is the experimentally difficult to accurately determine the values of ξ and ω_n by a time domain analysis. In order to provide a simple description of the transient behavior of a second order system most often in addition to the gain other parameters are given that can be easily measured, such as rise time, overshoot, peak time and settling time.

- From the overshoot the damping factor can be calculated:
  P.O. = $100 \exp(-\xi\pi/\sqrt{1 - \xi^2})$ (P.O. overshoot in %).
- The frequency of 'ringing' (= damped vibration = sinusoidal component) is approximately the natural frequency of the system. The amplitude of this damped sinusoidal component decreases exponentially as a function of time with a time constant = $1/\xi\omega_n$.
- The amplitude of the output in steady state enables to calculate the gain.

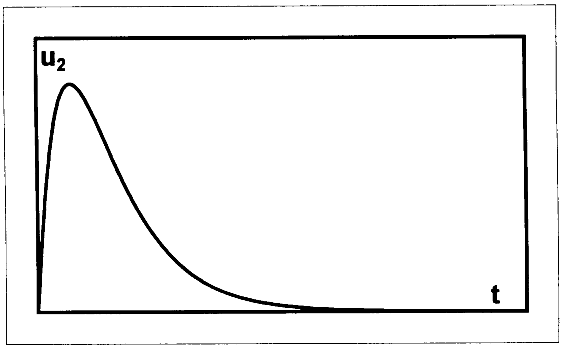


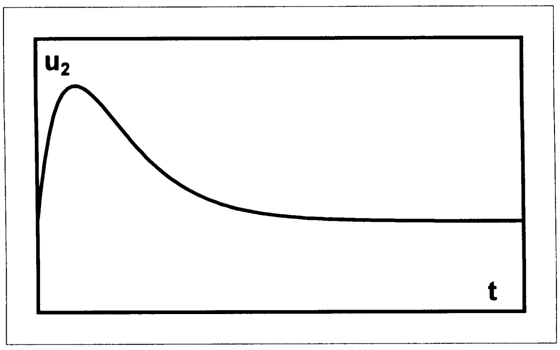
Step response of a control system



3. System of order -1.

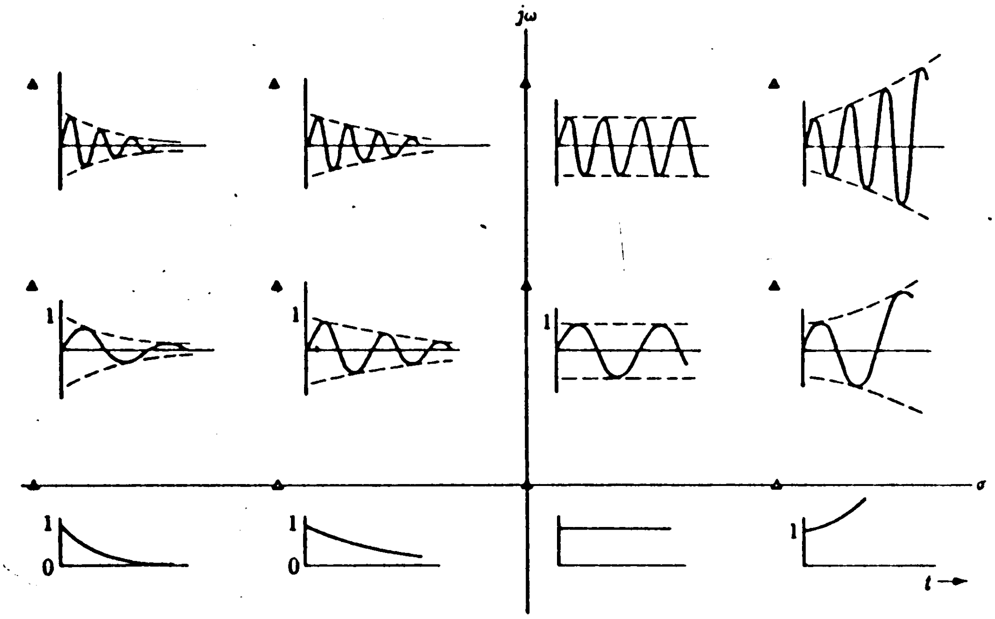
This kind of system is described by an equation that contains the time integral of the output signal. However, by making the time derivative of both sides of the equation one obtains an equation not containing integrals, but which contains the time derivative of the input signal in the right side of the equation. Such a system can thus be considered speedometer. The response to a step function is a theoretically infinitely high, infinitely small peak. Limited response velocities of real physical systems cause a rounding of the peak. When the right side of the equation contains in addition to a differentiating also a proportional component, the output will not decrease to a value of zero in steady state (compare phasic and tonic response of sensory receptors).





4. System of order n.

The transients of time dependent processes that can be described by a linear diff. eq. with constant coefficients are determined by the general solution of the characteristic equation (this is the differential equation in which the right side of the equation is set to zero (homogeneous equation)). These transient components are a sum of n complex exponentials $u_2 = \sum e^{st}$ with $s = \sigma + j\omega$ the roots of the equation. In the solution of the differential equation, the odd order terms produce the real components, which will result in exponential behavior and thereby will cause damping by energy dissipation of the system when σ is negative, or an exponential increase when σ is positive. The even order terms in the diff. eq. are responsible for the imaginary components ($j\omega$) that always occur in pairs with the same amplitude and opposite sign, and which will cause sinusoidal behavior with energy conversions. The transient behavior of a system of higher order can therefore always be more or less considered as a combination of systems of first and second order, this is as a sum of k exponentials (real roots) and of n-k terms (complex roots) that represent sinus waves. The time course of the response depends on the sign and amplitude of the real component, and of the amplitude of the imaginary component of the different roots of the characteristic equation. The real roots and the real components of the complex roots always have to be negative, otherwise one obtains increasing exponentials in the solution of the diff. eq., and thus an unstable system.



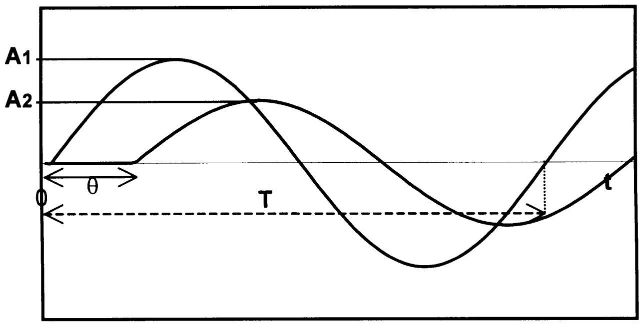
Impulse response for various root locations in the s-plane. (The conjugate root is not shown.)

B. <u>Frequency domain analysis.</u>

Frequency domain analysis studies a system based on the *" steady state" response of the system to sinusoidal input signals of different frequencies:* $u_1 = A_1 \sin \omega t$. The solution of linear diff. eq. with constant coefficients with sinusoidal signals as input u_1 consists of two parts:

- transient terms (a number of exponential functions)
- *steady state term of the form* $A_2 \sin (\omega t + \varphi)$.

In steady state, the output of a linear system has a sinusoidal time course with the same frequency as the input, but with a phase shift with respect to the input signal.



A_1 and A_2 represent the amplitude of the sinus of the input and output signals and we define the gain A as $A \equiv A_2/A_1$

T : period

ν : frequency

ω : pulsation of corner frequency in radians / second

φ : phase shift between output- and input sinus.

Relation between period, frequency and pulsation:

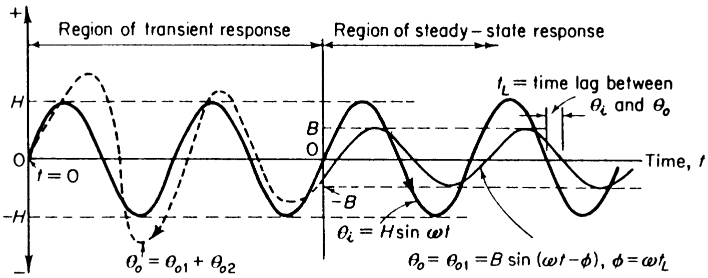
$T = 1/\nu$ and $\omega = 2\,\pi\,\nu$

Relation between phase shift and delay (θ):

$\varphi = 2\pi\,\theta / T = \theta\, 2\,\pi\nu = \theta\,\omega$

In steady state the terms A_2 and φ are dependent on ω, and the shape of the two functions $A_2\,(\omega)$ and $\varphi\,(\omega)$ enables to completely determine the properties of the system. Only the steady state term is considered in frequency domain analysis.

In an experimental analysis of a system, application of an input signal always results in transient phenomena of the output. For frequency domain analysis, after application of the sinusoidal input signal one has to wait until the transient have become negligible and the system has come to a stationary regime, before the parameters of the output are measured.



Typical system response to sinusoidal input: θ_{o2} is transient term of response; θ_{o1} is steady-state term of response.

1. First order system.

a) Differential equation.

$$\tau \cdot \frac{du_2}{dt} + u_2 = \alpha \cdot A_1 \sin \omega t$$

b) Solution.

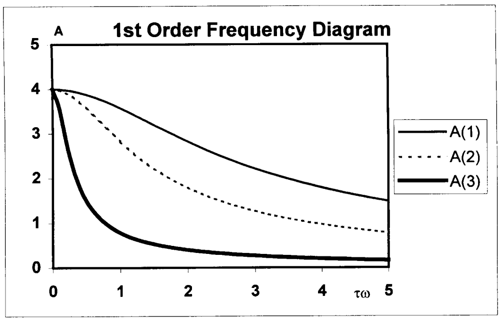
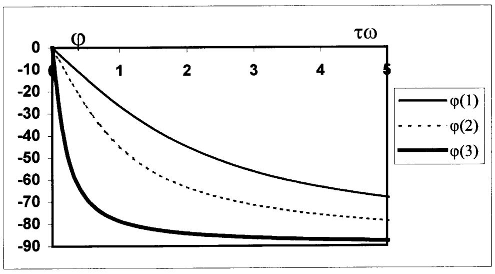
$u_2 = C \cdot e^{-t/\tau} + A_2 \cdot \sin(\omega t + \varphi)$

with $A_2 = \alpha A_1 / \sqrt{(1 + \tau^2 \omega^2)}$

and $\varphi = \mathrm{arctg}\,(-\,\tau\omega)$

The gain A (= A_2/A_1) as well as the phase shift φ are frequency dependent.

c) Graphical Representation.

 	$\omega = 0 \;\Rightarrow\; A = \alpha$ $\varphi = 0$ $\omega = \infty \;\Rightarrow\; A = 0$ $\varphi = -90^o = -\pi/2$

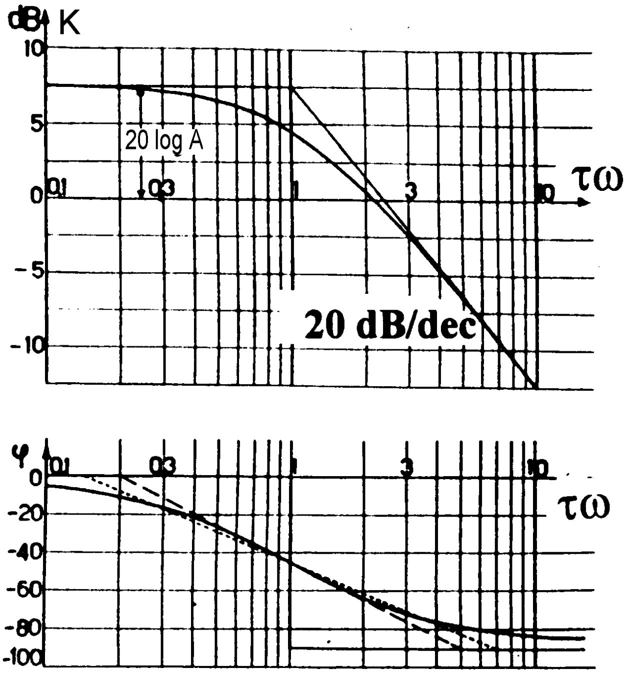
The steady state sinusoidal component of the output contains all information about the system. In a frequency domain analysis of the system, the gain and phase shift A and φ are plotted in a frequency diagram as function of ω:

The figure shows that the gain becomes lower while the phase shift becomes higher with increasing frequency. The maximal phase shift of a 1st order system amounts to -90^o. The minus sign indicates that u_2 lags with respect to u_1.

These two linear frequency diagrams contain all the information about the system. Frequency domain analysis can be made easier however by introducing two variants of these diagrams: the Bode diagram and the polar diagram.

(1) Bode diagram.

A Bode diagram is a combination of an amplitude diagram and a phase diagram, where frequency and gain are plotted on logarithmic scale.



Bode diagram of 1st order system

Convention: in amplitude- and phase diagrams of frequency domain analysis it is so customary to use the Bode diagram convention, that often it is not even indicated that logarithmic axes are used for the frequency and the gain (can be seen from the scale). Most often the axes are also normalized, often without explicit indication.

(a) The K(ω) diagram (gain-frequency plot).

Definition: K ≡ 20 log A

K is the logarithmic gain and is expressed in decibel.

When $A = 10^n$ then $K = n \times 20$ dB.

Mathematical expression of solution of 1st order system:

$$K = 20 \log \frac{\alpha}{\sqrt{(1 + \tau^2\omega^2)}}$$

K is plotted as function of log ω

Note: While in the Bode diagram K is plotted as function of log ω, most often for notation on the ordinate A and for notation of the abscissa ω is used. The logarithmic scale of the axes makes it clear that Bode diagram convention is used. For the ordinate, sometimes the symbol "dB" is used instead of the logarithmic axis.

Analysis of K (ω) diagram: application of asymptotic approximation.

For $\tau^2\omega^2 \ll 1$ (thus for $\omega \ll 1/\tau$):

$K = 20 \log \alpha$ thus *horizontal line indicating the gain of the system in dB*

For $\tau^2\omega^2 \gg 1$ (thus for $\omega \gg 1/\tau$):

$K = 20 \log \alpha/\tau\omega = 20 \log \alpha/\tau - 20 \log \omega$

this is the equation of a *decreasing straight line with a slope of 20 dB/decade* (when frequency is enhanced by a factor 10, the K decreases by 20 dB (A becomes a factor 10 smaller)). The pulsation ω_b where the two straight lines intersect is called break pulsation or 3 dB point of the system.

ω_b is the boundary between the two asymptotic approximations. Thus $\omega_b = 1/\tau$.

For $\tau^2\omega^2 = 1$ (thus for $\omega = 1/\tau$):

since $K_{\omega o} = 20 \log \alpha$ and $K_{\omega b} = 20 \log \alpha/\sqrt{2}$:

$K_{\omega o} - K_{\omega b} = 20 \log \sqrt{2} = 10 \log 2 \cong 3$ dB

The intersection of the two approximations is therefore often called the *3 dB point*. The maximal error that is made by applying the asymptotic approximation is the error at $\omega = \omega_b$, and this amounts to 3 dB. The frequency of the 3 dB point is $\nu_b = 1/2\pi\tau$ or $\omega_b = 1/\tau$.

For $\omega = \omega_a \Rightarrow K = 0$ of $A = 1$

ω_a is also called the unity gain pulsation, and ν_a is the unity gain frequency or unity gain bandwidth. It can be easily demonstrated that in a first order system $\omega_a = \alpha\omega_b$

(b) The φ(ω) diagram (phase-frequency plot).

$\varphi = \text{arctg}(-\tau\omega)$

Analysis

Phase shift of the ideal first order system.

$\omega = 0 \Rightarrow \varphi = 0^o$

$\omega = \infty \Rightarrow \varphi = -\pi/2 = -90^o$

$\omega = \omega_b = 1/\tau \Rightarrow \varphi = -\pi/4$

An *ideal first order system is thus characterized by a maximal phase shift of* -90^o.
The minus sign of the phase shift indicates that the output signal lags with respect to the input.

Phase shift in the presence of transportation lag.

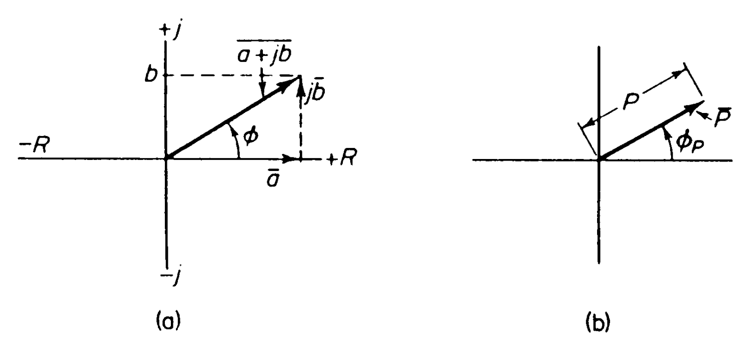
$\varphi_{tot} = \varphi + \varphi_\theta$ with φ the phase shift caused by transformations in the (ideal) process,
and $\varphi_\theta = 2\pi\theta/T = 2\pi\nu\theta = \theta\omega$ the phase shift by transportation lag

The phase shift by transportation lag increases proportionally to the frequency (or pulsation). A transportation lag in a system therefore implies that the system has no maximal phase shift. The phase diagram can therefore not always be used to determine whether a system is 1st order. A better method is to determine the maximal slope of the decreasing straight line in the asymptotic approximation of the gain-frequency plot.

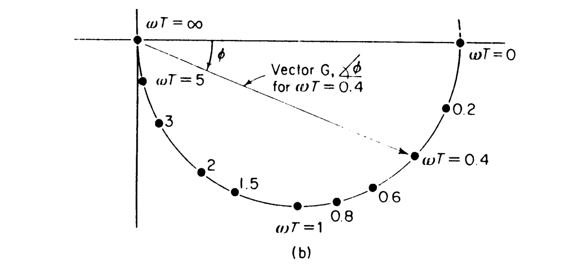
(2) Polar diagram.

The polar diagram uses the phasor representation of sinusoids, where a sinus is represented by a vector with length equal to the amplitude and angle equal to the phase shift.

Frequency domain analysis of a process gives for each frequency of the input sinus, a value of A and φ. For representing the result in a *polar diagram* (also called Nyquist plot), *a vector* is constructed for each *frequency*. This vector starts from origin of the coordinate system and has a *magnitude equal to the gain* A, *and an angle with the abscissa equal to the phase shift* φ. When the end points of all vectors for all frequencies of the input signal are connected, one obtains a curve that gives the system polar diagram. In this diagram, A (in linear scale, thus not K in dB!) is thus plotted in polar coordinates as function of φ.



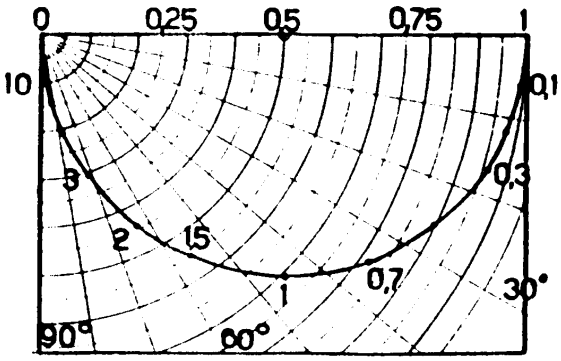
Phasor representations.



ωT =	0	0.2	0.4	0.6	0.8	1	1.5	2	3	5	10	∞
G =	1	0.98	0.925	0.855	0.78	0.71	0.55	0.45	0.32	0.20	0.10	
φ, degrees =	0	-11.3	-21.8	-31	-38.7	-45	-56.3	-63.4	-71.6	-78.7	-84.3	-90

General frequency response characteristics of $1/(1 + TD)$ in polar form.

For the 1st order system this becomes:



$$A = \frac{\alpha}{\sqrt{1 + tg^2\varphi}} = \frac{\alpha \sqrt{\cos^2\varphi}}{\sqrt{\cos^{+2}\varphi + \sin^{+2}\varphi}} = \alpha \cos\varphi$$

$\omega = 0 \Rightarrow A = \alpha$ and $\varphi = 0$

$\omega = \omega_b \Rightarrow A = \alpha/\sqrt{2}$ and $\varphi = -\pi/4$

$\omega = \infty \Rightarrow A = 0$ and $\varphi = -\pi/2$

The equation of the polar diagram is obtained by expressing A as function of φ, and thus by eliminating ω in the two functions A (ω) and φ (ω). For the first order system this becomes $A = \alpha \cos \varphi$. This is the equation of the circle in polar coordinates (since a triangle formed by a point on the circumference of the circle and the end points on a diameter of the circle is always a rectangular triangle).

The figure is normalized (Thus A=1 corresponds to A = α, and the numbers near the circle are expressed in units of τ; thus ω = 1 corresponds to $\omega = \omega_b = 1/\tau$)

2. Second order system.

a) Differential equation.

$$\frac{1}{\omega_n^2} \cdot \frac{d^2u_2}{dt^2} + \frac{2\xi}{\omega_n} \cdot \frac{du_2}{dt} + u_2 = \alpha \, . \, A_1 \sin \omega t$$

b) Solution.

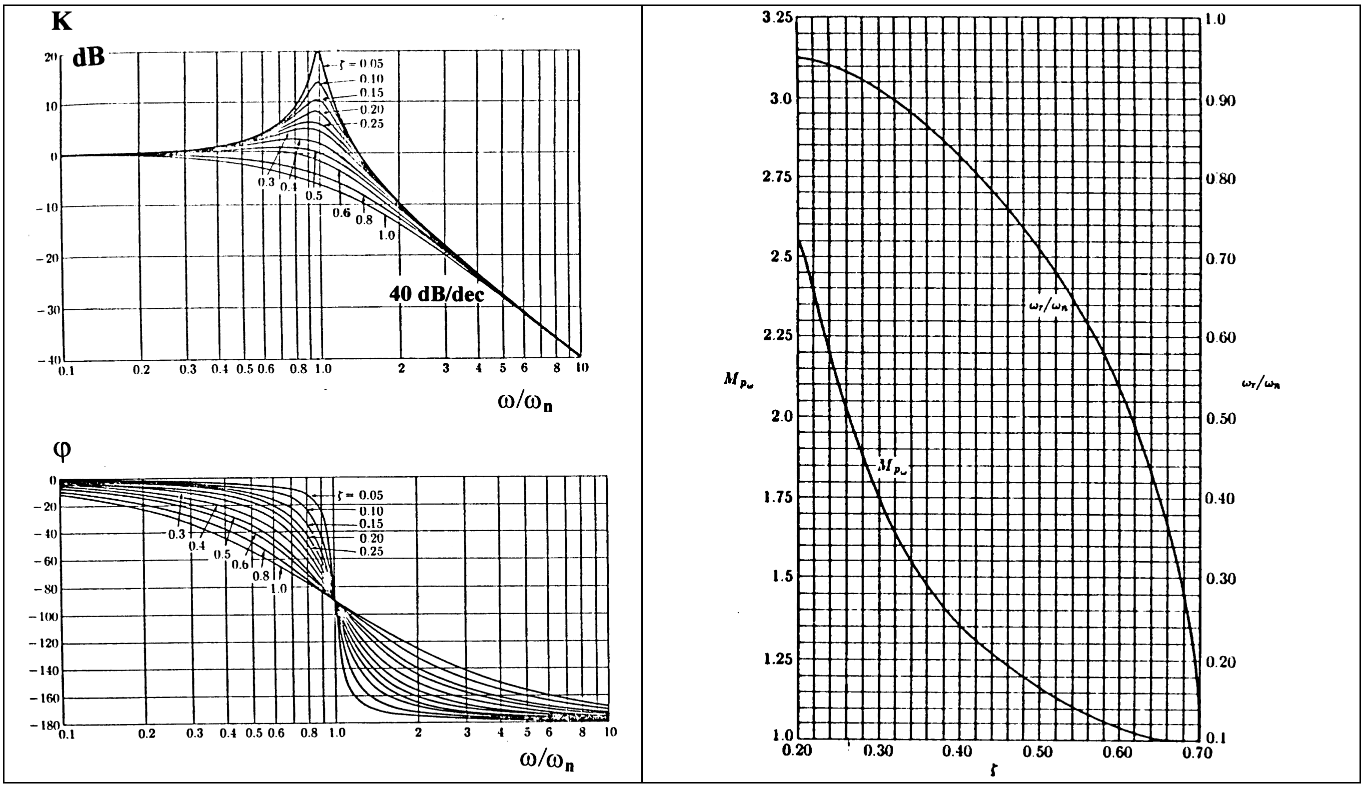
$u_2 = A_2 \sin (\omega t + \varphi)$

$$A = \frac{\alpha}{\sqrt{\left((1-(\omega/\omega_n)^2)^2 + 4\xi^2(\omega/\omega_n)^2\right)}} \quad \text{(with } A \equiv A_1/A_2\text{)}$$

$$\text{tg}\ \varphi = \frac{-2\xi\omega/\omega_n}{(1-(\omega/\omega_n)^2)}$$

c) Graphical representation.

(1) Bode diagram.



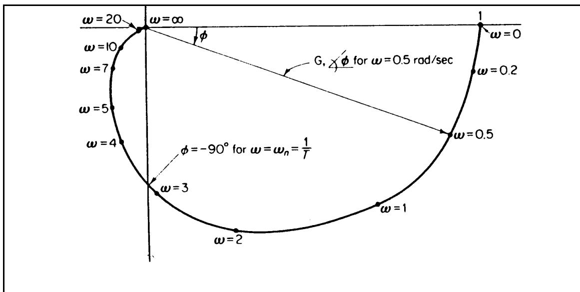
- for $\omega = 0 \Rightarrow K = 20 \log \alpha$
- for $\omega = \infty \Rightarrow$ decreasing asymptote with slope 40 dB/dec
- for $\omega = \omega_n$ and $\xi < 1$ the gain-frequency plot shows a resonance peak
- $\varphi_{max} = -180° = -\pi$ rad
- $\varphi_{\omega=\omega n} = -90° = -\pi/2$ rad

The graph on the right side gives amplitude and pulsation of the resonance peak as function of the damping factor.

Meaning of the resonance peak.

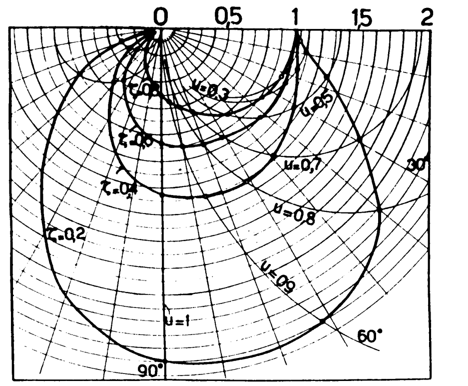
A 2nd order system has a natural frequency (also called eigen frequency). This is a frequency at which the system can spontaneously oscillate, when no damping takes place. In this situation, energy is not dissipated, but only internal conversions of energy from one form to another form take place. In a damped system this vibration will finally extinguish since damping corresponds to loss of energy. In a damped system the energy loss -E is proportional to the rate of change of the output du_2/dt (compare to friction, which has loss of energy proportional to velocity). When energy is supplied in phase with the energy loss (du_2/dt) of a damped system, the energy loss due to damping will be compensated, and the system will keep oscillating. The resonance peak thus occurs at a frequency ω_n.

(2) Polar diagram.



ω, rad/sec	0	0.2	0.5	1	2	3	4	5	7	10	20	100	∞
G	1	0.994	0.96	0.87	0.65	0.49	0.35	0.26	0.16	0.09	0.024	~0	0
φ, degrees	0	−8	−19.5	−38	−67	−87	−102	−113	−128	−142	−160	−176	−180

Frequency response of the system $\theta_o/\theta_i = 1/(1 + 0.7D + 0.1D^2)$



$\varphi_o = 0$

$\varphi_{\omega n} = -\pi/2$

$\varphi_\infty = -\pi$

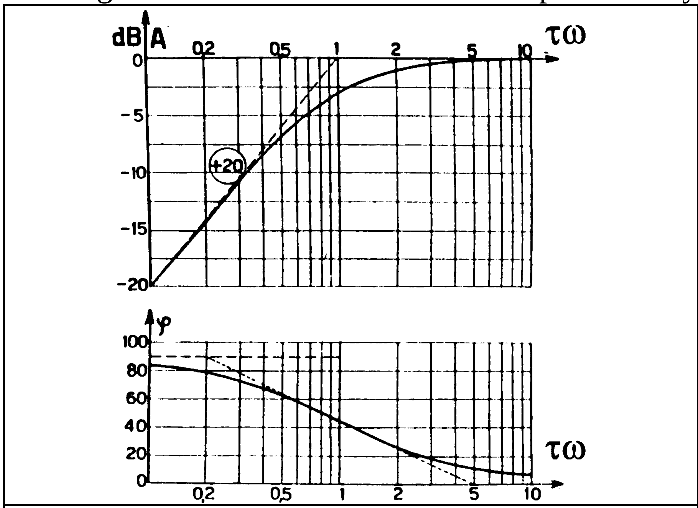
with $u \equiv \omega/\omega_n$

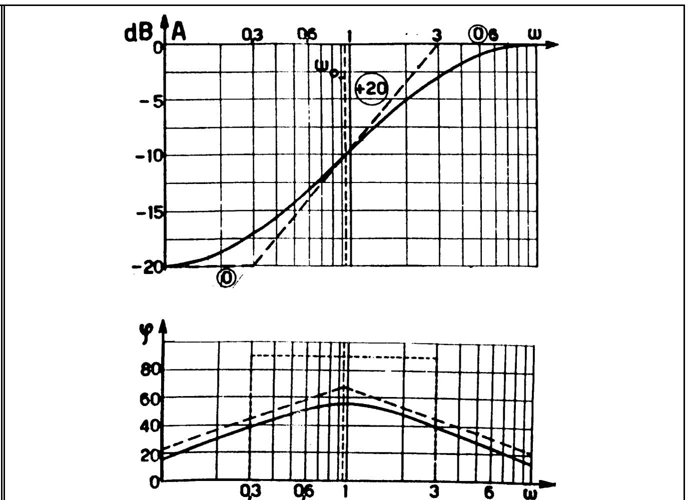
3. System of order -1.

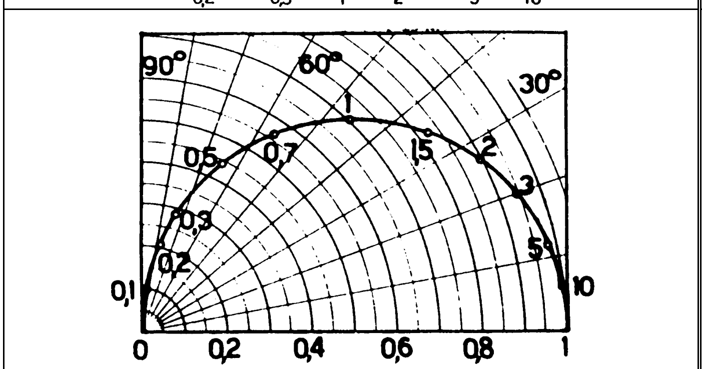
A system of –1st order (differentiator) has an increasing asymptote with slope +20 dB/decade, and a phase shift that maximally amounts to +90° (phase lead).

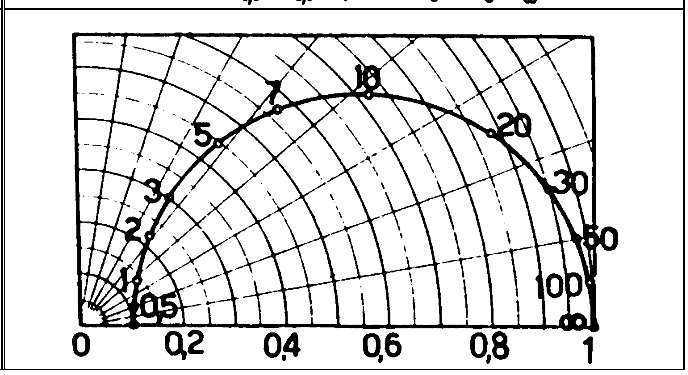
N.B. Phase lead does not mean that the output already starts earlier than the input signal. Frequency domain analysis only shows the result in steady state, thus after the transients have declined. Phase lead of +90° means that the output signal is proportional to the slope of the input signal.

The finite response rate of physical systems limits the differentiation at the high frequencies. The range of differentiation can also be limited at low frequencies. A differentiator with two sided frequency range limitations is often used as compensation system.





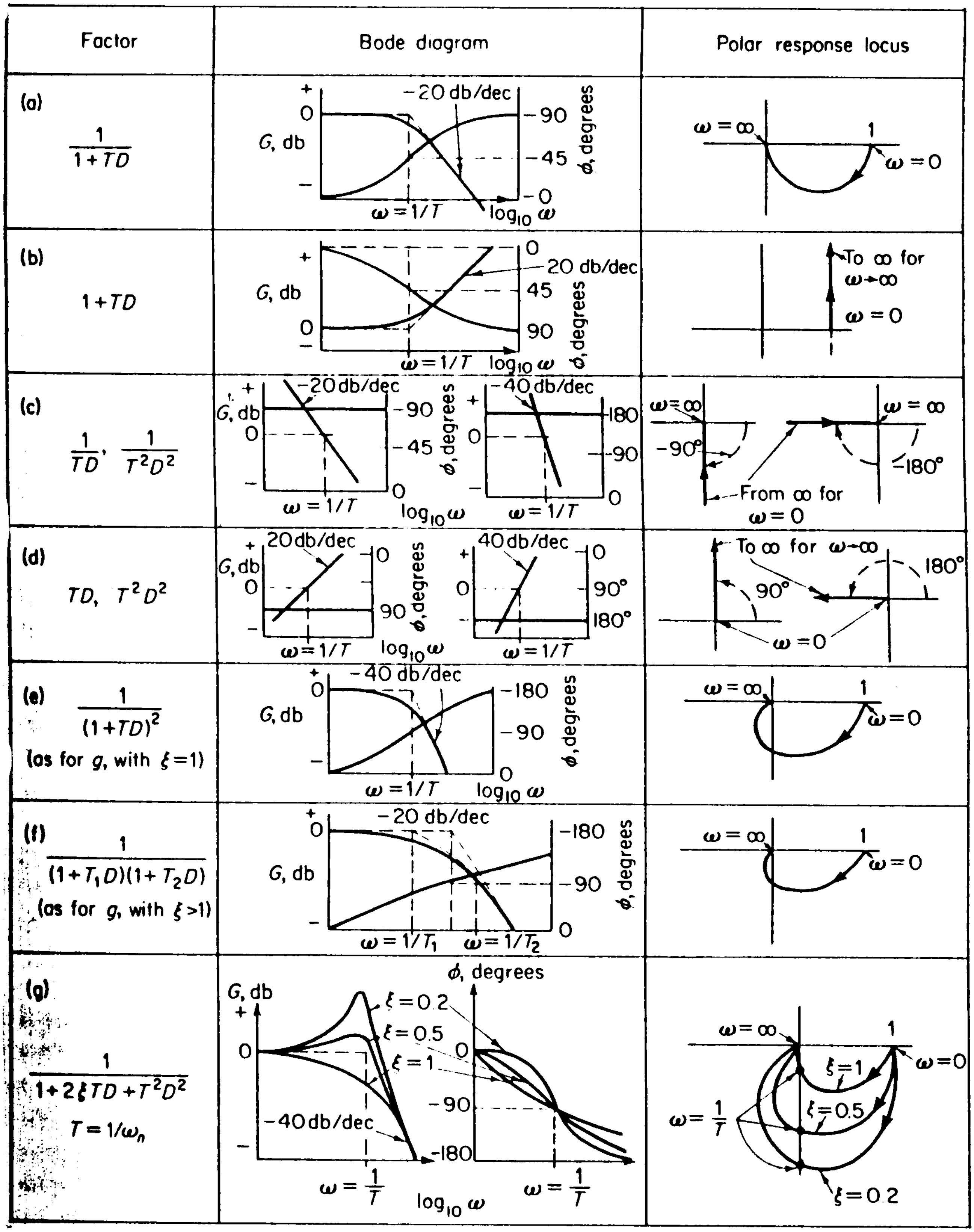




4. General properties of systems of order n.

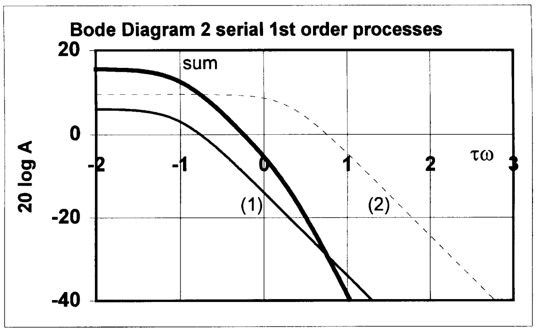
- Maximal slope of the decreasing asymptote in the Bode diagram is n.20 dB/dec
- $\varphi_\infty = -n.\pi/2$

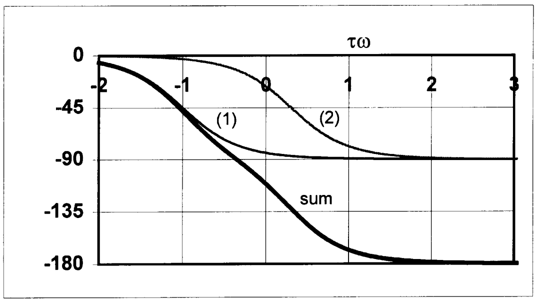
Following figure shows transfer function, Bode diagram and polar diagram for a number frequently occurring processes.



Frequency response characteristics of the factors which commonly appear in or as transfer functions: Bode and polar forms.

Appendix 1. Bode diagram of 2 systems in series.





$T = T_2\ T_1$

from there it follows that

$A = A_{(1)} \,.\, A_{(2)}$

$\varphi = \varphi_{(1)} + \varphi_{(2)}$

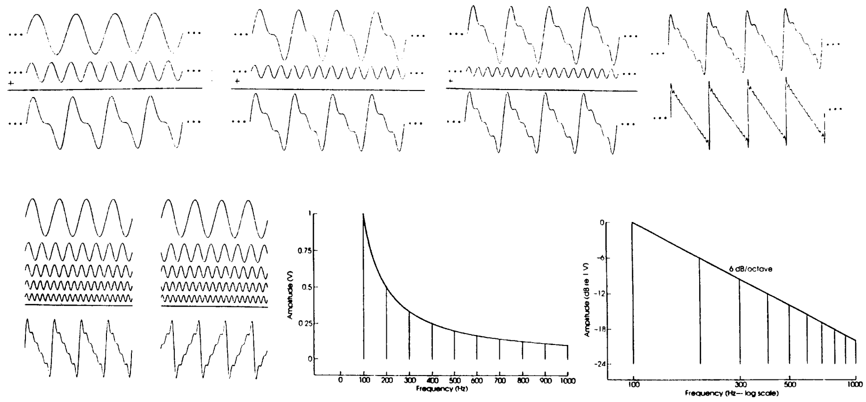
$K = 20 \log A = 20 \log A_{(1)} \,.\, A_{(2)}$

$K = 20 \log A_{(1)} + 20 \log A_{(2)}$

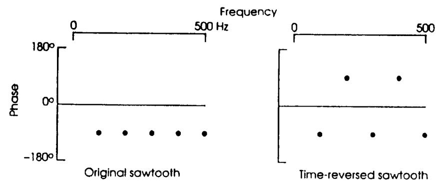
$K = K_{(1)} + K_{(2)}$

Appendix 2. Relation between time domain analysis and frequency domain analysis.

Fourier's theorem from mathematics states that (most) periodic mathematical functions can be approximated by a Fourier series; this is a sum of sinuses of different frequencies (the frequencies are integer multiples of a basis frequency (in analogy, aperiodic functions can be approximated by Fourier integrals).

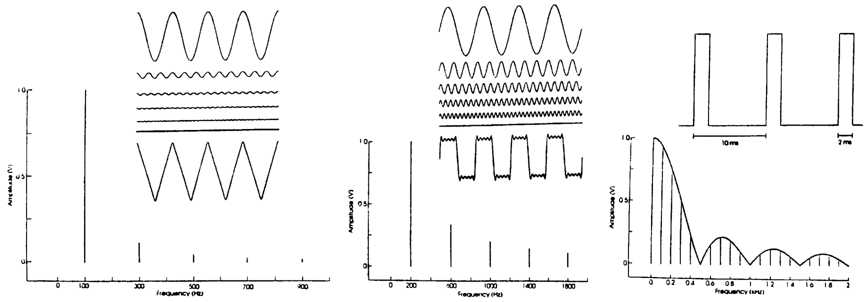


Frequency (Hz)	Amplitude (V)
100	1
200	1/2
300	1/3
400	1/4
500	1/5
600	1/6



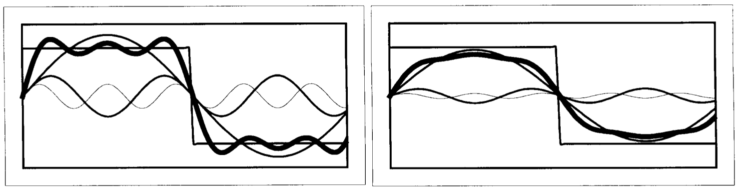
Examples of Fourier series.

- $y = -a$ for $-\pi < x < 0$
  $y = a$ for $0 < x < \pi$
  Fourier series:
  $y = 4a/\pi \,.\, \sum \sin((2i+1)x))/(2i+1)$
  with $i = 0 \rightarrow \infty$
- $y = x$ for $0 < x < 2\pi$
  Fourier series:
  $y = \pi - 2.\sum \sin(ix)/i$
  with $i = 1 \rightarrow \infty$
- $y = |x|$ for $-\pi < x < \pi$
  Fourier series:
  $y = \pi/2 - 4/\pi.\sum \cos((2i+1)x))/(2i+1)^2$
  with $i = 0 \rightarrow \infty$

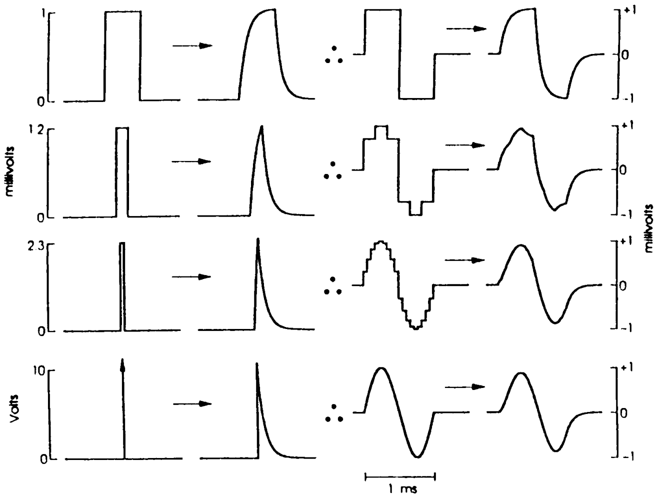


A particular function u_1 applied to the input of the system can thus be approximated as a sum of sinuses of different frequencies. For each sinus frequency, the system has a certain K and φ so that for each sinus the corresponding output signal can be calculated. The sum of the responses to the different sinus components of the output signal can then be added and this summation gives the output signal. Similarly, the response to a sinus can be described by means of step functions.

Example: 1st order system with u_1 = step function (left)



Fourier analysis decomposes u_1 in a number of sinuses, some of low frequency, others of high frequency (these last ones correspond to the fast change at $t = 0$). For the 1st order system, the gain is low at high frequencies. The steep slope of the output therefore disappears and the system thus requires a certain amount of time to reach steady state.
Example: 1st order system with u_1 = sinus function (right panel) described as sum of step functions.



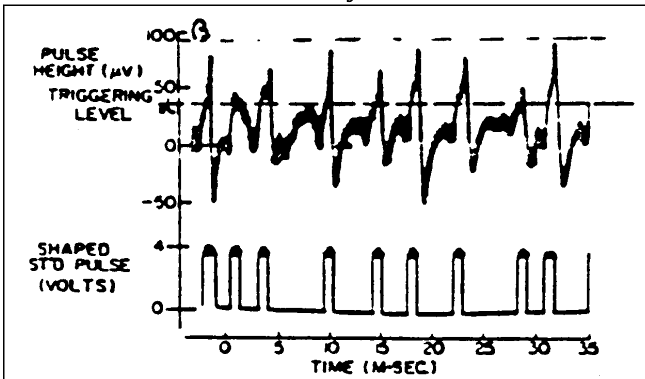
<u>Application: The photoreceptor ganglion of the lobster.</u>

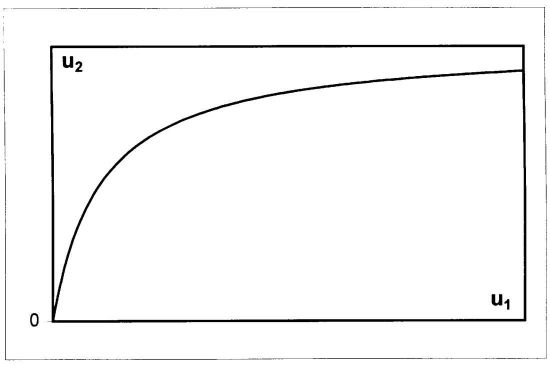
The photoreceptor ganglion of the lobster consists of a number of light sensitive neural cells. The activity of these cells influences the light avoidance reflex of the lobster.

Experimental set-up.

The ganglion is illuminated by a light source with variable intensity and the electrical activity of the afferent neuron is measured as a function of time. Illumination of the ganglion results in a series of electrical discharges in the form of spikes. The number of discharges per unit of time is counted, but only discharges with amplitude within certain limits are taken into account, since these responses appear to originate from neurons with the same rate of propagation of the action potential. The relation between firing rate u_2 and the light intensity u_1 is measured.

1. Determination of the steady state transfer function.



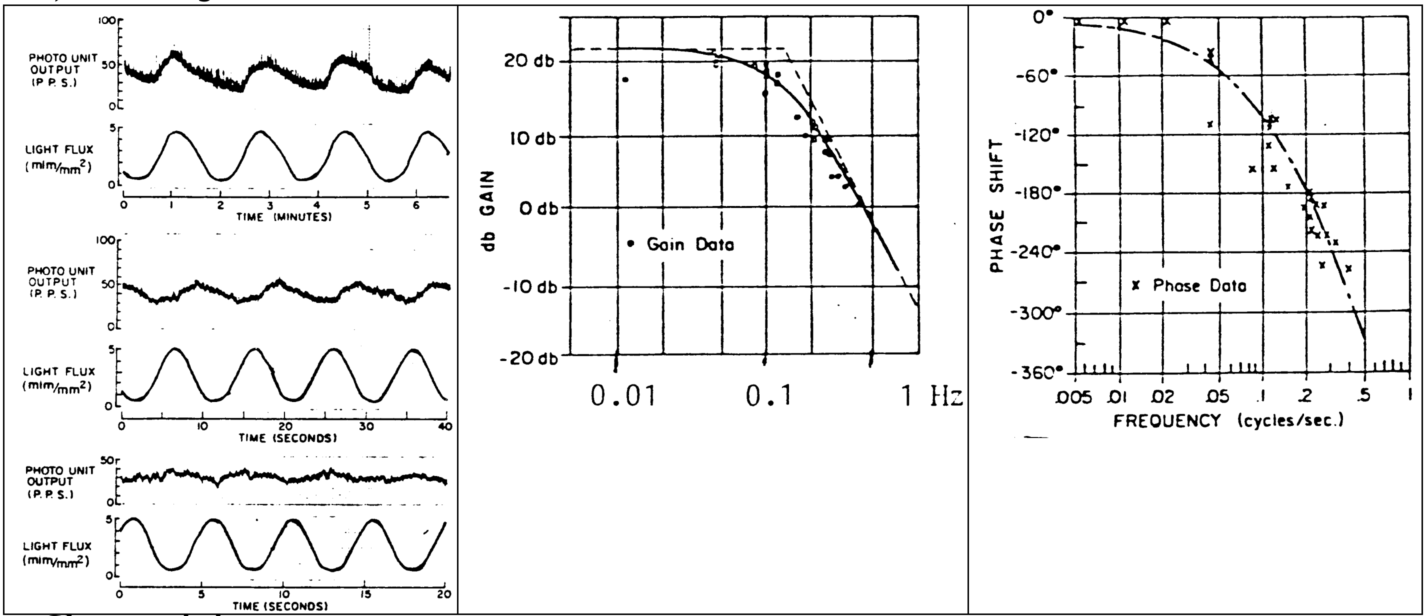


The light intensity of the source is kept at a constant level until the rate of discharge of the neuron has reached a steady state. When this condition is fulfilled, the rate of discharge (spike frequency) is measured. This experiment is repeated for different light intensities. The results of this experiment demonstrate that increasing the light intensity causes an increase of the rate of discharge until a saturation level is reached. The system is thus nonlinear. For further analysis of the dynamics of the system, the light intensity changes are made small enough to enable linear approximation.

2. Frequency domain analysis.

The light intensity is varied in a sinusoidal way: the steady state frequency of discharges of the nerve is measured. This rate of discharge (output) will also vary sinusoidally. The gain and the phase shift of the system are plotted in a Bode diagram as a function of frequency of the input signal.

a) Bode diagram.



Characteristics:

K_o = 22 dB

Slope of the decreasing asymptote: 20 dB/dec

$\nu_b = 0.12/s$

No maximal phase shift is found

$\omega = \omega_b \Rightarrow \varphi_{tot} = -66^o$

Analysis:

Slope of 20 dB/dec $\Rightarrow$ 1st order system

$\omega_b = 2\pi\, \nu_b = 1/\tau$; thus $\tau = 1/2\pi\, \nu_b = 1.3$ s

Phase shifts larger than -90^o: an important transportation lag is present

$\varphi_\theta = \varphi_{tot} - \varphi$

for $\omega = \omega_b$ $\varphi_\theta = 66^o - 45^o = 21^o$

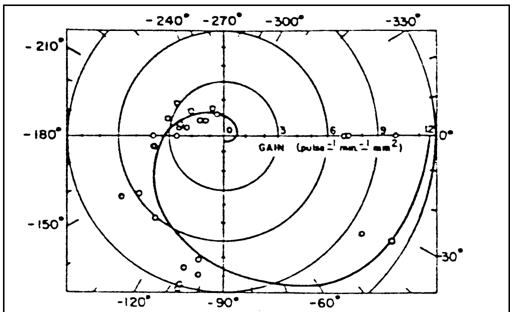
$\varphi_\theta = 2\pi\theta/T = \theta \,.\, 2\pi\nu$

$21^o = \theta \,.\, 2\pi.\ 0.12$ rad $= \theta \,.\, 360° \,.\, 0.12$ s

$\theta = 21^o / (360 \,.\, 0.12) = 0.5$ s

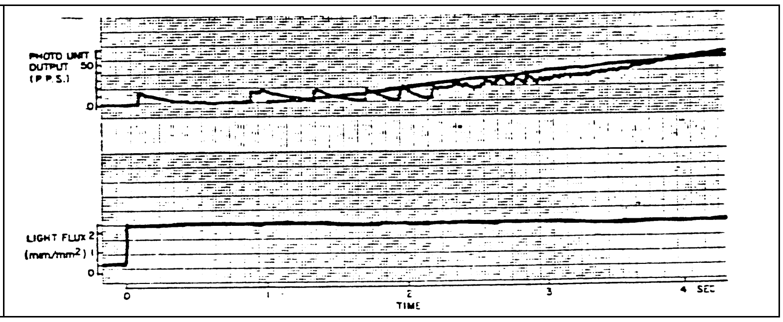
b) Polar diagram.

3. Time domain analysis.



$\theta = 1$ s

$\tau = 2$ s

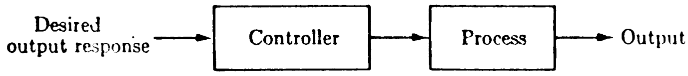


III. ANALYSIS OF CLOSED LOOP SYSTEMS.

A. Introduction.

In order to make sure that the output of a process has the desired value, one can use an “open loop” control system. An “open loop” control system makes use of a regulator (“controller”, “actuating device”) to obtain the desired response. In such a system, disturbances in the system will influence the value of the output.

Process to be controlled.

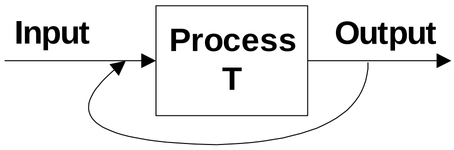


Open-loop control system.

Errors due to disturbances (that could be internal or external to the system) can be avoided by controlling the process by means of “closed loop” control systems (“feedback” systems).

1. What is a closed system?

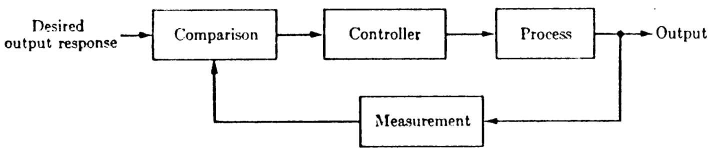
A closed loop system is a system in which the input of the process is influenced by the value of the output variable.



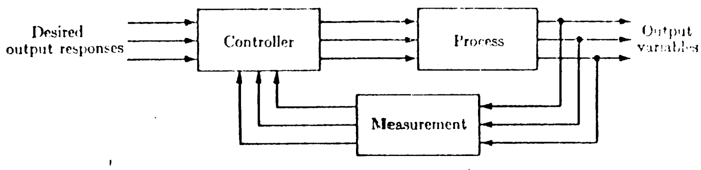
Such a feedback of output to input of the system will thereby cause a change of the output. Depending on the sign of the feedback of the output signal, two types of feedback can be distinguished.

a) Negative feedback.

The value of the output is compared to the desired value (“reference signal”, “command signal”, “input signal”); the difference (“the error signal” = input – measured output) causes a change in the regulating action in such a direction that the value of the output will become closer to the desired value. Systems with negative feedback are called closed loop control systems (or simply “control systems”).



Closed-loop feedback control system.



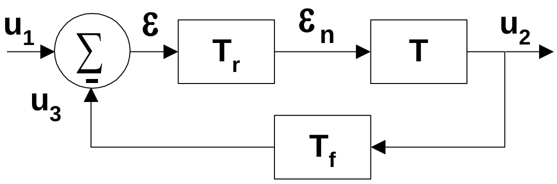
Multivariable control system.

b) Positive feedback.

When the measured output is added to the input, the feedback causes a regulating action in a direction that makes the difference between input and output larger. In positive feedback systems the output increases until the saturation level is reached, and such a system is therefore not a control system. Positive feedback is however often used to ensure a very fast all or none behavior (e.g. action potentials).

Control theory deals with the theory of steering of processes by means by means of closed loop negative feedback systems. The rest of the course will therefore mainly deal with systems with negative feedback.

2. Schematical representation of a control system.



The output u_2 of the process is measured and after eventually processing, the value u_3 (which is the measured (and eventually processed) value of u_2) is fed back into the *comparator*. This comparator compares u_3 to the desired value u_1 by making the difference $\boldsymbol{\varepsilon = u_1 - u_3}$. This error signal ε is used as the direct input signal for the regulator. The regulating ε_n is a function T_r of the input signal ε. The output value of the process will keep changing until $\varepsilon_n \to 0$.

E.g.: regulation of a valve by a servomotor. The transfer function of the servomotor determines the relation between the angular position of the valve and the voltage applied to the motor.

3. Types of regulators.

a) Continuous regulators.

ε_n is a continuous function of ε.

(1) Proportional regulators (P-regulators).

$\varepsilon_n = k \cdot \varepsilon$ (k dimensionless)

(2) Differentiating regulators (D-regulators).

$$\varepsilon_n = k \cdot \frac{d\varepsilon}{dt} \quad \text{(k has dimension of time)}$$

(3) Integrating regulators (I-regulators).

$\varepsilon_n = k \cdot \int \varepsilon dt$ (k has dimension of $time^{-1}$)

(4) Combinations (PI, PD, PID) of these regulators.

b) Discontinuous regulators.

(1) Two or more state regulators.

ε_n can only have a finite number of different values dependent on the value of ε.

E.g. thermostat.

(2) Regulators with constant velocity.

$$\frac{d\varepsilon_n}{dt}$$ can only have two different values: 0 or 1.

4. Types of feedback.

a) Proportional feedback.

$u_3 = k \cdot u_2$

b) Differentiating feedback.

$$u_3 = k \cdot \frac{du_2}{dt}$$

c) Integrating feedback.

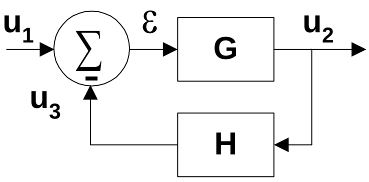
$u_3 = k \cdot \int u_2 dt$

d) Combinations.

B. <u>System analysis.</u>

1. General properties of control systems.

a) Schematic representation.



b) Equation of closed loop systems.

$u_2 = G \cdot \varepsilon$
$\varepsilon = u_1 - u_3$
$u_3 = H \cdot u_2$
Elimination of u_3 and ε provides the relation between u_2 and u_1:

$$\boxed{\mathbf{u_2 = \frac{G}{1 + GH} u_1}}$$

$$\boxed{\boldsymbol{\varepsilon} = \mathbf{\frac{1}{1 + GH} u_1}}$$

c) Properties.

(1) Gain.

The gain of the closed system is smaller than the gain of the system without the feedback (open system): $G/(1 + GH) < G$

(2) Sensitivity of systems to parameter variation.

A disturbance causes a change of the relation between input and output, and thus of the transfer function of the process. In order to study the influence of a disturbance on the output, we can thus investigate how the output of the system changes for a change of particular parameters of the transfer function.
The sensitivity of a system to a particular parameter x, represented as S_X, is defined as the ratio of the relative change of the output to the relative change of the parameter.

$$\mathbf{S_x = \frac{\delta u_2/u_2}{\delta x/x}}$$

(a) Sensitivity of the open system.

Let G be the system transfer function, and we calculate the sensitivity of the output to variation of G.

$u_2 = G\, u_1 \;\Rightarrow\; u_2 = \delta G\, u_1$

$$\frac{\delta u_2}{u_2} = \frac{\delta G u_1}{u_2} = \frac{\delta G u_1}{G u_1} = \frac{\delta G}{G}$$

$$S_G = \frac{\delta u_2/u_2}{\delta G/G} = 1$$

(b) Sensitivity of the closed system.

$$u_2 = \frac{G}{1 + GH} u_1$$

(i) Sensitivity to G.

$$\delta u_2 = \frac{\delta G u_1}{1+GH} - \frac{GH\, u_1 \delta G}{(1+GH)^2} = \frac{\delta G}{(1+GH)^2}$$

$$\frac{\delta u_2}{u_2} = \frac{1}{1+GH} \cdot \frac{\delta G}{G}$$

$$\mathbf{S_G = \frac{\delta u_2/u_2}{\delta G/G} = \frac{1}{1+GH} \cong \frac{1}{GH}}$$

The sensitivity of a control system to variations occurring in the direct process (G) is decreased by the presence of the negative feedback. The decrease is dependent on the so-called "*open loop transfer function*" GH. A system with negative feedback and large open loop gain becomes rather insensitive to a change of parameters of the direct process.

E.g.: G = 100 and H = 1

10% change of G will change u_2 by 0.1 %.

(ii) Sensitivity to H.

$$\delta u_2 = \frac{- G^2\, \delta H u_1}{(1+GH)^2}$$

$$\frac{\delta u_2}{u_2} = \frac{-G}{1+GH} \cdot \delta H = \frac{-GH}{1+GH} \cdot \frac{\delta H}{H}$$

$$S_H = \frac{\delta u_2/u_2}{\delta G/G} = \frac{-GH}{1+GH}$$

for GH >> 1 $\Rightarrow$ $\mathbf{S_H \cong -1}$

While a negative feedback system becomes insensitive to variations of parameters of the direct process, it is sensitive to variation of parameters of the feedback.

However, since the feedback process is only an information transmission process, its structure can be much simpler than the structure of the direct process which it helps controlling, so that the parameters can much more easily be kept accurately constant (for adequate control of a complex machinery, only a good measuring device is needed).

(3) Follower behavior for H = 1.

For H = 1

$$u_2 = \frac{G}{1 + G} u_1$$

When G >> 1 then: $\mathbf{u_2 \cong u_1}$. In this case, the output will approximately follow the input.

u_2 is not identical to u_1 but deviates somewhat from the input, because of the finite value of the amplitude of G. The deviation Δ of a follower is defined as $\Delta \equiv (u_2 - u_1)/u_1 = \varepsilon/u_1$.

$\Delta = (Gu_1/(1 + G) - u_1)/u_1 = -1/(1 + G)$

Δ is dependent on the gain of the transfer function G, and is approximately:

$\Delta \cong -1/G$

(4) Approximation.

For GH >> 1, the term 1 in the numerator of the transfer function of the control system can be neglected. In this case the following approximation for the closed system is valid:

$\mathbf{u_2 \cong H^{-1}.u_1}$

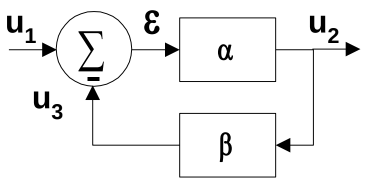
This can be easily seen, since for $G \to \infty \Rightarrow \varepsilon \to 0$. The action of the feedback makes the error signal very small, and since $\varepsilon = u_1 - u_3 = u_1 - Hu_2$ and ε goes to zero, $H.u_2 \to u_1$; thus $u_2 \to H^{-1}.u_1$. The transfer function of a good feedback system (high open loop gain) is approximately equal to the inverse of the transfer function H of the feedback chain (independent of the transfer function G of the direct process)!

2. Applications to different systems.

a) Proportional regulators ($\varepsilon_n = k\varepsilon$).

(1) Proportional feedback.

(a) Zero order process.



(i) Equation.

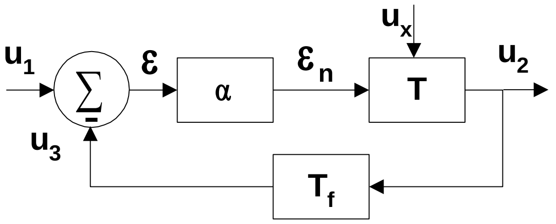
$$u_2 = \frac{\alpha}{1 + \alpha\beta} u_1$$

(ii) Properties.

- The gain of the closed system is smaller than the gain of the open system.
  When α and β > 0 ⇒ $\alpha/(1 + \alpha\beta) < \alpha$
- The influence on the output of disturbances acting on the amplifier α decreases by negative feedback. Also the influence of nonlinearity decreases. This is because the sensitivity to the process in the direct chain is decreased.
  $S_\alpha = 1/(1 + \alpha\beta)$
- For β = 1 the system acts as a follower.
  $u_2 = u_1/(1 + 1/\alpha)$
  for $\alpha \to \infty \Rightarrow u_2 \to u_1$
  steady state error of the follower: $\Delta = -1/(1 + \alpha)$
  approximation: $\Delta \cong -1/\alpha$ for α >> 1
- For β < 1 the system acts as a follower with gain, while for β > 1 the system acts as a follower with attenuation.
  for $\alpha \to \infty \Rightarrow u_2 \to \beta^{-1}.u_1$

(b) First order process (with disturbance).

For the proportional feedback we assume a unity gain feedback (choosing a $\beta \neq 1$ would only change the scale of the amplitude of u_2).



(i) Differential equation.

Differential equation open system (1)

$$\tau . \frac{du_2}{dt} + u_2 = u_1 + u_x$$

Differential equation open system with amplification (2)

$$\tau . \frac{du_2}{dt} + u_2 = \alpha . u_1 + u_x$$

Differential equation closed system (3)

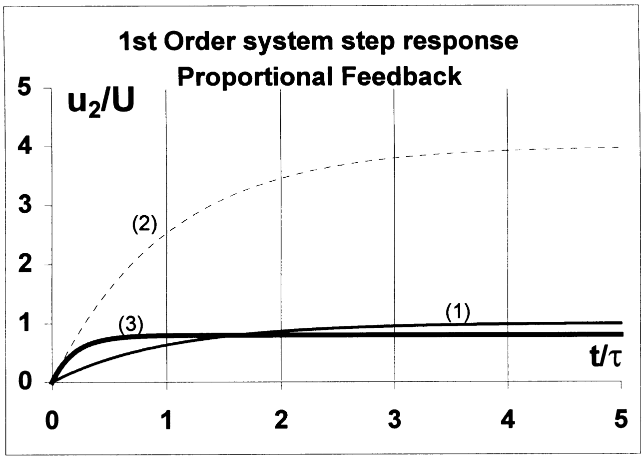
$$\tau . \frac{du_2}{dt} + u_2 = \varepsilon_n + u_x$$

$$\varepsilon_n = \alpha \, \varepsilon$$

$$\varepsilon = u_1 - u_2$$

$$\frac{\tau}{1+\alpha} . \frac{du_2}{dt} + u_2 = \frac{\alpha}{1+\alpha} . u_1 + \frac{1}{1+\alpha} . u_x$$

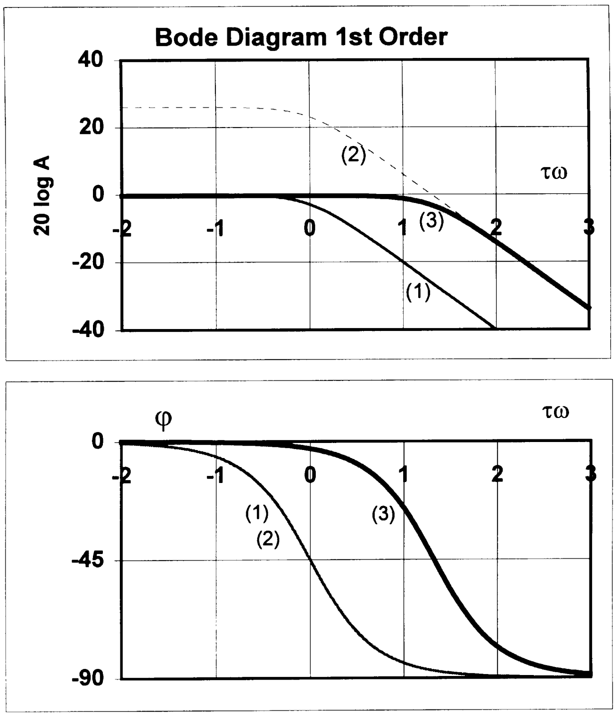
(ii) Properties.



1 Open

2 Open with extra gain

3 With unit feedback



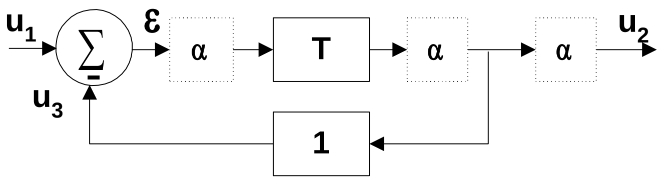
The closed system can now be considered as a 1st order process with a transfer function T'. The 3 systems thus have a similar type of behavior (order of diff. eq.); only the parameters are different.

	System 1	System 2	System 3
Time constant τ'	τ	τ	$\tau/(1+\alpha)$
Gain α'	1	α	$\alpha/(1+\alpha)$
Gain of disturbance α'_x	1	1	$1/(1+\alpha)$

Comparison between system (3) and (1)

- For $\alpha >> 1 \Rightarrow \alpha/(1+\alpha) \cong 1$
  The closed system thus has approximately *the same gain* as the open system (1). The system has a steady *state error* $\Delta = -1/(1+\alpha)$
- The *gain of the disturbance* in the closed system is *lower* than in the open system. The influence of the disturbance on the closed system will thus be smaller than on the open system.
- The *time constant* of the closed system is *lower* than of the open system. The closed system will thus react faster to changes of the input and the dynamic behavior will be better than that of the open system!

In order to improve the dynamic behavior of a process and to decrease the influence of disturbance, one can thus make use of negative feedback. Since negative feedback causes a decrease of the gain of the system, extra amplification needs to be applied to maintain the total gain at the desired level. The extra amplification could in principle be introduced at three different locations, but for a good feedback system the extra amplification needs to be inserted immediately behind the comparator (amplify the error signal that controls the regulatory action!). The thus added extra gain is at the basis of the improved response.



- **The enhanced speed of the response of the system.**

The extra amplification in the system would cause the output upon application of a step function to go towards a higher amplitude, but the negative feedback will limit the increase of the amplitude. However, immediately or shortly after the application of the step function, the output signal is still very small, because of the finite time constant of the system. Therefore, the signal fed back to the comparator is still very small immediately after start of the step input, and the closed system initially approximates the open system with extra gain. When gradually the output increases, also the signal fed back to the comparator increases, which causes a limitation of the increase of the output. The result is a faster evolution to the steady state value.

With low frequencies sinusoidal input signals, the effect of the extra gain will be compensated by the negative feedback. However at the higher frequencies the output amplitude is small and the feedback will thus exert less influence, and at these frequencies the Bode diagram will be approximate Bode diagram of the open loop system with extra gain.

- **Decrease of sensitivity to disturbance.**

The negative feedback causes a decrease of the amplitude of the output signal. However, the attenuation of the effect of the input signal is compensated by the extra amplification in the direct chain. When this amplification is inserted between comparator and process, it will not affect the disturbance, and the negative feedback will therefore decrease net effect of the disturbance.

(c) Second order process.

(i) Differential equation.

Open system:

$$\frac{1}{\omega_n^2}\cdot\frac{d^2u_2}{dt^2}+\frac{2\xi}{\omega_n}\cdot\frac{du_2}{dt}+u_2=u_1$$

Closed system (with extra gain):

$$\frac{1}{\omega_n^2}\cdot\frac{d^2u_2}{dt^2}+\frac{2\xi}{\omega_n}\cdot\frac{du_2}{dt}+u_2=\alpha\,.\,\varepsilon$$

$\varepsilon = u_1 - u_2$ (we assume unity gain feedback)

$$\frac{1}{(1+\alpha)\omega_n^2}\cdot\frac{d^2u_2}{dt^2}+\frac{2\xi}{(1+\alpha)\omega_n}\cdot\frac{du_2}{dt}+u_2=\frac{\alpha}{1+\alpha}\,.\,u_1$$

(ii) Properties.

The closed system of 2nd order with unity gain feedback thus has a transfer function (T') of second order with parameters:

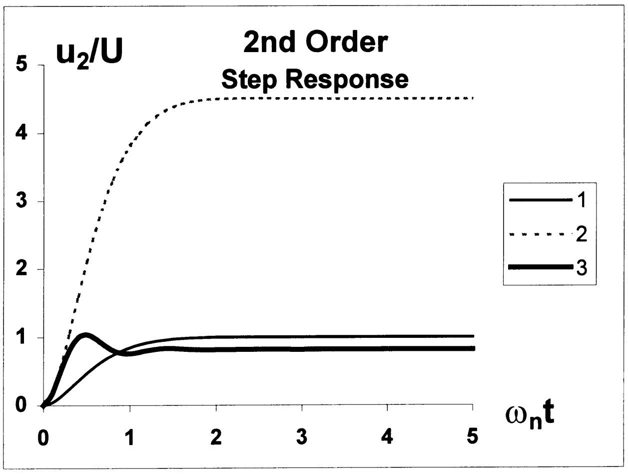
$\omega'_n = \omega_n.\sqrt{(1+\alpha)}$

$\xi' = \xi/\sqrt{(1+\alpha)}$

$\alpha' = \alpha/(1+\alpha)$

1. *the system acts as a follower;* there is a steady *state error;*
2. *break pulsation* is *larger*;
3. *damping factor* is *smaller.*

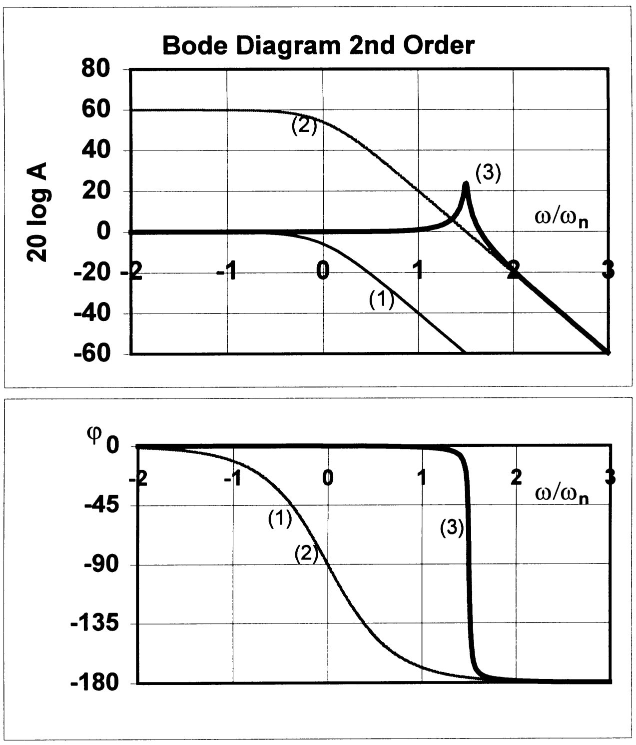
For very large values of α the system would behave as an ideal follower. However, increasing α also causes a decrease of the damping factor ξ, resulting in underdamped oscillations, eventually decreasing the stability of the system. Therefore α may not be too large. When $\alpha \to \infty \Rightarrow \xi \to 0$: the system would become unstable because of undamped oscillations.



1 Open

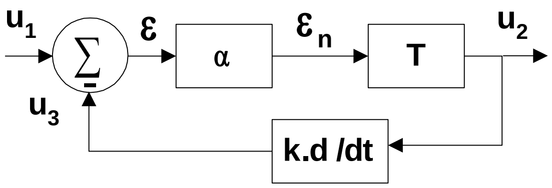
2 Open with extra gain

3 With unit feedback



(2) Differentiating feedback.

Second order process.



(i) Differential equation.

$$\frac{1}{\omega_n^2}\cdot\frac{d^2u_2}{dt^2}+\frac{2\xi}{\omega_n}\cdot\frac{du_2}{dt}+u_2=\alpha\cdot\varepsilon$$

$\varepsilon = u_1 - u_3$

$u_3 = k\cdot\frac{du_2}{dt}$ (k has dimension of time)

$$\frac{1}{\omega_n^2}\cdot\frac{d^2u_2}{dt^2}+\frac{2(\xi+\omega_n\alpha k/2)}{\omega_n}\cdot\frac{du_2}{dt}+u_2=\alpha\cdot u_1$$

(ii) Properties.

$\omega'_n = \omega_n$

$\xi' = \xi+\omega_n\alpha k/2$

$\alpha' = \alpha$

The damping factor is larger, with as a consequence: less danger of instability by resonance.

N.B. when k is very large, most terms can be neglected

$u_1 \cong k\cdot\frac{du_2}{dt}$

$u_2 = \frac{1}{k}\cdot\int u_1 dt$: the system becomes an integrator

(3) Integrating feedback.

(i) Differential equation.

$$\frac{1}{\omega_n^2}\cdot\frac{d^2u_2}{dt^2}+\frac{2\xi}{\omega_n}\cdot\frac{du_2}{dt}+u_2=\alpha\cdot\varepsilon$$

$\varepsilon = u_1 - u_3$

$u_3 = k.\int u_2 dt$ (k has dimension of time^{-1})

$$\frac{1}{\omega_n^2}\cdot\frac{d^2u_2}{dt^2}+\frac{2\xi}{\omega_n}\cdot\frac{du_2}{dt}+u_2+\alpha k\cdot\int u_2 dt=\alpha\cdot u_1$$

Differentiation of both sides of the equation:

$$\frac{1}{\omega_n^2}\cdot\frac{d^3u_2}{dt^3}+\frac{2\xi}{\omega_n}\cdot\frac{d^2u_2}{dt^2}+\frac{du_2}{dt}+\alpha k\cdot u_2=\alpha\cdot\frac{du_1}{dt}$$

(ii) Properties.

We thus obtain a diff. eq. of higher order.

We do not give the solution of the system but only want to mention that this type of system has no steady state error.

N.B.: when k is large the system behaves as a differentiator.

$$u_2 \cong \frac{1}{k} \cdot \frac{du_1}{dt}$$

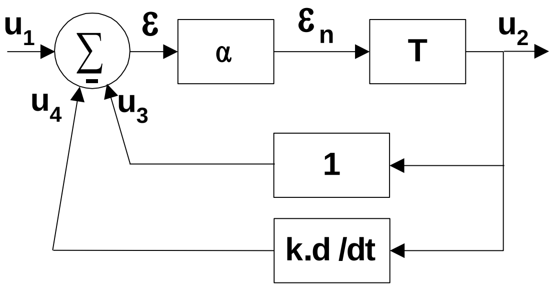
(4) Combined feedback.

Most control systems make use of proportional feedback. Pure proportional feedback however has some disadvantages:

- danger of instability by resonance (only for systems of 2nd or higher order).
- steady state error.

These problems can eventually be corrected by combining proportional feedback with differentiating and/or integrating feedback, since differentiating feedback increases the damping factor, while integrating feedback eliminates the steady state error.

As an example we show proportional feedback with differentiating correction in a second order system.



(i) Differential equation.

$$\frac{1}{\omega_n^2} \cdot \frac{d^2u_2}{dt^2} + \frac{2\xi}{\omega_n} \cdot \frac{du_2}{dt} + u_2 = \alpha \cdot \varepsilon$$

$\varepsilon = u_1 - u_3 - u_4$

$u_3 = u_2$

$$u_4 = k \cdot \frac{du_2}{dt}$$

$$\frac{1}{(1+\alpha).\omega_n^2} \cdot \frac{d^2u_2}{dt^2} + \frac{2(\xi+\omega_n\alpha k/2)}{(1+\alpha).\omega_n} \cdot \frac{du_2}{dt} + u_2 = \frac{\alpha}{1+\alpha} \cdot u_1$$

(ii) Properties.

The transfer function T' of the closed system thus has as parameters:

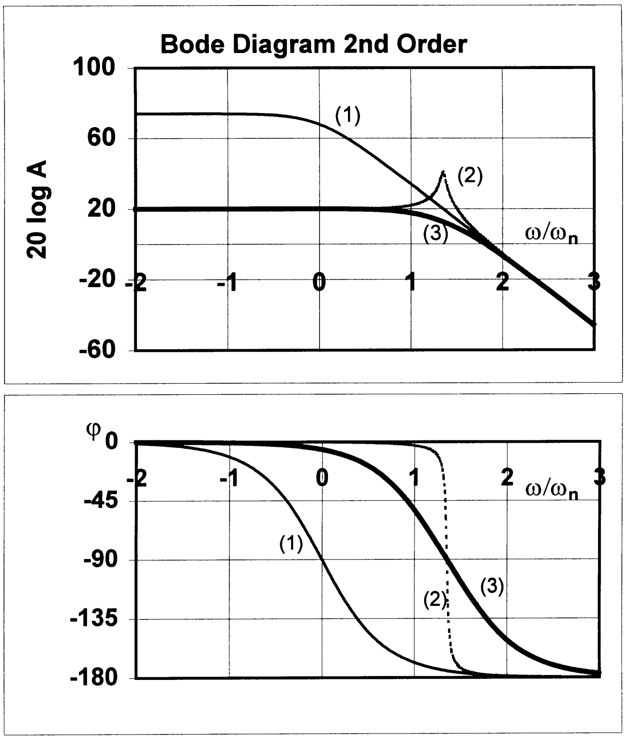
$\omega'_n = \omega_n . \sqrt{(1+\alpha)}$

$\xi' = \xi/\sqrt{(1+\alpha)} + \omega_n\alpha k/2\sqrt{(1+\alpha)})$

$\alpha' = \alpha/(1+\alpha)$

We can see that the differentiating feedback has no influence on the natural frequency or gain, since these parameters are identical in the combined feedback and in the purely proportional unity gain feedback, but that it slows the closed loop system by increasing the damping factor.

Bode diagram



1. Open.
2. Proportional feedback.
3. Proportional feedback with differentiating correction.

b) Differentiating regulators (**ε_n = k . dε/dt** ***).***

u_2 is only in steady state when $\varepsilon_n = 0$

$\varepsilon_n = k.d\varepsilon/dt = 0 \Rightarrow \varepsilon = \text{constant}$ (k has dimension of time)

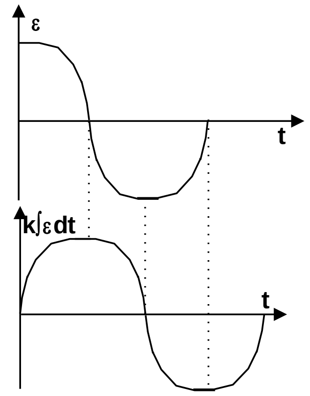
A differentiating regulator allows a constant error; only changes in the error are corrected. When the error ε changes very fast, ε_n is very large, and a large regulating action will ensure a fast correction. With very slow changes, ε_n is so small that the regulator does not detect ε_n so that no correction follows. A system with a pure D-regulator therefore can drift slowly without any correcting action. For this reason, a D-regulator is normally used in combination with a P-regulator, where it improves the transient behavior, by providing a phase lead that compensates phase lags in the system, and thereby can enhance the bandwidth of the system.

c) Integrating regulators (**ε_n = k . ∫ εdt** ***).***

u_2 is only in steady state when $\varepsilon_n = 0$.

$\varepsilon_n = k \,.\int \varepsilon \, dt = 0$ (k has dimension of time^{-1})

When an error is present, a correction will follow. As long as some error remains, the correcting action increases. The correcting action only does not change when ε = 0, thus when $u_2 = u_1$, and therefore steady state errors are excluded in this case. This is however not a sufficient condition for steady state:



A constant ε error produces an increasing correcting action ε_n. This causes ε to decrease; however as long as ε > 0, the correcting action increases, eventually making ε < 0. When ε becomes negative, the correcting action starts decreasing but initially is still positive so that ε keeps decreasing, until the correcting action becomes zero. At that moment the error is negative and the process of correction will occur in reverse direction. Therefore, an integrating regulator always keeps chasing the error.

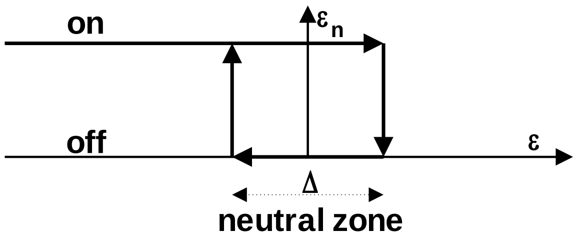
A second disadvantage of the I regulator that it is slow, so that also the integrating regulator is most often used only in combination with a P regulator, to enhance the low frequency gain of the system.

Combination of differentiating and integrating regulators is often used in addition to a proportional regulator to improve system performance. These regulators are called PID regulators.

d) All or none regulator (ε_n = 0 or 1).

E.g.: thermostat.

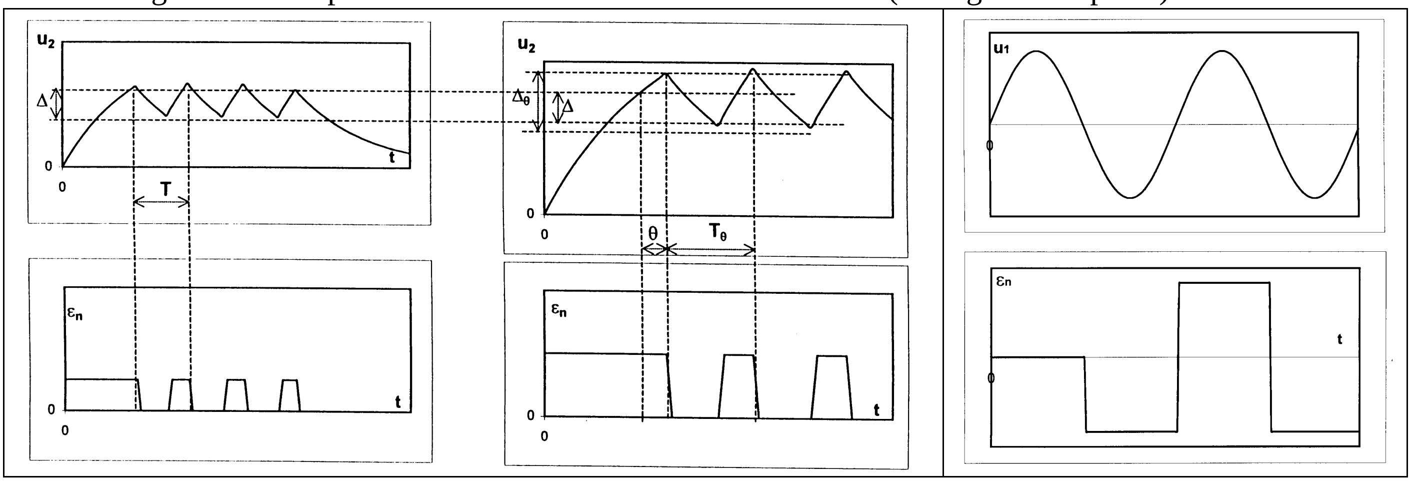
Transfer function of the regulator.



The transfer function of an all or none regulator shows hysteresis. If this would not be the case, an infinitely small error ε would already result in correcting action switching the system between 0 and 1 at an infinite rate. Most processes allow a small error in the (neutral zone). When ε varies within the neutral zone, no correction takes place, so that the switching frequency of the process is decreased.

Application: temperature regulation by thermostat.

When the temperature in the room is too low, a large negative ε will be present. The output of the regulator will thus be in "on" position and the heating element will be produce heat. This will cause a rise of the room temperature and ε becomes less negative and eventually even positive. When $\varepsilon = + \Delta /2$, the regulator switches to its "off" position. This causes the heat flow to stop and the room temperature decreases due to heat losses: ε becomes less positive, reaches zero and subsequently becomes negative. At the moment that $\varepsilon = - \Delta /2$ the regulator switches to the "on" position and the room temperature will increase again. The temperature thus varies within the neutral zone (see figure: left panel).



Influence of transportation lag θ.

A transportation lag causes a delay of the regulating action, so that the output exceeds the desired neutral zone. As a consequence, the effective neutral zone is wider, and the switching rate is decreased.

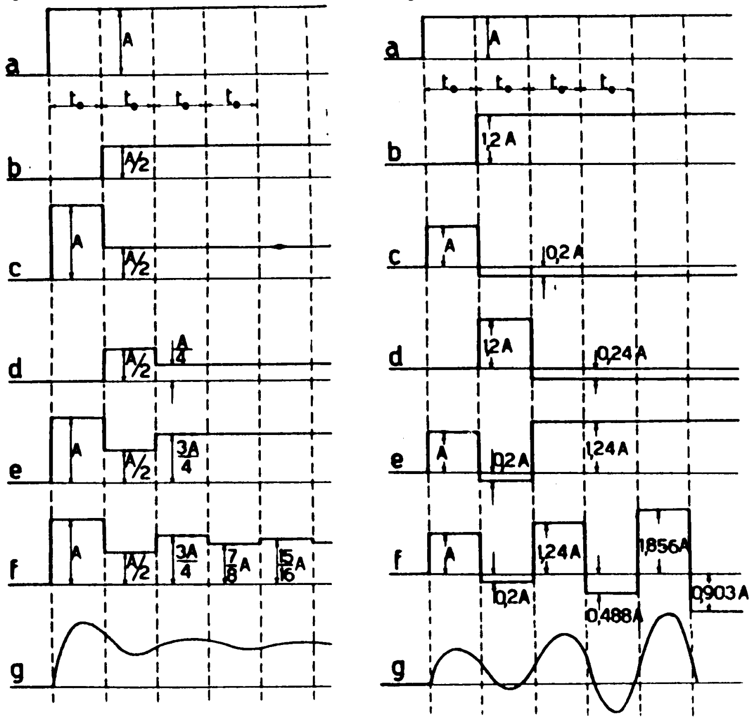
When fast temporary changes of u_2 occur, it is possible that the system only reacts after the temporary disturbance has already disappeared. If this is the case, the correcting action will produce a late change in the opposite direction, which itself will cause a correcting action (see right panel figure). It is thus possible that a system with a transportation lag will oscillate in response to a disturbance.

IV. STABILITY OF CONTROL SYSTEMS.

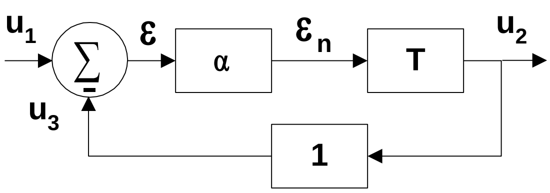
A. Problem.

Systems of higher order can contain sinusoidal transient components, but in open systems the amplitude of these vibrations will generally diminish with the time because of energy losses (friction). In systems with feedback, this is not always the case. In these systems a disturbance can eventually cause vibrations with increasing amplitude (until the saturation level of the system is reached). Such a system is thus unstable. Instability is a property of closed systems.

N.B. Internal interactions in "open systems" can cause hidden internal feedback within the system, so that apparently open systems can also show instability.



A closed system can oscillate spontaneously in the absence of an input signal. An output signal is present with parameters only dependent of the system: the effect of a brief disturbance keeps influencing the system indefinitely. To give a correct non-mathematical physical representation of the mechanism of the generation of instability is not easy. Following simplified representation should make clear the fundamental reason. As an example we start from a system T of 3rd order.



The maximal phase shift of the system T is -270^o, and therefore there is a certain ω_k at which $\varphi_{\omega k} = -180^o$. A phase shift of 180^o is however a sign inversion. Thus, when $u_1 = A_1.\sin\omega_k t$ we have:

$u_2 = A_2.\sin(\omega_k t - 180^o) = -A_2.\sin\omega_k t$

$u_3 = u_2$

$\varepsilon = u_1 - u_3$

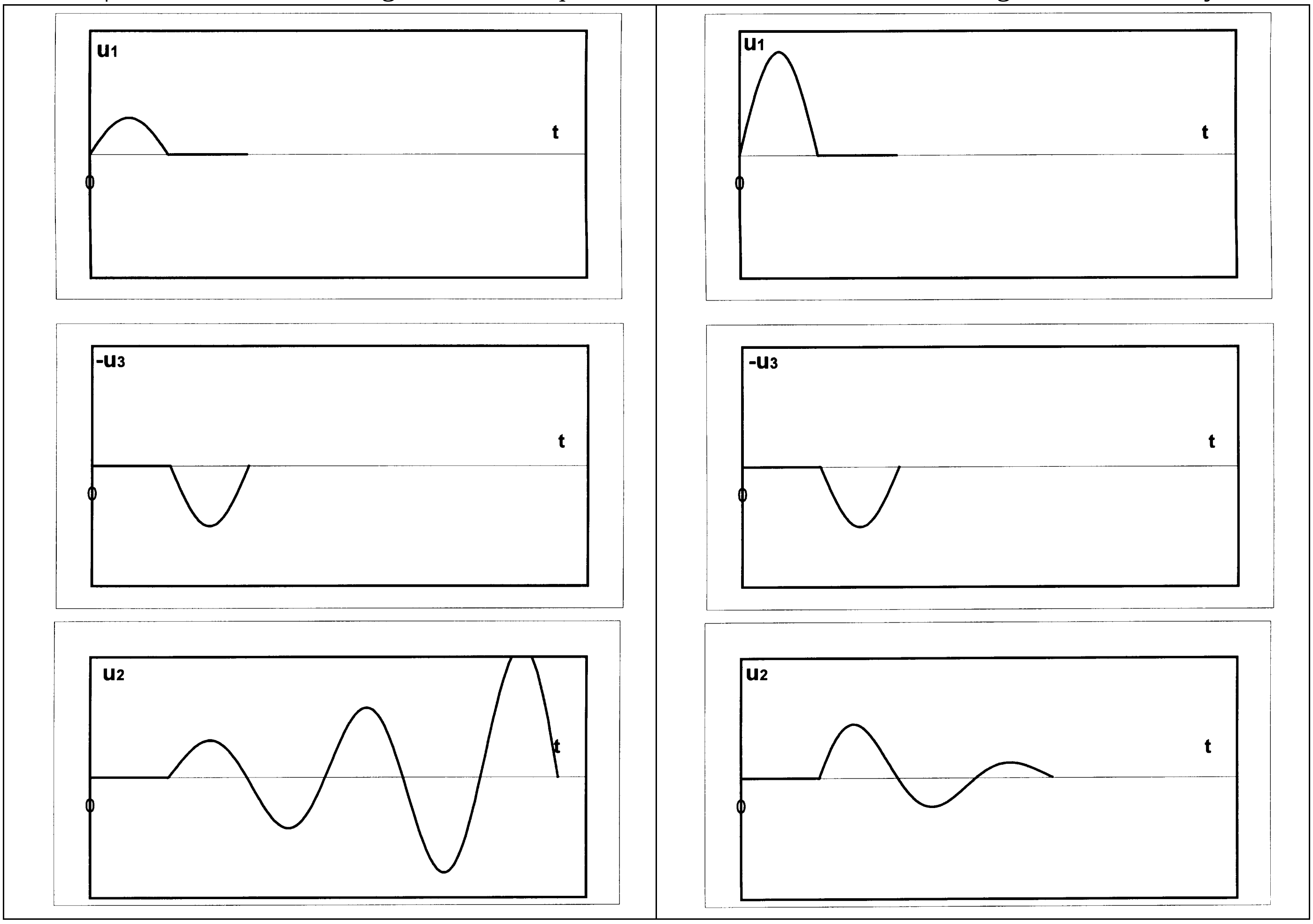
$\varepsilon = A_1.\sin\omega_k t - (-A_2.\sin\omega_k t)$

$\varepsilon = (A_1 + A_2).\sin\omega_k t > A_1.\sin\omega_k t$

Therefore the feedback causes ε to increase rather than decrease. Instead of a negative feedback we now have a positive feedback, which will lead to saturation of the output.

B. Origin of the resonance.

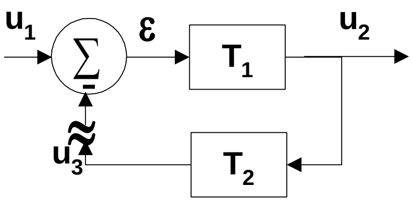
Assume a brief disturbance that can be approximated as half of a period of a sinus with frequency ω_k where $\varphi_{\omega k} = -180^o$. We investigate the consequences of such a disturbance in a negative feedback system.



The disturbance is amplified by the system an fed back to the input, but is half a period delayed because of the -180^o phase shift. The new input is thus $-u_3$. Similarly this new input is again amplified and fed back; as final result, we get a sinus wave with amplitude increasing with time. A small temporary disturbance will thus produce a vibration with continuously increasing amplitude, which is only limited by the saturation levels of the system. The characteristics of this vibration are only dependent on the system. The vibration has a sinusoidal shape. The frequency of the vibrations is ω_k, which is exactly the frequency at which the phase shift is 180^o. This frequency is called the natural frequency of the system. The amplitude will reach the saturation level of the system. With these two properties of the open system the shape of eventual oscillations in the closed system can be predicted.

The amplitude of the sinusoidal vibration, however, does not always increase with time. Suppose that the gain $\alpha < 1$, then $u_2 = \alpha.u_1 < u_1$. Thus $u_3 < u_1$, and therefore the amplitude of the input ε now decreases with time, and we now obtain a damped sinusoidal vibration. Thus in this case the influence of a brief disturbance decays with time, and the system does not oscillate, since for oscillation it is required that the system sustains the vibration.

The condition for oscillations in a control system is therefore that $|u_3| \geq |u_1|$ when u_3 has a phase shift of 180^o with respect to u_1. When this condition is not fulfilled the vibration decays.



Summarizing: the system will be unstable when the gain of the open loop > 1 at the frequency at which $\varphi = 180^o$. In this context "open loop" means the open system with as input u_1 and as output u_3, thus with $u_3 = T_2\, T_1\, u_1$ (thus after interrupting the connection u_3 from the feedback process to the comparator).

C. Nyquist Criterion.

The stability condition can also be directly derived from the transfer function of the closed system.

$$u_2 = \frac{G}{1 + GH} u_1$$

A finite amplitude of the output requires that the denominator (1+GH) of the transfer function must be different from zero. When GH = -1, the denominator is equal to zero and therefore $u_2 = \infty$. GH is the open loop transfer function of the system, and in terms of frequency domain analysis, GH = -1 is obtained when the open loop gain is equal to 1 and the open loop phase shift equal to 180°. Whenever at a particular frequency (ω_a) the gain in the open loop is equal to 1 (0 dB) and the phase shift 180°, then, according to the theory, at this frequency (ω_a) the amplitude of the output of the closed loop will be infinitely large (in practice limited by saturation). The system will thus oscillate at the frequency ω_a.

Somewhat simplified, the **Nyquist** s**tability criterion of** states "**A system with negative feedback is unstable when the gain of the open loop ≥ 1 at the frequency where the open loop phase shift $\varphi = 180^o$.**"

From the Nyquist criterion it can easily be determined by frequency domain analysis of T_1 and T_2 whether a closed system will stable.

In the most cases (when gain and phase shift monotonously decrease with increasing frequency) the Nyquist criterion is equivalent to a somewhat less general formulation of the stability criterion: "The closed system will be unstable when the open-loop system has a phase shift at the "unity-gain" frequency that is larger or equal to 180^o."

1. Nyquist criterion and Bode diagram.

We construct the open loop Bode diagram.

$u_3 = T_2.T_1\, u_1 \quad T = T_2.T_1$

$K = K_1 + K_2$

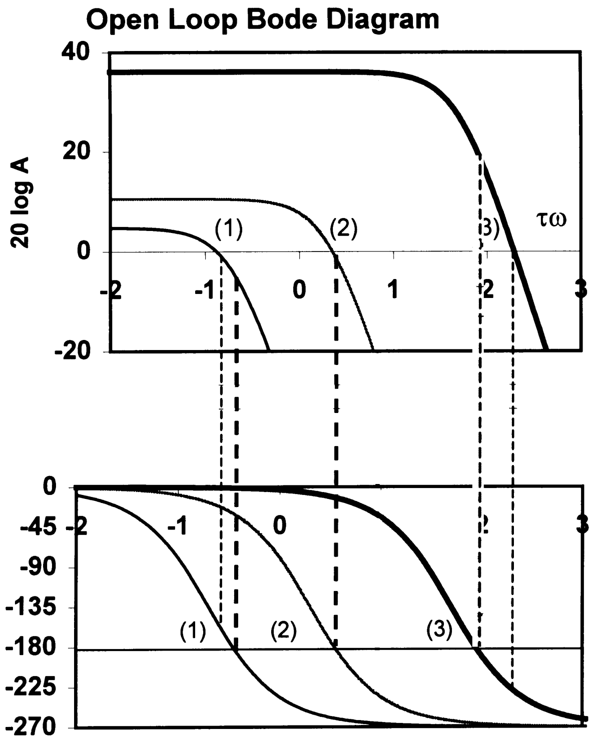
$\varphi = \varphi_1 + \varphi_2$

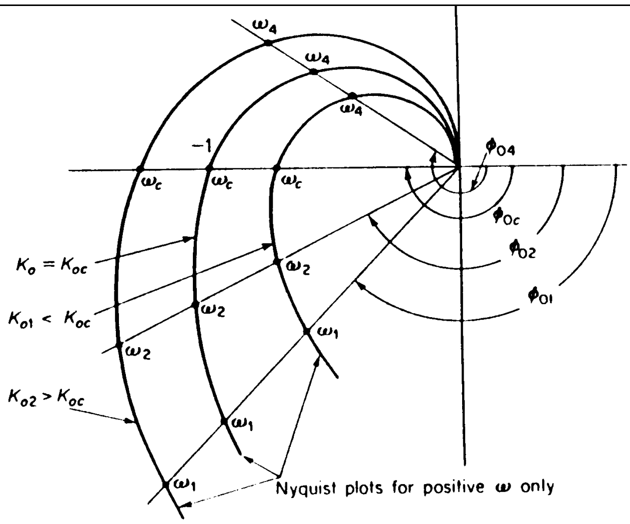
Application of the Nyquist criterion in its first form requires to verify whether the open loop gain K< 0 when the open loop phase shift $\varphi = -180^o$. Application of the criterion in its second form means to verify whether the open loop phase shift is smaller than -180^o at the frequency where K is smaller than zero.

The figure shows open loop Bode diagrams of three different systems. System 1 will result in a stable feedback system, while both other closed systems will be unstable. This is illustrated using both forms of the Nyquist criterion.

A the frequency where $\varphi_{\omega a} = -180^o$, only system 1 has an open loop gain smaller than 1 (K<0).

Let ω_a be the "unity gain" frequency ($K = 0$ for $\omega = \omega_a$). In system 1: $\varphi_{\omega a} < -180^o$. The system is thus stable. Systems 2 ($\varphi_{\omega a} = -180^o$) and 3 ($\varphi_{\omega a} > -180^o$) will be unstable when the loop is closed.





Effect of K_o on Nyquist plots of the form

2. Nyquist criterion and polar diagram.

We construct the polar diagram of the open loop system.

$u_3 = T_2.T_1\, u_1 \quad T = T_2.T_1$

$A = A_2.A_1$

$\varphi = \varphi_1 + \varphi_2$

We now apply the Nyquist criterion in its first form. The *system is unstable* when the open loop vector has an amplitude equal to 1 or larger than 1, at a frequency where the phase shift equals -180^o. When *the curve crosses or surrounds the point* -1 *on the abscissa* the closed system unstable (curve 2).

Only the closed system with open loop characteristic corresponding to the inner curve 1 will be stable.

D. <u>Phase and amplitude margin.</u>

While a system that fulfills the Nyquist criterion will be stable in absolute terms, its dynamic behavior may be close to instability. Suppose e.g. that $u_3 = 0.99u_1$ at the frequency ω_k where $\varphi_{\omega k} = 180^o$. In this case vibrations will be damped, but the damping will be very weak, so that it will take a considerable time before they have decayed. Therefore it is useful to introduce a measure of relative stability. For this purpose the concepts of phase margin and gain margin have been defined.

1. Concept of phase margin.

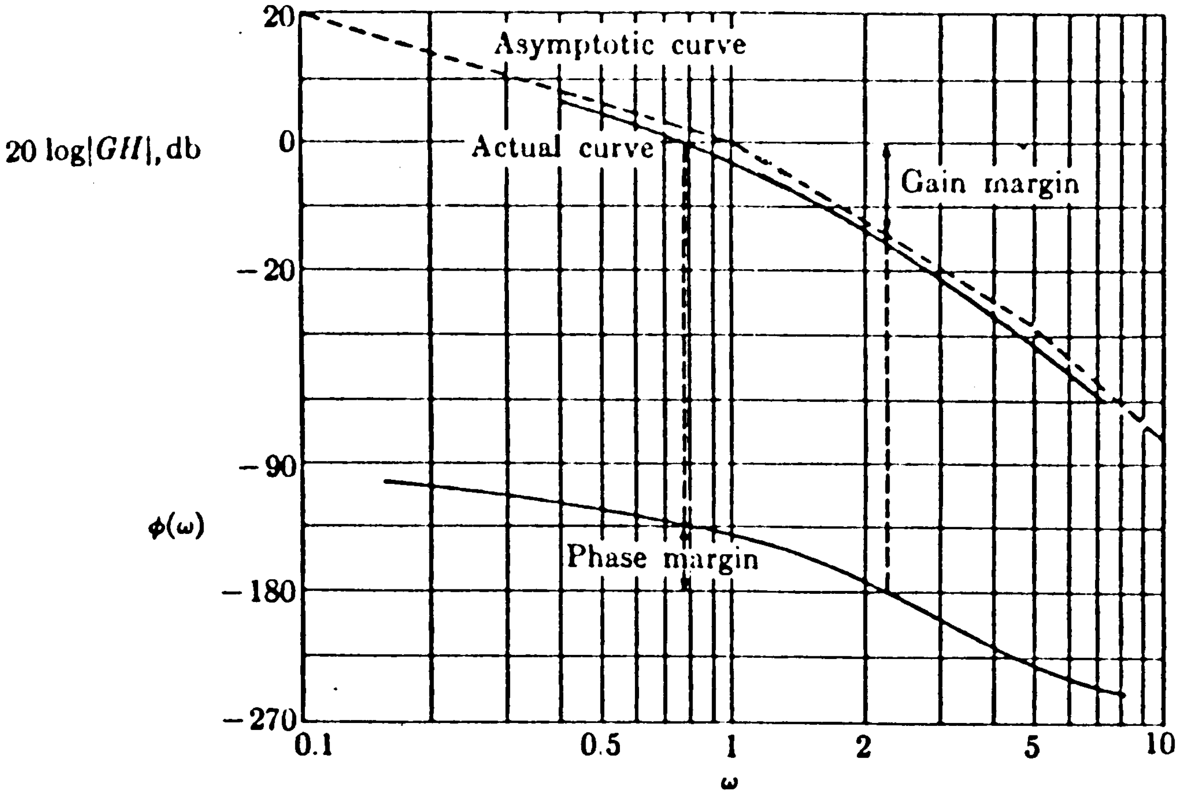
Definition: Let ω_a be the "unity gain" frequency of the open loop (thus $K_{\omega a} = 0$), then $F \equiv 180^o - \varphi_{\omega a}$ is called the *phase margin* of the system.

When the system stable, the phase shift at the unity-gain frequency is smaller than -180^o. The phase margin is the extra phase shift that can be introduced in the open loop before the closed system would become unstable. A large phase margin thus means a very stable system.

2. Concept of gain margin.

Let ω_k be the frequency where the phase shift of the open loop $\varphi_{\omega k} = -180^{\circ}$, then $V \equiv -K_{\omega k}$ is called the *gain margin* of the system.

The gain margin is the supplementary amplification (in dB) that can be introduced in the open loop system before the closed loop system would become unstable. A large amplitude margin therefore means a very stable system.



V. PRACTICAL METHODS.

A. Practical methods for studying control systems.

1. Synthesis of control systems.

The basic problem for the synthesis of a control system is how a particular process can be controlled to ensure than the output of the system fulfills certain requirements. In practical terms this most often means that one has to determine how the system wil behave when feedback T_2 is added to the process T_1.

- **Stability analysis.**

The transfer function of the "open loop" system is determined: $u_3/u_1 = T_2T_1$. The stability of the system can then be determined by application of the Nyquist criterion or other methods (e.g. root locus method).

- **Determination of the transfer function of the closed system.**

In order to know beforehand the exact behavior of the control system, transfer function T_1 of the process must be known in addition to the open-loop transfer function T_2T_1. Using e.g. frequency domain analysis T_1 is determined. From both transfer functions the transfer function of the closed system can then be calculated:

$$u_2 = \frac{T_1}{1 + T_1T_2} u_1$$

2. Analysis of control systems.

The purpose of analysis of a closed system is to determine the properties of the closed system and of its different components.

- Measurement of the transfer function of the closed system (e.g. with frequency domain analysis)
- Opening of the system and application of system analysis to the different processes of the system.

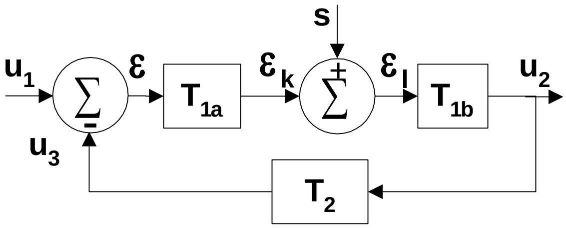
B. Study of regulator systems.

Definition: a regulator system is a system where the input is kept at a constant value, and the system attempts to keep the output constant despite disturbances.

E.g.: thermostat, homeostatic regulation of ion concentrations in the blood.

1. The effect of disturbances.

The main purpose of the study of regulator systems is to determine in how far the output of the system is kept at the desired level, in other words, how the system output changes with external disturbances. The relation between the output of the system and the disturbance has therefore to be investigated. Therefore “an other” system is therefore defined with as output u_2 and as input the disturbance s. Thus $u_2 = f(s)$.



The disturbance is then changed sinusoidally and the output is measured. The analysis is then not different from the analysis described previously. A difficulty however is to determine the exact site in the system where the disturbance acts as an input. When this is known, the process T_1 is split in two sub processes

T_{1a} and T_{1b} in series, so that: $T_1 = T_{1a}T_{1b}$ and that are defined in such a way that the disturbance acts as an input to T_{1b}. The disturbance can then be summed with the output ε_k of T_{1a}, and the input ε_1 of the closed system then becomes:

$\varepsilon_l = s + \varepsilon_k = s + T_{1a}\varepsilon = s + T_{1a}u_1 - T_{1a}u_3$

$\varepsilon_l = s + T_{1a}u_1 - T_{1a}T_2\, u_2$

$u_2 = T_{1b}\varepsilon_l = T_{1b}s + T_1u_1 - T_1T_2\, u_2$

$(1 + T_1T_2)\, u_2 = T_1u_1 + T_{1b}s$

$$u_2 = \frac{T_1}{1 + T_1T_2}\, u_1 + \frac{T_{1b}}{1 + T_1T_2}\, s$$

2. The set point.

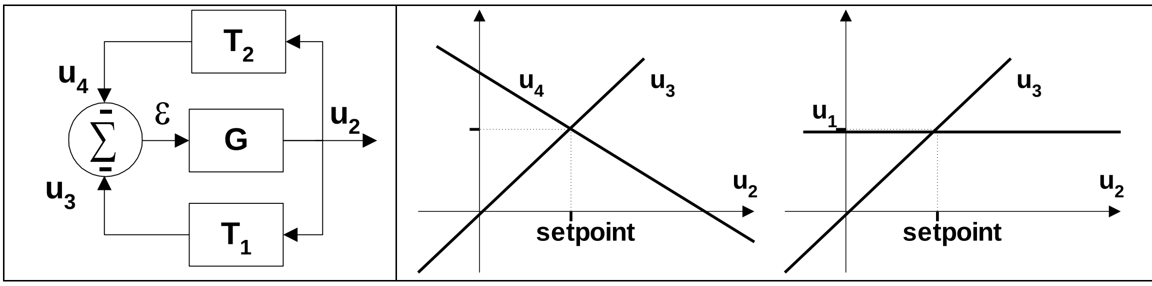
For many regulator systems (especially in living organisms) defining the input (reference) signal u_1 is not easy. A separate input signal is however not required to ensure a stable output level. Most often the result is determined by a number processes that keep each other in balance.

E.g. The normal heart rate is determined by a balance of sympathetic activity (adrenaline causes a rise of heart rate) and of parasympathetic activity (acetylcholine decreases the rate) of the autonomic nervous system.

A stable equilibrium of a parameter can in principle be obtained by two processes that are dependent on this parameter, but have an opposite effect on the parameter.

As an example we consider a simple system with two proportional feedback processes (T_1 and T_2) in parallel. For simplicity, we further assume that the open loop gain of the system is large, so that we can neglect the term 1 in the denominator of the transfer function of the closed system. We also assume that

- the two processes cause effects in opposite direction (slope A and –B with A and B positive); one process thus causes a negative and the other a positive feedback;
- the output of process u_3 (u_4) reaches a stable value at particular value C (K) of its input (u_2).



$u_2 = G\varepsilon$

$\varepsilon = -(u_3 + u_4)$

$u_3 = Au_2 - C$

$u_4 = -Bu_2 - K$

$$u_2 \cong \frac{C+K}{A-B}$$

The output (u_2) of the system will reach a stationary value determined by the constants and slope factors of both processes. The equilibrium level can thus be influenced by a shift or by a change in slope of each of the two curves.

It is clear that, when B and C are equal to zero (thus $u_3 = Au_2$ and $u_4 = -K$), u_4 obtains the meaning of a reference input ($u_1 = K$) since ($u_2 = K/A$). A reference input u_1 of a regulator can then be considered as a special case, when one of the processes that influence the direct system (G) is independent of the output.

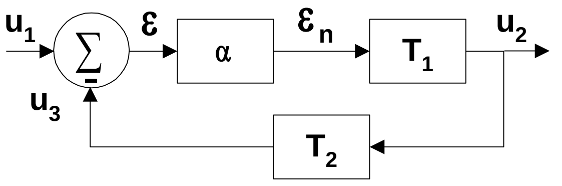
E.g.: Temperature regulation.

A rise of body temperature causes activation of processes that result in heat loss (e.g. sweating, peripheral vasodilatation), and inhibition of heat production (e.g. shivering). These processes are controlled by heat sensitive and cold sensitive neurons in the hypothalamus. The normal body temperature is the temperature at which both types of processes exactly balance each other.

An infection will result in secretion of pyrogens by white blood cells (pyrogens are substances that affect the activity of temperature sensors in the brain), e.g. interleucine-1, which is secreted by macrophages. The balance between heat production and heat loss will hereby be changed, resulting in fever. Fever is the consequence of a change of the set point of the thermoregulation. This explains a number op apparently paradoxical phenomena.

- A feeling of cold (cold shivering) and peripheral vasoconstriction (looking pale) occurs when the fever starts and the temperature rises, while the body temperature at that moment is higher than normal. The infection has increased the set point of the thermoregulation (e.g. from 36° to 39°), but as long as the body temperature did not yet reach this new set point, the sensor will indicate that the real body temperature is lower than the set point, and thus heat production will be stimulated and heat loss inhibited.
- Sweating is also often the first signal of recovering from the infection (as is known already for a very long time by conventional wisdom). When the patient is recovering from the infection, the thermostat returns to its normal level of 36°; however the body temperature is still temporary too high, so that heat loss processes (sweating, vasodilatation) are stimulated. Since sweating is the consequence, not the cause, of the healing process, it should be clear that this phenomenon should not lead to the conclusion that one must make a patient sweat to support the healing process (as is often thought).

C. <u>Fast approximate analysis of control systems.</u>



$$u_2 = \frac{\alpha T_1}{1 + \alpha T_1 T_2} u_1$$

Since $\varepsilon = u_2 / \alpha T_1$:

$$\varepsilon = \frac{1}{1 + \alpha T_1 T_2} u_1$$

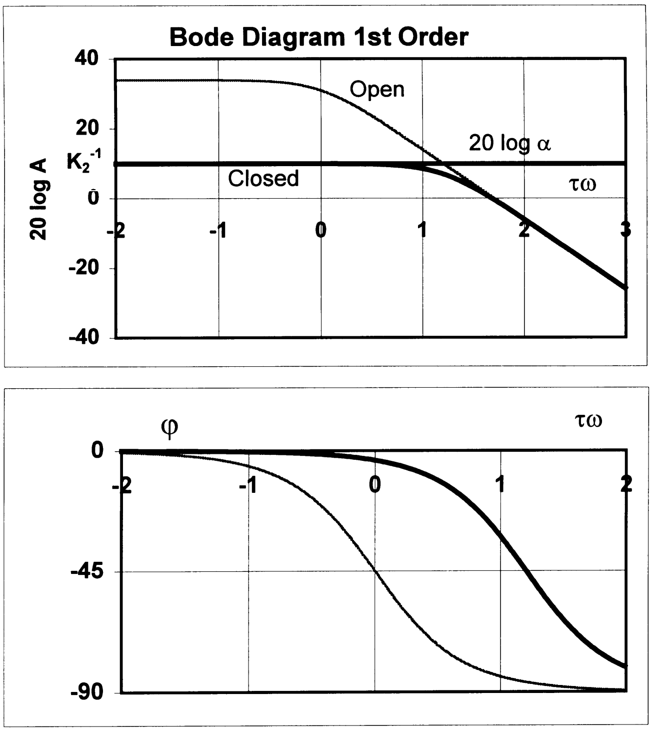
For $\alpha T_1 T_2$ very large, ε thus becomes very small; for an ideal system with $\alpha \to \infty$ then $\varepsilon \to 0$. For good control systems we can therefore approximate ε as zero. Since $\varepsilon = u_1 - u_3$, we thus obtain $u_3 \cong u_1$. Since $u_3 = T_2 u_2$ it follows that $T_2 u_2 \cong u_1$ and thus

$$\mathbf{u_2 \cong T_2^{-1} u_1}$$

We see that α and T_1 do not anymore play a role in the relation between output and input of the closed system, and that the output u_2 only depends on the input and on the feedback process. Disturbances in the direct chain do not influence u_2, provided that α is very large. *The closed loop system behaves as a (open) system with transfer function that is the inverse of the transfer function of the feedback process*!

When can this approximation be applied?

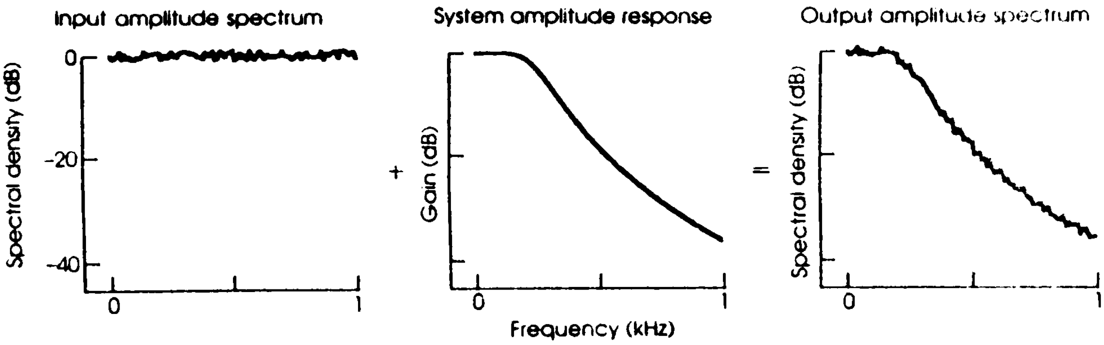
The approximation is determined by the open loop transfer function. When the open loop gain decreases the error ε will increase. Since the gain of the open loop transfer function decreases at high frequencies, the gain of the closed system will also decrease for higher frequencies. Whenever the feedback is negative, the Bode diagram of the closed system has a lower gain than the open loop gain. The closed loop gain will therefore be limited at high frequencies by the (frequency-dependent) gain of the open loop Bode diagram.



Since the closed system cannot deliver more output than the open system, the approximation is only valid up to frequencies where the curve of the approximate behavior of the closed system intersects the curve of the open system. In the range of higher frequencies, the gain of the open system is too small for valid application of the approximation. The curve of the closed loop gain is only valid in the range where it lies below the curve of the open loop gain.

D. Practical Fourier analysis.

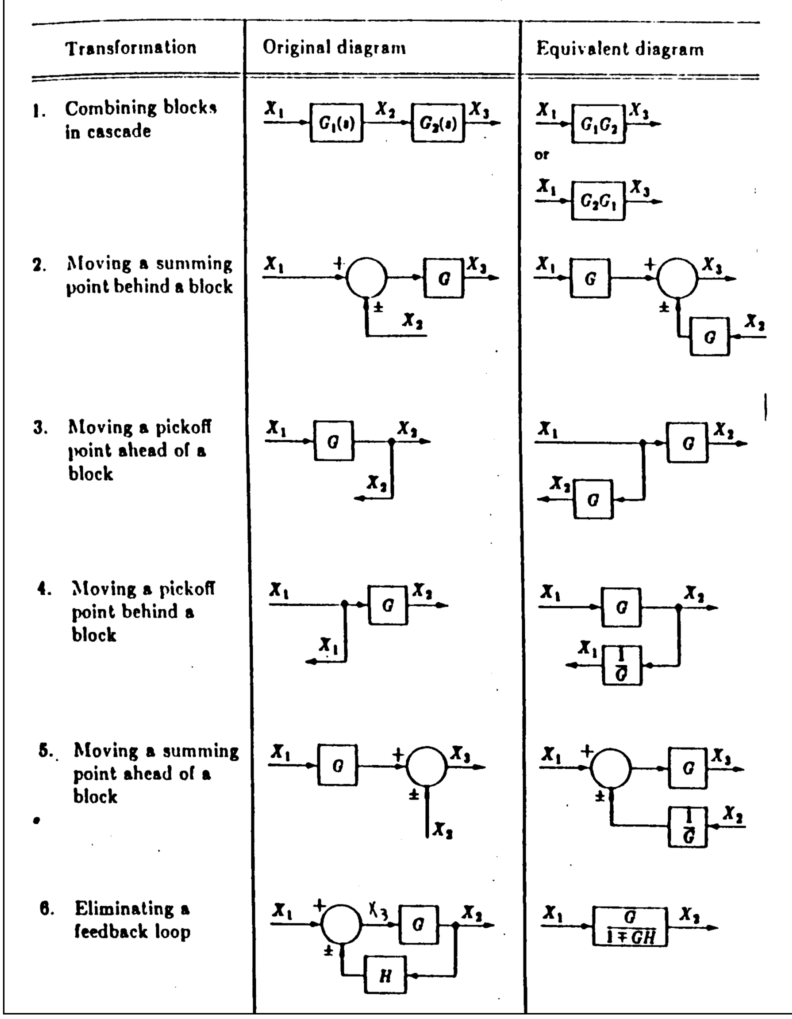
The method of Fast Fourier Transformation (FFT) enables a faster way to do a frequency domain analysis of a system. In principle the sinusoidal input signals need not be applied separately, but can be combined into a single signal applied to the input of the system. The output of the (linear) system will then also contain all these frequencies. In principle, these frequency components can be separated by sending the output to set of parallel filters that each pass only a single frequency (in reality a narrow frequency band). In practice the separation is realized by means of the Fourier analysis technique. The output is sampled at different times and these values are sent to a computer or Fourier analyzer. The Fourier transformation of the discrete data is done using an FFT digital program. This way one obtains a spectrum of amplitudes as function of frequency. When the input of the system under investigation has flat frequency spectrum, the output spectrum gives thus the frequency characteristic of the system. Therefore, as input "white noise" is used, which has the property that the power is constant for all frequencies.

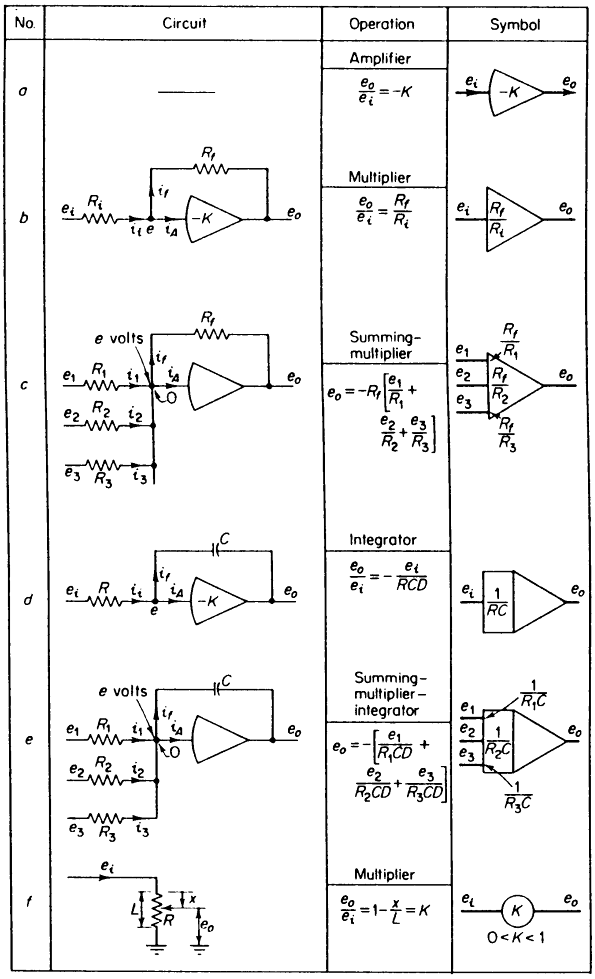


E. Block diagram transformations.

Block diagrams of complex systems with many components can most often be largely simplified using simple transformation techniques. The transfer functions are temporarily represented by a symbol. The structure of the system is simplified by combining processes with simple combination rules (e.g. two processes in series are replaced by one process with transfer function equal to the product of the two transfer functions). Only after such simplification of the system (often to a single block), the symbols for the different processes are replaced by the corresponding transfer functions.

Block Diagram Transformations





Analogue computer circuits, operations, symbols.

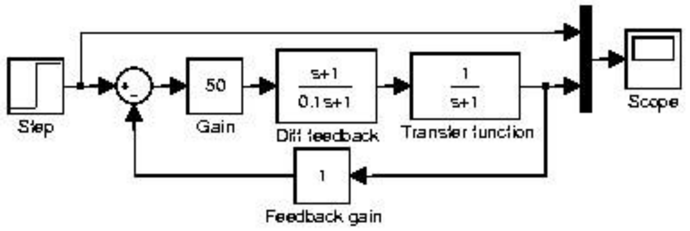
F. Simulations.

A system can be simulated by using analog or digital computers.

An analog computer consists of a series summing amplifiers, integrators, multipliers (see figure) and function generators. These functions are most often sufficient to simulate the differential equation. All variables are hereby simulated using electrical voltages.

With the advent of fast digital computers, most often digital computers are used for simulations. A computer program for solving the differential equation can be developed in a common computer language Assembler, Basic, Fortran, Pascal, C, Java, Lisp, Prolog etc.. An important aid is of course the existence of libraries of sub-programs that can be used as part of the program. Since several years however, also computer languages or software libraries have been especially developed for system simulation. With such software, the different operators and components of the system are represented by a symbol. With these symbols, only the elements of the system and their connections, the input and the desired output need to be symbolically defined, often by using a graphical user interface (GUI). The software (e.g. LABVIEW, or MATLAB and its toolbox SIMULINK) takes care of the generation of the system equations and the solution of the differential equations, and for printing and plotting of the results and curves.

```
tau_p = 1
gain_r = 50
tau_l_r = 1
tau_h_r = 0.1
gain_f = 1
sys_r = tf(gain_r*[tau_l_r 1],[tau_h_r 1]);
sys_p = tf([1],[tau_p 1])
sys_d = series(sys_r,sys_p);
sys_f = tf([gain_f],[1]);
sys_c = feedback(sys_d,sys_f)
step(sys_c);
```



Matlab simulation of step response of first order system with proportional feedback and PD regulator.
Left: Matlab m file code.
Right: Matlab Simulink symbolic simulation using graphical interface with predefined blocks.

Advantages of simulation can be summarized as follows:

- sufficient understanding of a system requires quantitative and therefore mathematical description of the system and its components, to enable to make quantitative predictions. In cases where analytical solution of the equations is impossible or too complex, simulation is the only alternative to test our knowledge of the system;
- testing the system behavior via simulation helps to understand the behavior of the system in response to certain changes of the parameters: because of the many interactions the response of a complex system can often be very different from intuitive expectations;
- simulations allow to study system behavior in conditions that could not be realized on real systems (e.g. destructive test);
- important time and money saving can be obtained by limiting experiments with the real system to very critical experiments.

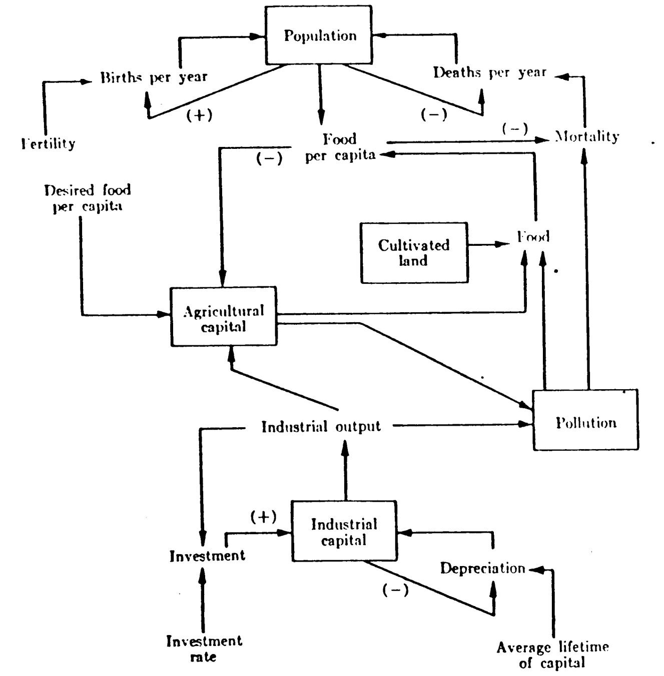
It is however also important to realize the limitations of simulations:

- description of systems often requires a large number of differential equations with many parameters;
- the equations are often nonlinear;
- often part of the elements in the system are not sufficiently known;
- some of the parameters are often not known with sufficient accuracy;
- simulations can provide new insight, but no new knowledge, as is often naively suggested.

It is thus almost always necessary to do a very large number of simulations at different combinations of the values of these parameters. A good analysis by simulation always requires to investigate the sensitivity of the output to the different parameters, and it is very important to perform such a sensitivity analysis at different combinations of the values of the parameters, since a change of a particular parameter could eventually strongly influence the sensitivity of the system to another parameter. Processes and parameters that under certain circumstances are relatively unimportant for the behavior of the system, can become very dominant when lowering or increasing some parameter of another process in the system. This is especially the case with non-linear systems, some of which can result in chaotic behavior.

Simulation of the world environment in the report of the Club of Rome.

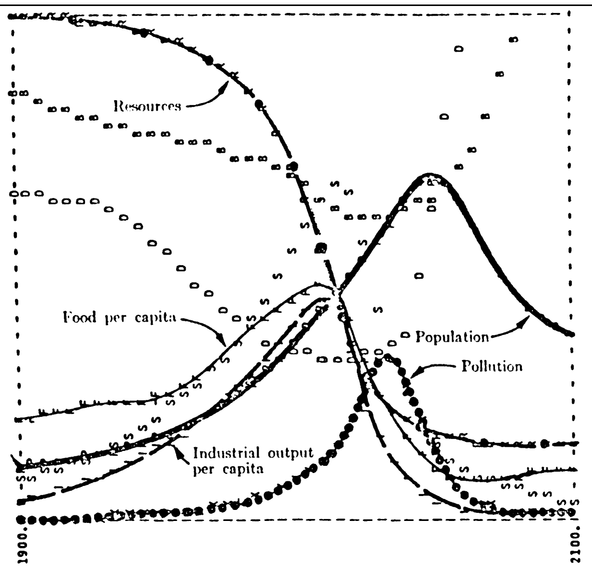
This simulation is a classical example of a large simulation with surprising results.



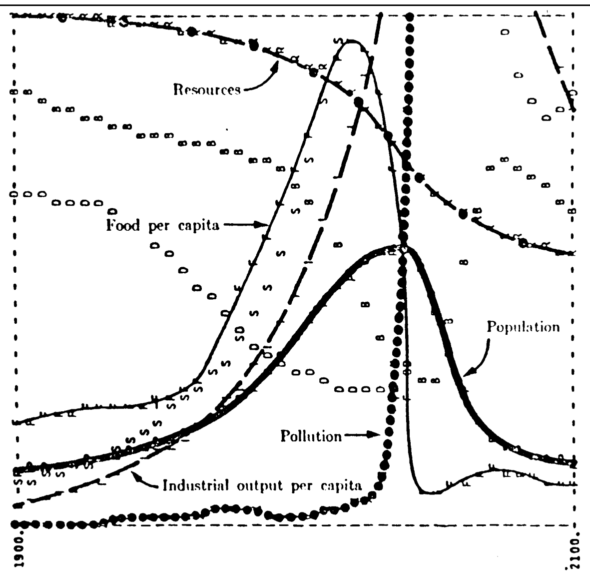
The feedback loops of population, capital, agriculture, and pollution. Each arrow indicates a causal relationship which may include a time delay. (By permission from *The Limits to Growth*, by Donella H. Meadows, Dennis L. Meadows, Jørgen Randers, and William H. Behrens, 3rd. A Potomac Associates book, published by Universe Books, New York, 1972.)

In a first simulation (the standard model) it was accepted that the trends between 1900 and 1970 would continue. The simulation predicts a world crisis (decrease of the population), which is mainly the result of depletion of resources. By the delays in the system the population and pollution will at first temporarily increase, and all these factors lead to food shortage, resulting in huge increase of mortality.

Against expectations, a simulation with doubling of the available resources and agricultural yields, and reduction of pollution per unit production, results in an even sharper crisis. The combination of these favorable circumstances takes away so many limits to growth that the production and the population increase very strongly. Despite the smaller pollution per unit production, this population boom causes a very large production, resulting in an enormous pollution crisis, that brings an end to growth and a dramatic drop in population.



The "standard" world model run assumes no major changes in the physical, economic, or social relationships that have historically governed the development of the world system. All variables plotted here follow historical values from 1900 to 1970. Food, industrial output, and population grow exponentially until the rapidly diminishing resource base forces a slowdown in industrial growth. Because of natural delays in the system, both population and pollution continue to increase for some time after the peak of industrialization. Population growth is finally halted by a rise in the death rate due to decreased food and medical services. (By permission from *The Limits to*



The world model simulation with doubled resource reserves and the pollution per unit of industrial and agricultural output reduced to one-fourth of its 1970 value. Also, the average land yield is doubled in 1975. The combination of these three policies removes so many constraints to growth that population and industry reach very high levels. Although each unit of industrial production generates much less pollution, total production rises enough to create a pollution crisis that brings an end to growth. (By permission from *The Limits to Growth*, by Donella H. Meadows, Dennis L. Meadows, Jørgen Randers, and William H. Behrens, 3rd. A Potomac

VI. DYNAMICS OF COMPLEX SYSTEMS.

Dynamics deals with the time evolution of the state of systems. Simple time invariant linear systems can be easily treated by using the transfer function method. Many systems are however nonlinear, and linearity is always limited by saturation and very often also to relatively small inputs. A single nonlinear component in the system can make the behavior of the total system nonlinear. Systems with many interacting components (complex systems), such as many biological systems, are thus often nonlinear. In complex systems, often components in the system strongly interact. Mutual influences can form either negative of positive feedback loops. The interactions can lead to the appearance of structures that have new properties (emergent properties) that are characteristics of a group of components, but are meaningless for individual components (compare with holism; reductionism implies linearity and additivity).

A. State space description of systems.

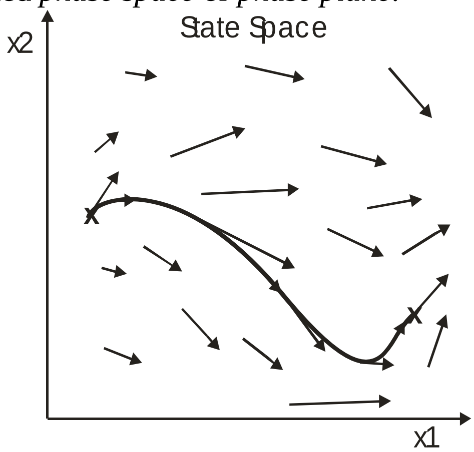
For most nonlinear complex systems no universal simple methods are available to solve the equations analytically. In 1892 H. Poincaré introduced a method for treating such systems, by applying a geometrical approximation and making abstraction of the specific physical properties of systems. The method enables to make important conclusions about the behavior of the system, without a complete solution of the equations. According to this method the dynamic behavior of a system is described in terms of "state space".

In order to describe the state of a system, the system needs to be characterized by a number of properties that are described by the so-called **state variables**. The state of the system is the smallest set of variables (the state variables) that contain sufficient information about the system to make it possible to calculate all future states of the system, given the equations that describe the dynamics of the system and all external inputs (forces, controls, disturbances...).

E.g., the state of a pendulum at a certain moment in time is completely described by giving the position and the speed at that moment. For three-dimensional motion of a particle, the state is described by 6 numbers (3 positions and 3 velocities).

The **state space** of a system is an n-dimensional space (with n the number of system state variables), in which the state of the system at each moment can be represented as a single point $\mathbf{x} \equiv (x_1, x_2, \ldots x_n)$.

A two dimensional state space, in which the second variable (x_2) is the rate of change of the first variable ($x_2 = dx_1/dt$), is also called *phase space* or *phase plane*.



A point in the state space completely determines the state of the system and, given system dynamics and system inputs, also the future time course of the system.

The ***initial state*** in which the system starts is called "***initial condition***".

The ***time evolution of the*** system can be regarded as a succession of states. The solution of the equations of motion gives the evolution of the state, and thus forms a curve in state space; this curve is called a "***trajectory***" ("orbit") in the state space. Different trajectories can form closed loops, can diverge or can converge to each other; however they can never intersect.

The rules (functions f) that indicate how the system can evolve from one state to another state describe the dynamics of the system. These rules assign to each point in state space a vector, the ***flow vector*** (F), which gives the rate of change (per unit of time) of the state variables (thus to which point in state space the system will move at the next moment: $F \equiv d\mathbf{x}/dt$). This vector is dependent on the dynamics f of the system, on the state of the system at that moment ($\mathbf{x}$ = state vector), and on eventual external inputs ($\mathbf{u} \equiv (u_1, u_2, \ldots u_m)$) (e.g. external forces). Thus $F \equiv d\mathbf{x}/dt = f(\mathbf{x},\mathbf{u})$, where the components of f are algebraic functions. The algebraic expression for the flow vector is obtained by rewriting the differential equations of the system as a set of first order differential equations; this is called reducing the equations tot the normal form (see example). The flow vector in an arbitrary point of the state space represents the change per unit of time of the state variables, and thus forms the tangent to the trajectory at that point.

Example: Description of a second order system in normal form (m. b. v. state variables).

$d^2v/dt^2 + a.dv/dt + b.v = 0$

Introducing the variable w, defined as the rate of change of v, thus $w \equiv dv/dt$, the differential equation can be written as $dw/dt + a.w + b.v = 0$. The second order differential equation is then reduced tot a set of two first order differential equations:

$dw/dt = -a.w – b.v$

$dv/dt = w$

This set of two equations is the normal form of the system equations. In these equations w and v are the state variables, and the vector $F \equiv \{x_1 \equiv dw/dt,\ x_2 \equiv dv/dt\}$ is the flow vector, with a value at each point (w,v) in state space equal to (-a.w – b.v, w).

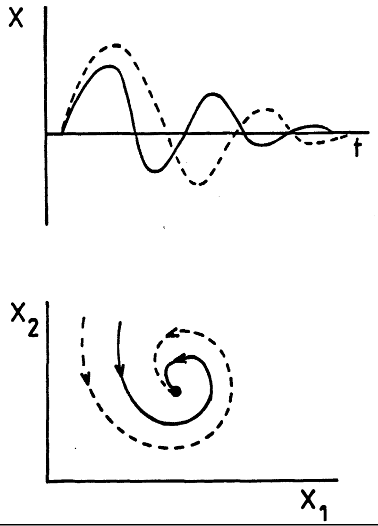
B. State space analysis of systems.

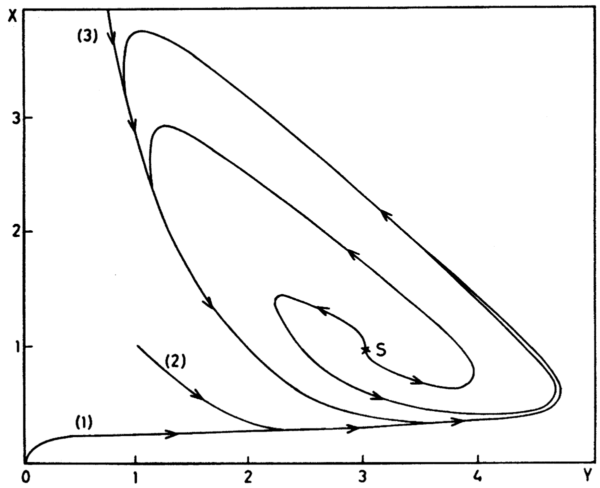
In general, the equations of more complex systems cannot be analytically solved, but it is not too difficult to get a good qualitative picture of the behavior of the system. This can be done by calculating and drawing the flow vectors in a number of points in state space (the components of the flow vector are algebraic functions). Since these flow vectors are tangent to the trajectory, the trajectories can then be easily qualitatively reconstructed, and this way the global behavior of the system can be determined when a sufficient number of flow vectors have been calculated.

A number of tools (such as calculation of the isoclines (= constant slope) and zero clines) make it possible to quickly calculate analytically some of the flow vectors. Zero clines are curves in the state space where the time derivative of a component of the flow vectors is zero (e.g. $dx_1/dt = 0$); isoclines are curves where the direction of the flow vectors (slope S) in a plane in the state space is constant (e.g. $S = dx_2/dx_1 = c$). The equations of the isoclines and zero clines are algebraic equations in the state variables, and the roots of these equations can thus often be calculated relatively easily.

Starting from an arbitrary state, the system evolves to a particular long-term behavior. The ***long-term behavior*** of a system is called an ***attractor***. An attractor is a state or set of states that have a vicinity in state space with the property that a system starting in a state within this vicinity of the attractor will finally evolve towards this attractor (compare with a valley). Many nonlinear systems have different attractors:

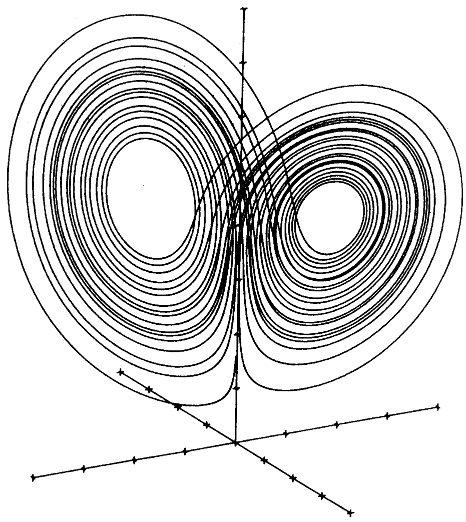
the initial condition determines to which state (attractor) the system will evolve. Four types of attractors can be distinguished, each one corresponding to a different type of dynamic behavior of the system. The first three types of attractors can also be found in linear dynamical systems, while the last one is specific for nonlinear systems:

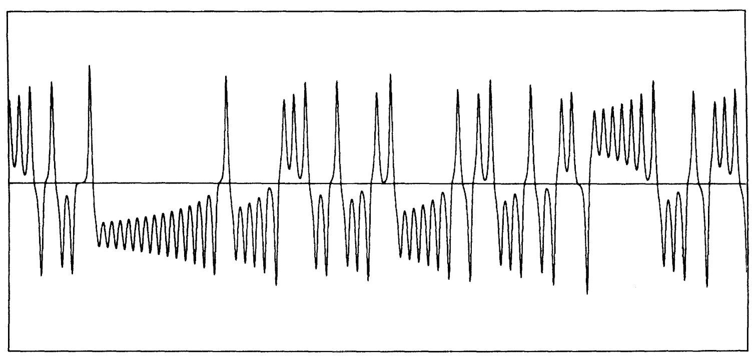




Fixed-point attractor in time domain and state space (left).
Limit cycle attractor in state space (right).

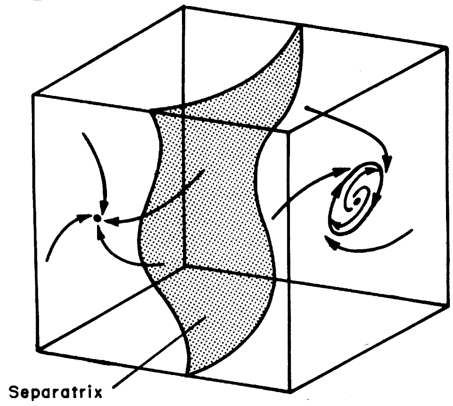
- *fixed-point attractor*: (steady state reached by a stable system);
- *limit cycle attractor* (oscillations around a particular point in a “unstable” system);
- *quasi-periodic trajectories* (consisting of at least two sinusoidal components, the ratio of the frequencies of which is an irrational number);
- *chaotic attractor* (unpredictable behavior);
  - ⇒ only in nonlinear systems,
  - ⇒ the cause of the unpredictability is that neighboring trajectories strongly diverge,
  - ⇒ small amounts of noise (or small uncertainty in start condition) are thus strongly amplified and thus determine the long term behavior,
  - ⇒ the attractors are very orderly fixed, and often geometrically elegant structures,
  - ⇒ some chaotic attractors show a fractal pattern; they are called “strange attractors” (note that also fractal non-chaotic attractors).



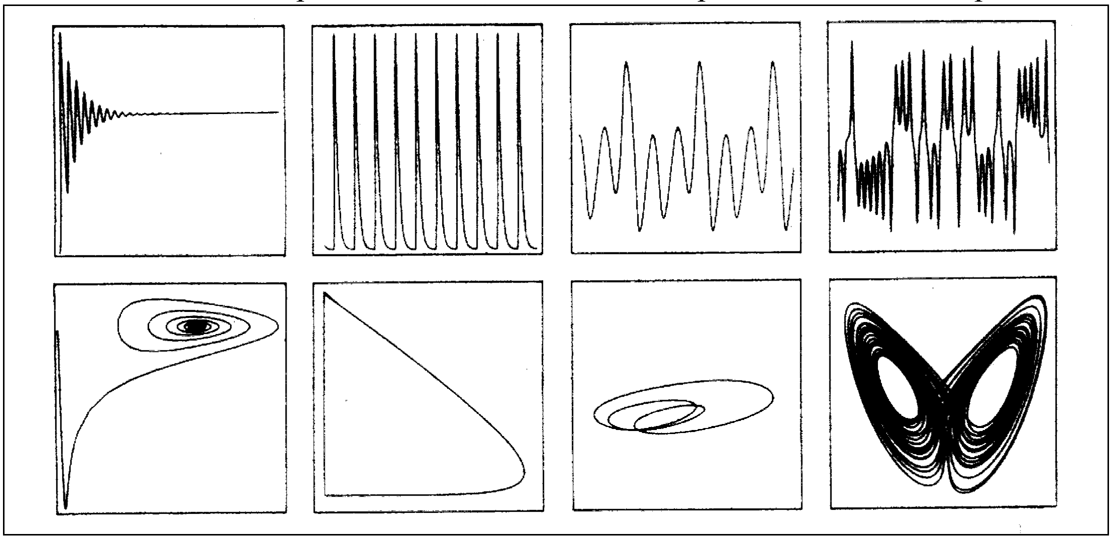


The Lorentz attractor.
In state space (left), and as a function of time (right).

The set of initial states that evolve to a particular attractor is called the "**basin**" of the attractor. The boundary between basins is called "**separatrix**".



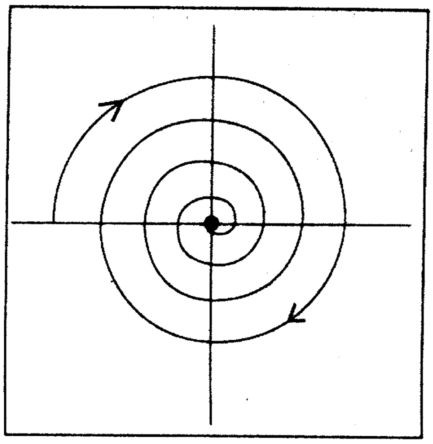
State space-analysis considers the set of states and how the system evolves between these states. The purpose of dynamical analysis is thus to detect and analyze the different types of trajectories. A dynamic analysis is in a certain sense complete when the "attractor-basin portrait" of the state space is known.



Behavior of different systems as function of time (upper panels), and in phase space (lower panels).

C. Examples of analysis of predictable systems.

1. Representation of the damped oscillation in state space.



2. Predator - prey system: the Volterra equations.

Error! Not a valid link.	**Block diagram of the Predator - Prey system described by the Volterra equations.** $dP/dt = k_1P - k_2PR$ with R = Predator, P = Prey, $dR/dt = k_2PR - k_3R$ Prey population increases proportionally to the number (P), but decreases when the probability of an encounter Predator - Prey (P.R) becomes larger. Predator population increases with probability of an encounter Predator - Prey, but decreases proportional to the death rate, which is proportional to number of Predators (R).

Determination of the two zero clines (dR/dt = 0 of dP/dt = 0).

When the number of Predators is small, the Prey population will increase, while for large numbers of Predators the Prey population will decrease. The horizontal line of separation between these two areas represents the Prey zero cline:

$dP/dt = 0 \Rightarrow R = k_1 / k_2$.

When Prey is abundant, the Predator population will increase, while the Predator population will decrease when Prey is scarce. The vertical line of separation between these two areas represents the Predator zero cline:

$dR/dt = 0 \Rightarrow P = k_3 / k_2$.

Equilibrium state (dR/dt = 0 and dP/dt = 0).

This is thus the intersection of the two zero clines.

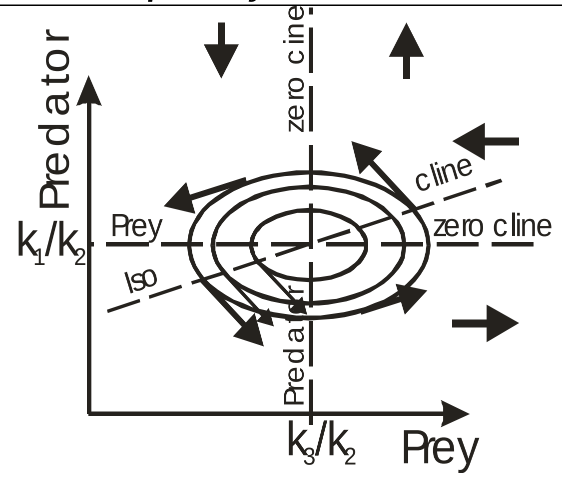
Determination of the isoclines (dR/dP = c).

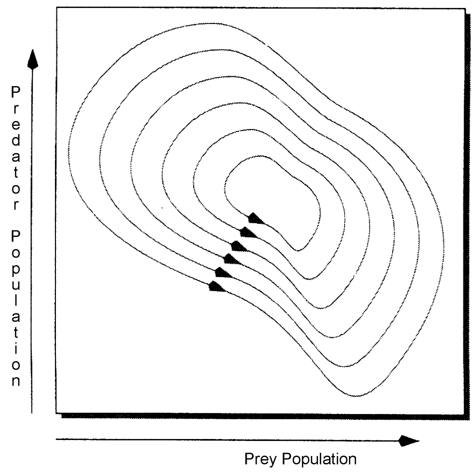
$dR/dP = dR/dt / dP/dt = (k_2PR - k_3R) / (k_1P - k_2PR) = c$

Solving R in function of P gives the equation of the curve for the set of points in the state space where the slope of the flow vector is equal to c (thus with angle α = arctg c):

$$R = \frac{c.k_1}{(c+1).k_2 - k_3/P}$$

Construction of the trajectories.



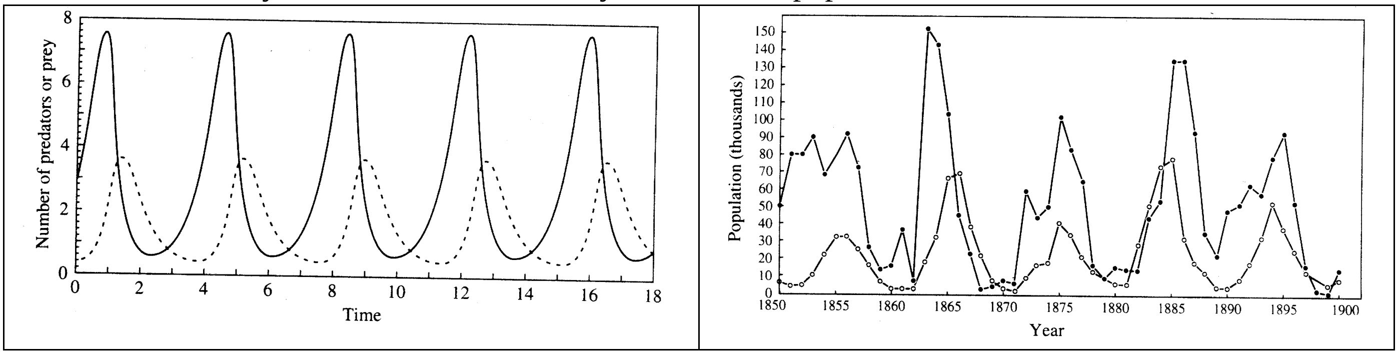


Combining these data in the state diagram results in a general picture of the evolution of the flow vectors, and of the trajectories.

Starting at a point with much more Prey than Predators, the number of Prey will increase. However, also the number of Predators will increase. At a certain point the number of Predators is so large that the number of Prey will decrease, which will continue until so little Prey is left that many Predators starve and their number dramatically decreases. At some moment the number of Prey can rise again: the cycle starts again.

The equations thus predict the presence of limit cycles. Different begin states results in different cycles. A disturbance will change the time course to another limit cycle. Very large disturbances can causes extinction of a population (R=0 or P=0). At one specific initial state, the birth rate of Prey and Predators is exactly equal to the death rate: this point is an equilibrium state.

From the state analysis the time course of Prey and Predator population can be calculated.



Simulation of the Volterra equations (left). Data about fluctuations of the population of hare and lynxes, according to the catches for the Hudson Bay Company (right).

3. Free damped pendulum.

$$\frac{1}{\omega_n^2} \cdot \frac{d^2\varphi}{dt^2} + \frac{2\xi}{\omega_n} \cdot \frac{d\varphi}{dt} + \sin\varphi = 0$$

State equation.

In the equation φ is the angle of the position of the pendulum with the vertical. We introduce the angular velocity as state variable: $v = d\varphi/dt$; the second order differential equation can then be rewritten as a set of two first order differential equations.

$$\frac{d\varphi}{dt} = v$$

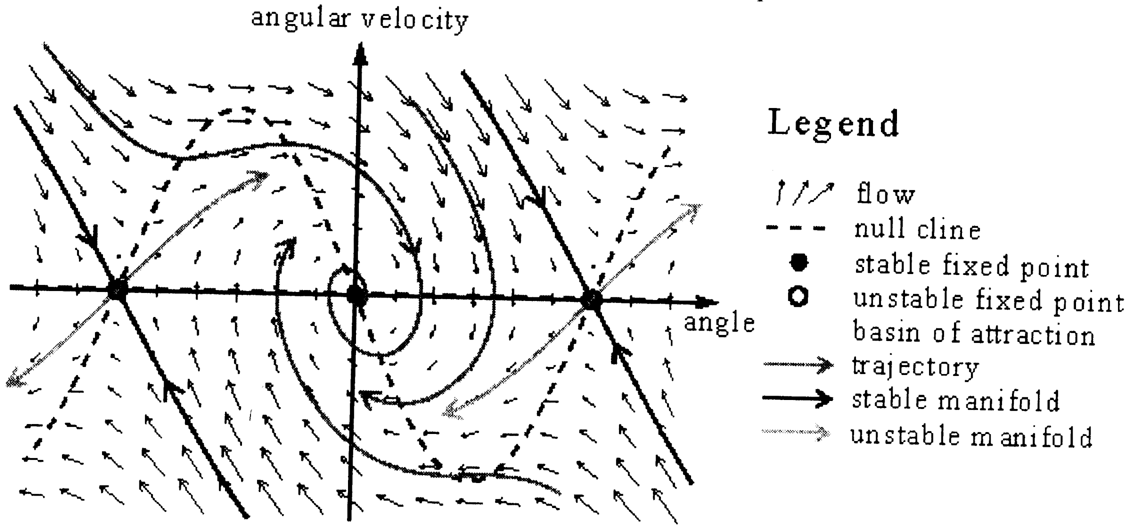
$$\frac{dv}{dt} = -2\xi\omega_n \cdot v - \omega_n^2 \cdot \sin\varphi$$

The state space (phase space) then consists of the φv-plane. The flow vector F is defined by its two components $F \equiv \{d\varphi/dt, dv/dt\}$.

Analysis.

We first determine the zero clines. Since $d\varphi/dt = v$, we obtain $d\varphi/dt = 0$ for $v = 0$. This means that for all points on the axis $v = 0$, no change of φ occurs with time. At all points on the φ axis, the flow vector is thus perpendicular to the φ axis. This axis thus forms the first zero cline. The sense of the flow vector is determined by the sign of dv/dt. Inserting $v = 0$ in the second equation demonstrates that the sign of the flow vectors at points of this axis is negative when $\sin\varphi$ is positive, thus for φ between 0 and 180°. In this range the flow vector thus points downward.

The second zero cline is determined via the second equation ($dv/dt = -2\xi\omega_n \cdot v - \omega_n^2 \cdot \sin\varphi$).
$dv/dt = 0$ for $v = -\omega_n \sin\varphi / 2\xi$; this equation is thus the equation of the second zero cline.
At the intersection of the two zero clines, the derivatives of both state variables are zero, and this point is thus a steady state solution (both state variables do not change with time). The intersection, which is found at the point $\varphi = 0$, $v = 0$, is a stable equilibrium point (fixed-point attractor). For the undamped vibration ($\xi = 0$) we then obtain: $dv/dt = 0$ for $\varphi = 0$. The second zero cline is then formed by the $\varphi = 0$ axis, this is the v-axis.
By determining the zero clines we identified all locations where the flow vectors are perpendicular to the axes. In order to obtain a finer image we can eventually analytically determine a number of isoclines, nl. every location where the slope of the flow vector has a certain value ($dv/d\varphi$ = const), and we can then determine the direction of the flow vectors at different locations by entering the coordinates v and φ in the equations and calculating dv/dt and $d\varphi/dt$. The angle of the flow vector α is equal to $\alpha = \text{arctg}(dv/dt / d\varphi/dt)$. From the picture of the flow vectors we can then outline the trajectories from any arbitrary point.



N.B. For small values of φ the approximation $\sin\varphi = \varphi$ can be used, and the system becomes a linear system. The trajectories in the linear approximation are then circles: the velocity is zero at maximal angle and the velocity is maximal at zero angle. The system keeps oscillating sinusoidally and thus forms a limit circle attractor (see undamped linear second order system).

D. <u>Chaotic systems.</u>

Around 1980 the phenomenon of chaos was discovered: simple but nonlinear systems can result in very complex apparently random behavior.
Chaos is generation of unpredictable behavior that is the consequence of simple but nonlinear rules (independent of noise, randomness of probability!). Even with very small uncertainty about the starting conditions the long-term behavior is difficult to predict because the rules have to be applied repeatedly and thereby the uncertainty increases with each step.

1. Properties.

Chaos reveals itself in a number of features.

- Even in a completely deterministic systems (like Newtonian mechanics) long term prediction is impossible.

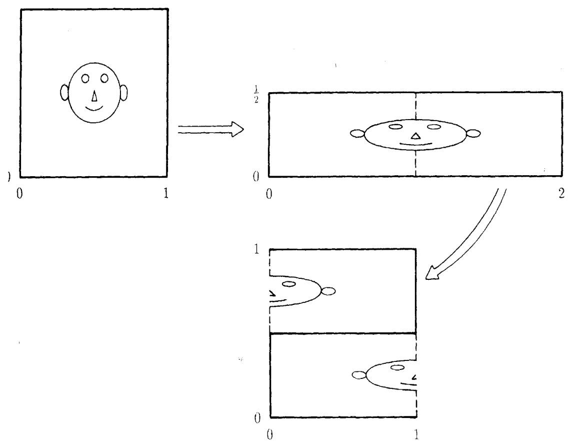
- The unpredictability of long term behavior is the direct consequence of nonlinearity: small causes can have large consequences:
  - ⇒ large sensitivity of the system to the initial conditions (e.g. the weather). Starting from almost identical initial conditions the time evolution finally becomes very different direction. While accurate determination of the state at a certain moment, will enable to determine the immediate time evolution of the system, the uncertainty grows with time, and it is impossible to predict the state of the system after longer times
  - ⇒ exponential amplification of errors: small disturbances can have enormous influence ("The fluttering of the wings of a butterfly somewhere near the Amazon can cause a hurricane in Kansas.".
- Local instability versus global stability. Large amplification of small errors and noise implies local instability: trajectories of neighboring states can diverge. In order to have a consistent stable behavior in the long run the system has to return to the same pattern of the system. This is expressed by the occurrence of chaotic attractors.

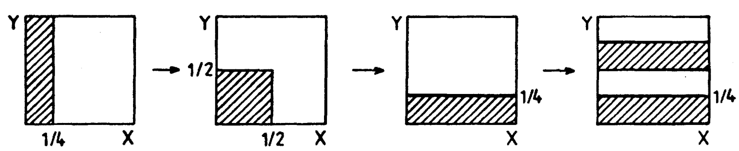
Unpredictability does not imply that the system is not deterministic (as often is erroneously assumed in philosophic discussions (e.g. free will)). The classical chaos theory only deals with chaos in deterministic systems.

Examples:

Baker's transformation.

Chaos in dynamical systems can be nicely illustrated by the so-called baker's transformation. A dynamical transformation can be described as a projection of a point in the state space to another point. In the baker's transformation the state space only has two dimensions: it has a side of 1. The baker's transformation is a discrete transformation where the square is stretched in one direction (stretching coordinate) and contracted in the other direction (contracting coordinate) to a rectangle with basis 2 and height ½. Then the right half is folded on top of the left half to obtain a new square (compare roll out and folding of dough by a baker). This transformation is quantitatively described by two simple mathematical equations. It can be repeated several times, and is reversible. The transformation is an example of a deterministic chaotic system. When we choose an arbitrary small area in the state space, then after a number of transformations this small area will get split up: two points that were originally arbitrarily close, will have reached totally different destinations after a number of steps. The transformation maintains the volume of the original area (theorem of Liouville), but finally splits this volume over a very large number of very small areas distributed over the whole state space.





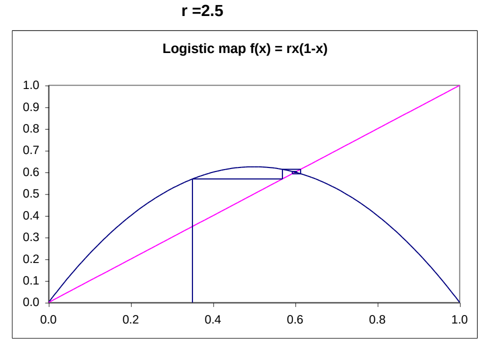
The baker's transformation, and the evolution of the location of an area in space after 1 and 2 transformations.

Logistic equation.

The logistic equation describes a growth process (reduced to its most elementary form) whereby the growth rate is limited by external factors. This equation is the simplest one-dimensional nonlinear system, and was therefore thoroughly investigated in order to analyze the general and fundamental properties of nonlinear systems.

Assume a growth process whereby the number of individual elements in the next generation x_{n+1} is a linear function of the current number of elements x_n. The growth rate is determined by a parameter r, the growth parameter (r >1). Then $x_{n+1} = r \,.\, x_n$. The number of elements will rise exponentially in function of the number of generations: $x_{n+1} = r^n \,.\, x_0$.

Most often however, the growth is limited by the availability of resources. In this case the growth rate will decrease with increasing number of elements, and r thus becomes dependent on the number of elements. The simplest way to describe such a type of growth rate limitation is by replacing r by r . (1 - x_n). The equation of limited growth then becomes the so-called "logistic equation" or "logistic parabola":



Logistic parabola.

$x_{n+1} = r \,.\, x_n \,.\, (1 - x_n)$ (with $0 \le x \le 1$), (in discrete form), or

$f(x) = r \,.\, x \,.\, (1 - x)$ (in continuous form).

For 1 < r < 3 (e.g. for x = 2.0) with repeated application of the transformation, starting from an arbitrary value of x, the value of x will converge to a fixed equilibrium point. However, when the value of r in the equation is somewhat larger than three (e.g. for r = 3.2) x will alternate between two different values: the system evolves to a limit cycle with period 2. When r is further enhanced (e.g. to r = 4.0) then regular behavior is replaced by chaos. The values of x in consecutive generations fluctuate in a wildly irregular way: the fluctuations are aperiodic, resulting in unpredictable behavior. Two different initial starting values of x can result in strongly different values of x after a number of iterations, even when the difference in starting values is extremely small.

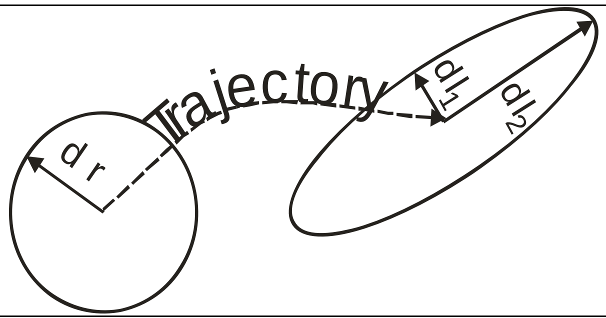
(a) $r = 2.3$		(b) $r = 3.2$		(c) $r = 4.0$	
x_0 = 0.23456000	x_0 = 0.78234000	x_0 = 0.23456000	x_0 = 0.78234000	x_0 = 0.98765000	x_0 = 0.98765001
x_1 = 0.41294568	x_1 = 0.39165349	x_1 = 0.57453314	x_1 = 0.54490922	x_1 = 0.04878991	x_1 = 0.04878987
x_2 = 0.55756954	x_2 = 0.54800037	x_2 = 0.78222340	x_2 = 0.79354612	x_2 = 0.18563782	x_2 = 0.18563768
x_3 = 0.56737721	x_3 = 0.56970071	x_3 = 0.54511986	x_3 = 0.52425816	x_3 = 0.60470568	x_3 = 0.60470532
x_4 = 0.56455871	x_4 = 0.56382615	x_4 = 0.79348544	x_4 = 0.79811693	x_4 = 0.95614689	x_4 = 0.95614718
x_5 = 0.56541399	x_5 = 0.56563030	x_5 = 0.52437216	x_5 = 0.51560414	x_5 = 0.16772007	x_5 = 0.16771899
x_6 = 0.56515831	x_6 = 0.56509312	x_6 = 0.79809919	x_6 = 0.79922083	x_6 = 0.55836020	x_6 = 0.55835733
x_7 = 0.56523510	x_7 = 0.56525463	x_7 = 0.51563799	x_7 = 0.51349406	x_7 = 0.98637635	x_7 = 0.98637769
x_8 = 0.56521207	x_8 = 0.56520621	x_8 = 0.79921745	x_8 = 0.79941731	x_8 = 0.05375220	x_8 = 0.05374697
x_9 = 0.56521898	x_9 = 0.56522073	x_9 = 0.51350053	x_9 = 0.51311767	x_9 = 0.20345161	x_9 = 0.20343294
x_{10} = 0.56521690	x_{10} = 0.56521638	x_{10} = 0.79941675	x_{10} = 0.79944937	x_{10} = 0.64823620	x_{10} = 0.64819192
x_{11} = 0.56521753	x_{11} = 0.56521768	x_{11} = 0.51311874	x_{11} = 0.51305625	x_{11} = 0.91210412	x_{11} = 0.91215662
x_{12} = 0.56521734	x_{12} = 0.56521729	x_{12} = 0.79944928	x_{12} = 0.79945451	x_{12} = 0.32068079	x_{12} = 0.32050770
x_{13} = 0.56521740	x_{13} = 0.56521741	x_{13} = 0.51305642	x_{13} = 0.51304639	x_{13} = 0.87137848	x_{13} = 0.87113005
x_{14} = 0.56521738	x_{14} = 0.56521737	x_{14} = 0.79945450	x_{14} = 0.79945533	x_{14} = 0.44831210	x_{14} = 0.44904993
x_{15} = 0.56521738	x_{15} = 0.56521738	x_{15} = 0.51304642	x_{15} = 0.51304481	x_{15} = 0.98931345	x_{15} = 0.98961636
x_{16} = 0.56521738	x_{16} = 0.56521738	x_{16} = 0.79945533	x_{16} = 0.79945547	x_{16} = 0.04228940	x_{16} = 0.04110327
x_{17} = 0.56521738	x_{17} = 0.56521738	x_{17} = 0.51304481	x_{17} = 0.51304456	x_{17} = 0.16200405	x_{17} = 0.15765516
x_{18} = 0.56521738	x_{18} = 0.56521738	x_{18} = 0.79945546	x_{18} = 0.79945549	x_{18} = 0.54303495	x_{18} = 0.53120005
x_{19} = 0.56521738	x_{19} = 0.56521738	x_{19} = 0.51304456	x_{19} = 0.51304452	x_{19} = 0.99259197	x_{19} = 0.99610623
x_{20} = 0.56521738	x_{20} = 0.56521738	x_{20} = 0.79945549	x_{20} = 0.79945549	x_{20} = 0.02941259	x_{20} = 0.01551444
x_{21} = 0.56521738	x_{21} = 0.56521738	x_{21} = 0.51304452	x_{21} = 0.51304451	x_{21} = 0.11418998	x_{21} = 0.06109498
x_{22} = 0.56521738	x_{22} = 0.56521738	x_{22} = 0.79945549	x_{22} = 0.79945549	x_{22} = 0.40460251	x_{22} = 0.22944952
x_{23} = 0.56521738	x_{23} = 0.56521738	x_{23} = 0.51304451	x_{23} = 0.51304451	x_{23} = 0.96359727	x_{23} = 0.70720976
x_{24} = 0.56521738	x_{24} = 0.56521738	x_{24} = 0.79945549	x_{24} = 0.79945549	x_{24} = 0.14031027	x_{24} = 0.82825647

2. Lyapunov exponent.

Chaos is thus not lack of order: deterministic chaotic systems are very orderly (e.g. in the logistic equation the irregular sequence of points that occurs for $r > 3$ is described by a very simple equation, and thus has low algorithmic complexity), and even predictable for relatively short time scales. The long-term unpredictability is the consequence of the large sensitivity of the system to initial conditions, in addition to the limited accuracy of our description of the state of the system.

A quantitative measure for the sensitivity of a system to initial conditions is the Lyapunov exponent. Lyapunov exponents measure the rate at which adjacent trajectories in state space converge or diverge. A system has as many Lyapunov exponents as the number of dimensions of its state space, but the largest Lyapunov exponent is generally the most important one (when 'the' Lyapunov exponent is mentioned, it always refers to the largest one). One can largely define the (largest) Lyapunov exponent λ as the rate constant is ($\lambda = 1/\tau$, thus λ is the inverse of τ the time constant) in the expression that represents the time course of the distance (d) between two adjacent trajectories: $d = \exp(\lambda \, . \, t)$. When λ is negative, the trajectories converge with time, and the system is insensitive to initial conditions. When λ is however positive, the distance between two adjacent trajectories will exponentially increase with time, and the system is sensitive to initial conditions.

Imagine a small sphere with radius dr around an initial state. Due to the time evolution of the system, all trajectories that pass through this initial sphere will after some time t pass through some volume (e.g. a ellipsoid with half length of the longest axis dl_2). The Lyapunov exponent is then $\lambda = \lim (1/t) \ln (dl_2(t)/dr)$ for t going to infinity.



Meaning of the Lyapunov exponent

The Lyapunov exponent is important, since it determines the prediction horizon. The prediction horizon t_H is the time range outside of which it is impossible to making even qualitative predictions. Given the uncertainty (error) ε on the initial state and the Lyapunov exponent λ of the system, the prediction horizon (Lyapunov time) is given by $t_H = \ln \varepsilon / \lambda$.

In deterministic systems the concept of 'cause' is narrowly connected to the concept of 'identical' or 'the same'. To be able to speak about a 'cause' implies that "in identical circumstances, identical causes have identical consequences". Since a description of a state has a finite accuracy, this implies that a description can never be a 'point' in the state space, but must be a small volume. For dynamic systems with $\lambda > 0$, the trajectories will diverge. For times large than the prediction horizon different points within this small initial volume represent different states; the states within this initial volume cannot be considered as 'identical' for predictions of future states to times $t > t_H$. Therefore, causes that appear to be identical have different consequences even in otherwise identical conditions.

The prediction horizon can be prolonged by enhancing the precision of the description of the current state of the system (this way reducing the number of different states that are considered 'identical'). However in order to increase this Lyapunov time with a factor 10, the precision of the description must be enhanced by a factor e^{10} (about 22000)!

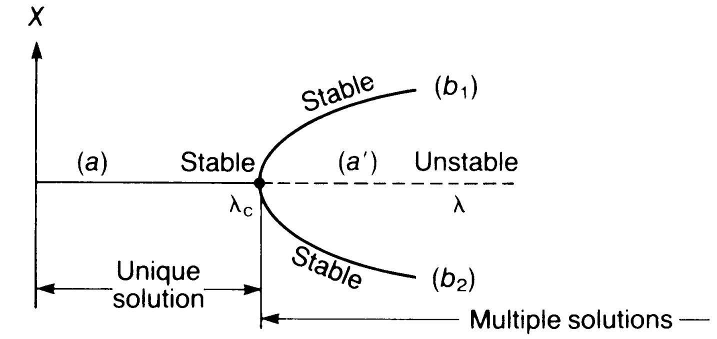
In classical dynamic systems without chaos: $\lambda \leq 0$. The prediction horizon is then infinite. States that are close together then remain close together forever (their trajectories are parallel or converge), and the area in state space occupied by a description of the state of the system within our finite precision, can be approximated as a single point in state space. In this case, identical causes have identical consequences.

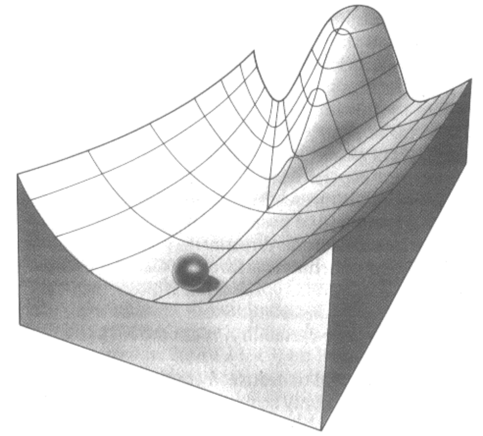
Lyapunov exponent of the baker's transformation.

The value of the Lyapunov exponent of this transformation can easily be calculated. With each stretching the horizontal distance is doubled. After n stretchings this means a factor 2^n . The effect of n stretchings is thus a change of horizontal distance $2^n = e^{(n \ln 2)}$. The number of n in a discrete transformation is the equivalent of time in continuous transformations. The Lyapunov exponent therefore is equal to ln 2 (with the contracting coordinate this exponent corresponds to –ln 2). The largest Lyapunov exponent is thus larger than 1, which results in a long-term unpredictable behavior in this deterministic system.

3. Bifurcations.

The number of attractors can change when a system parameter is modified. This change in number of attractors is called a bifurcation. Bifurcations are accompanied by changes in the stability of attractors.





Bifurcations with change of a parameter (λ) and visualization of the meaning of a bifurcation.

Logistic equation.

$x_{n+1} = r \,.\, x_n \,.\, (1 - x_n)$ (with $0 \leq x \leq 1$) or in continuous version:

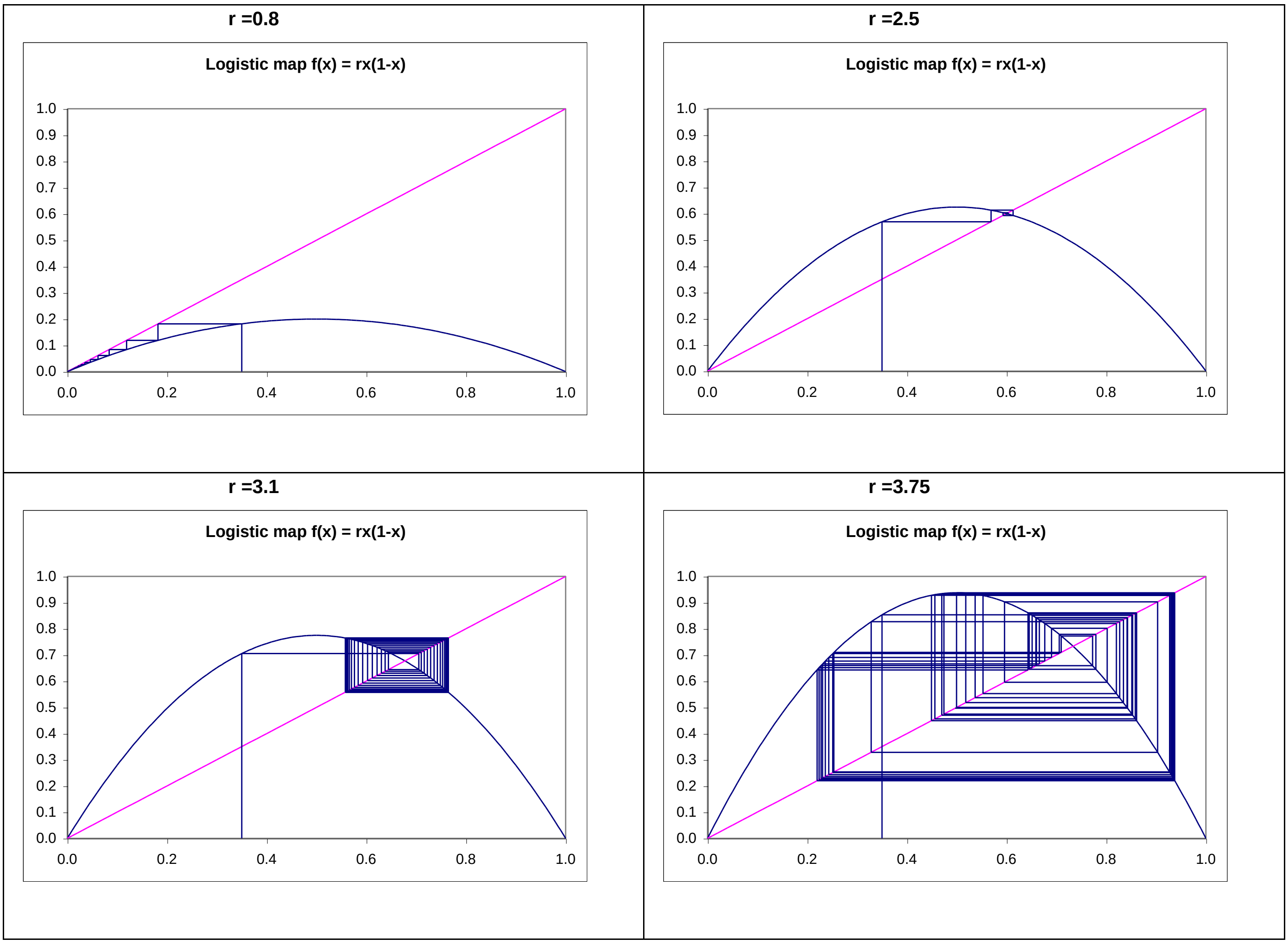
$f(x) = r.\, x.\, (1\text{-}x)$

Investigation of fixed points.

A fixed point is a point x* that is not changed by the transformation, thus for which $f(x) = x$. Fixed points are stable when $|df(x)/dx| < 1$.

Thus $x^* = r \,.\, x^*.\, (1\text{-}x^*)$. This equation gives two fixed points $x^* = 0$ and $x^* = 1 - 1/r$. For $r = 2$ a fixed point is found at $x = 0.5$. For $r < 3$ these fixed points are stable.

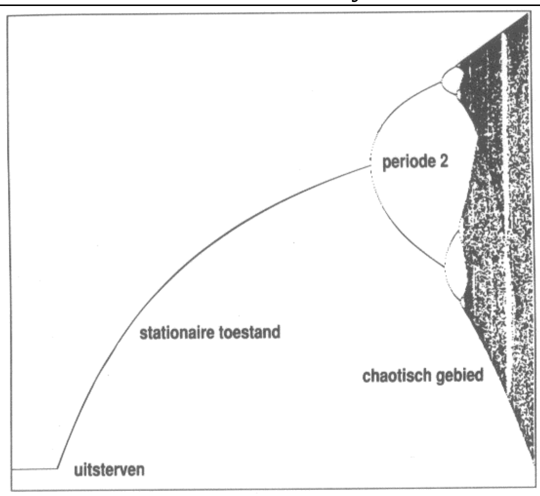
For $r > 3$ bifurcations are found. The equilibrium point is split at a bifurcation in two points. When r is increased (e.g. tot $r = 3.54$), the cycle of two is replaced by a cycle of period 4. Further increase of r leads to a cascade of bifurcations, which come closer and closer together. Beyond a certain value of r, the accumulation point, regularity is replaced by chaos. For $r > 3.57$ some values of r lead to periodic regimes, while other values cause irregular aperiodic oscillations whereby the behavior of the system is unpredictable.

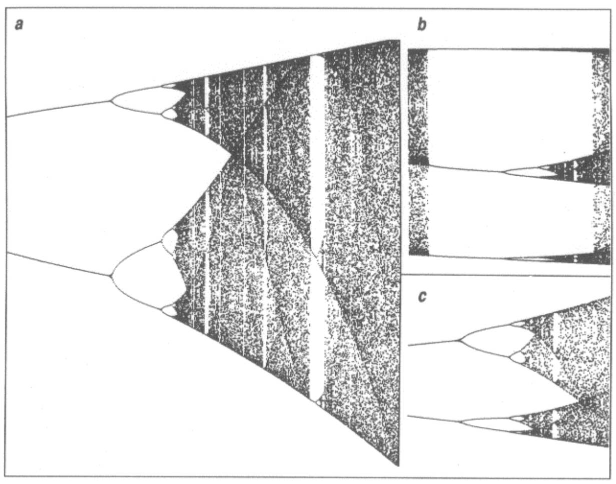


Iterations of the logistic equation.

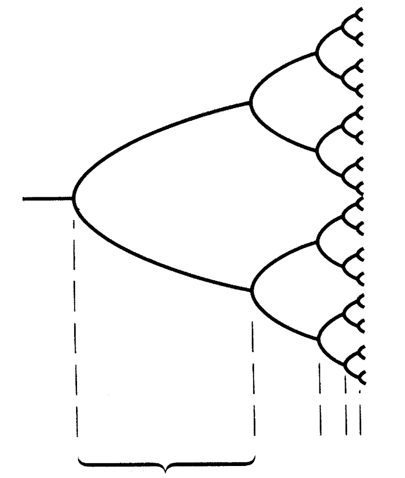
Starting from an arbitrary x (here $x_0 = 0.35$) f(x) is determined, and the value of f(x) is taken as the new value of x used as starting point for the next iteration (here graphically by transposing the value from ordinate to abscissa via the construction line y=x).

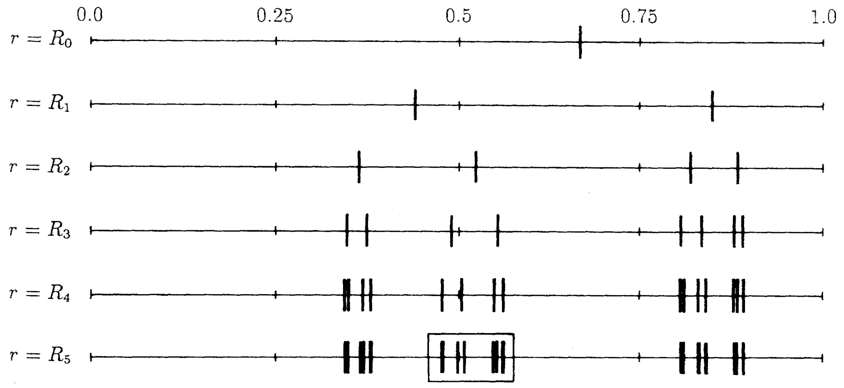
A bifurcation diagram shows how a change of a parameter influences the final behavior of the system. In this diagram the values of the parameter (r) are plotted on the horizontal axis. The final values of the states of the system that are obtained (at a particular value of the parameter r), are plotted as points on a vertical (that intersects the horizontal axis at the value of r). Such a diagram clearly demonstrates the faster and faster succession (with increasing r) of period doublings, leading to chaotic behavior: some parts of the graph are filled up with points. Nevertheless, within these area of chaos at certain values of r, areas can be found with stable cycles.



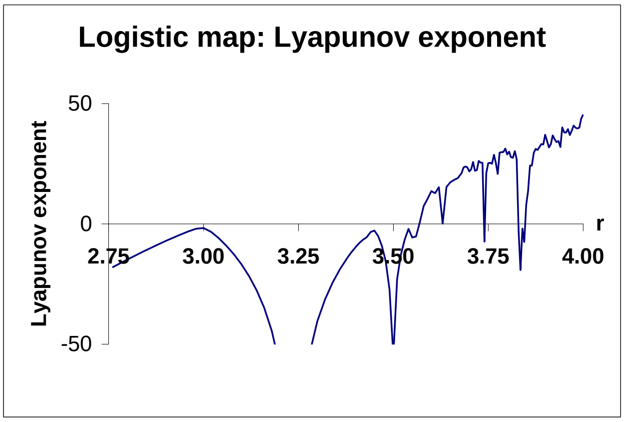


The question of course arises at which values of r bifurcations are found. Feigenbaum was able to demonstrate that strong regularity can be found in the succession of bifurcations: the difference between two successive values of r at which bifurcations occur, decreases with a factor that converges towards a constant $\delta \approx 4.6692016091029...$, and also the successive values of the centers of the oscillations show a regular pattern: a similar pattern repeats itself at smaller and smaller scales. In addition, Feigenbaum could prove that this constant (δ) is not specific for the logistic function, but that it is universal for all nonlinear functions that have a maximum with a minimal steepness to the extent that the existing uncertainty increases with time (number of iterations), in other words that the Lyapunov exponent is positive. Self-similarity thus occurs in chaos. Self-similarity is also the reason that chaos, which appears to be very "complex", can often be described by simple laws, and therefore can have low algorithmic complexity.





The indicated area in R_5 has a similar structure as R_3.



M. Escher: Angels and devils.
Illustration of Self-similarity

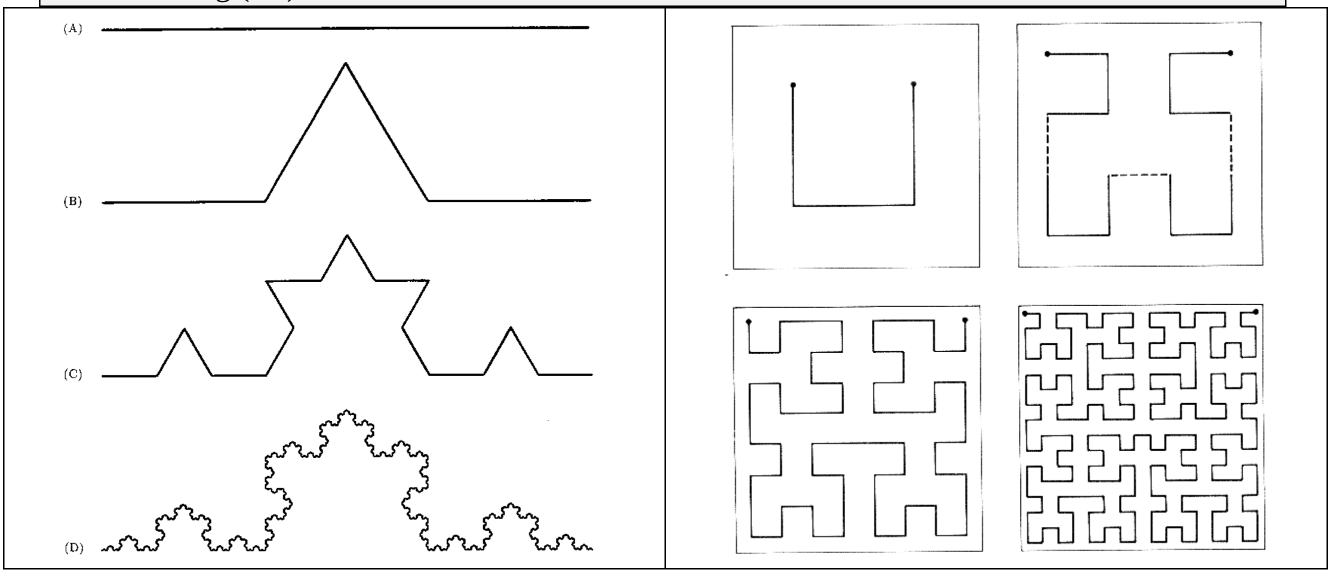
4. Fractals.

The property of repeating pattern occurring at ever-decreasing scale (self-similarity)) is the essence of fractals, the theory of which was developed by B. Mandelbrot. Essential in the theory is the extension of the concept of "dimension" to non-integer numbers (fractal dimension, Hausdorff dimension).

What is the length of the coast of England?

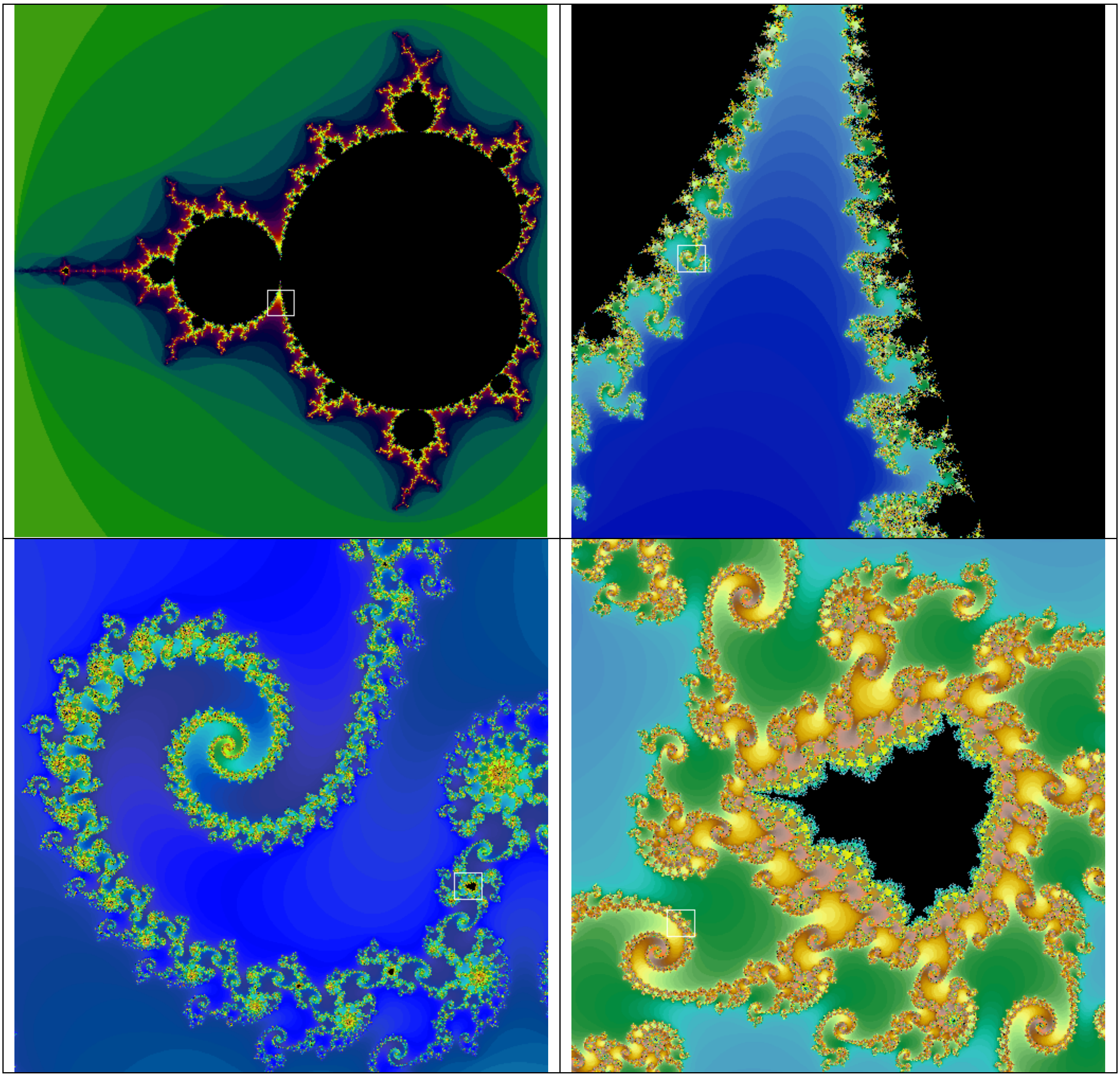
In normal Euclidic spaces, the product of the measure N (volume, area, length) of an object times the D-th power of the unit of length (s) is constant (with a small unit of length one obtains a large measure): $N_1.s_1^D = N_2.s_2^D$ of $N_1/N_2 = (s_2/s_1)^D$, whereby D the dimension of the space is a constant integer number. This can be generally written as: $N = (1/s)^D$. In this equation D defines the concept of dimension.
The equation can however be rewritten as $\log (N) = -D \log (s)$ and in this form D does not necessarily have to be an integer, but can be a fraction. On this bases F. Hausdorff generalized the concept of dimension to non-integer numbers (so called fractal dimensions):
$D= \log (N) / \log (1/s)$.
When the length of a coastline or any very irregular object is measured, the value obtained will depend on the resolution of the measurement. At finer resolution (use of a smaller gauge or scale) more details will be visible and taken into account, so that a larger value for the length is obtained, even when all the values are afterwards transformed to the same units (e.g expressed in km). The concept of length therefore appears to be meaningless. Richardson demonstrated this effect for the coastlines of different countries, and obtained e.g. for the length of the coast of England $L(s) = 5012 . s^{-0.24}$, with L the length, and s the length of the scale. Plotted on double logarithmic scale one thus obtain a straight line with negative slope: $\log L(s) = -0.24 . \log (s) + 3.7$. Mandelbrot used the generalized concept of dimension and applied it to the slope of the line. The length of the coast as a function of the unity of measure used then becomes $\log (L(s)) = (1-D) \log (s) + c$. The fractal dimension of the coast of Engeland is then D =1.24. For smooth regular lines, such as e.g. the coast of South Africa, the slope of the straight line is almost zero, and the dimension approximates a values D = 1.
Mandelbrot apllied the concept of fractal dimension to self similar objects and developed on that basis the theory of fractals. Fractal dimension is then defined as:

$$D = \lim_{s \to 0} \frac{\log (N)}{\log (1/s)}$$



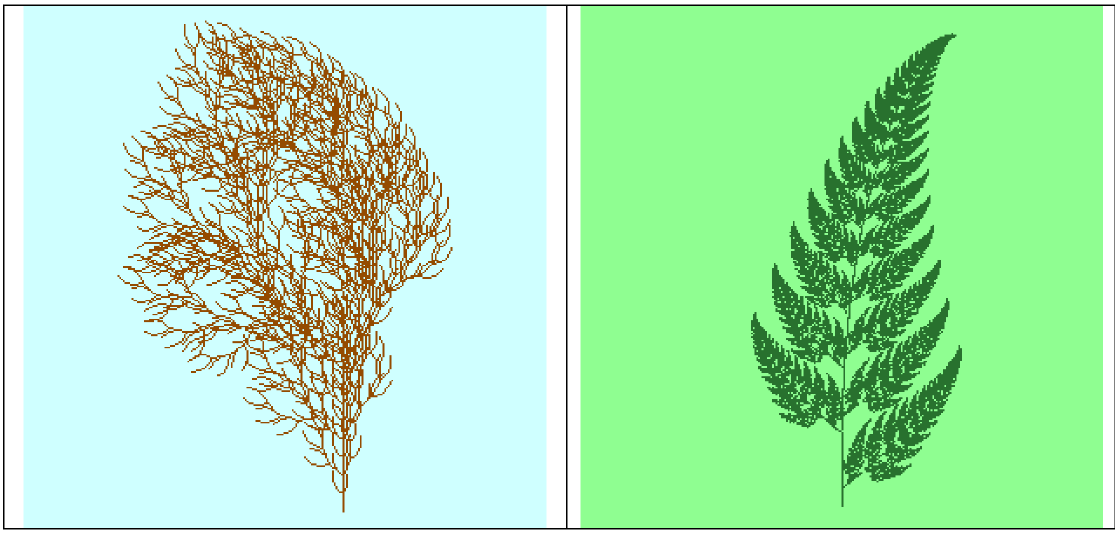
Koch curve and Hilbert curve. The Koch curve (links) has dimension D_H = log 4 /log 3 = 1.26.... For the Hilbert curve D_H = 2. This curve finally fills the whole plane. Adjacent points on the curve are also adjacent in the plane, but adjacent points in the plane are not necessarily adjacent points of the curve.

The attractors of chaotic systems are non-periodic and evolve most often in a limited part of the state space, while the trajectories of points that are arbitrarily close to each other in the state space diverge. Such attractors cannot be described by continuously differentiatable curves with integer dimensions, but must be described using fractal dimensions. Hence, the name ‘strange’ or ‘fractal’ attractors.



Graphical representation of the Mandelbrot set and three details at increasing magnification:
$z_{n+1} = z_n^2 + c$ with z and c complex numbers; iteration with z starting from 0, and c varying.

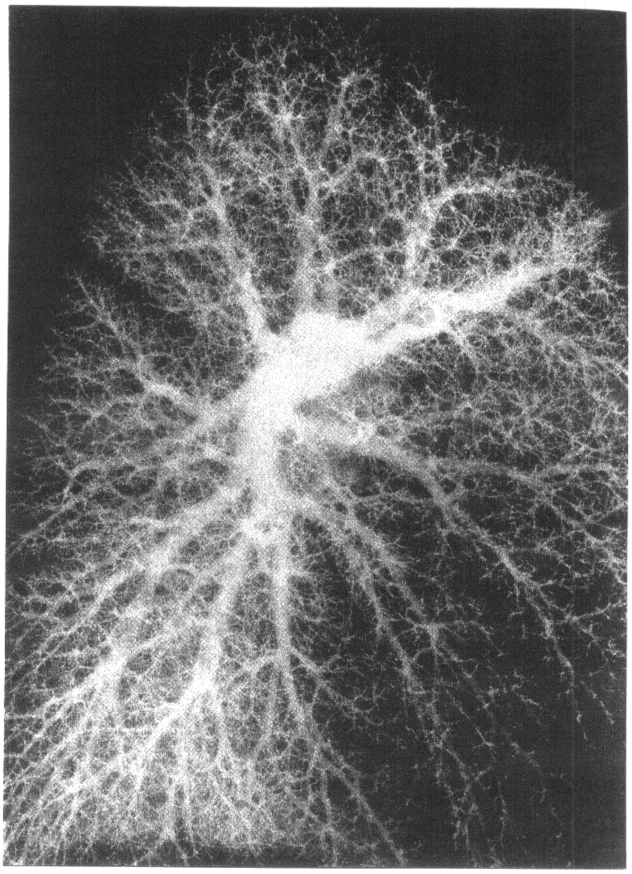
5. Chaos and self-similarity in nature.

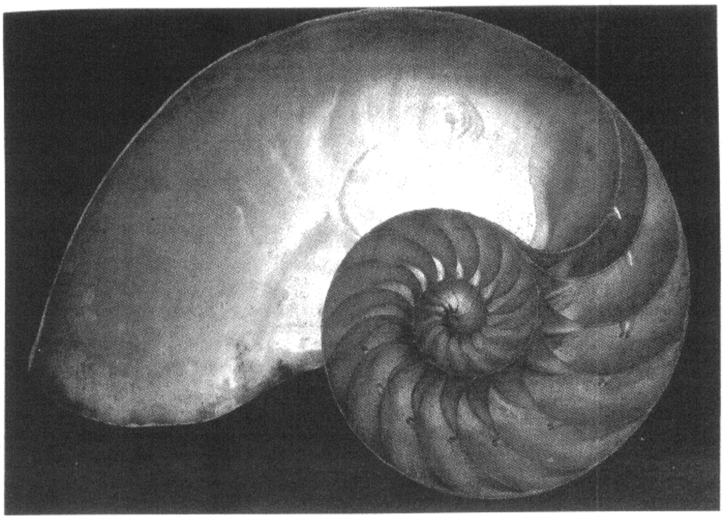


Fractal art: illustration of the fractal structure of branches and leaves.

Chaos often occurs in nature. In addition to pure physical (dripping valve, three-dimensional pendulum) and chemical processes (Belousov – Zhabotinski reaction) chaos was described in biology (growth processes), ecology, demography and epidemiology (fluctuation of populations), in physiology (ion channel kinetics, cardiac rhythm and arrhythmias) and neurophysiology (brain waves), in different behavioral situations, in the economy (fluctuations of the stock exchange).

Self-similarity and fractals are at the basis of many growth and branching processes in nature (trees and plants, bronchial tree, arteries and veins). The empirically found self-similarity can be explained on the basis of the fact that these fractal structures are the result of the shortest possible growth algorithm: "Each step repeats the previous step on a smaller scale". The decrease of scale is even not necessarily included in the algorithm, but appears to be the logical consequence of continuity- or conservation laws (sizes that have to match and are therefore limiting, or the availability of some resources).





Self-similarity in nature: structure of bronchial tree (left), and of Nautilus (right).

E. <u>Extension of the concept of stability.</u>

In contrast to linear systems, the stability of nonlinear systems depends in general on the input, and also on the initial state of the system. It is therefore necessary to more accurately define the concept of stability (we limit ourselves to systems with a constant input, since for arbitrary inputs the theory of stability is not yet fully developed).

Local versus global stability.

Local stability: the system contains a singular point x^0, so that, when a small disturbance occurs, the system remains in a small well defined area R around the point x^0.

Global stability: the area R contains the whole state space, so that also for large disturbances the system is stable.

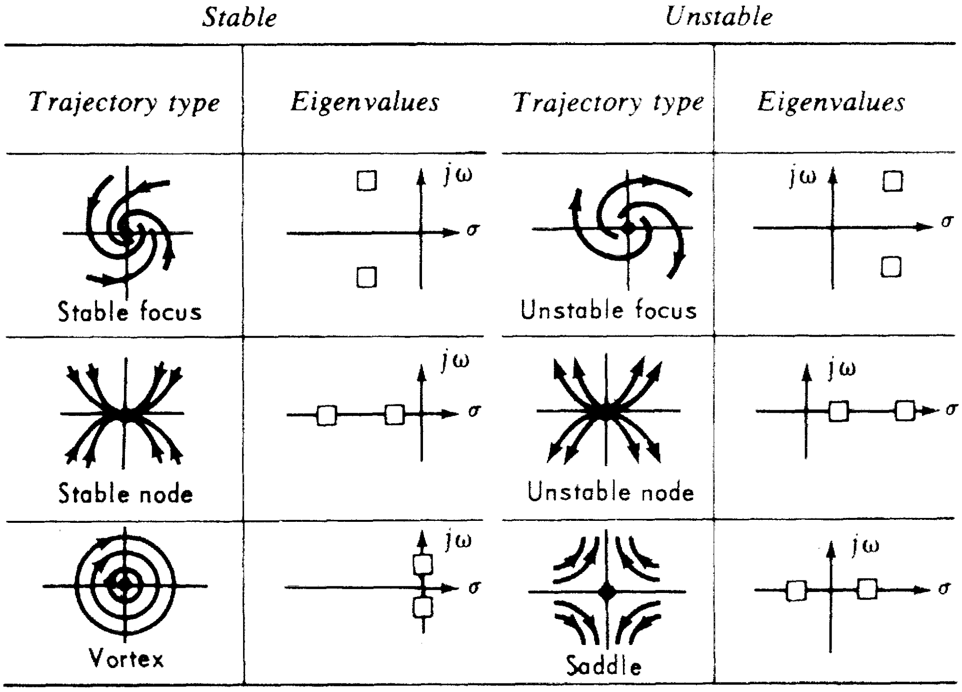
Asymptotic stability versus Lyapunov stability.

Lyapunov stability: each trajectory that starts within a certain area R remains in R.

Asymptotic stability: when in a system with Lyapunov stability, there exists a finite area R around an attractor, so that when the state of the system starts from any point within R, it returns to this attractor. Asymptotic stability can take two forms:

- Limit cycle;
- Stable focus or node: an equilibrium state is a fixed-point attractor.

  Condition for a fixed-point attractor: $K \equiv d\mathbf{x}/dt = 0$



Types of singularities.
Trajectories and representation of the correspondering roots of the system equations in the complex plane

REFERENTIES

BIBLIOGRAPHY.

Control theory

Dorf RC, Bishop RH. Modern Control Systems. Addison Wesley Menlo Park (1998).

Milhorn H. The Application of Control Theory to Physiological Systems. W.B. Saunders Comp. Philadelphia, London (1966).

Dransfield P. Engineerings Systems & Automatic Control. Prentice Hall Inc. Englewood Cliffs (1968).

Cool J, Schijff F, Wiersma T. Regeltechniek. Agon Elsevier, Amsterdam, Brussel (1969).

Kalmus H. Regulation and Control in Living Systems. John Wiley & Sons, London (1966).

Rosen S, Howell P. Signals and Systems for speech and hearing. Acad. Press London (1991).

Complex systems

Acheson D. From calculus to chaos. An introduction to dynamics. Oxford Univ Press 1977.

Bassingthwaighte JB, Liebovitch LS, West BJ. Fractal physiology Oxford Univ Press 1994.

Coveny P, Highfield R. Frontiers of complexity. Fawcett Columbine NY (1995)

Flake GW. The computational beauty of nature. Computer explorations of fractals, chaos, complex system, and adaptation. MIT Press Cambridge Massachusetts (1999).

Gleich J. Chaos. Contact Amsterdam 1993.

Liebovitch L.S. Fractals and chaos simplified for life sciences. Oxford Univ Press 1998.

Prigogine I, Stengers I. Between tijd en eeuwigheid. Bert Bakker Amsterdam 1989.

Schroeder M. Fractals, chaos, power laws. Freeman & Co New York 1991.

Internet sites over nonlinear dynamics, chaos and fractals.

http://amath.colorado.edu/appm/faculty/jdm/faq.html
http://mitpress.mit.edu/books/FLAOH/cbnhtml/index.html
http://www.calresco.force9.co.uk/intro.htm
http://www.cna.org/isaac/Glossb.htm
http://www.jesus.ox.ac.uk/~dacheson/programs/index.html
http://www.lut.ac.uk/departments/ma/gallery/index.html
http://www.math.vt.edu/people/hoggard/FracGeomReport/node1.html
http://www.non.com/new s.answers/fractal-faq.html
http://www.phys.hawaii.edu/~teb/java/pendulum/pendulum/lroom.htm
http://www.vanderbilt.edu/AnS/psychology/cogsci/chaos/workshop/Fractals.html