

Sensorics

by

Prof. Dr.-Ing. Oliver Nelles

Contents

A: Measurement Techniques

1. Introduction to Measurement Techniques
2. Measurement of Electrical Quantities **End B**
3. Measurement of Non-Electrical Quantities
4. Digital Measurement Techniques
5. Measurement Errors and Statistics **End A**
6. Static and Dynamic Behavior of Sensors

B: Signal Processing

7. Introduction to Signal Processing
8. Time-Discrete Systems and Signals
9. Transformation into the Frequency Domain (Discrete Fourier Transform)
10. Filters
11. Selected Methods in Signal Processing

A: Measurement Techniques

1. Introduction to Measurement Techniques

Contents of Chapter 1

1. Introduction to Measurement Techniques

1.1 Historical Issues

1.2 SI: International System of Units

1.3 Relevance of Measurement Techniques

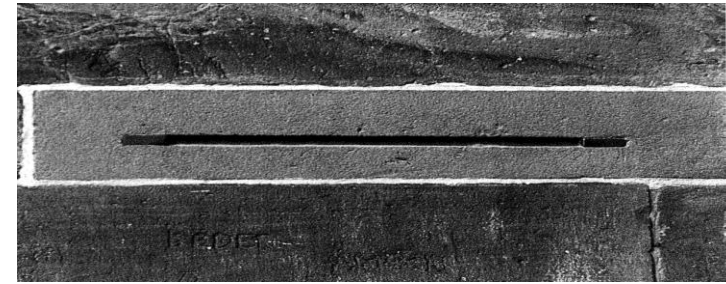
1.4 Basics

1.5 Literature

1.1 Historical Issues

Historical Milestones

- From 3000 B.C. first descriptions of length and weight measures have been found.
- During the medieval, trade and jurisdiction concentrated on the environment around churches. First accepted standards have been established like the Freiburger “Elle”.
- Each trade center defined individual standards. At the end of the 18th century 118 different definitions of an “Elle” and 80 different definitions of a “Pound” have been common.
- 1791 world-wide valid and accepted standards and 1875 the metric system was established in Paris. But in some countries it is not used until today (e.g. USA).
- Since 1889 measurement standards made from platinum and iridium for the original “m” and “kg” are displayed in Paris.



In the tower of the Freiburger Münster a stainless metal bar is built in. Its length is one “Elle“ (54 cm), a 1/20 of an “Elle“ equals one inch.

Schneider Böck in den Streichen von Max und Moritz



Schnelle springt er mit der Elle
über seines Hauses Schwelle, ...

(Wilhelm Busch)

1.2 SI: International System of Units

MKS System

- 1791 the units *Meter*, *Kilogram* and *Second* were established world-wide for the next 200 years in Paris, the so-called **MKS system**. From this system many important units could be derived.
- The *Meter* was defined as the 40 millionth fraction of the circumference of earth (orthogonal to the equator).
- The *Kilogram* was defined as the mass of 10 cubic centimeter (cm³) of water with maximal density (at 4°C).

To complete the MKS system and to improve the accuracy and generality of the units by the help of modern physics, 1960 the **SI system** (*Système International d'Unités*) consisting of 7 units was founded.

- All units can be derived from the basic 7 SI units.
- The definition are mainly based on physical constants.
- In principle, these units could be understood by aliens!

1.2 SI: International System of Units

SI base units^{[11][12]}

Unit name	Unit symbol	Quantity name	Quantity symbol	Dimension symbol
metre	m	length	l (a lowercase L), x , r	L
kilogram ^[note 1]	kg	mass	m	M
second	s	time	t	T
ampere	A	electric current	I (an uppercase i)	I
kelvin	K	thermodynamic temperature	T	Θ
candela	cd	luminous intensity	I_v (an uppercase i with lowercase non-italicized v subscript)	J
mole	mol	amount of substance	n	N

1.2 SI: International System of Units

SI System

From the 7 basic SI units for example the following important units can be derived:

Speed: $v = \frac{s}{t} \rightarrow [v] = \frac{\text{m}}{\text{s}}$

Acceleration: $a = \frac{2s}{t^2} \rightarrow [a] = \frac{\text{m}}{\text{s}^2}$

Force: $F = ma \rightarrow [F] = \frac{\text{kg m}}{\text{s}^2} = \text{N}$

Torque: $M = Fs \rightarrow [M] = \frac{\text{kg m}^2}{\text{s}^2} = \text{N m}$

Energy: $E = Fs \rightarrow [E] = \frac{\text{kg m}^2}{\text{s}^2} = \text{J}$

Power: $P = \frac{E}{t} \rightarrow [P] = \frac{\text{kg m}^2}{\text{s}^3} = \frac{\text{J}}{\text{s}} = \text{W}$

Magnetic Field: $H = \frac{I}{s} \rightarrow [H] = \frac{\text{A}}{\text{m}}$

Electric Voltage: $U = \frac{E}{Q} = \frac{E}{It} \rightarrow [U] = \frac{\text{kg m}^2}{\text{As}^3} = \text{V}$

Torque and *Energy* have identical units!
Does this mean they are the same?

Torque throughout this script is named with *M*, not *T* as usual in English, because this is the common German abbreviation for “Moment”.

1.3 Relevance of Measurements

Measurement Techniques are the Foundation of Science

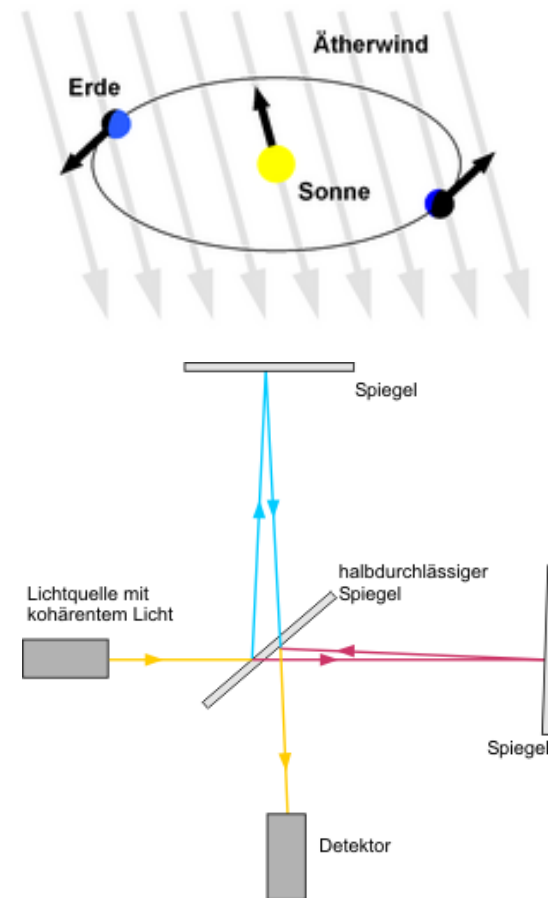
- The foundation of science are observations. Science comes up with theories that aim to explain existing observations and predict future ones. If theory and observations contradict each other, either the observations are flawed or the theory is wrong (falsification). The more independent observations support a theory, the more likely it is true. But, in principle, it can never be proven (verification)!
- Very concrete and quantitative observations are **measurements**. Mainly with their help sciences progresses, in particular natural sciences.
- Discovered patterns within the measurements often lead towards a theory that is coherent with them.
- New technological possibilities often have supported or refuted theories. Example: The measured spectrum of black body radiation was in contradiction to the classical theory. The introduction of quantization of the emitted frequencies by Planck in 1900 could align theory and observations. This was the birth of quantum mechanics (which ironically Planck never accepted).

1.3 Relevance of Measurements

Example: Interferometer Experiment by Michelson-Morley

The experiment by Michelson in 1881 and a refined version by Morley in 1887 tried to prove that some “stuff” in vacuum exist (German: “Äther”) that transmits light waves. The theory at this time stated that any wave needs a medium for its transportation in order to propagate the energy. Examples are water waves or sounds with air as the medium. That vacuum is simply empty was unimaginable because light can travel through space. So what is the medium that light needs?

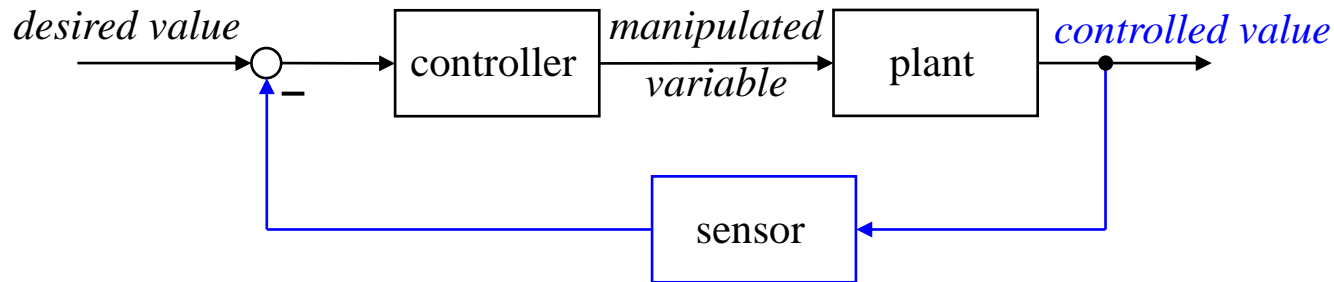
This obscure stuff was not found and called “Äther”. The interferometer experiment was designed to find out whether it exists. If the earth travels through “Äther” then turning the interferometer changes the speed direction and should lead to a phase shift because light should be faster or slower depending on the relative “Äther” speed. But *nothing* happened! Light speed always is c in vacuum. No “Äther” exists. This was explained by *Einstein’s* theory of special relativity in 1905.



1.3 Relevance of Measurements

Measurement for Feedback Control

Control is based on measuring the quantity that shall be controlled. Without the measurement there is no feedback possible, no comparison between desired and actual value.



In many applications in signal processing a delay is not very tragic. If you see a football goal 100 ms delayed because of computations in your digital TV this is no significant drawback.

This is different in feedback control! The controlled variable must be fed back to the comparison of desired with actual value immediately. Any delay due to a slow sensor or filtering or other signal processing techniques deteriorates the control performance.

You can never make up for a delay in a subsequent step!

1.4 Basics

Measuring

Definition:

Measuring means comparing with an agreed unit.

A measurement consists of a number and a unit. The number describes which multiple of the unit is assigned:

$$\text{measurement} = \text{number} \cdot \text{unit}$$

Examples: Speed = 3 m/s = 3 m·s⁻¹, Mass = 4 kg, Force = 5 kg·m/s² = 5 N

Requirements:

1. The quantity to be measured must be qualitatively uniquely determined.
2. The standard unit must be defined by a convention.

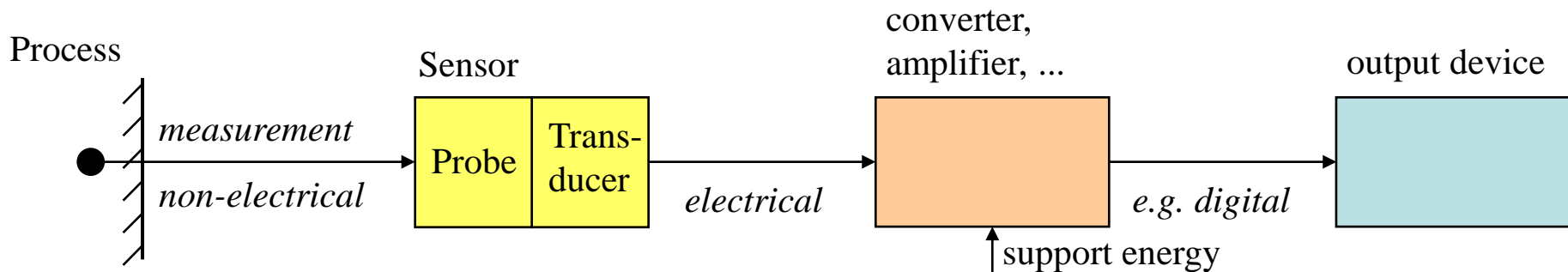
These requirements are *not* met by many quantities in our everyday lives, like wellness, beauty, intelligence.

1.4 Basics

Measurement Setup

A measurement setup typically consists of 3 blocks:

1. The quantity to be measured by a sensor is converted into an electrical signal. Recently the term *smart sensor* has become popular. This means sensors that incorporate an intelligent signal processing that carries out tasks inside the sensor like filtering, data reduction, extraction of features, combining different physical principles, ...
2. The electrical signal is converted into another electrical signal which is e.g. of higher power and/or digital, etc.
3. The amplified and possibly digitized signal is outputted to a display, printer, plotter or only saved.



1.4 Basics

Measurement Method [1]

- *Deflection Method:* The measured quantity is directly converted into the output, e.g. a display. No support energy is needed from outside. The required energy for the conversion is taken from the medium or the environment (e.g. gravitation).
Examples: spring balance, expansion thermometer.
- *Difference Method:* The measured quantity is compared with a quantity from outside. This quantity for comparison stays constant during the measurement. The difference between both is the output. *Example:* volume measurement (displaced liquid).
- *Compensation Method:* A quantity opposed to the measured quantity is applied. A zero indicator determines whether both quantities are equal. If so the compensation quantity is a measure for the original one. The compensation quantity can be of other kind than the original one.
Examples: Equal-armed balance with weights as compensation quantity (same kind) or with an electro-magnet induced force (different kind)



http://en.wikipedia.org/wiki/File:Balance_scale_IMGP9755.jpg

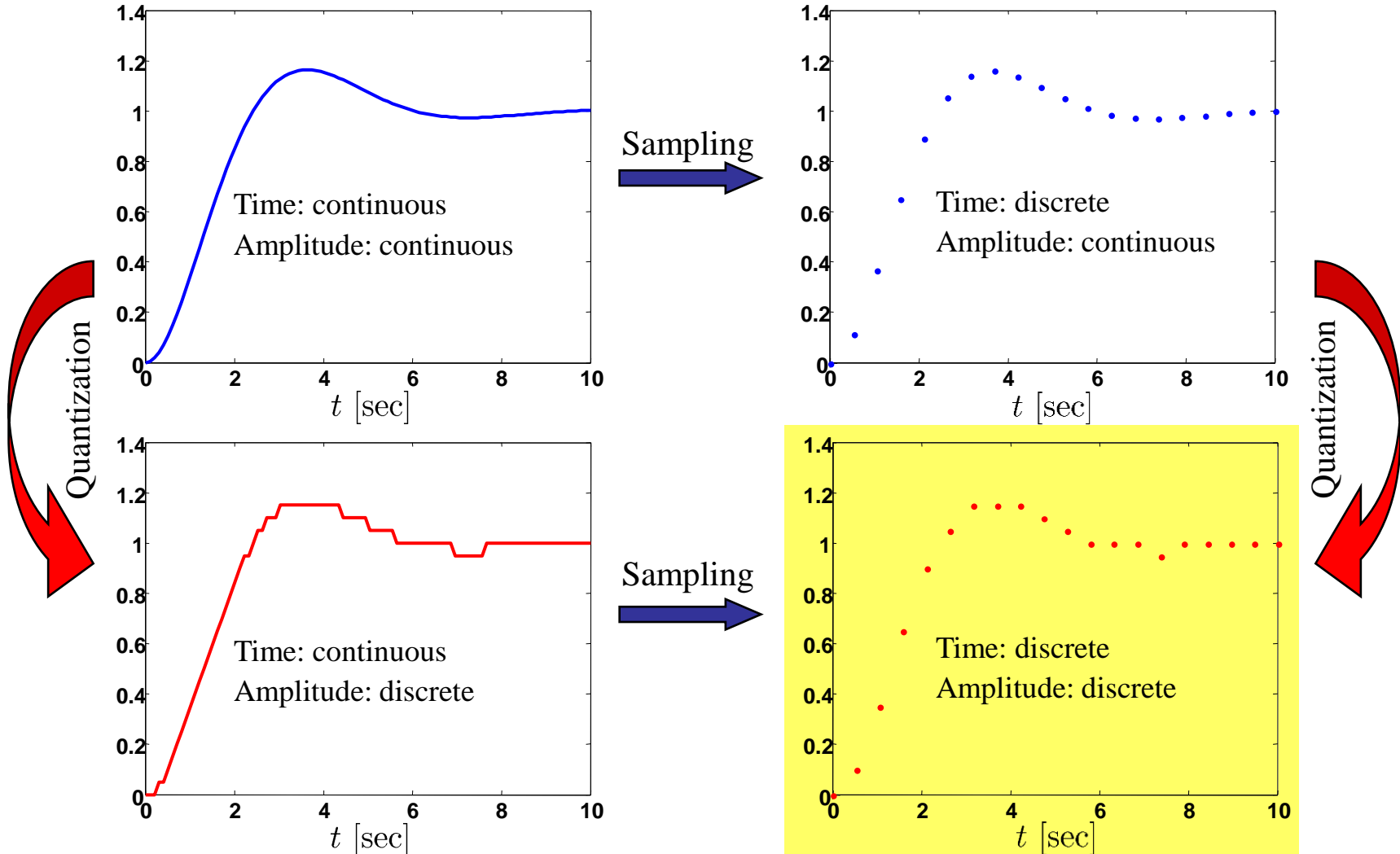
1.4 Basics

Measuring Technique [1]

- *Direct Measurement:* Comparison with a gauge. The most fundamental technique.
Example: Length measurement with a ruler.
- *Indirect Measurement:* The quantity to be measured is determined by other relevant quantities. *Examples:* Determination of pressure by measuring force and dividing by the area. Determination of power by measuring voltage and current and multiplying them. Determination of speed by measuring distance and time and dividing them. Determination of acceleration by measuring speed and differentiating.
- *Incremental Measurement:* From a reference point, increments (= smallest change) are added or subtracted to determine the actual value. Typically, equidistant markings are scanned (optically or magnetically or otherwise). *Examples:* Measuring angles or displacements.

1.4 Basics

Analog and Digital Measurement Processing



1.5 Literature

These books are the main basis for these lecture notes:

1. J. Hoffmann: “Taschenbuch der Messtechnik“, 4. Aufl., Hanser, 2004
2. J. Niebuhr, G. Lindner: “Physikalische Messtechnik mit Sensoren“, 5. Aufl., Oldenbourg, 2005.
3. E. Schrüfer: “Elektrische Messtechnik: Messung elektrischer und nichtelektrischer Größen“, 7. Aufl., Hanser, 2001
4. U. Kiencke, R. Eger: “Messtechnik“, 6. Aufl., Springer, 2005.

A reference:

Mayer, J.R. Rene: “Measurement, Instrumentation and Sensors Handbook“, CRC Press, 1999

A good book in English:

Morris A.S., Langari, R.: “Measurement and Instrumentation: Theory and Application”, Academic Press, 2012

4. Digital Measurement Techniques

Contents of Chapter 4

4. Digital Measurement Techniques

- 4.1 Discretisation of Amplitude and Time
- 4.2 Sampling Theorem
- 4.3 Quantization
- 4.4 A/D and D/A Converters
- 4.5 Measurement of Frequency

4.1 Discretization of Amplitude and Time

Advantages of Digital Measurement Techniques

- Digital electronics is insensitive with respect to environmental influences (temperature).
- Digital electronics becomes more powerful, cheaper, smaller, more robust. It can be integrated together with the sensor (smart sensor).
- Documentation and archiving purposes require or favor a digital form.
- Digital signal processing is much more powerful and flexible than analogue electronic circuits:
 - Digital and adaptive filtering.
 - Nonlinear transformation/inversion.
 - Transformation of signals into the frequency domain (fast Fourier transform, FFT).
 - Parameter estimation, supervision, diagnosis.
 - Sensor fusion.
 - Storage of data on a digital storage medium (hard disk, flash).
 - Transmission of data without any information loss.
 - Powerful display technologies.

4.1 Discretization of Amplitude and Time

sampled

digital = time-discrete & quantized

Here: Focus on digital signals

- Difference equation and sums are simpler to manage and understand than differential equations and integrals.
- Digital realizations replace analogue circuits because it is
 - cheaper in most cases (especially for high quantities),
 - easy to implement,
 - more flexible: faster and cheaper to change (even afterwards with updates),
 - more robust and durable with respect to environmental influences (wear, temperature, humidity).

amplitude in
e.g., 8 or 16 bits

Focus of this lecture

- Development of an understanding for the methods and their potential applications.
- No implementation details and tricks.
- No programming of digital signal processors (DSPs).
- More width than depth.

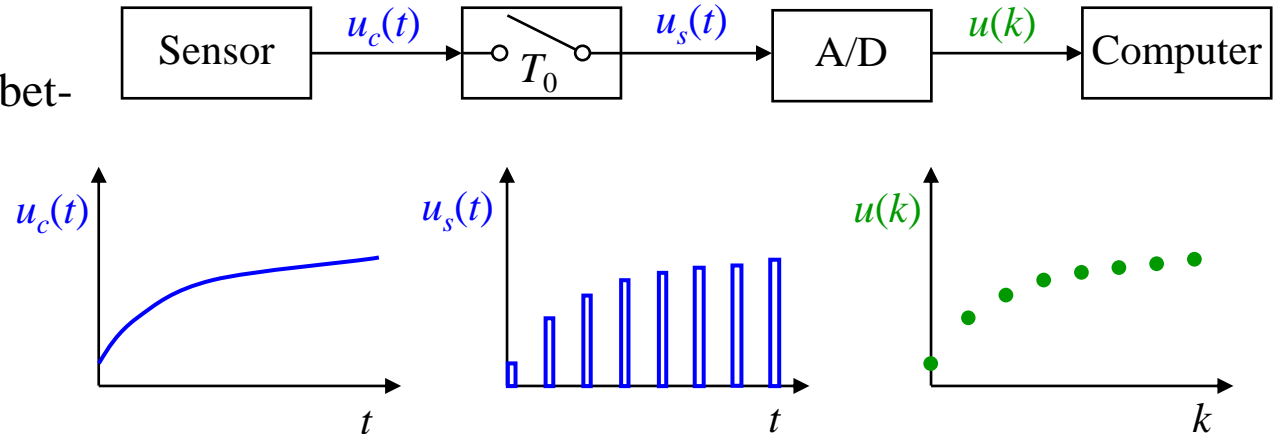
4.1 Discretization of Amplitude and Time

Abbreviation: $u(k) = u_c(kT_0)$
 $y(k) = y_c(kT_0)$

Analog/Digital and Digital/Analog Conversion

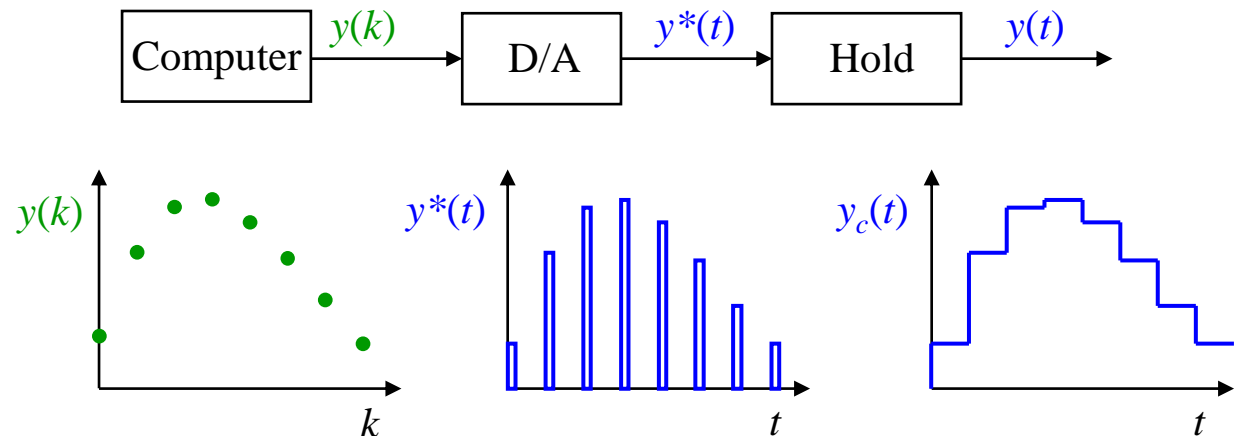
A/D Convertor

- Sampling time T_0 can be between μsec (signal proc.) and hours (thermal, biological processes)
- Amplitude resolution of 8, 12 or 16 bit.



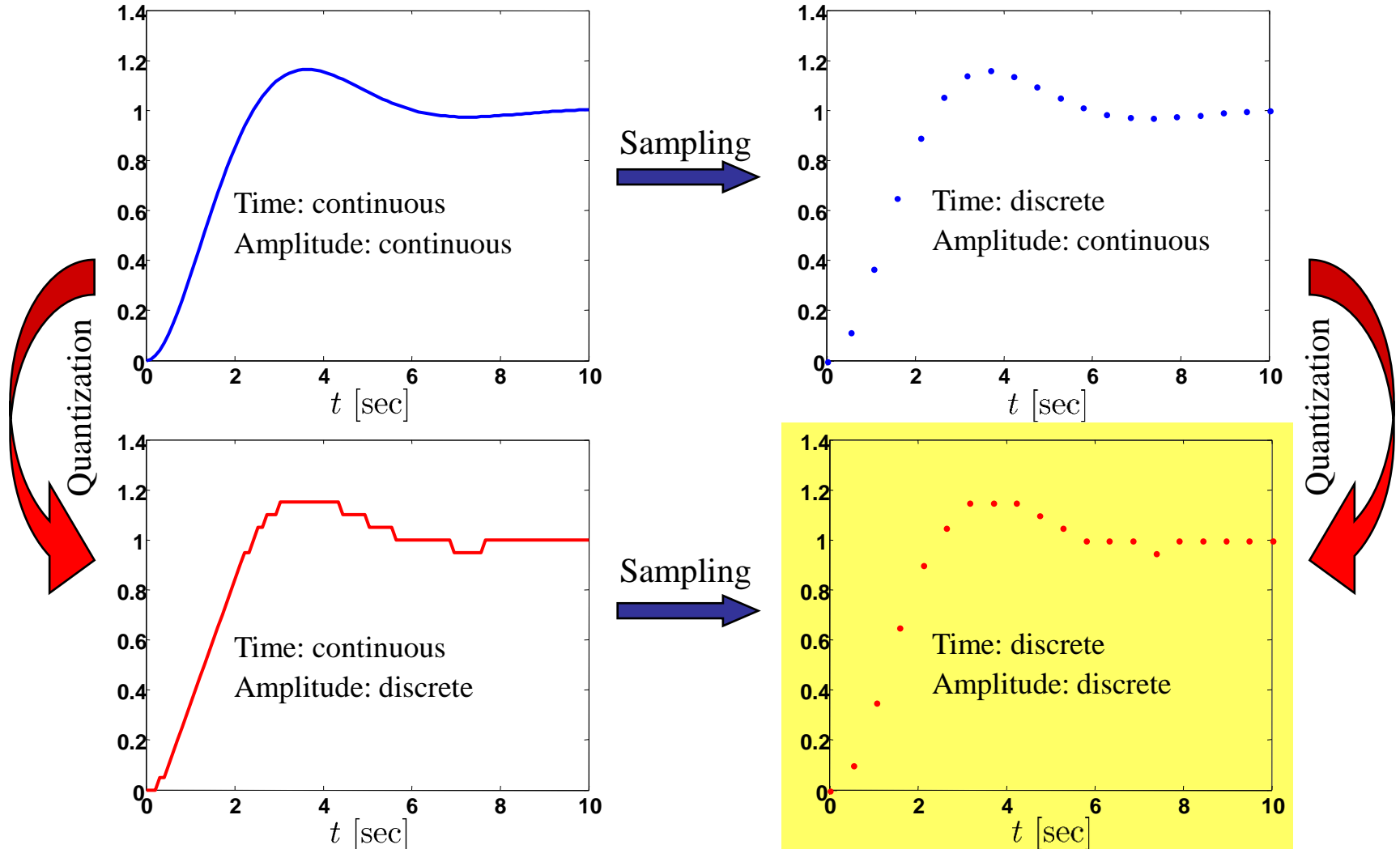
D/A Convertor

- Computer handles time-discrete series.
- Hold of 0. order generates piece-wise constant signals.



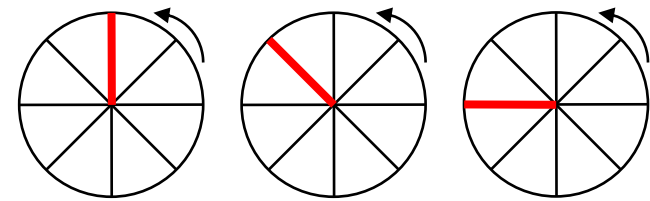
4.1 Discretization of Amplitude and Time

Analog and Digital Measurement Processing



4.2 Sampling Theorem

Wheel seems
to freeze!

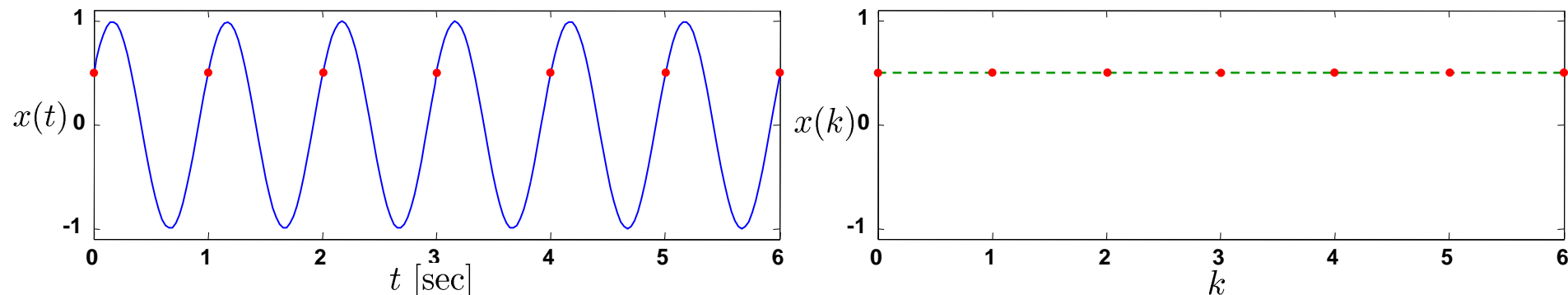


Sampling of a Continuous-Time Signal

Everybody has seen spoke wheels of a starting carriage or car – at least in the movies. First the accelerating wheel can be observed. With a certain speed or angular velocity of the wheel, it suddenly changes direction and seems turn the other way round although the carriage further accelerates. Further on the wheel slows down before it finally stands still. That is obvious contradiction to the faster and faster carriage.

This strange effect can be explained by the so-called **Aliasing**. It exists for all time-discrete and therefore sampled systems. Obviously problems occur, if the signal is sampled too slowly for its velocity (or more precisely frequency). This effect becomes prominent, if we approach half of the sampling frequency. The *movie* plays the role of the sampler with a sampling frequency of $f_0 = 24$ Hz or 25 Hz, i.e., the refresh rate.

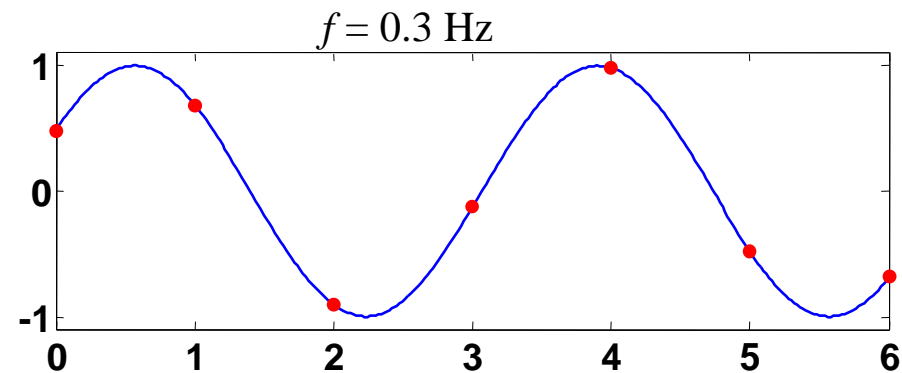
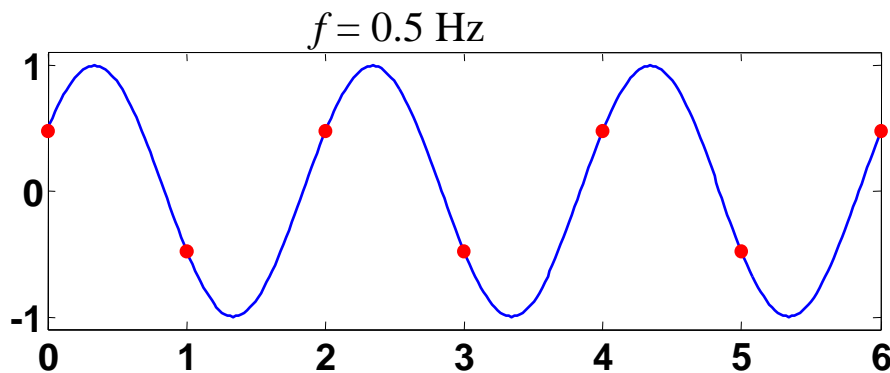
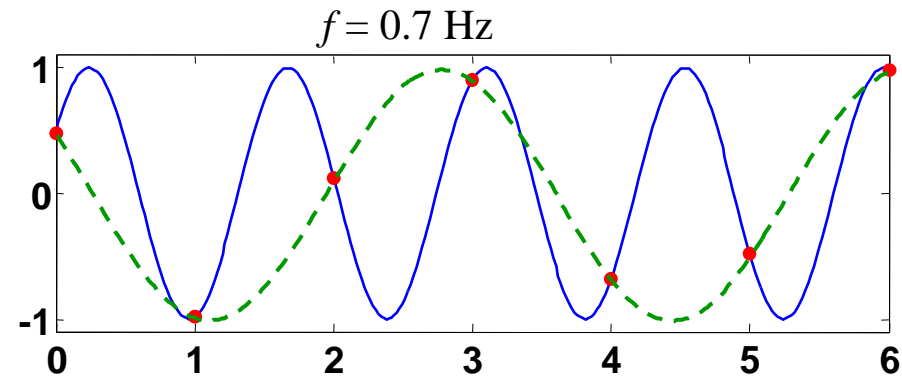
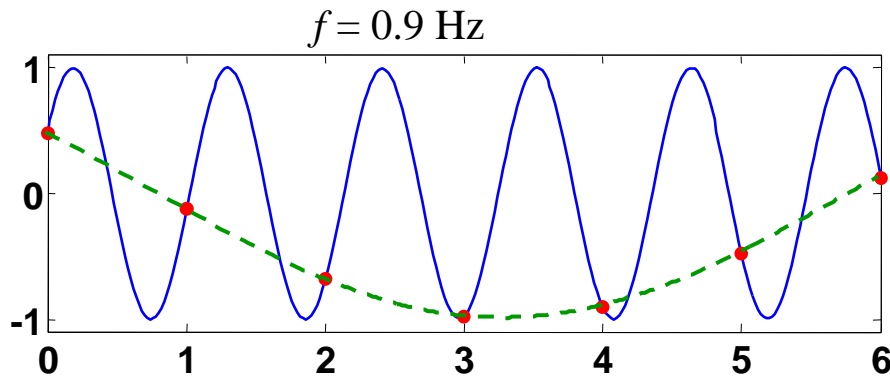
What happens if we sample a signal of frequency $f = 1$ Hz with $f_0 = 1$ Hz?



4.2 Sampling Theorem

Aliasing

Obviously the oscillation is completely gone! We get a signal of frequency zero (a dc value). This happens independently of the phase orientation of the sampler (only the value of the dc value depends on it). For illustration some further examples with $f = 0.9$ Hz, 0.7 Hz, 0.5 Hz and 0.3 Hz sampled with $f_0 = 1$ Hz.





4.2 Sampling Theorem

Sampling Theorem

From the examples on the previous slide we see, that at least the *double* of the sampling frequency is required to reconstruct the original signal from its sampled version ($f = 0.5$ Hz sampled with $f_0 = 1$ Hz). Real signals consist of many (typically infinite many) frequencies. Then, this requirement relates to the highest contained frequency f_{\max} .

Very entertaining podcast:
Fritterin' Away Genius
Cautionary Tales with Tim Harford

Shannon's Sampling Theorem

The signal $x(t)$ shall be sampled. The highest significant frequency component of $x(t)$ is at f_{\max} . Then the sampling frequency has to be at least *twice* this highest frequency component of $x(t)$:

$$f_0 > 2f_{\max}$$

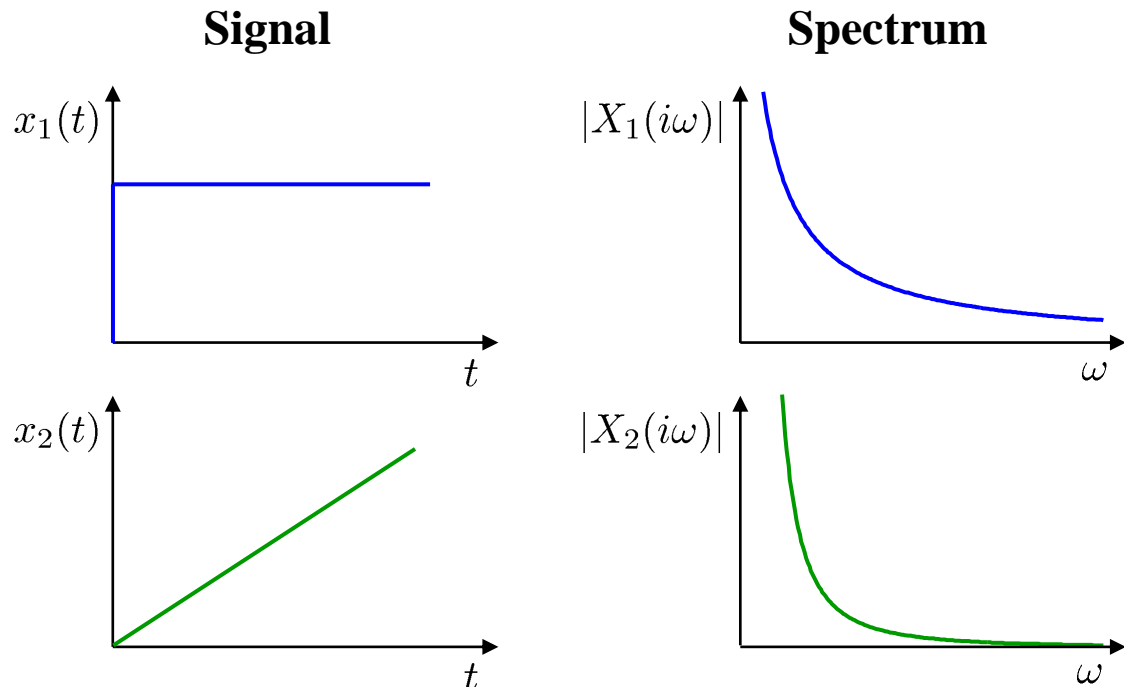
If this theorem is violated, **aliasing** occurs, i.e., frequency components above the half sampling frequency ($f > \frac{1}{2}f_0$) are mirrored into a lower frequency range. By this effect high frequency noise can disturb the signal in any frequency range. Thus **aliasing** should be avoided or at least kept to a minimum.

It is practice to choose $\sim f_0 = 5 \dots 10 f_{\max}$

4.2 Sampling Theorem

Illustration of the Sampling Theorem and the Aliasing Effect

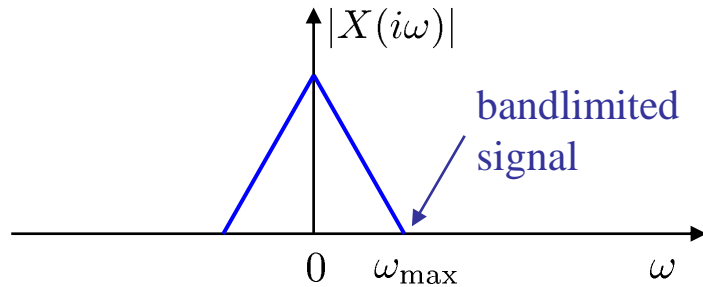
If the sampling theorem is met, it is possible to reconstruct the original signal from its sampled version, i.e., no information loss takes place. However, in reality most signals are not bandlimited. This means they have frequency components up to infinity, i.e., no upper bound exists ($f_{\max} = \infty$). Typical signals like steps, ramps, rectangular shapes stretch their spectrum between zero and infinity. Such signals cannot be reconstructed *perfectly*.



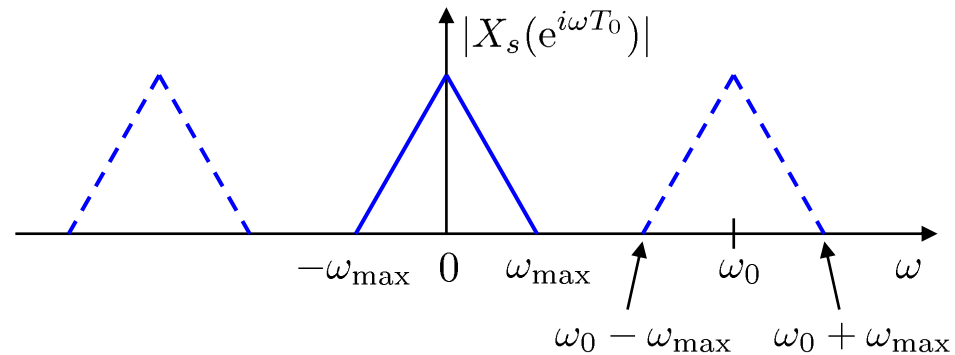
4.2 Sampling Theorem

Illustration of the Sampling Theorem and the Aliasing Effect

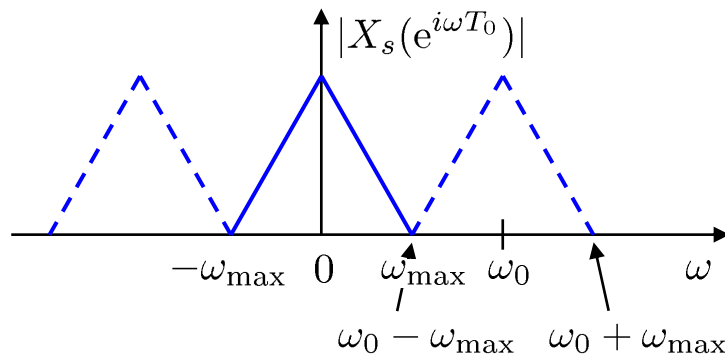
Spectrum of the continuous signal



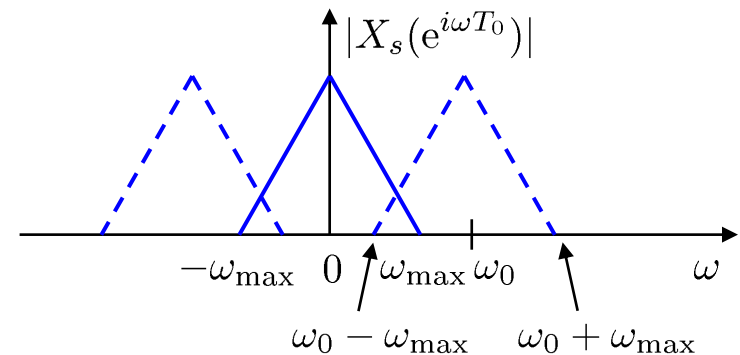
Spectrum of the sampled signal $\omega_0 > 2\omega_{\max}$



Spectrum of the sampled signal $\omega_0 = 2\omega_{\max}$

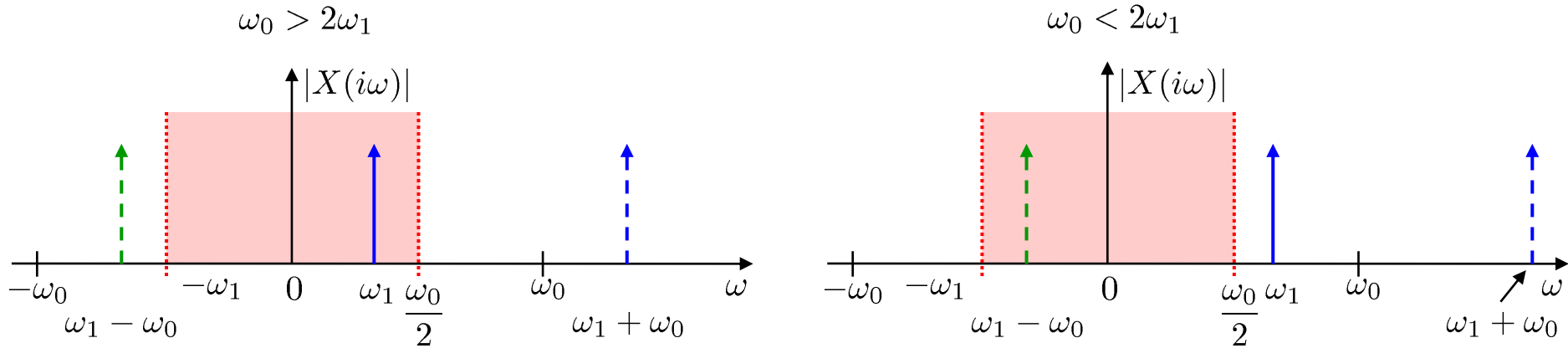


Spectrum of the sampled signal $\omega_0 < 2\omega_{\max}$



4.2 Sampling Theorem

Aliasing for Sampling a Sin-Signals With Angular Frequency ω_1



Each signal component of frequency ω_1 is mirrored through the sampling process to:

$$\omega_l = \omega_1 + l\omega_0 \quad \text{mit } l = \dots, -2, -1, 0, 1, 2, \dots$$

As long as ω_1 lies inside the red area (solid), i.e., the sampling theorem is not violated, the mirrored components (dashed) keep lying outside the red area (left figure).

As soon as ω_1 lies outside the red area (solid), i.e., the sampling theorem is violated, the mirrored components (dashed) lie inside the red area (right figure). **Aliasing occurs!**

If a component changes from ω_1 to ω_0 , a mirrored alias component at $\omega = 0$ is created.

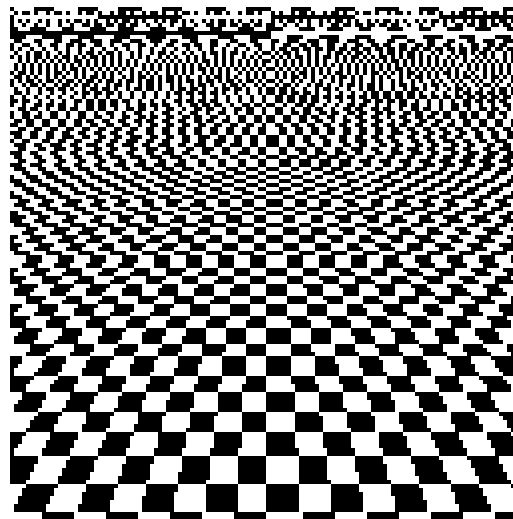
4.2 Sampling Theorem

Aliasing in Image Processing

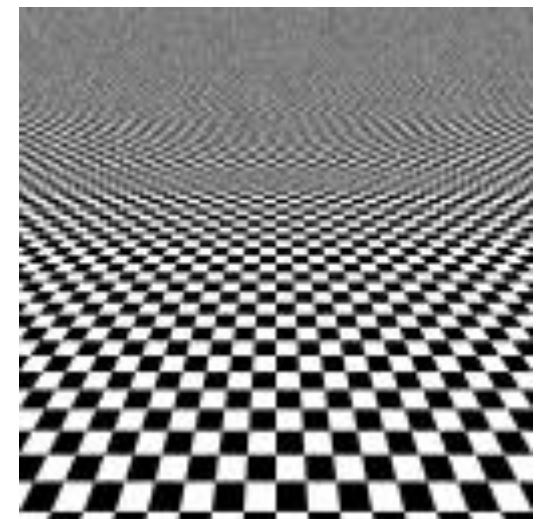
Signal processing is relevant not only for signals over *time*. It is also important for signals over space like pictures/photos(2-D: columns & rows) or a combination of both in videos (3-D). For such *spatial* signal the same laws and relationships hold. Signals over time can be filtered, so can signals over space.

Image processing therefore also has to deal with the **aliasing effect**. A high spatial frequency corresponds to alternating points of black and white (or differently colored). Without a special so called **anti-aliasing filter**, such components of high frequency can significantly disturb the picture. It is particularly prominent for tiny *checkered patterns* and known as the **Moiré effect**. A low-pass anti-aliasing filter prevents such destructive effects. Every digital photo and video camera has build in such a filter.

Without Anti-Aliasing



With Anti-Aliasing



4.3 Quantization

Quantization Error

Any digital value is quantized in its amplitude. A continuous value has to be mapped to a discrete value via the A/D converter. This means that each interval in the continuous range corresponds to some integer number. All values inside of such an interval are indistinguishable after the A/D conversion.

If we quantize a continuous value in the range from x_{\min} to x_{\max} into n bits, 2^n intervals or **quantization levels** exist. In such a quantization the maximum error can be calculated as

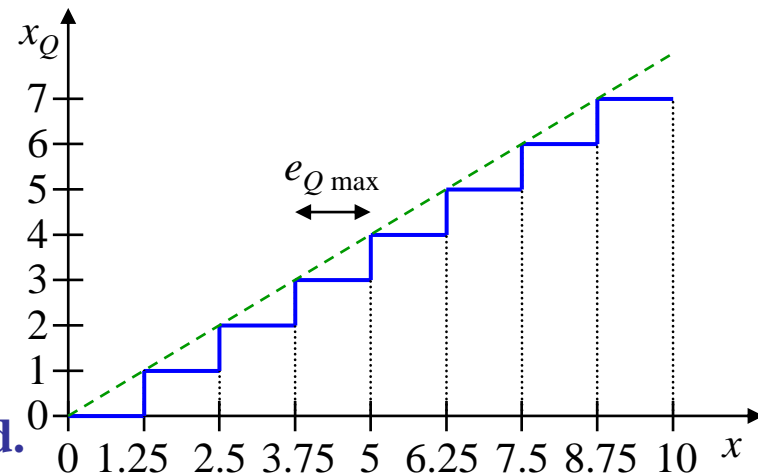
$$e_{Q \max} = \frac{x_{\max} - x_{\min}}{2^n}$$

because this is the interval width. The quantization error with this approach is always positive because the green line (dashed) always is above the blue one (solid).

Example: $x_{\min} = 0$, $x_{\max} = 10$, $n = 3$ Bits

$$e_{Q \max} = 10 / 8 = 1.25$$

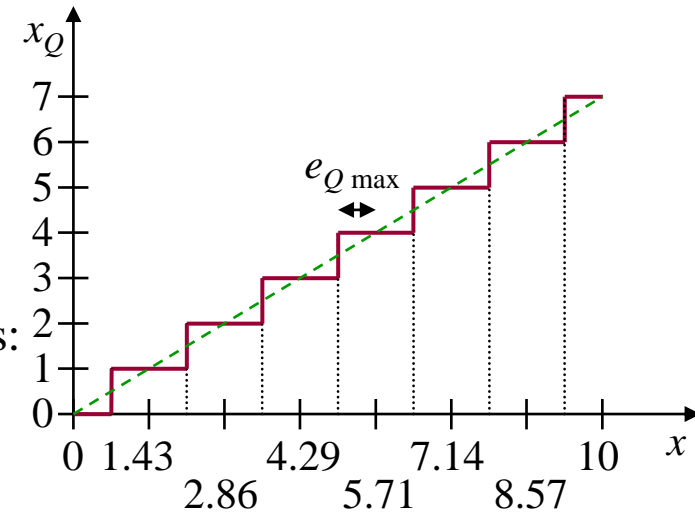
In practice 8, 12, 16 bit A/D converters are standard.



4.4 A/D and D/A Converters

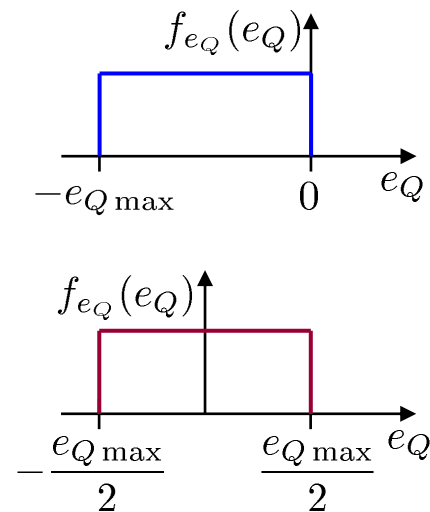
It is possible to improve this quantization error by almost a factor of 2. Instead of always rounding down, we can draw the green (dashed) line through the average by shifting it $e_Q/2$ to the right. Now x_{\min} and x_{\max} are in the medium values of the intervals, not their limits. This sacrifices one interval. The maximum quantization error is:

$$e_{Q \max} = \frac{1}{2} \frac{x_{\max} - x_{\min}}{2^n - 1}$$



Quantization Noise

Although the quantization error is caused *systematically*, it appears to be of random nature. Thus, one speaks of **quantization noise** that any A/D conversion creates in principle. Since all values are of equal probability, it can be modeled by an equal probability distribution. In old synthesizers or CD players quantization noise could be heard for low volume sounds.



4.4 A/D and D/A Converters

A/D Converters: Fundamentals

The three main characteristics of A/D converters are:

- Resolution
- Speed
- Realization effort / price

These characteristics are in conflict with each other. E.g. a high resolution implies a low speed or high effort/price (or both).

With **resolution** we mean the number of bits n which results in 2^n quantization levels. It is not reasonable to request a much higher resolution from the A/D converter than the measurement noise or other disturbances have as mean amplitude since the accuracy of the signal then is limited to this value anyway. Otherwise the lowest significant bits are determined by noise and carry no information.

The **speed** (bandwidth) determines how fast the A/D conversion is performed and therefore how fast the sampling is possible (maximum sampling frequency). The **effort** typically shows directly in the **price**.

A **low sensitivity** with respect to **environment conditions** is also an important criterion.

4.4 A/D and D/A Converters

A/D Converter: Parallel Principle or Flash Converter [1]

The voltage which shall be converted U_E is directly compared with n different reference values. For any of the existing $2^n - 1$ quantization levels one comparator is required.

Properties: Very fast
(10 MHz), low resolution (8 bit).

Application Field: Video.

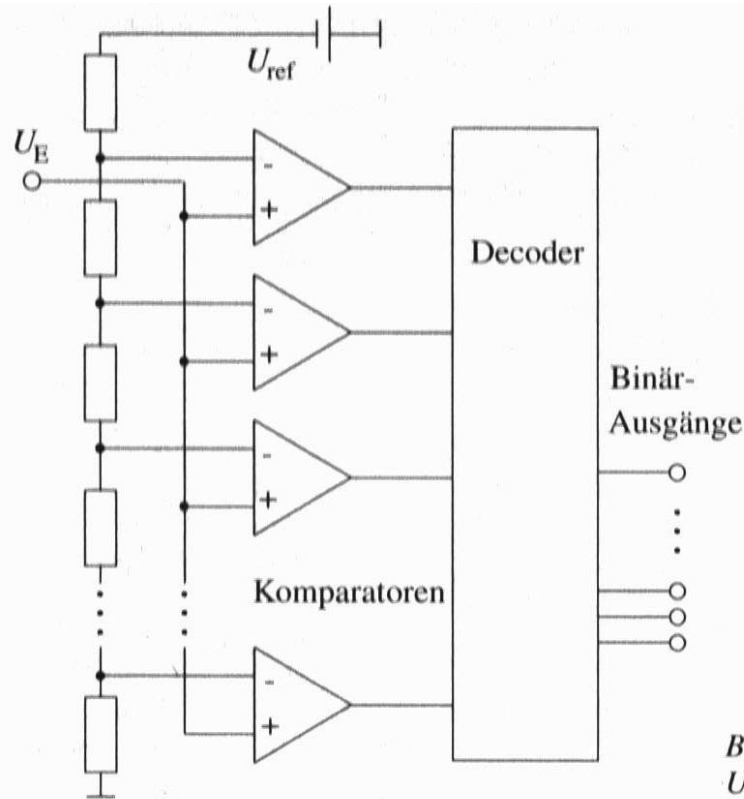


Bild 5.5 Parallel-A/D-Umsetzer

4.4 A/D and D/A Converters

A/D Converter: Successive Approximation or Weighting Method [1]

The procedure is identical to the weighting with a beam balance where the available weights are $1, \frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{2^n}$. A combination of these weight represents the quantization levels. One starts with the highest weight and adds or removes weights in descending order to balance the beam. At the end we have n steps (n times a weight is added and possibly removed). The remaining weight represent “1”, the removed “0” in the converted value. Weights are realized by voltages, the beam balance is realized by comparators.

Properties: Medium speed (1 MHz), medium-high resolution: 12, 16, even 24 bit.

Application Field: Computer plug-in A/D converter cards for measuring signals.

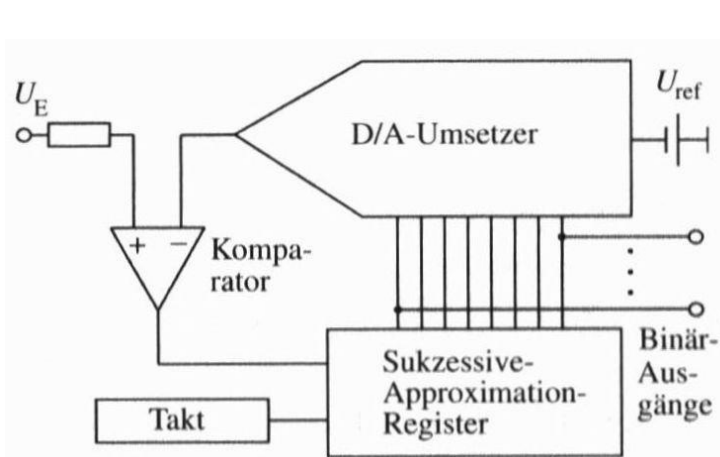


Bild 5.6 Sukzessive-Approximation-A/D-Umsetzer

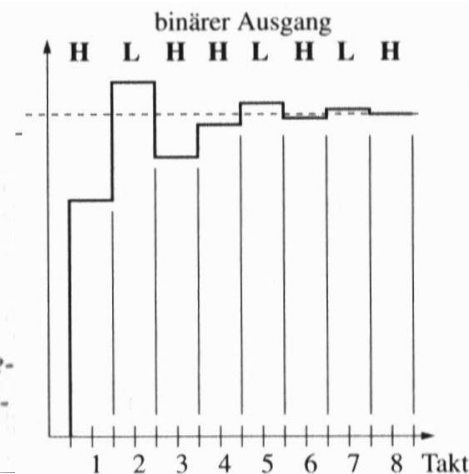


Bild 5.7 Aufbau des binären Ausgangssignals beim Sukzessive-Approximation-A/D-Umsetzer

4.4 A/D and D/A Converters

A/D Converter: Servo Principle [1]

Constantly the difference of the voltage U_E to be converted and the output of the A/D converter which is converted back into an analogue signal is compared like in a control system. If this difference is equal to zero, then the A/D conversion is correct. A positive or negative difference triggers a count which is counted up or down (feedback!). Because the clock has a certain speed, the conversion needs a lot of time that depends on the size of the difference; this is similar to an integrative controller. However, if the difference is small

because the signal hardly changes (no steps or impulses) the converted voltage follows closely.

Properties: Speed depends on the size of steps.

Application Field: continuous conversion, slowly changing signals.

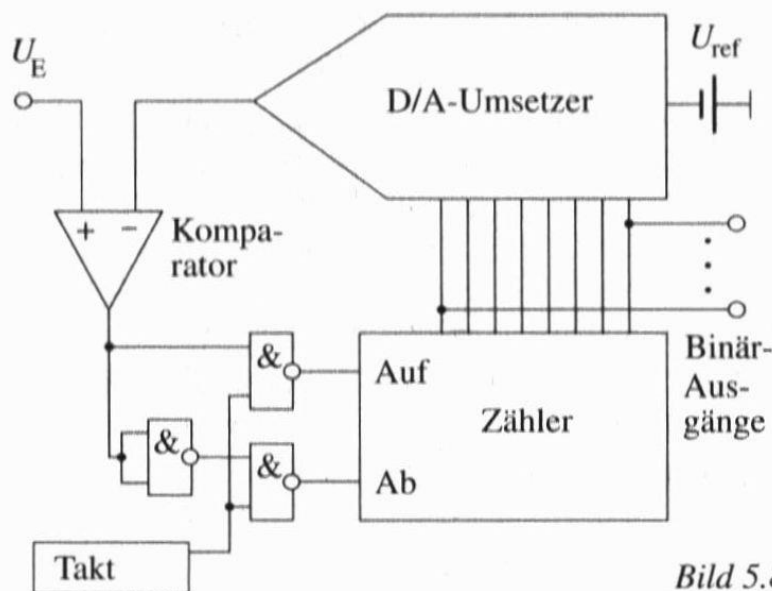


Bild 5.8 Nachlauf-A/D-Umsetzer

4.4 A/D and D/A Converters

A/D Converter: Dual Slope Principle [1]

The dual slope converter uses an extended **ramp method**. The input voltage U_E is integrated over a fixed period of time t by an integrator circuit. Subsequently, the integrated voltage is integrated down again until zero by some reference voltage U_{ref} of opposite sign. During the latter time period a counter runs whose counting then is proportional to the original input voltage U_E .

Properties: Excellent quality and suppression of unwanted influences. It is almost independent of material properties, temperature changes, etc. because those effects cancel each other during up- and down-integration. Slow speed since integration takes a lot of time.

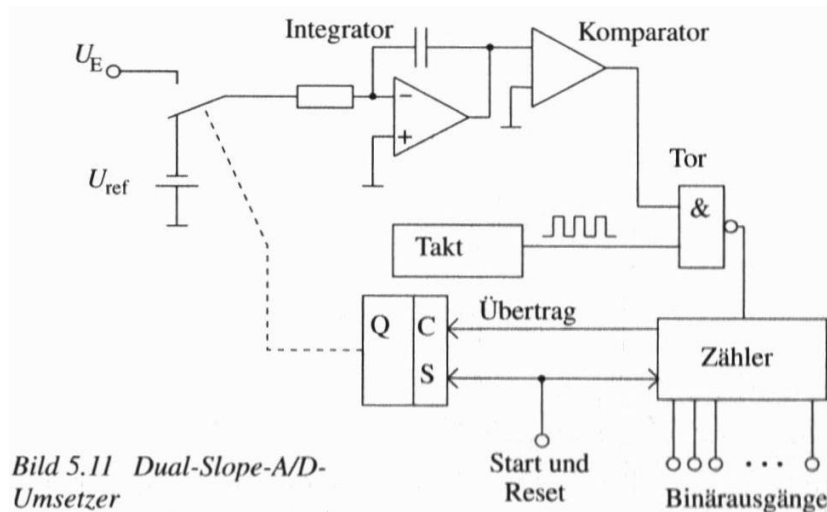


Bild 5.11 Dual-Slope-A/D-Umsetzer

Application Field: Digital volt meter.

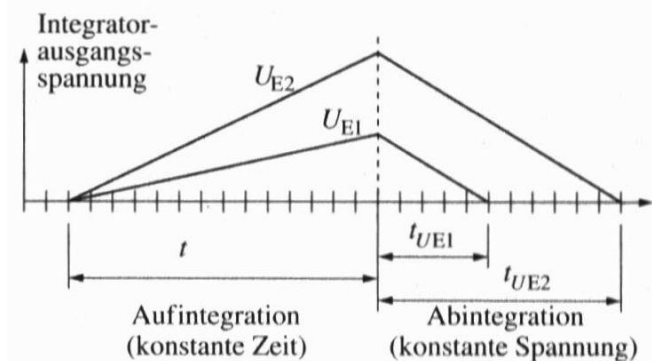


Bild 5.12 Funktionsweise des Dual-Slope-A/D-Umsetzers

4.4 A/D and D/A Converters

A/D Converter: Sigma-Delta- or Charge-Balance- or 1-Bit-Method [1]

In the first part of a sigma-delta-converter a bit stream is generated whose average value is proportional to the input voltage U_E that shall be converted. This is achieved through a control loop in which the difference between U_E and a positive and negative reference voltage is fed to a comparator. For $U_E = 0V$ the up- and down-integration phases are equally long.

In the second part the bit series in the bit stream is counted and converted into a digital value.

Properties: very high resolution (24 bit), medium speed.

Application Field: audio, instrumentation.

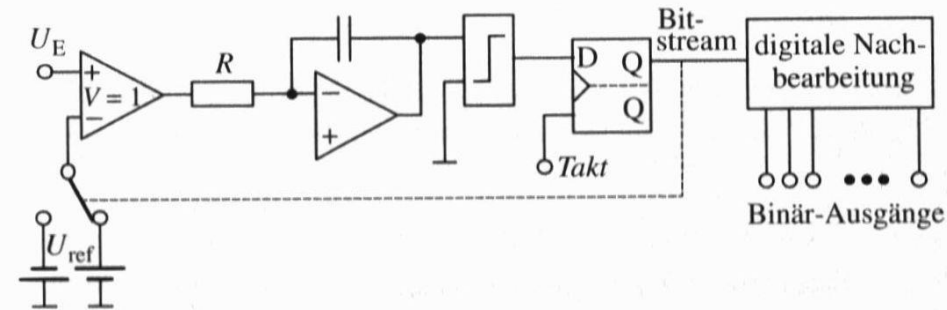


Bild 5.15 Delta-Sigma-Umsetzer

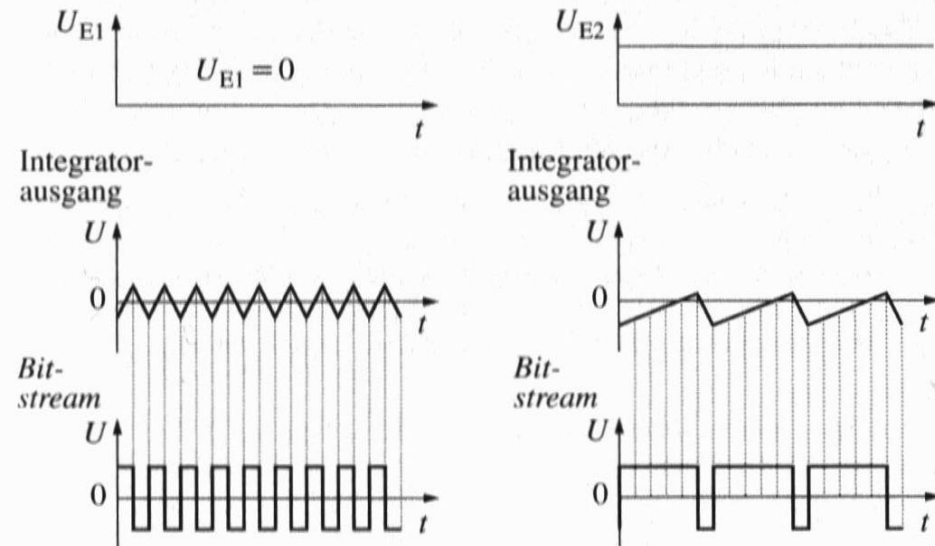


Bild 5.16 Funktionsweise des Delta-Sigma-Umsetzers

4.4 A/D and D/A Converters

D/A Converter: Current Weighted Principle [1]

In comparison to A/D conversion is the way back quite simple.

One possibility is to drive a constant current through a number of resistors with geometrically ordered resistances, i.e., R , $2R$, $4R$, $8R$, ... The voltage drop over each resistor corresponds to a bit in the digital value ("1" for "on" and "0" for "off"). The sum of these voltages then corresponds to overall value, e.g., the bit series

0001 1010

gives the analogue voltage

$$U = (16R + 8R + 2R)I = 26RI$$

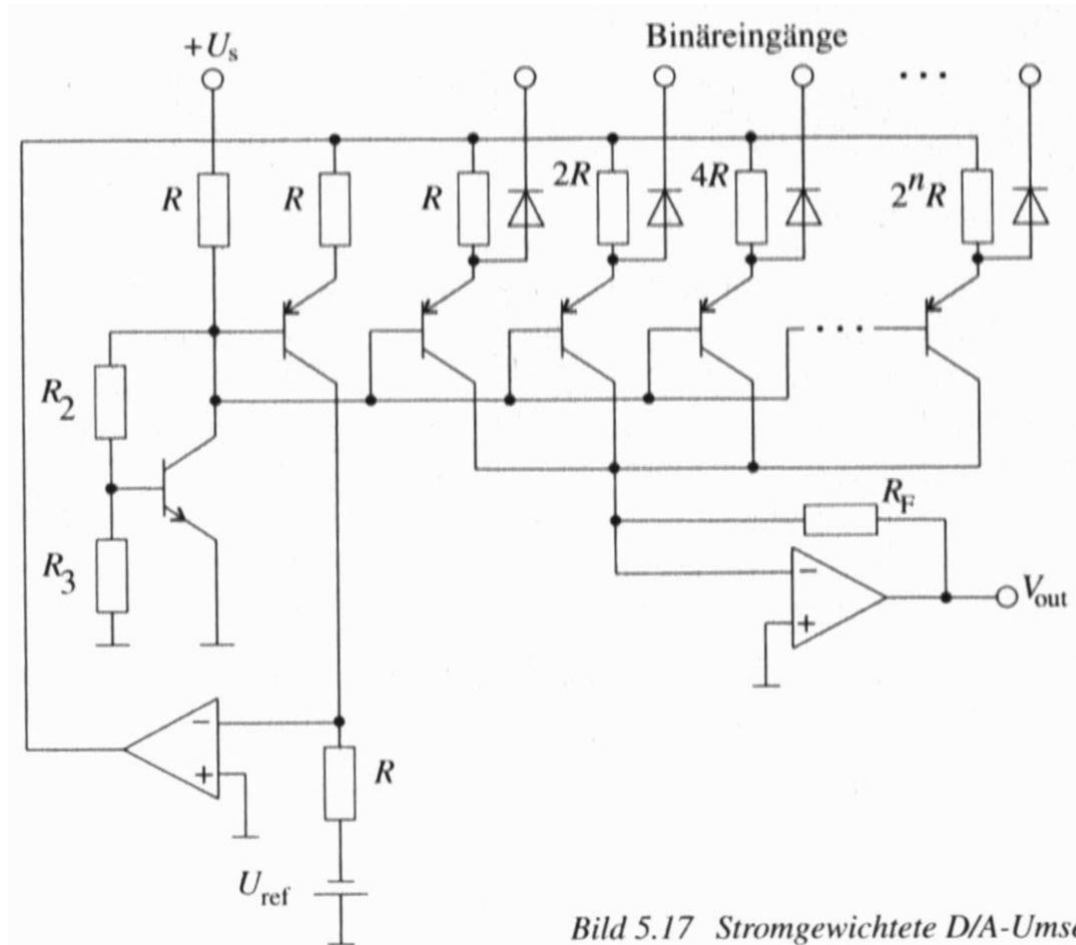


Bild 5.17 Stromgewichtete D/A-Umsetzer

4.4 A/D and D/A Converters

D/A Converter: R - $2R$ Principle [1]

The R - $2R$ converter divides a current in each knot into 2 halves (factor 2). One half drives a resistor with resistance $2R$ and thereby creates a proportional voltage drop. The other half is again divided into 2 halves etc. The main advantage compared to the method explained on the last slide is that *only two kinds of resistors* R and $2R$ are required. They are much easier and cheaper to manufacture in high quality (low temperature dependence) than all the different kinds for the current weighted principle R , $2R$, ..., $1024R$ (for a 10 bit converter).

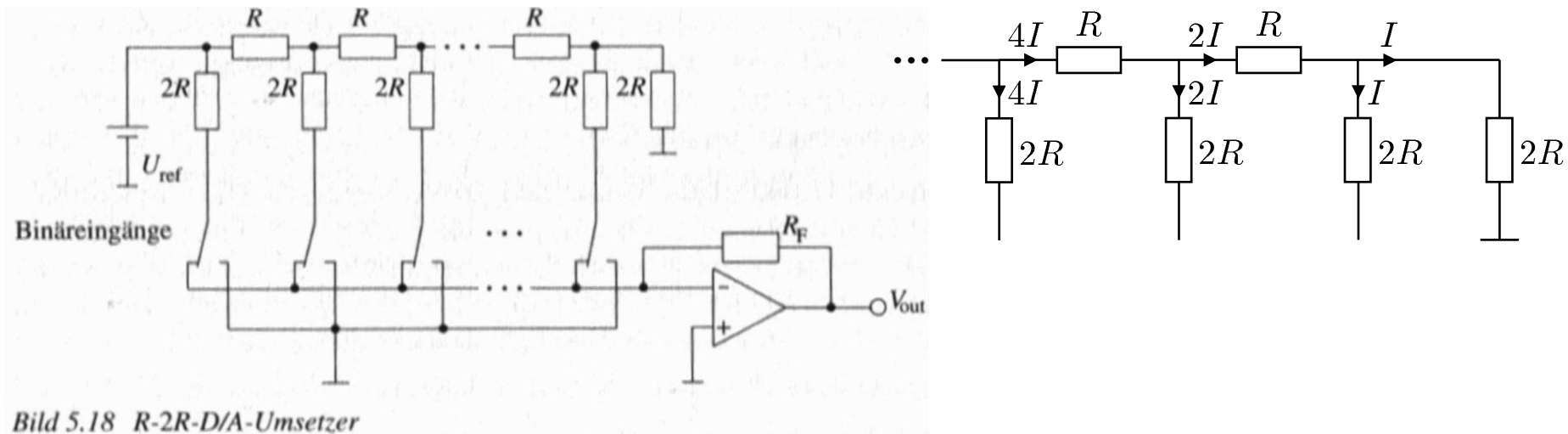


Bild 5.18 R - $2R$ -D/A-Umsetzer

4.5 Measurement of Frequency

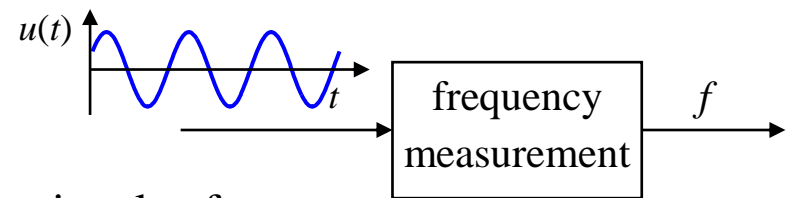
Fundamentals of Frequency Measurement

In the discussion of velocity and angular velocity measurements in Chapter 3.3 it is explained how such a measurement can be transformed into a voltage signal of same frequency. The last step that still is open, is to determine this frequency! The reason for this is that frequency measurement is typically done *digitally* – thus it has been postponed up to here.

The task here is therefore to determine the frequency f of a given voltage signal $u(t)$.

Two alternative approaches are presented:

- *Measurement of the cycle duration (period):* For signals of *low frequency* it makes sense to measure the time for one (or even half of an) oscillation T_p and calculate the frequency from $f = 1/T_p$.
- *Counting the number of cycles within one time interval:* For signals of *high frequency* it makes sense to measure the number of oscillations within a given time interval and to count them. The frequency can be determined by $f = \text{number of oscillations} / \text{time interval}$.



4.5 Measurement of Frequency

Measurement of a Period [4]

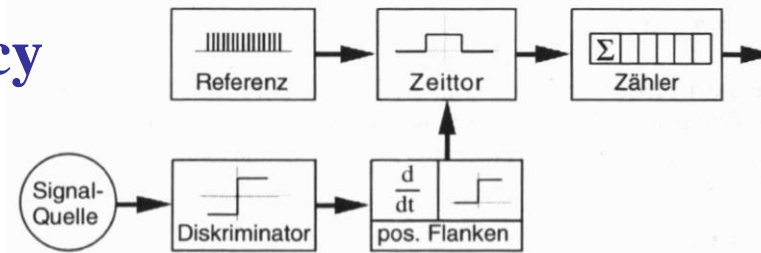


Abbildung 8.6. Strukturbild zur Periodendauerermessung

Well-suited for low-frequency signals

Gate time is one or one half of an oscillation of the original signal.

Reference frequency is artificially generated.

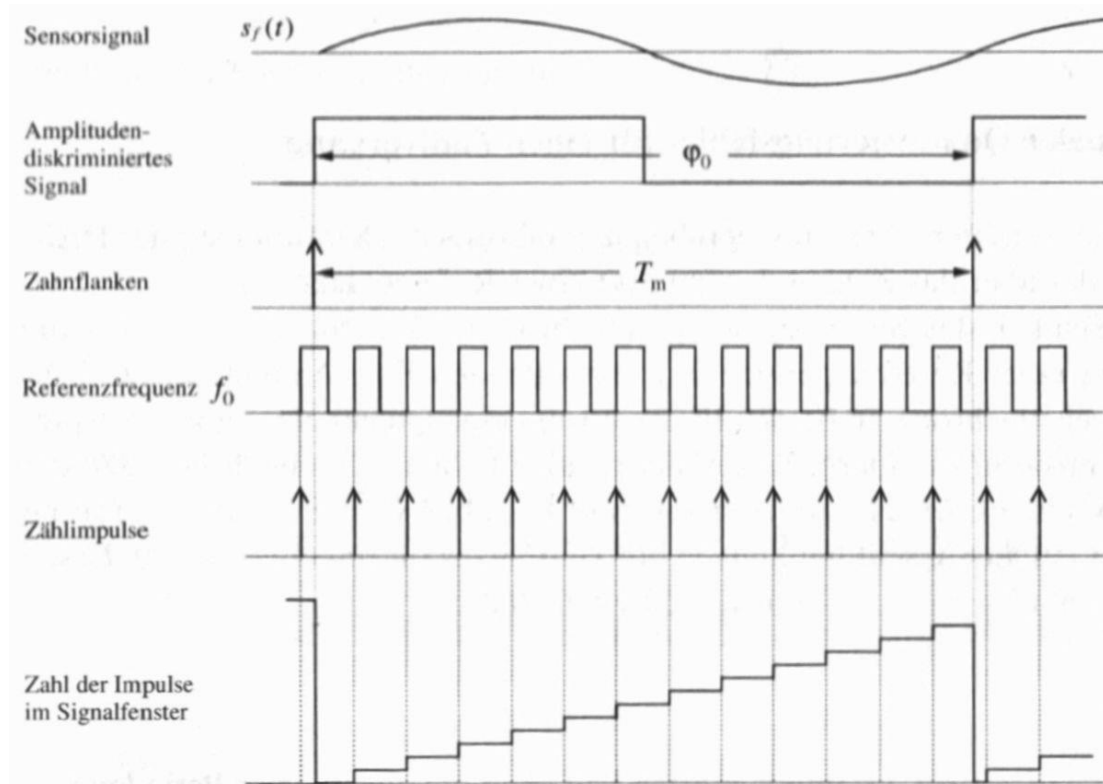
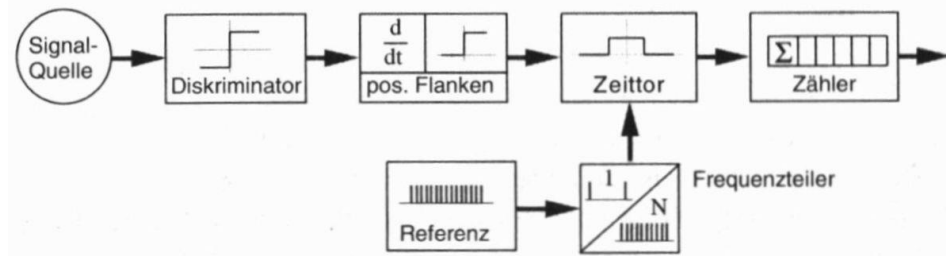


Abbildung 8.7. Digitalisierung frequenzanaloger Signale durch Periodendauerermessung (winkelsynchrone Erfassung)

4.5 Measurement of Frequency



Counting of Many Periods [4]

Abbildung 8.8. Strukturbild zur Frequenzzählung

Well-suited for high-frequency signals.

Frequency come from the original signal.

Gate time is generated arbitrarily (the long the more accurate but slower).

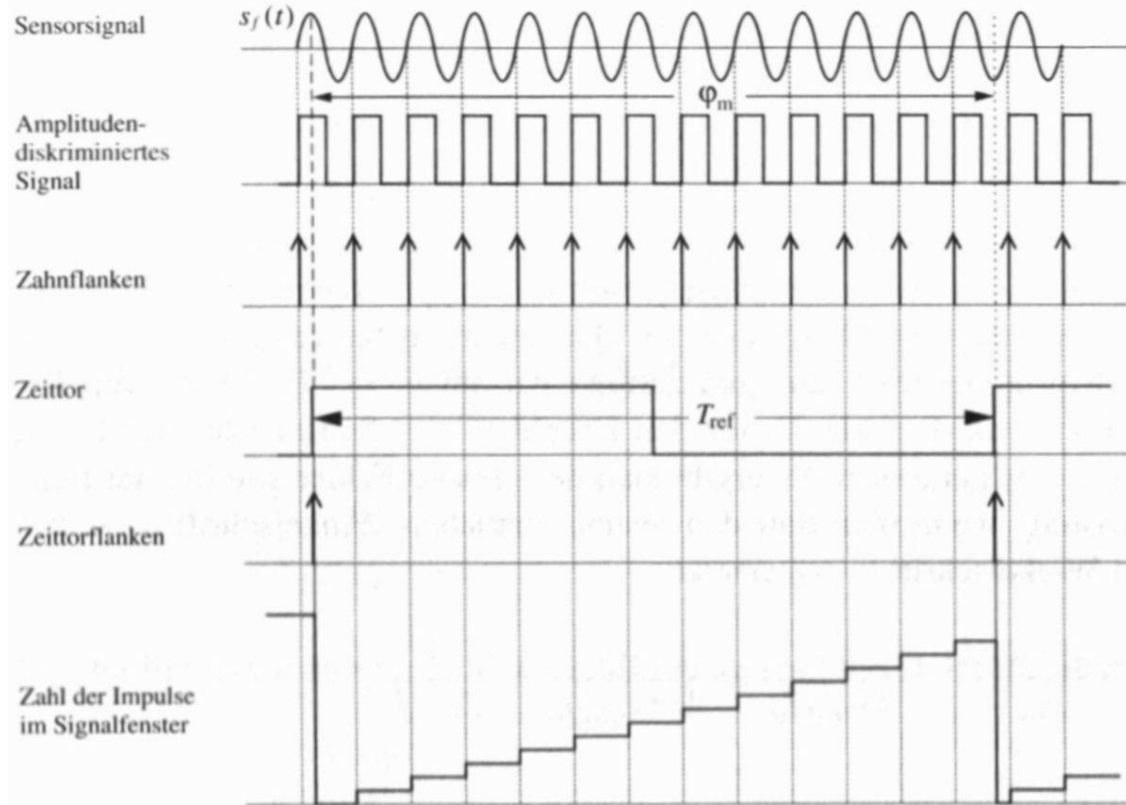


Abbildung 8.9. Digitalisierung frequenzanaloger Signale durch Frequenzzählung (zeitsynchrone Erfassung)

B: Signal Processing

7. Introduction to Signal Processing

Contents of Chapter 7

7. Introduction to Signal Processing

7.1 What for?

7.2 Deterministic and Stochastic Signals

7.3 Application Examples

7.4 Literature

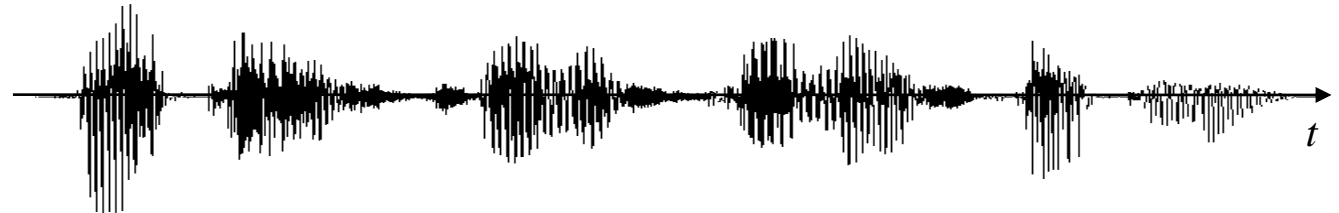
7.1 What For?

What are Signals?

- Signals transfer information.
- Signals are functions, typically of time.
- Signals are measured with sensors and can be available in every physical form like pressure, temperature, voltage, ...

Some Typical Signals

- Speech, music
- Pictures, videos
- EKG, EEG, signals of CT, MRT or PET → image processing, conversion in pictures, ...
- Distance measurement with laser, ultrasound or radar, echo lot, GPS, seismic signals, ...
- Data streams via telephone lines, cable TV, satellite, cell phone, bluetooth, internet, ...
- All kinds of measurements at machines, machines in factories, ...
- Pressure in cylinders of a combustion engine
- Stock prices, number of unemployed people, development of populations, ...



7.1 What For?

What is signal processing?

The *analysis, manipulation* and *integration* of signals

Application areas of signal processing?

- Storage, reconstruction
- Separation of desired signal and disturbance (signal-to-noise ratio)
- Compression
- Feature extraction (pre-stage of every classification)

Method/Tools of signal processing

- Transformation, correlation
- Filtering, disturbance suppression
- Detection, classification, pattern recognition
- Identification, estimation
- Compression, integration, fusion

7.1 What For?

Applications



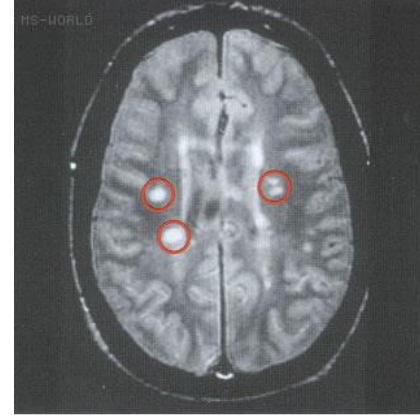
Camera



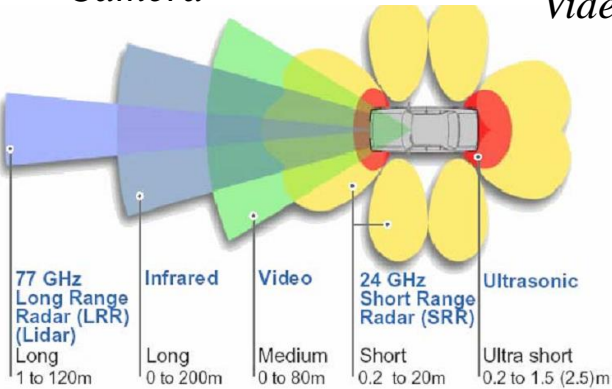
Video



Cell Phone



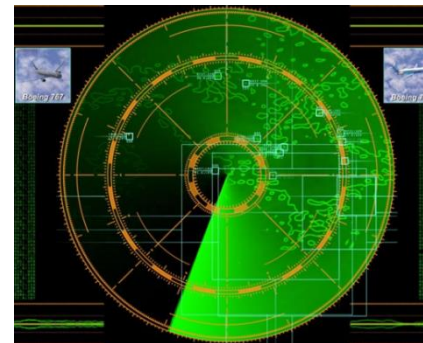
MRT



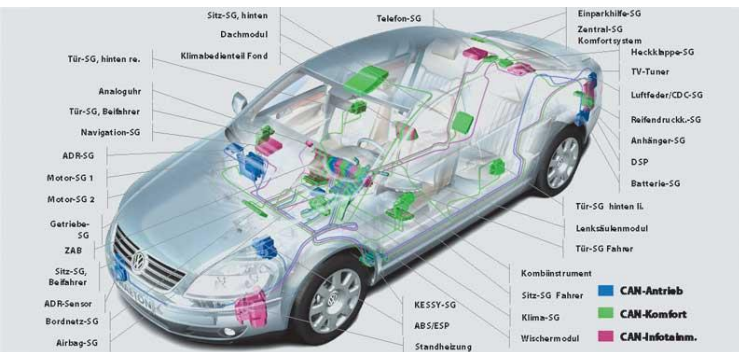
Driver Assistance



Messtechnik



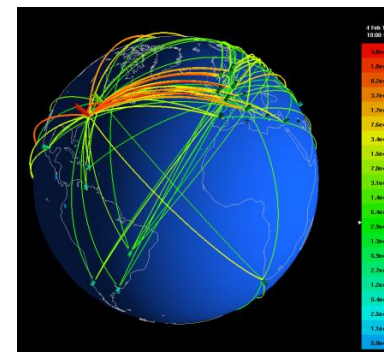
Radar



Integration Sensorics/Control Units



Night Vision



Internet



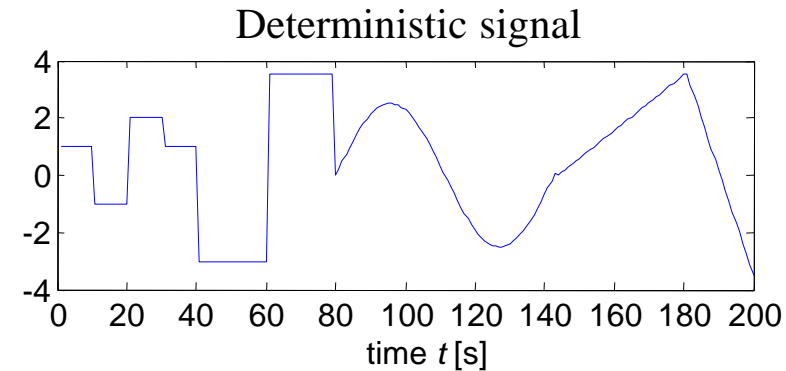
GPS

7. Introduction to Signal Processing

7.2 Deterministic and Stochastic Signals

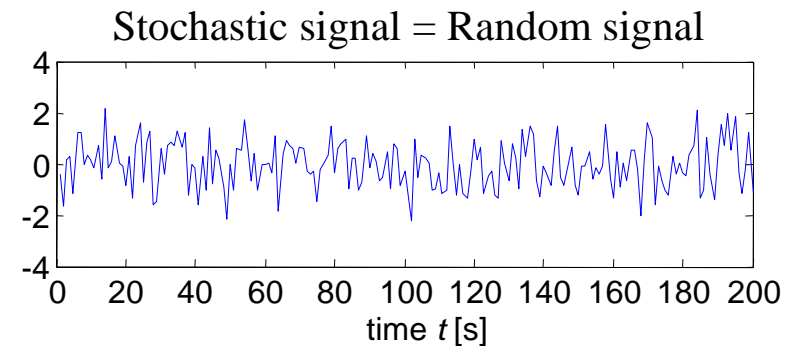
Deterministic Signals Do *Not* Depend on Randomness:

- Dirac impulse
- Step
- Ramp
- Periodic signals: sine, rectangular, ...



Stochastic Signals Depend on Randomness:

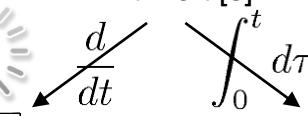
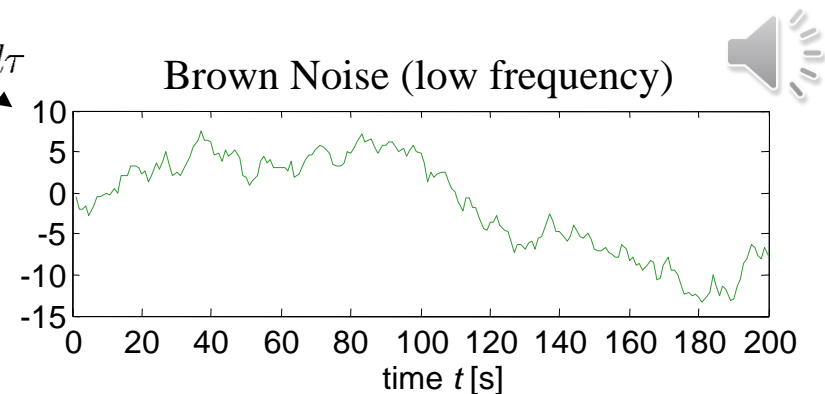
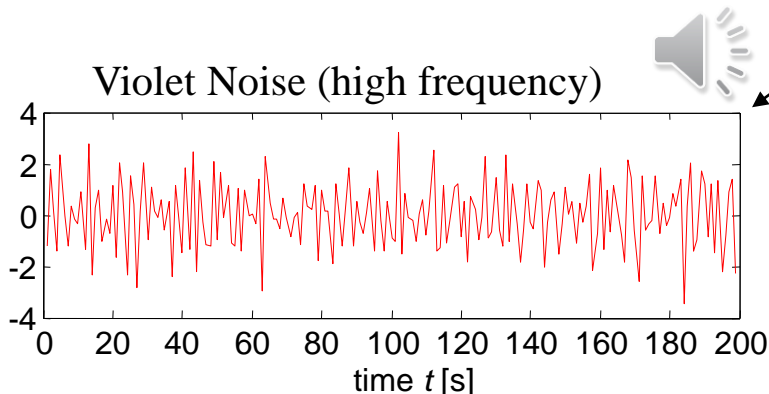
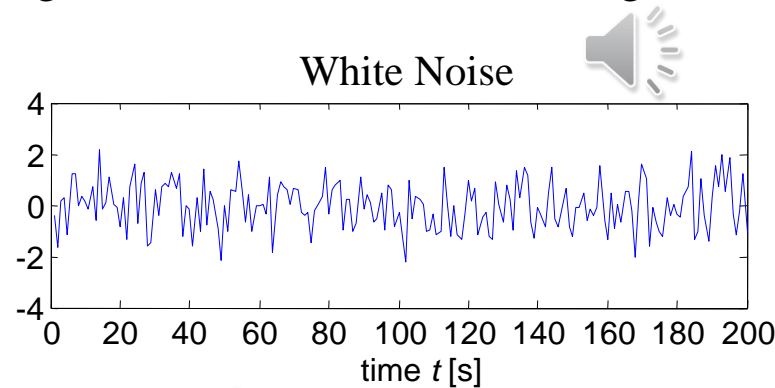
- Noise
- Distribution of amplitudes:
 - Gaussian,
 - uniform, ...
- Frequency characteristics:
 - white: all frequencies have the same power,
 - band limited: only a certain frequency range is present, ...



7.2 Deterministic and Stochastic Signals

Motivation for Using Stochastic Signals

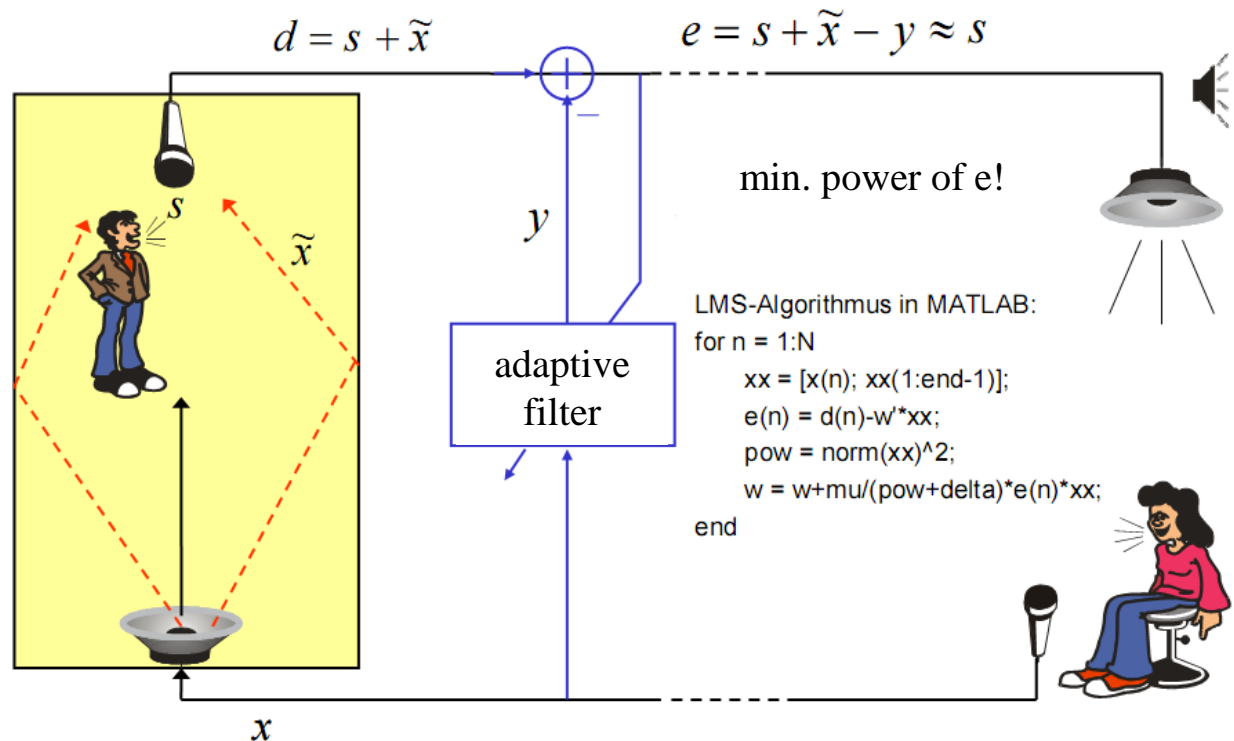
- Physical effects are truly random (e.g. radioactive decay).
 - Many tiny disturbances *appear* like random, but are of deterministic nature each if we look in close detail (what needs time and dedication).
- In both cases: Modeling of the effects as stochastic signal makes sense!



7.3 Application Examples

Acoustic Echo Compensation (Hands-Free Talking)

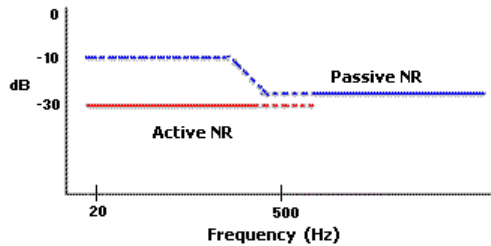
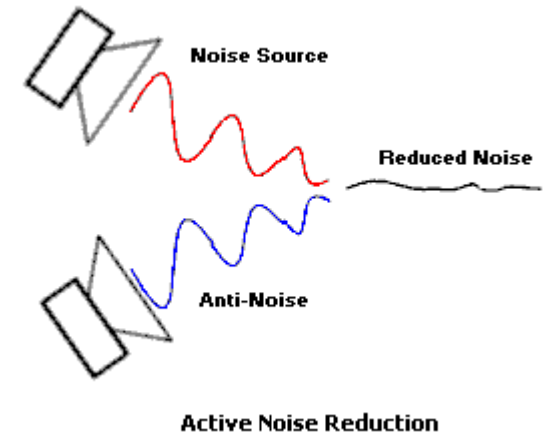
- Adaptive (automatically self-adjusting) filters eliminate disturbing and annoying feedbacks by modeling the transfer characteristics between speaker and microphone and subtracts this signal part from the overall signal.



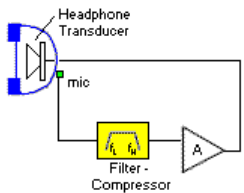
7.3 Application Examples

Active Noise Cancellation

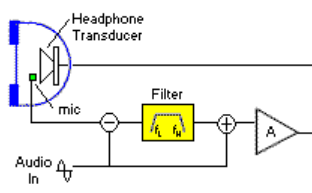
- By direct measurement of the noise and generation of a opposite phase signal (180° phase shift) *destructive interference* annihilates the noise or at least parts of it.
- Works well in the low-frequency range up to 1000 Hz.
- Damping (active + passive) up to -30 dB possible!



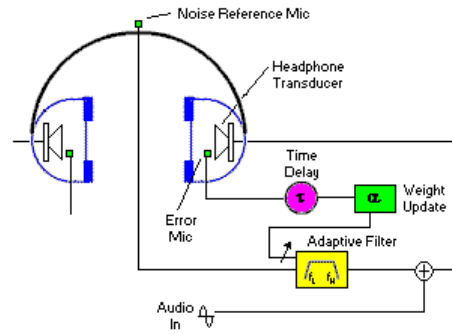
Operational Range of Active and Passive Noise Reduction



Noise-Filtering Headphone

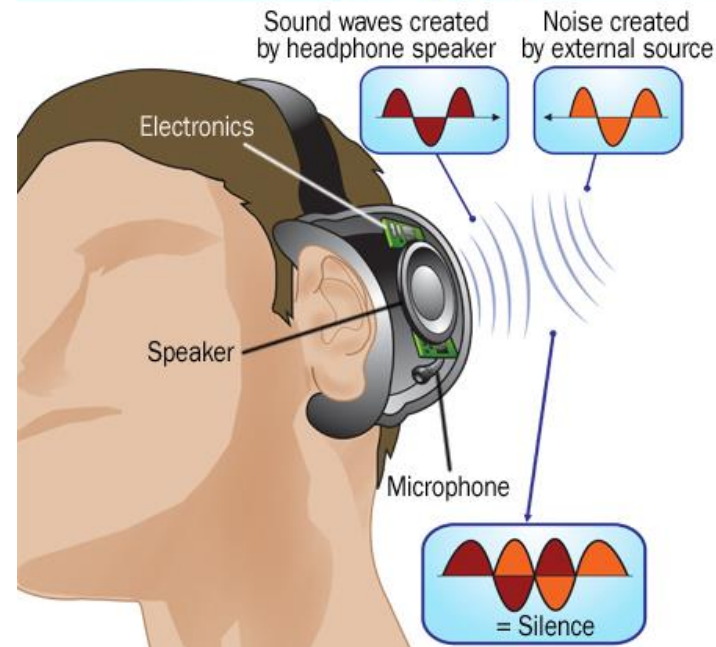


Closed-Loop ANR with Internal Mic



Adaptive Noise Cancellation Headphones

Inside noise-canceling headphones



7.3 Application Examples

Active Noise Cancellation with *Sony Xperia Z2*

- Works via cell phone, not with headphones alone.
- Use processor and battery of cell phone.

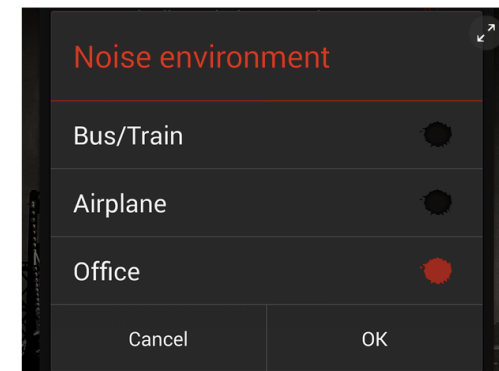
Source: <https://www.theguardian.com/technology/2014/apr/17/sony-xperia-z2-review-phone-android>:

Noise cancelling

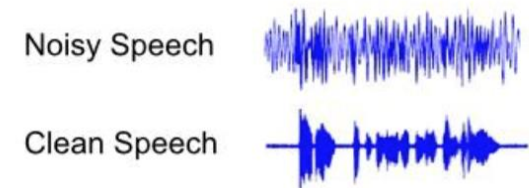
The Xperia Z2 is the first smartphone with active noise-cancelling technology integrated into its body for use with a special headset. The headset contains microphones that monitor incoming noise and send it back to the Z2, which then blends in noise cancellation to whatever is being played.

Active noise-cancelling is not a new thing, but normally it requires headsets with a bulky battery and electronics pack attached to the headphone wire. Sony has squeezed the circuitry and software into the Z2, removing the bulk that normally makes noise-cancelling earphones bulky or heavy.

Because the noise cancelling control system is built into the phone, you can select an appropriate profile for the noise to block out - the options are planes, trains, buses and the office - which makes the technology much more effective. I found the office setting to be particularly effective at blocking out the hubbub of an open-plan office, much more so than most other noise-cancelling ear or headphones.



Sony MDR-NC31EM Digital Noise Cancelling Headset (Amazon UK)

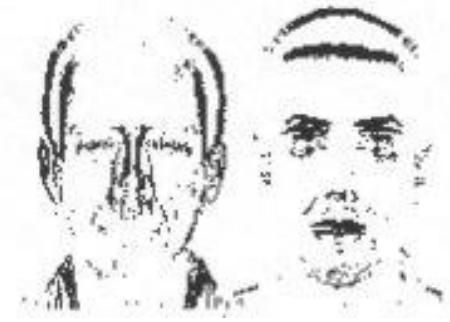


Source: <http://www.techradar.com/news/phone-and-communications/mobile-phones/background-noise-reduction-one-of-your-smartphone-s-greatest-tools-1228924>

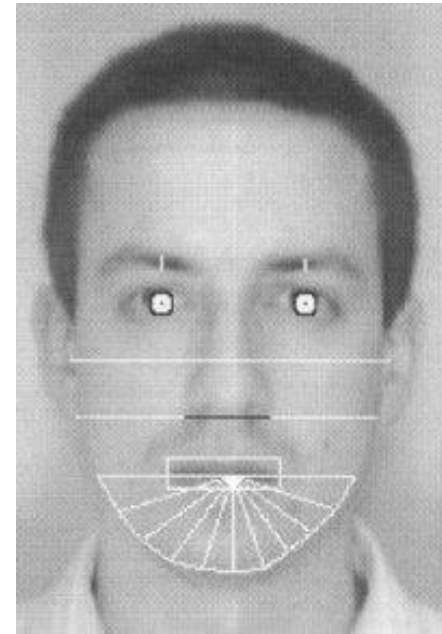
7.3 Application Examples

Face Detection

- By calculating the gradients in x - and y -direction, a vertical and horizontal edge-image can be generated.
- From these edge-image the features can be extracted more easily.
- This software extracts 22 features per face:
 - vertical position of nose and its width,
 - vertical position of mouth, its width, and its height,
 - vertical position and heights of eyebrows over eye center,
 - 11 radii that describe the form of the chin,
 - width of face at nose bottom edge,
 - width of face at center of eyes and nose.



Vertical and horizontal edge-image



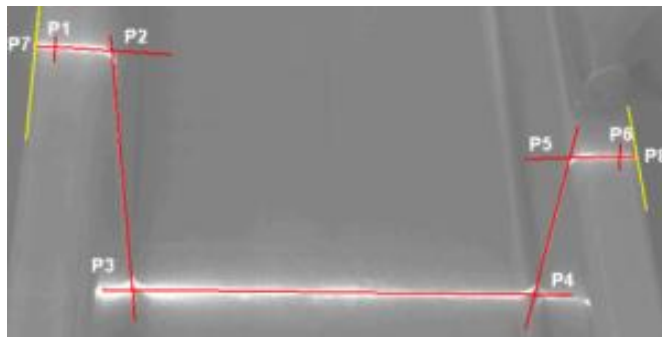
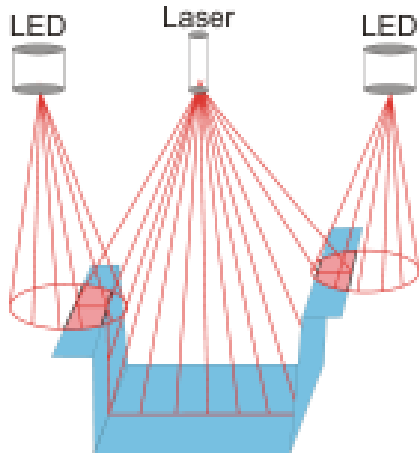
22 features used for face detection.

Quelle: www.markus-hofmann.de

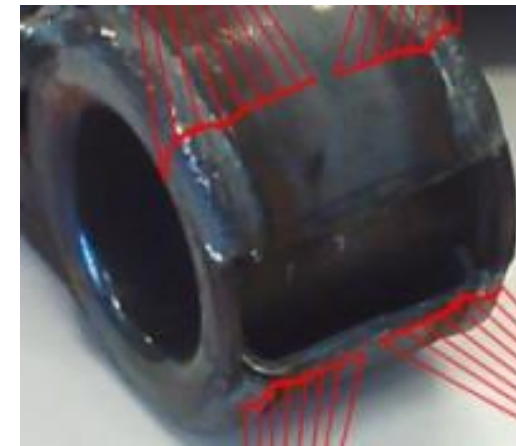
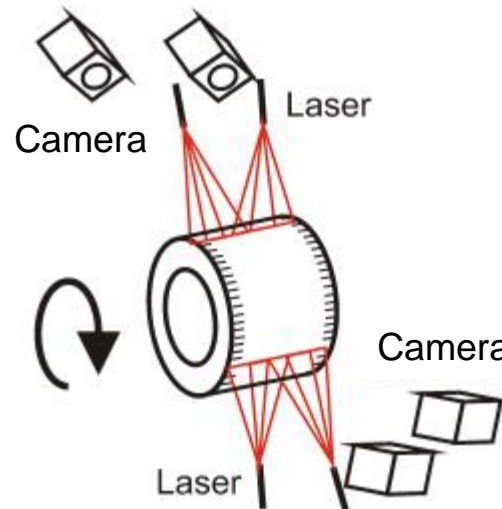
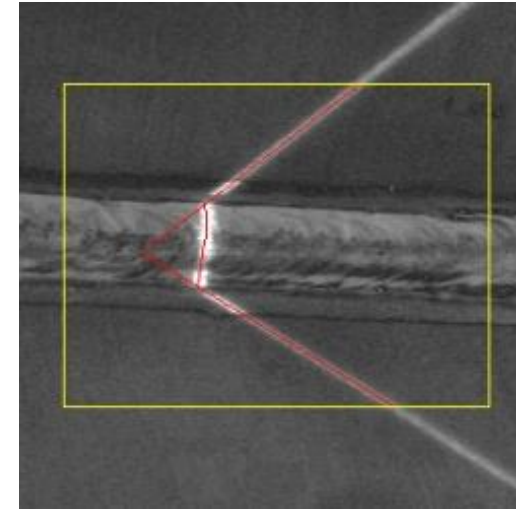
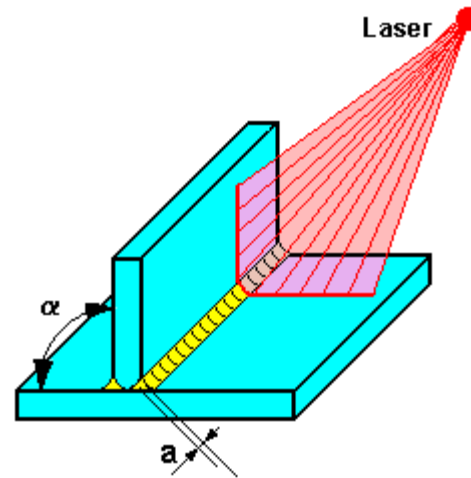
7.3 Application Examples

Industrial Image Processing

Component measurement to supervise tolerances



Supervision of quality welding line



7.3 Application Examples

Image Compression

Image 279 x 356 pixel: as *.tif (without loss): 394 kB

*.jpg (100%): 119 kB



*.jpg (60%): 22 kB



7.3 Application Examples

Image Compression

Image 279 x 356 pixel: as *.tif (without loss): 394 kB

*.jpg (100%): 119 kB

*.jpg (20%): 10 kB



7.3 Application Examples

Image Compression

Image 279 x 356 pixel: as *.tif (without loss): 394 kB

*.jpg (100%): 119 kB

*.jpg (10%): 5,4 kB



7.3 Application Examples

Image Compression

Image 279 x 356 pixel: as *.tif (without loss): 394 kB

*.jpg (100%): 119 kB

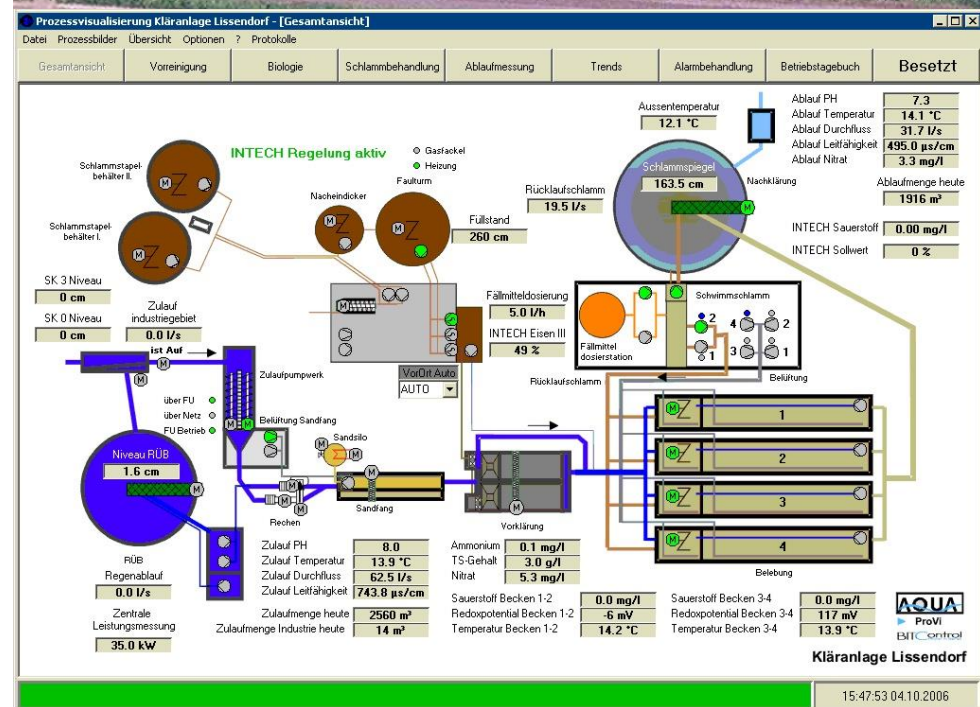
*.jpg (2%): 2,1 kB



7.3 Application Examples

Process Automat. for Waste Water Plants

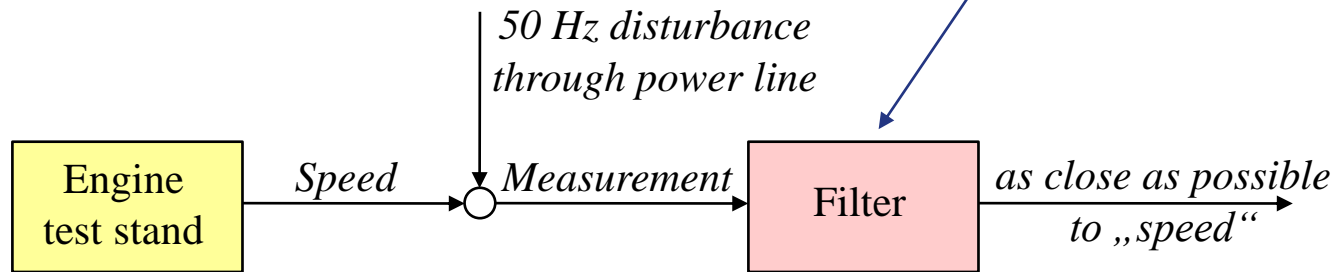
- Graphic description of the plant
- Measurement of many process quantities
 - temperatures
 - flow rates
 - concentrations
- Measurement of disturbances
- Logging of all value for measurements, manipulated and control variables
- Control of many quantities
- Supervision of limits
- Sensor fault diagnosis
- Optimizat. of profiles for desired values
- Manual fine tuning via control system



7.3 Application Examples

Suppression of Disturbances

Example:



Goal: Desired signal „speed“ can pass (almost) unchanged but disturbance is suppressed.

How to design a filter that fulfills its task (disturbance suppression) well?

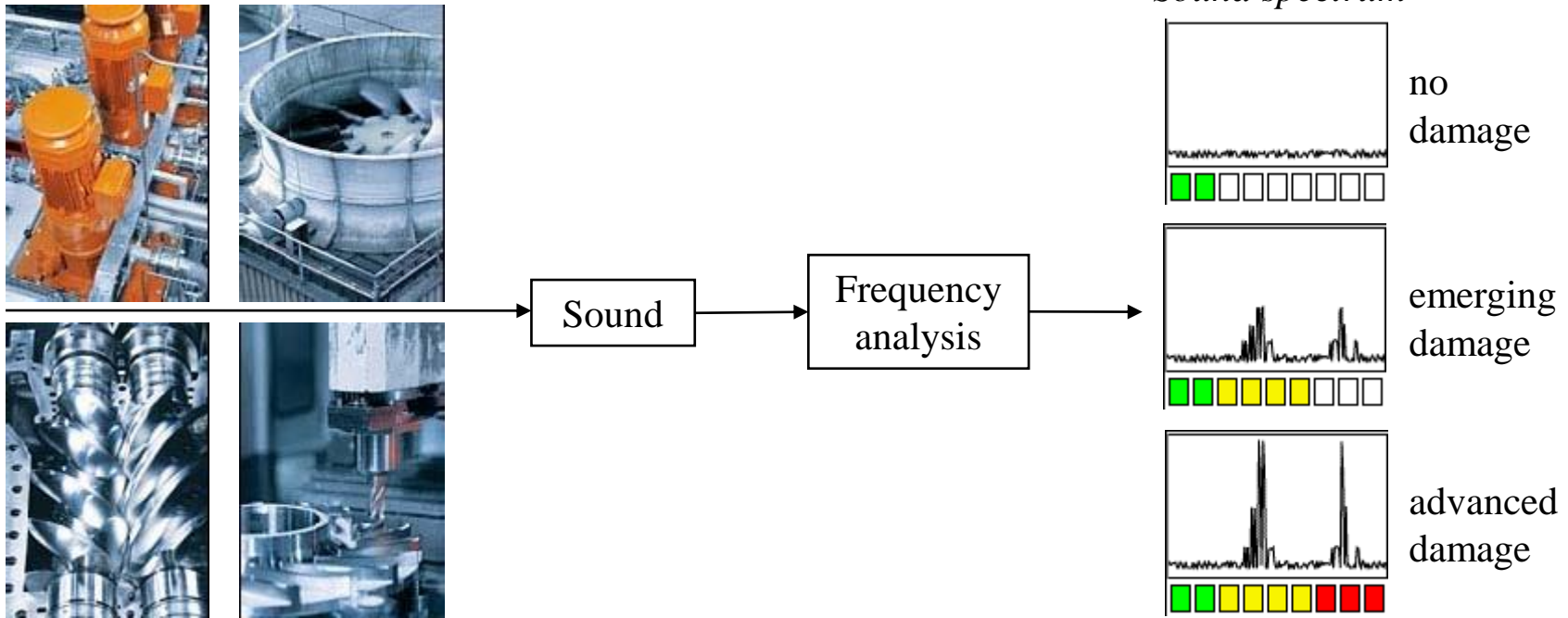
- What does “well” mean? → Criterion needed!
- Structure of the filters: linear/nonlinear, FIR/IIR, order, ... to be determined.
- Parameters of the filter to be determined.
- Prior knowledge about the disturbance is required:
 - kind: stochastic or deterministic
 - frequency range: single frequencies, certain frequency bands, ...

7.3 Application Examples

Detection of Damages in Bearings by Analysis of Structure-Borne Sound

- Humans/experts often are able to detect faults in machines by their sound. Even emerging faults can be detected early.
- Characteristic features can be found in the spectrum of the sound signal.
- Automatic methods for calculating and analyzing the sound spectrum are required!

Bearing damage?



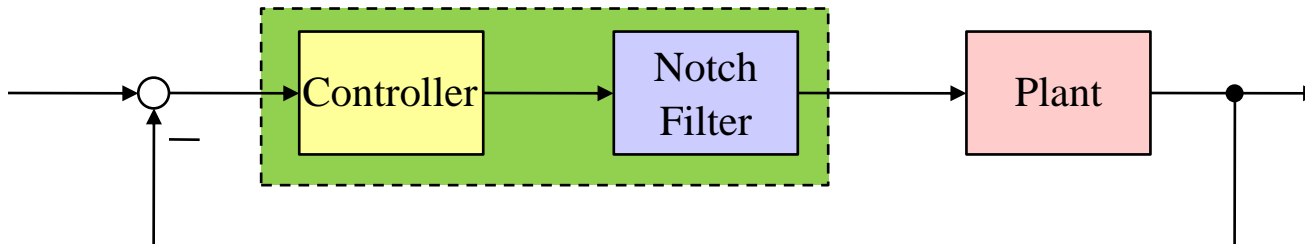
7.3 Application Examples

Notch Filter in Position Control in Aeronautics

Notch filter band-stop filter that address a very small frequency range. They are often used to remove frequencies that otherwise would harm the system., e.g.:

- Ship control: Elimination of disturbances caused by periodic waves.
- Control of planes, solar panels, and other weakly damped structures (light construction becomes more important in almost every application).
- TV- and radio receiver: Interfering and disturbing frequencies are filtered.

Control System With Incorporated Notch Filter for Damping of Ressonances

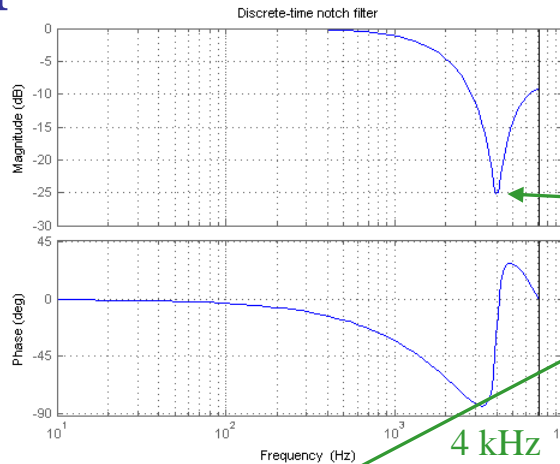


7.3 Application Examples

Example: Control of a Read/Write Head of a Hard Disk

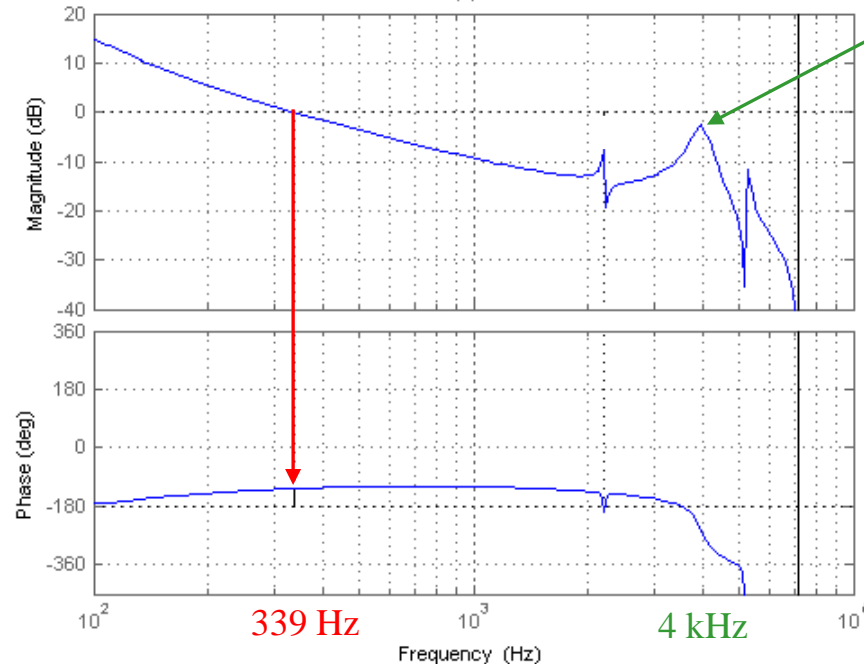
- Improvement of the frequency characteristics of the open loop.
- Notch filter at 4 kHz

Strong damping at 4 kHz pushes amplitude response down and increases the amplitude margin!

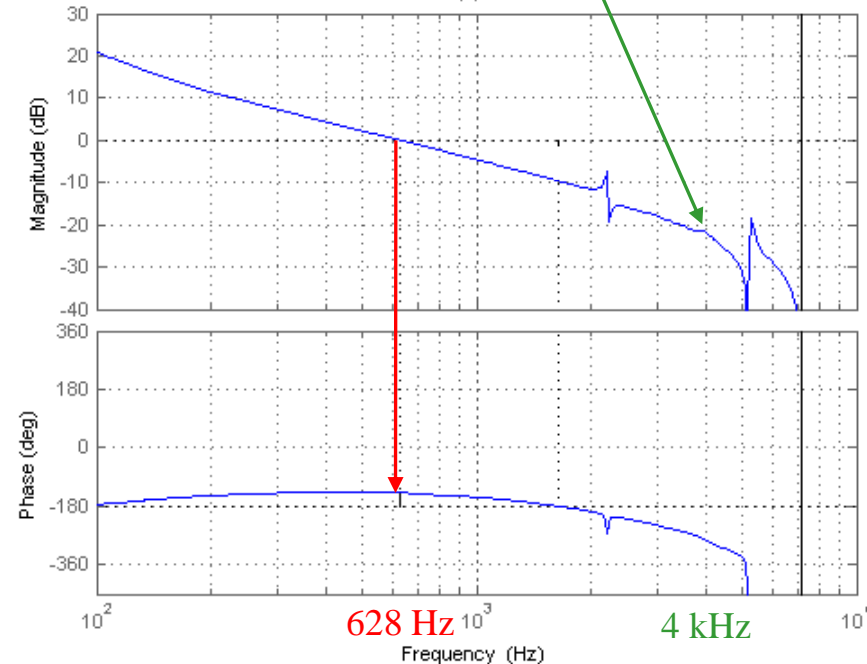


Bode Diagram
Gm = 7.07 dB (at 2.2e+003 Hz), Pm = 55.4 deg (at 339 Hz)
From: In(1) To: PES

Bode Diagram
Gm = 9.56 dB (at 1.64e+003 Hz), Pm = 41.7 deg (at 628 Hz)
From: In(1) To: PES



+ Notch
filter

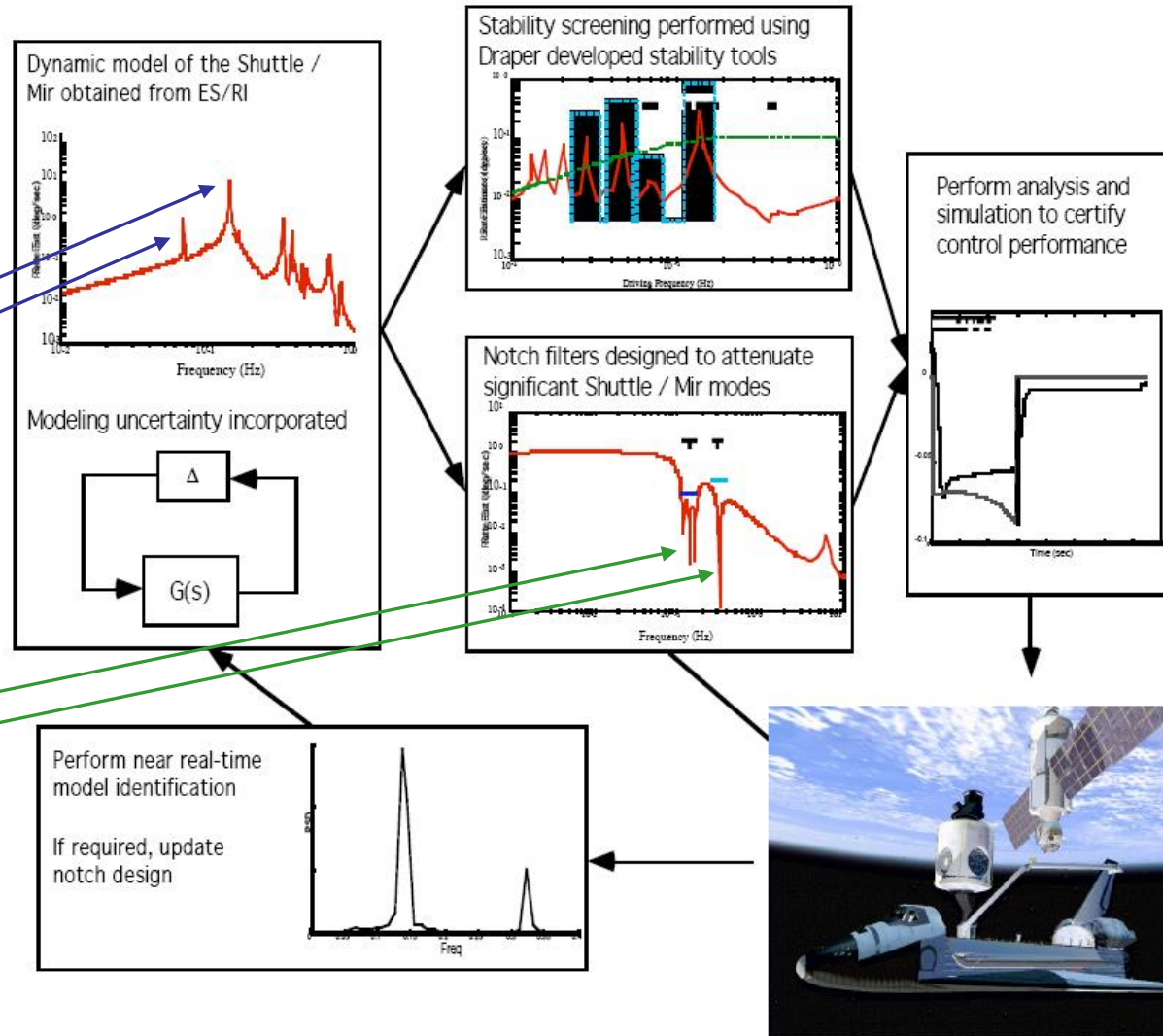


7.3 Application Examples

Example: Control of Space Shuttle

Resonances in the dynamics of the shuttle

Notch filter suppresses these frequencies



Source: „Flight Control Overview of STS-88, the First Space Station Assembly Flight“ by R. Hall, K. Kirchwey, M. Martin, G. Rosch, D. Zimpfer, AAS-99-371

Figure 7: Shuttle Gain Stabilization Design Process

7.4 Literature

In German

Wendemuth A.: „Grundlagen der digitalen Signalverarbeitung“, Springer, 2004, 268 S.

Werner M.; „Digitale Signalverarbeitung mit MATLAB: Grundkurs mit 16 ausführlichen Versuchen“, 10. Ed., Vieweg + Teubner, 2008, 294 S.

Oppenheim A.V., Schafer R.W., Buck J.R.: „timeDiscrete Signalverarbeitung“, Pearson, 8. Ed., 2004, 1040 S.

In English

Oppenheim A.V., Schafer R.W., Buck J.R.: „Discrete-Time Signal Processing“, Prentice-Hall, 9. Ed., 2008, 950 p.

Ifeachor E., Jervis B.: „Digital Signal Processing: A Practical Approach“, Prentice-Hall, 8. Ed., 2001, 960 p.

8. Time-Discrete Systems and Signals

Contents of Chapter 8

8. Time-Discrete Systems and Signals (Fundamentals: Mainly Home Study)

8.1 Time-Discrete Signals

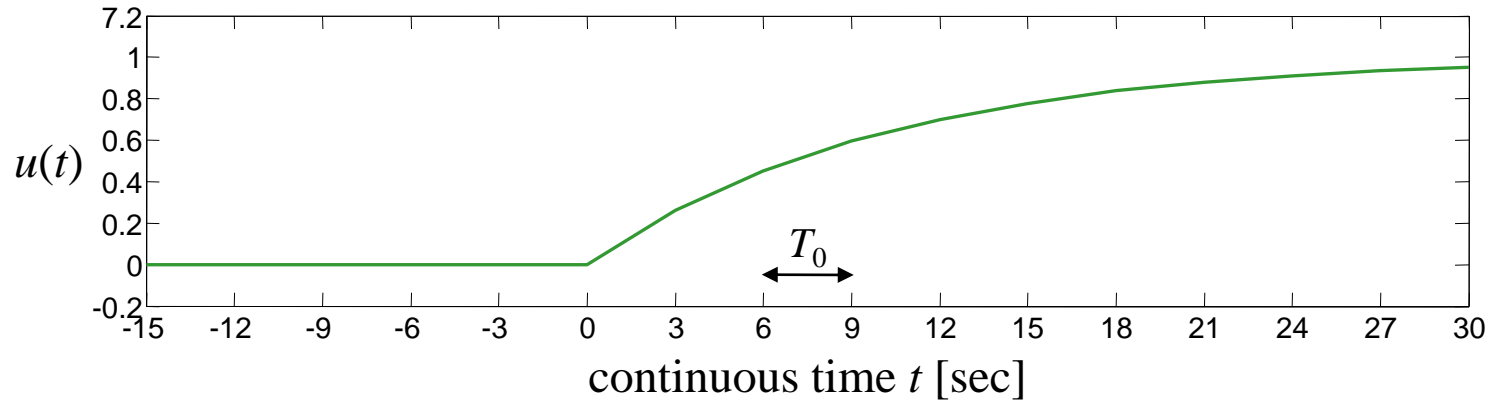
8.2 Difference Equations

8.3 Z-Transform

8.4 Transfer Functions

8.1 Time-Discrete Signals

Equidistant Sampling of a Time-Continuous Signal With Sampling Time T_0

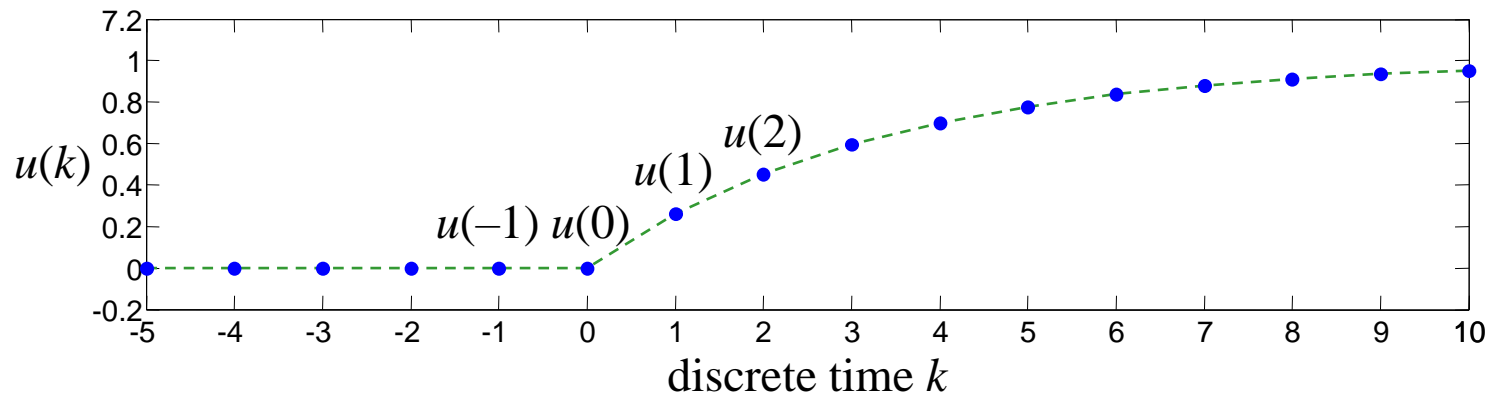


Sampling at the time steps

$$t = kT_0$$

$$T_0 = 3 \text{ sec}$$

$k = \dots, -2, -1, 0, 1, 2, 3, \dots$
Sequence: $\{u(k)\} = \{\dots, 0, 0, 0, 0.26, 0.45, 0.59, \dots\}$



8.1 Time-Discrete Signals

Leopold Kronecker, 1823-1891
(www.wikipedia.org)

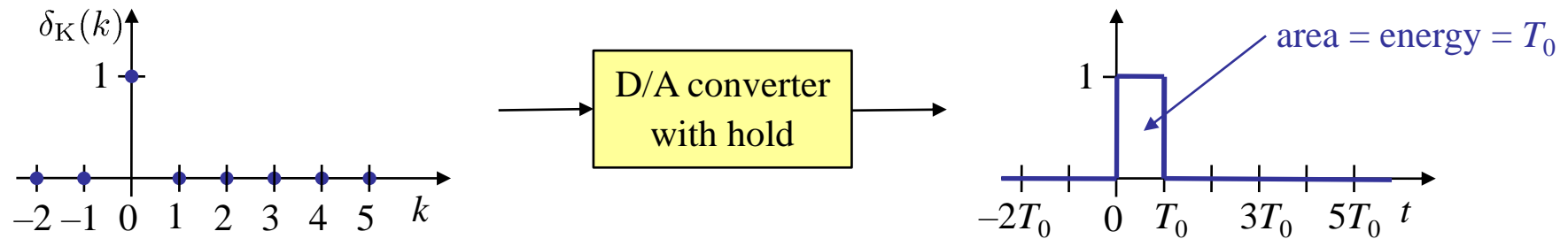


Paul Dirac, 1902-1984
(www.wikipedia.org)

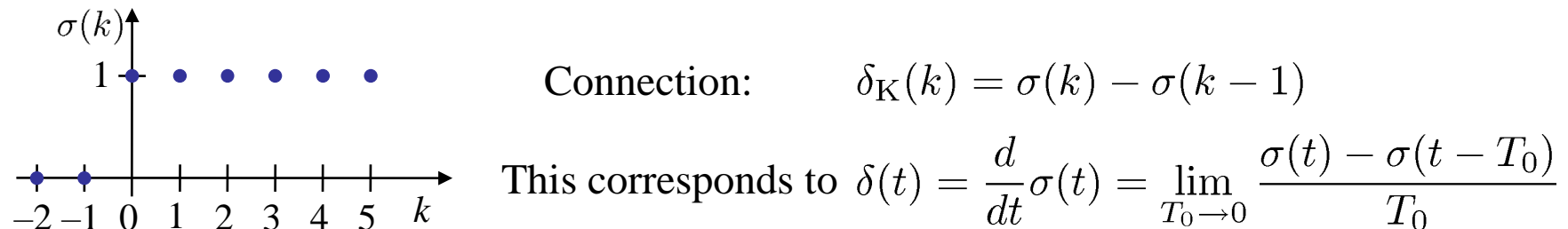


Unit Impulse and Unit Step

The unit impulse in discrete time is called *Kronecker delta* and has height 1. This is in contrast to the continuous-time *Dirac impulse* which has infinite height. Therefore the Kronecker delta can indeed be realized in practice, while the Dirac impulse is only a theoretical idealization (or limit). If a Kronecker delta is fed to a D/A converter the output's length is 1 sampling interval and its energy is proportional to T_0 .



The discrete-time unit step simply corresponds to the continuous-time unit step sampled with T_0 . During the 1. sample the unit step $\sigma(k)$ and the delta impulse $\delta_K(k)$ are identical!

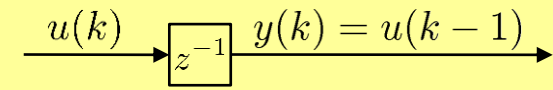


8.1 Time-Discrete Signals

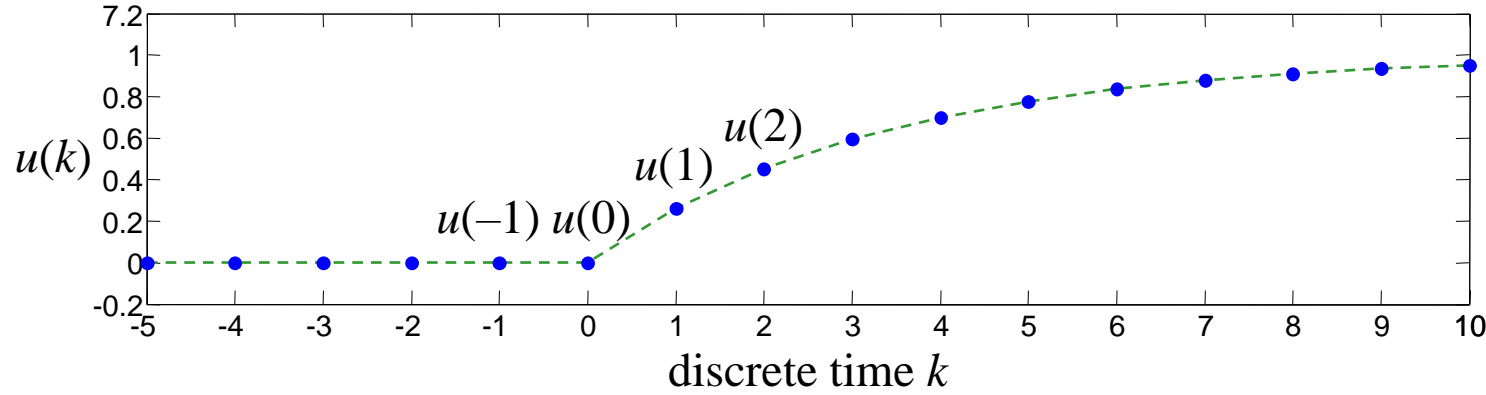
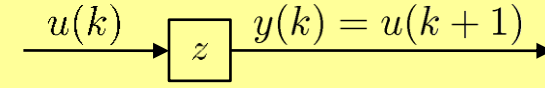
time shift operator

$$z^i = \begin{cases} i < 0 : & i \text{ steps of delay} \\ i = 0 : & \text{No delay} \\ i > 0 : & i \text{ steps of prediction} \end{cases}$$

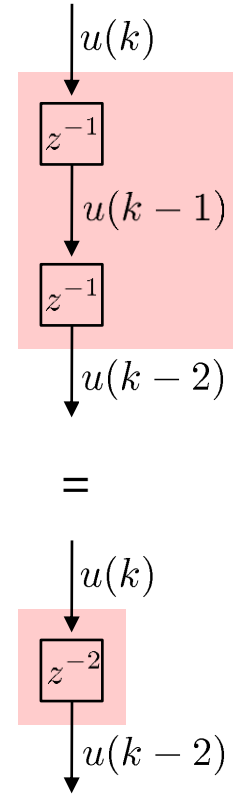
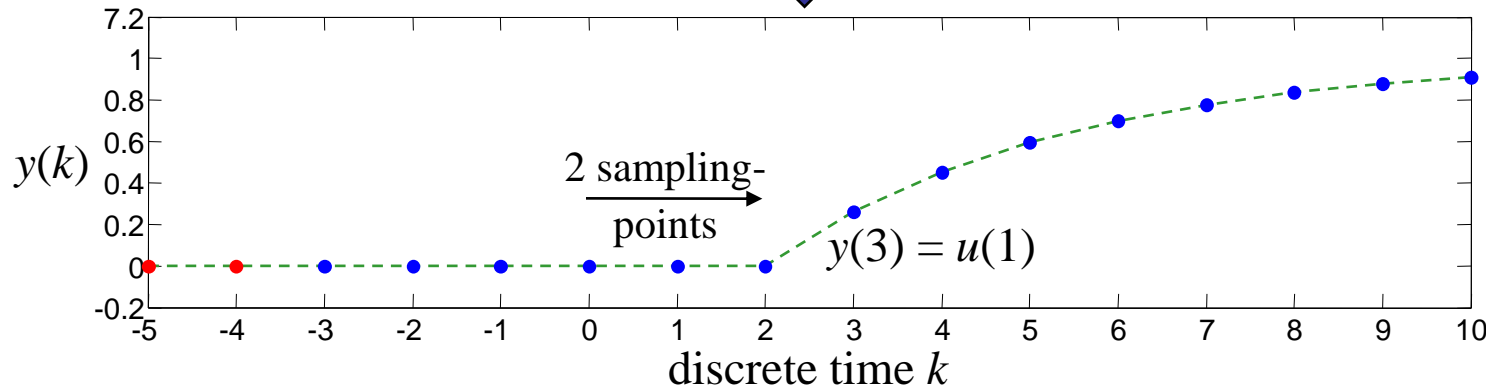
Backward shift:



Forward shift:



delay of 2 steps: z^{-2} \downarrow $y(k) = u(k-2)$

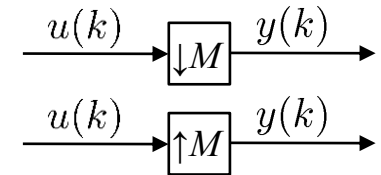


8.1 Time-Discrete Signals

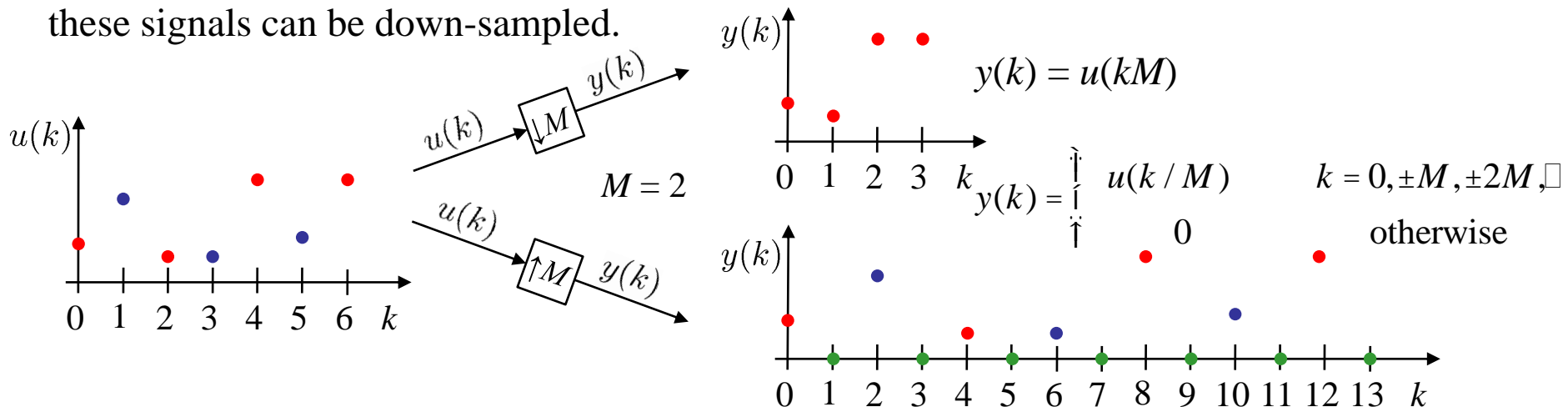
Up-Sampling and Down-Sampling

If the sampling rate of an already time-discrete sampled signal shall be changed the following operations are required:

- *Down-Sampling*: Increase of sampling time by a factor of M .
- *Up-Sampling*: Decrease of sampling time by a factor of M .



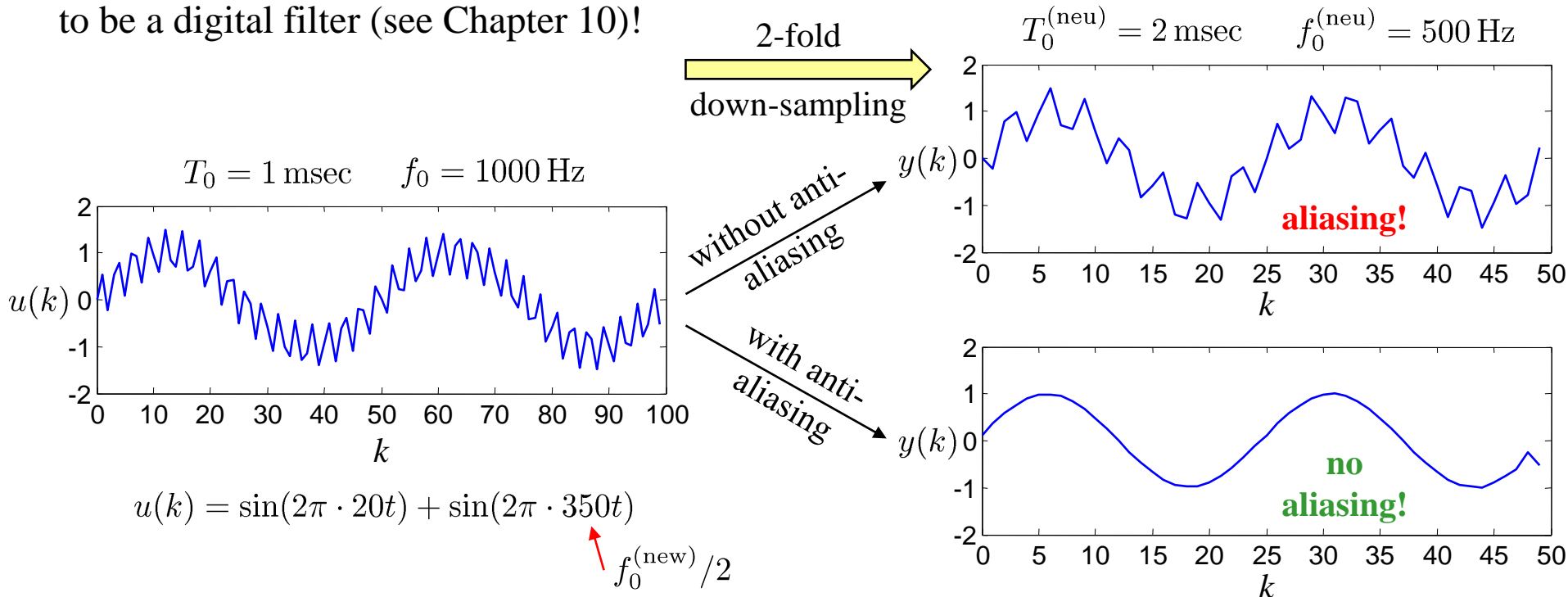
These operations are needed to work with differently sampled signals (multi-rate systems) in order to synchronize them compress the data. Commonly the sampling time is chosen very small to make sure that the sampling theorem is not violated. However such an approach creates huge amounts of data and causes problems with numerical accuracy, particularly in control. Therefore, in a second step, these signals can be down-sampled.



8.1 Time-Discrete Signals

Aliasing With Down-Sampling

During sampling of a continuous-time signal aliasing arises if the sampling theorem is violated, i.e., the sampling frequency f_0 is not larger than the maximal signal frequency f_{\max} . The same is true for sampling an already sampled signal, i.e., down-sampling. Thus, *before* down-sampling it is important to run an **anti-aliasing filter** that ensures no frequency component above $f_0/2$ (new f_0) is inside the signal. In this case, the anti-aliasing filter needs to be a digital filter (see Chapter 10)!



8.2 Difference Equations

Differential Equations → Difference Equations

For small sampling times $T_0 \rightarrow 0$ a differential equation can be approximated by a difference equation (*discretization*) by approximating a differential quotient by a difference quotient:

$$\dot{x}(t) \approx \frac{x(t) - x(t - T_0)}{T_0}, \quad \ddot{x}(t) \approx \frac{\dot{x}(t) - \dot{x}(t - T_0)}{T_0} = \frac{x(t) - 2x(t - T_0) + x(t - 2T_0)}{T_0^2}, \quad \dots$$

This approximation has significant drawbacks for $T_0 \gg 0$. A differential equation of order n ($m \leq n$)

$$y(t) + a_1 \dot{y}(t) + a_2 \ddot{y}(t) + \dots + a_n y^{(n)}(t) = b_0 u(t) + b_1 \dot{u}(t) + b_2 \ddot{u}(t) + \dots + b_m u^{(m)}(t)$$

corresponds to a difference equation of order n :

$$y(k) + a_1 y(k-1) + a_2 y(k-2) + \dots + a_n y(k-n) = b_0 u(k) + b_1 u(k-1) + b_2 u(k-2) + \dots + b_m u(k-m)$$

While the simulation of continuous-time systems requires integrations, a discrete-time system “only” needs the solution of algebraic equations, i.e., simply the isolation of $y(k)$:

$$y(k) = b_0 u(k) + b_1 u(k-1) + b_2 u(k-2) + \dots + b_m u(k-m) - a_1 y(k-1) - a_2 y(k-2) - \dots - a_n y(k-n)$$

Knowledge about the previous time steps $k-1, k-2, \dots, k-n$ is required.

8.2 Difference Equations

Moving Average (MA) System

The output is a weighted average of the previous *input* signal:

$$y(k) = b_0u(k) + b_1u(k - 1) + b_2u(k - 2) + \dots + b_mu(k - m)$$

Such a system is also called FIR (*finite impulse response*) because its output to an impulse inputs decays to *zero* after *m* steps.

Autoregressive (AR) System

The output is a weighted average the previous *output* signal

$$y(k) = -a_1y(k - 1) - a_2y(k - 2) - \dots - a_ny(k - n)$$

Such a system also called IIR (*infinite impulse response*) because its output to an impulse inputs *never* decays to zero.

Moving Average Autoregressive (ARMA) System

A combination of a MA and an AR system. This corresponds to the general linear form. Because it includes AR terms it possesses an IIR.

8.2 Difference Equations

Homogeneous Solution: Simulation for $u(k) = 0$

If the input is $u(k) = 0$ then the output depends only on the initial values. The most simple example is the following difference equation of first order with $b_1 = 0$:

$$y(k) = b_0 u(k) - a_1 y(k-1)$$

If the initial condition $y(-1)$ is known the output $y(k)$ can be calculated for all times k :

$$k = 0 : y(0) = b_0 u(0) - a_1 y(-1) = -a_1 y(-1)$$

$$k = 1 : y(1) = b_0 u(1) - a_1 y(0) = -a_1 y(0) = (-a_1)^2 y(-1)$$

$$k = 2 : y(2) = b_0 u(2) - a_1 y(1) = -a_1 y(1) = (-a_1)^3 y(-1)$$

⋮

$$k : y(k) = b_0 u(k) - a_1 y(k-1) = -a_1 y(k-1) = (-a_1)^{k+1} y(-1)$$

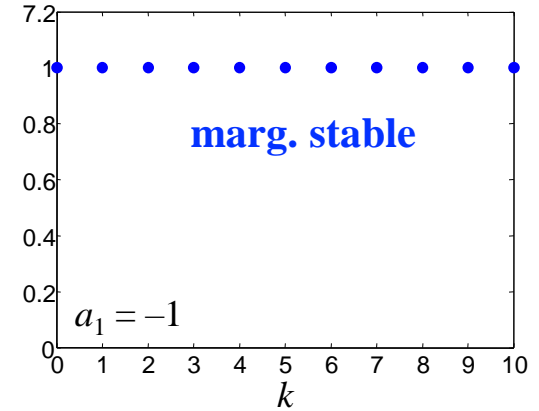
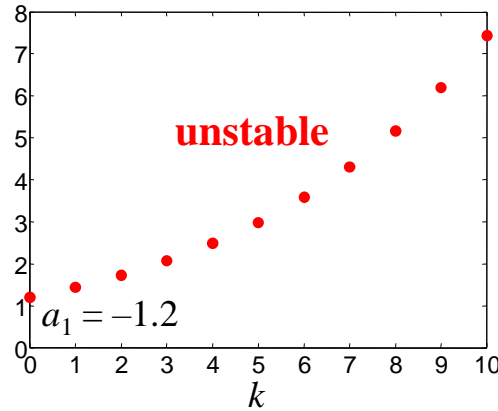
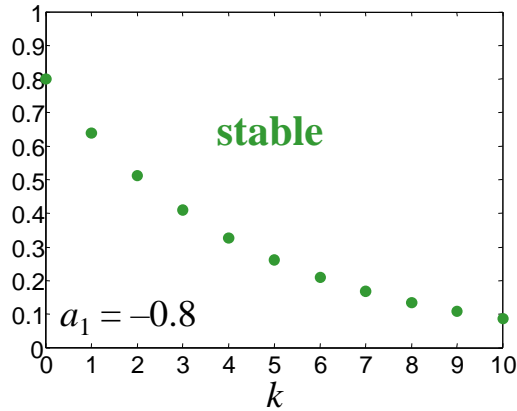
Stable: $|a_1| < 1$
Unstable: $|a_1| > 1$
Marg. stable: $|a_1| = 1$

For difference equations of order n with $n > 1$ it can be calculated correspondingly. However, in the general case n initial values $y(-1), y(-2), \dots, y(-n)$ are required because $y(k)$ depends on $y(k-1), y(k-2), \dots, y(k-n)$.

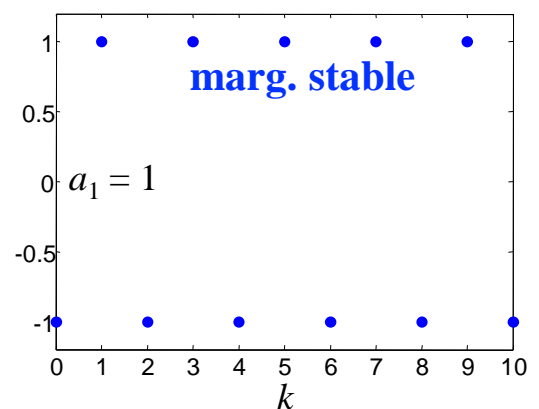
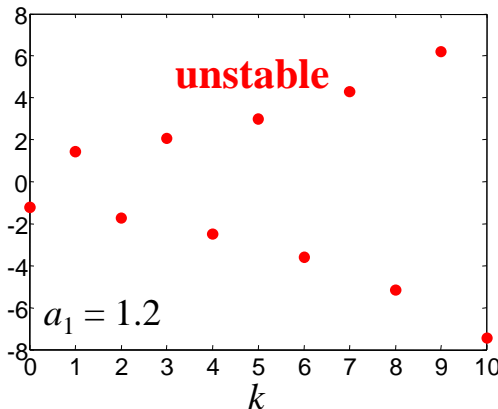
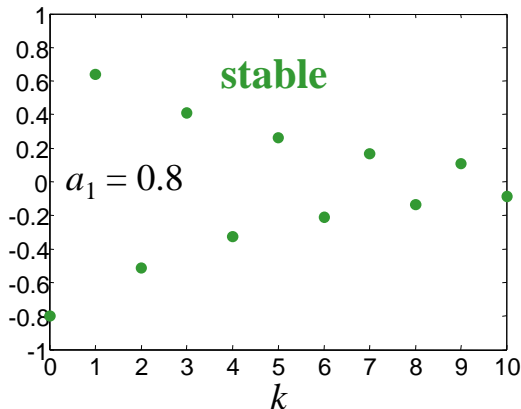
8.2 Difference Equations

Stability of a Difference Equation of 1. Order

We can distinguish between three cases:



If $a_1 > 0$ we obtain alternating (in turn positive and negative) solutions. It does not exist any analogue correspondence for time-continuous systems:



8.2 Difference Equations

$$u(k) = \delta_{\mathbb{K}}(k) = \begin{cases} 1 & k = 0 \\ 0 & \text{otherwise} \end{cases}$$

Impulse Response

For $u(k) = \delta_{\mathbb{K}}(k)$ the generated output $y(k)$ is called the **impulse response**. Like for time-continuous systems the impulse response characterizes completely the dynamic behavior of any linear system because the impulse contains all frequencies with equal power. In contrast to the continuous time case, it is a sequence not a continuous function. For simplicity, we assume all initial conditions are $= 0$, thus the homogenous solution part is zero. For a first order difference equation with $b_1 = 0$ we get:

$$y(k) = b_0 u(k) - a_1 y(k-1)$$

In the homogenous case we have $y(-1) = 0$ and thus the output $y(k)$ for all times k becomes:

$$k = 0 : y(0) = b_0 u(0) - a_1 y(-1) = b_0$$

$$k = 1 : y(1) = b_0 u(1) - a_1 y(0) = -a_1 b_0$$

$$k = 2 : y(2) = b_0 u(2) - a_1 y(1) = (-a_1)^2 b_0$$

⋮

$$k : y(k) = (-a_1)^k b_0$$

We obtain the same power law as in the homogenous case.

8.2 Difference Equations

$$u(k) = \sigma(k) = \begin{cases} 1 & k \geq 0 \\ 0 & k < 0 \end{cases}$$

Step Response

For $u(k) = \sigma(k)$ the generated output $y(k)$ is called the **step response**. Like for time-continuous systems the step response is the most intuitive way to find the picture the dynamics. For simplicity, we assume all initial conditions are $= 0$, thus the homogenous solution part is zero. For a first order difference equation with $b_1 = 0$ we get:

$$y(k) = b_0 u(k) - a_1 y(k-1)$$

In the homogenous case we have $y(-1) = 0$ and thus the output $y(k)$ for all times k becomes:

$$k = 0 : y(0) = b_0 u(0) - a_1 y(-1) = b_0 \quad (\text{identical with the impulse response})$$

$$k = 1 : y(1) = b_0 u(1) - a_1 y(0) = b_0 - a_1 b_0 = b_0 (1 - a_1)$$

$$k = 2 : y(2) = b_0 u(2) - a_1 y(1) = b_0 - a_1 b_0 (1 - a_1) = b_0 (1 - a_1 + a_1^2)$$

⋮

$$k : y(k) = b_0 (1 - a_1 + a_1^2 - \dots - (-a_1)^k) = b_0 \sum_{i=0}^k (-a_1)^i$$

8.2 Difference Equations

Relationship Between Impulse and Step Responses

Remember: In *continuous time* the following relationship holds between the impulse response $g(t)$ and the step response $h(t)$:

$$\delta(t) = \frac{d}{dt}\sigma(t) \quad \rightarrow \quad g(t) = \frac{d}{dt}h(t)$$

$$\sigma(t) = \int_0^t \delta(\tau)d\tau \quad \rightarrow \quad h(t) = \int_0^t g(\tau)d\tau$$

In discrete time the relationships are correspondingly:

$$\delta_K(k) = \sigma(k) - \sigma(k - 1) \quad \rightarrow \quad g(k) = h(k) - h(k - 1)$$

$$\sigma(k) = \sum_{i=0}^k \delta_K(k - i) \quad \rightarrow \quad h(k) = \sum_{i=0}^k g(k - i)$$

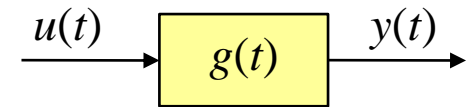
Difference replace differentials, sums replace integrals. In discrete time the handling is much simpler with the help of a computer. However, in this form, the number of sum terms (summands) increases with k ! Therefore we look for some other way to calculate the output of a discrete-time system.

8.2 Difference Equations

Convolution Sum

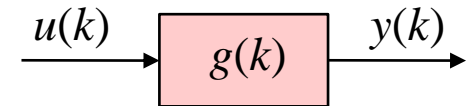
The impulse response sequence contains all properties of a linear dynamic system in discrete time. For the *time-continuous* case the output in response to an arbitrary input signal $u(t)$ can be calculated by the *convolution integral*:

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



In *discrete time* the corresponding expression is the *convolution sum*. With it the output $y(k)$ to every input signal $u(k)$ can be calculated:

$$y(k) = \sum_{i=-\infty}^{\infty} g(i)u(k - i) = \sum_{i=-\infty}^{\infty} g(k - i)u(i)$$



Usually we assume that for negative times the input is equal to zero, i.e., $u(k) = 0$ for $k < 0$. This means that the first sum must be calculated only up to $i = k$ or alternatively the second sum has to start at $i = 0$. Additionally, if the system is *causal*, i.e., $g(k) = 0$ for $k < 0$, then the first sum can start at $i = 0$ and the second sum run up to $i = k$.

8.2 Difference Equations

Convolution Sum (simplified)

With these simplifications the first sum can be written as:

$$y(k) = \sum_{i=0}^k g(i)u(k-i) = g(0)u(k) + g(1)u(k-1) + \dots + g(k)u(0)$$

In the second sum the order is reverse:

$$y(k) = \sum_{i=0}^k g(k-i)u(i) = g(k)u(0) + g(k-1)u(1) + \dots + g(0)u(k)$$

Obviously, both sums are identical! With the help of a computer the *sums* are very fast and easy to calculate. It is much easier than the convolution *integral* in the continuous-time case.

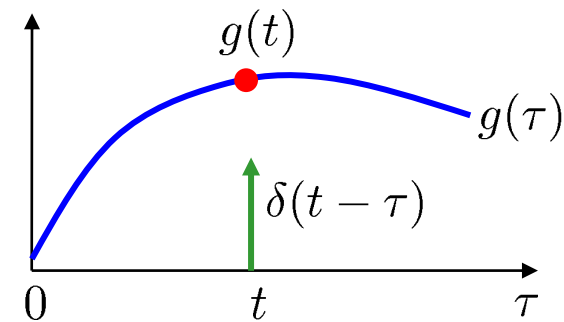
WARNING: With increasing simulation times $k \rightarrow \infty$ the number of terms in the sum increases linearly. If the impulse response $g(k)$ is of *infinite* length (IIR) then the computational and storage demand increases without limits! This means that we have to find out a way how to calculate the output of IIR systems in a more practical and efficient manner. For systems with *finite* impulse responses of length L (FIR) the number of terms in the sum is limited to L .

8.2 Difference Equations

Convolution with an Impulse

In continuous time the impulse $\delta(t)$ has the *sifting* property, i.e., a convolution with a Dirac impulse yields the signal itself. The Dirac impulse is the neutral element in a convolution like “0” in addition or “1” in multiplication. For the calculation of the impulse response we choose $u(t) = \delta(t)$ and this yields:

$$y(t) = \int_{-\infty}^{\infty} g(\tau)\delta(t - \tau)d\tau = g(t)$$

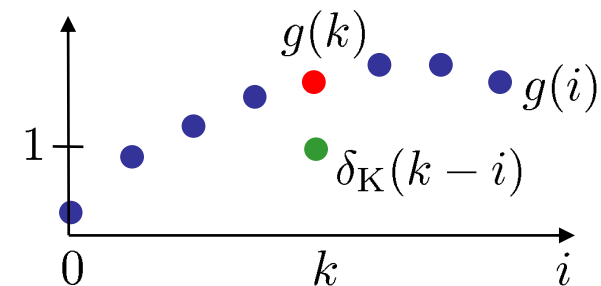


In discrete time we choose $u(k) = \delta_{\mathbb{K}}(k)$ and calculate with the convolution sum:

$$y(k) = \sum_{i=-\infty}^{\infty} g(i)\delta_{\mathbb{K}}(k - i) = g(k)$$

← = 1 for $k = i$

This is exactly the corresponding result as in the time-continuous case.

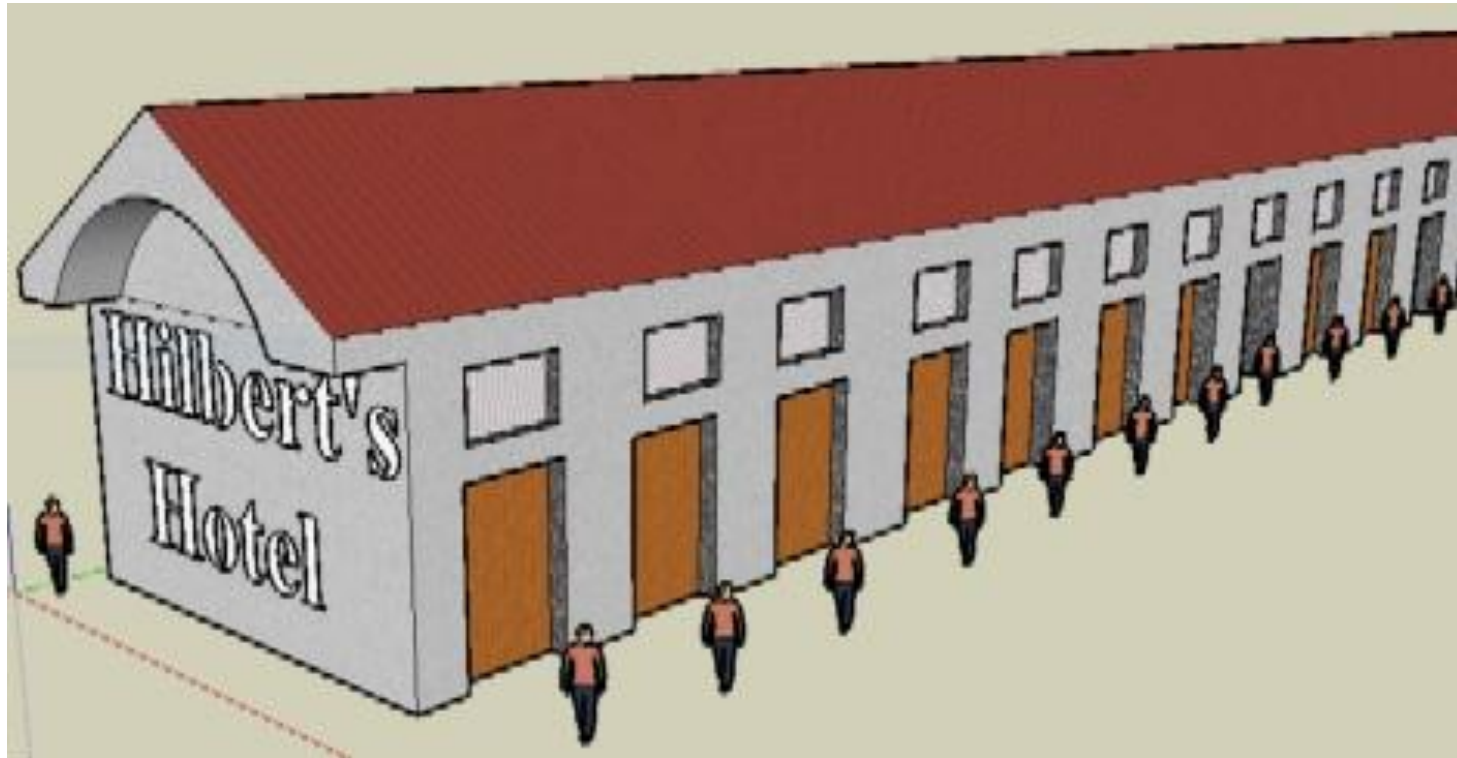


8.2 Difference Equations

Hilbert's Hotel

This hotel has an infinite number of rooms. It illustrates the understanding of infinite sets.

If all rooms are taken, is a room available for additional guests or for 2 of for ∞ ?



8.2 Difference Equations

Exponential Relationships: Not intuitive!

Q: If we fold a piece of paper (thickness = 0.1 mm) 50 times, doubling the thickness with each fold:

How high is the stack?

A: From Earth to Mars = 100 mio. meter.

Q: If we stack coins, one stack on each field of a chess board:

1 coin on chess field 1,

2 coin on chess field 2,

4 coin on chess field 3,

8 coin on chess field 4, ...

How high is the stack on chess field 64?

A: Up to α -Centauri = 4 light years.

A human can calculate these numbers but **cannot guess** them! Human intuition fails with exponential relationships. That make them potentially dangerous (extinction of species).



Source:

<http://www.wdr.de/tv/kopfball/sendungsbeitraege/2011/1120/papier-falten.jsp>



Source: <https://www.youtube.com/watch?v=0mOZZLJZwpw>

8.2 Difference Equations

Geometric Series

In the previous slides geometric sequences or series play an important role. A geometric series is a sum of exponentially staged numbers:

$$\sum_{k=0}^{\infty} x^k = x^0 + x^1 + x^2 + x^3 + \dots$$

The following trick allows to calculate this infinite sum exactly:

$$q = x^0 + x^1 + x^2 + x^3 + \dots$$

$$xq = x^1 + x^2 + x^3 + x^4 + \dots$$

$$q - xq = x^0 \rightarrow q(1 - x) = 1 \rightarrow q = \frac{1}{1 - x}$$

Thus, for $|x| < 1$ (for $|x| \geq 1$ the series diverges to infinity):

$$\frac{1}{1 - x} = \sum_{k=0}^{\infty} x^k = x^0 + x^1 + x^2 + x^3 + \dots$$

An extended formula can be derived for finite sums:

$$\frac{1 - x^{n+1}}{1 - x} = \sum_{k=0}^n x^k$$

8.3 Z-Transform

Abbreviation: $u(k) = u_c(kT_0)$

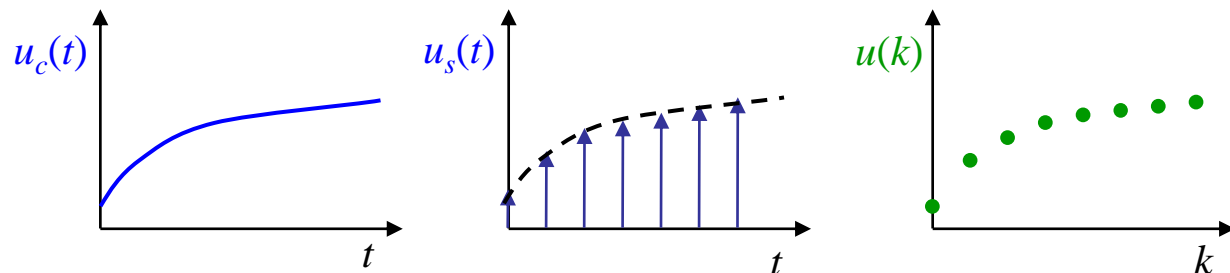
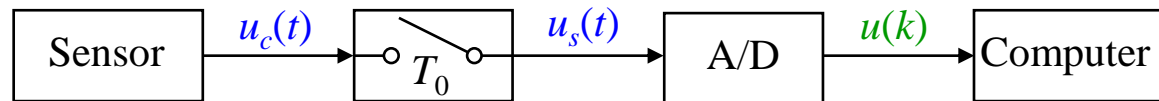
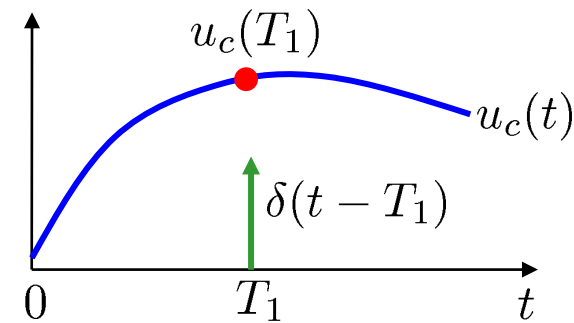
Description of Sampled Signals

An A/D converter samples a continuous-time signal $u_c(t)$ and thereby creates a time-discrete signal $u(k) = u_c(kT_0)$. The sampling is performed at time T_1 . It can be *mathematically modeled* as a multiplication of $u_c(t)$ with Dirac impulses at times T_1 , i.e., $\delta(t - T_1)$:

$$u_s(t) = u_c(t)\delta(t - T_1) = u_c(T_1)\delta(t - T_1)$$

If this sampling is performed periodically at the time steps kT_0 then the continuous signal $u_c(t)$ must be multiplied (modulated) with a *train* of impulses:

$$\begin{aligned} u_s(t) &= u_c(t) \sum_{k=-\infty}^{\infty} \delta(t - kT_0) \\ &= \sum_{k=-\infty}^{\infty} u_c(kT_0)\delta(t - kT_0) \\ &= \sum_{k=-\infty}^{\infty} u(k)\delta(t - kT_0) \end{aligned}$$



8.3 Z-Transform

Interpretation of the Train of Dirac Impulses

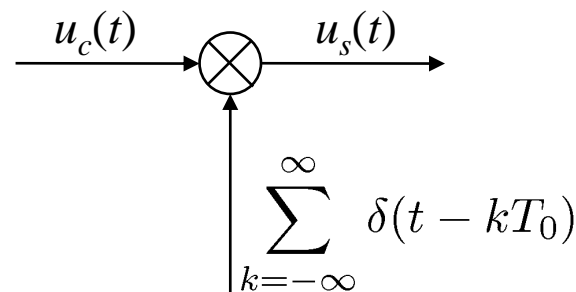
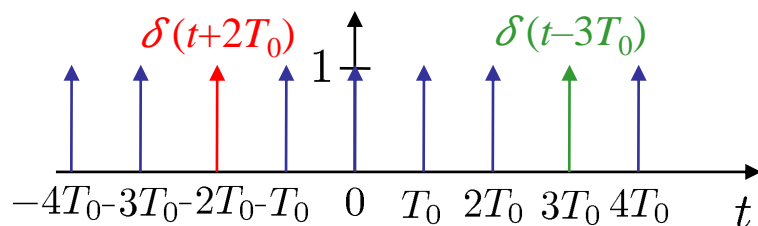
The continuous-time description of a sampled signal as modulated impulse train is given by:

$$u_s(t) = \sum_{k=-\infty}^{\infty} u(k)\delta(t - kT_0)$$

$$u_s(t) = \sum_{k=0}^{\infty} u(k)\delta(t - kT_0), \text{ if } u(k) = 0 \text{ for } k < 0$$

These formulas represent only a idealized model because in reality the impulses are not of infinite height, of course. These Dirac impulses do not exist in reality. But they associate a finite energy to each sampled signal point. Thus, also the multiplication with $u(k)$ makes sense.

Mathematical Model of the Sampling:



8.3 Z-Transform

Laplace Transform of the Sampled Signal

If we apply the Laplace transform to a sampled signal the so-called **z-transform** originates. The Laplace transform of a continuous-time signal $u(t)$ is defined as:

Laplace-Transformation:
$$U(s) = \int_0^{\infty} u(t)e^{-st} dt$$

If we choose for $u(t)$ a sampled signal, i.e., $u(t) = u_s(t)$ then we obtain:

$$\begin{aligned} U_s(s) &= \int_0^{\infty} u_s(t)e^{-st} dt = \int_0^{\infty} \sum_{k=0}^{\infty} u(k)\delta(t - kT_0)e^{-st} dt \\ &= \sum_{k=0}^{\infty} u(k) \underbrace{\int_0^{\infty} \delta(t - kT_0)e^{-st} dt}_{\mathcal{L}\{\delta(t - kT_0)\}} \end{aligned}$$

Remember:

$$\mathcal{L}\{\delta(t)\} = 1$$

$$\mathcal{L}\{\delta(t - kT_0)\} = e^{-skT_0}$$

This gives us:

$$U_s(s) = \sum_{k=0}^{\infty} u(k)e^{-skT_0} = \sum_{k=0}^{\infty} u(k) (e^{sT_0})^{-k}$$

8.3 Z-Transform

Laplace Transform \rightarrow z-Transform

With the abbreviation

$$z = e^{sT_0}$$

the Laplace transform of a sampled system is called the *z-transform* (the index “ s ” can be skipped because it is clear by the variable denotation “ z ” that we deal with discrete time):

z-Transform:

$$U(z) = \sum_{k=0}^{\infty} u(k)z^{-k}$$

Frequency Response

To calculate the frequency response of a *continuous-time system* the Laplace variable s is evaluated on the imaginary axis in the s -plane by setting $s = i\omega$ for $\omega = 0 \dots \infty$. The frequency response for a *discrete-time system* can be calculated in the same way. Correspondingly, the z -variable becomes $z = e^{i\omega T_0}$. For $\omega = 0 \dots \infty$ we run along the *unit circle* in the z -plane. It would be circled infinite many times. Thus the frequency response is periodic which is caused by the sampling! But according to the sampling theorem the frequency has to be limited to $\omega T_0 = \pi$. So we circle only once! (Symmetry with respect to $\pm\omega$!)

8.3 Z-Transform

Derivation of Periodicity of the Frequency Response

We want to consider the periodicity of the frequency response in more detail. The frequency response of a discrete time system $z = e^{i\omega T_0}$ is:

$$U(i\omega) = \sum_{k=0}^{\infty} u(k) (e^{i\omega T_0})^{-k}$$

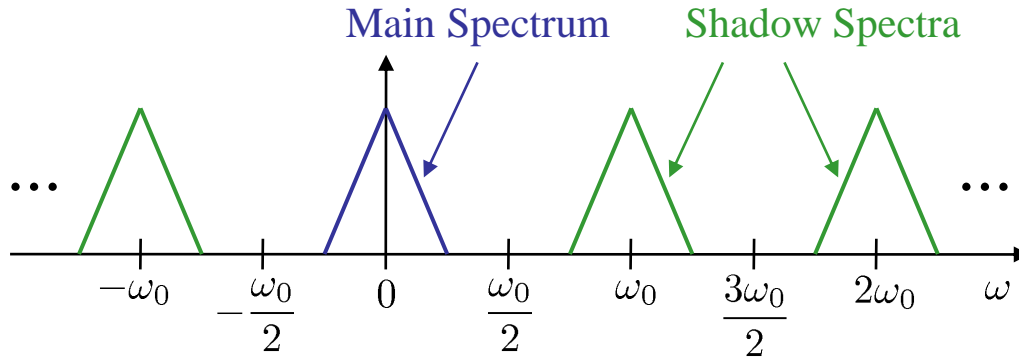
With the facts $e^{in2\pi} = 1$ for $n = 0, \pm 1, \pm 2, \dots$ and $\omega_0 T_0 = 2\pi$ we can show

$$\begin{aligned} U(i\omega) &= \sum_{k=0}^{\infty} u(k) (e^{i\omega T_0} e^{in2\pi})^{-k} = \sum_{k=0}^{\infty} u(k) (e^{i(\omega T_0 + n2\pi)})^{-k} \\ &= \sum_{k=0}^{\infty} u(k) (e^{i(\omega T_0 + n\omega_0 T_0)})^{-k} = \sum_{k=0}^{\infty} u(k) (e^{i(\omega + n\omega_0) T_0})^{-k} \end{aligned}$$

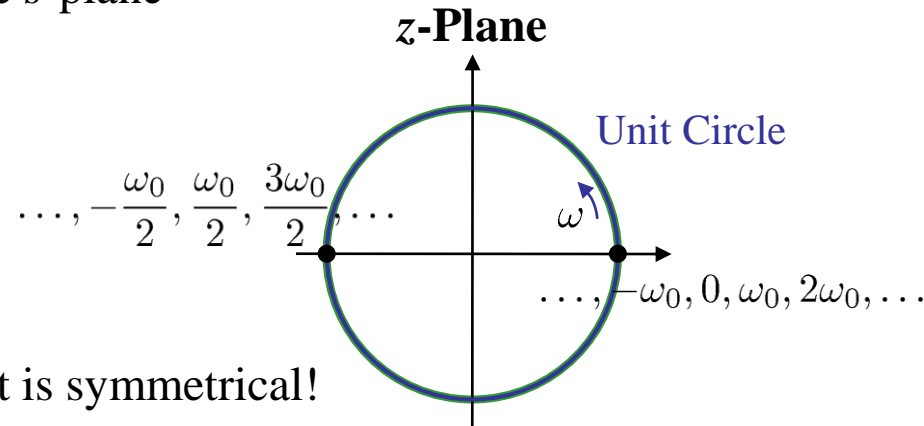
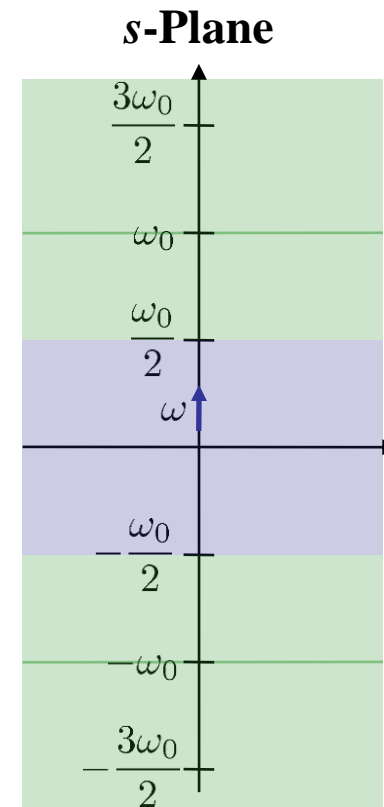
that the frequency response repeats all multiples of ω_0 (each time we circle around the unit circle in the z -plane). This means the frequency response is a periodic function. It is identical for: $\omega, \omega \pm \omega_0, \omega \pm 2\omega_0, \omega \pm 3\omega_0, \dots$

8.3 Z-Transform

Illustration of the Periodicity of the Frequency Response



- The shadows spectra around the multiples of ω_0 are created by the sampling with frequency ω_0 .
- The Im-axis between $-i\omega_0/2$ and $i\omega_0/2$ in the s -plane is mapped into the unit circle in the z -plane.
- The whole information in a time-discrete system is contained in the frequency response along the unit circle between the frequencies $\omega = 0$ and $\omega = \omega_0/2$; in the part of the unit circle $\omega = -\omega_0/2 \dots 0$ it is symmetrical!

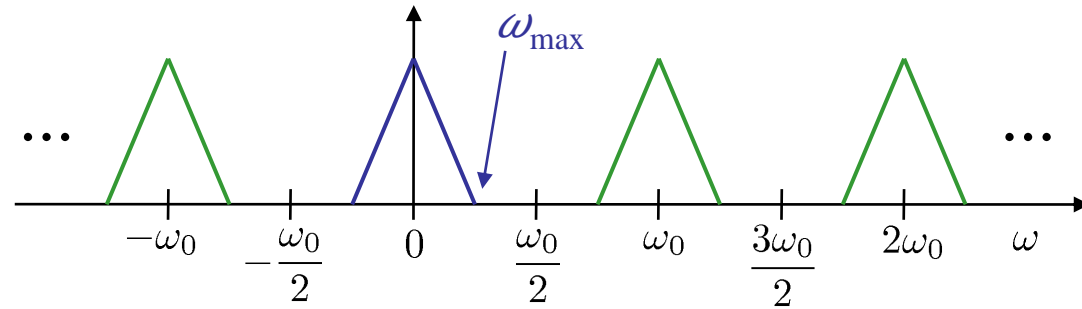


8.3 Z-Transform

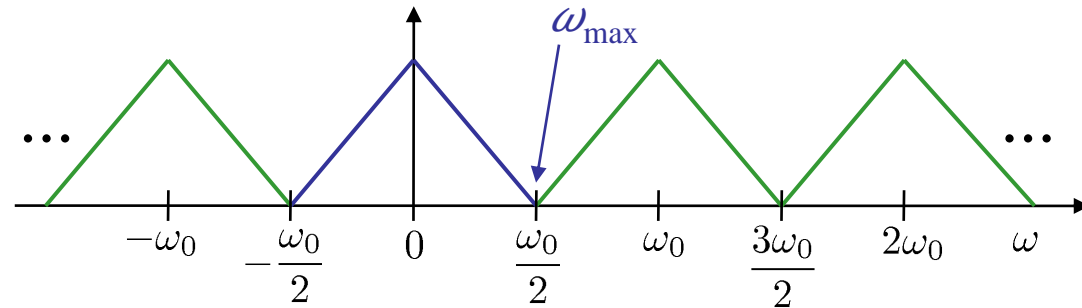
Sampling Theorem and Aliasing

- If the maximal signal frequency ω_{\max} is smaller than the half sampling frequency $\omega_0/2$, the continuous-time signal can be reconstructed perfectly from the sampled one. No information is lost because main and shadow spectra do not overlap. We have no aliasing.
- If $\omega_{\max} > \omega_0/2$ the main and shadow spectra overlap. We consequently have aliasing which deteriorates the original signal. A perfect reconstruction is impossible.

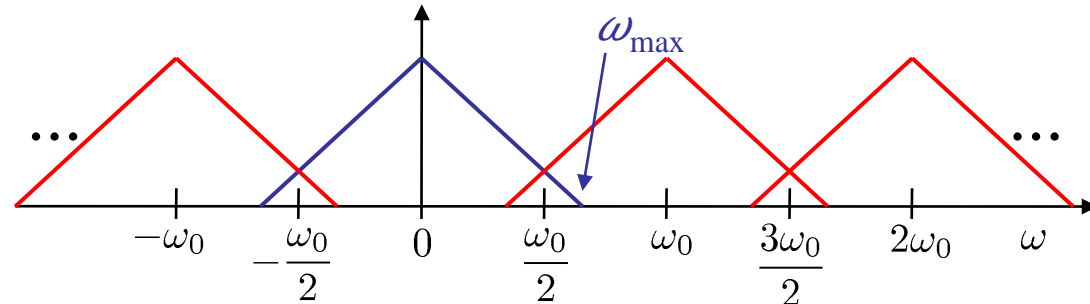
Sampling fast enough: no aliasing!



Limit case



Sampling too slow: **Aliasing!**



8.3 Z-Transform

Z-Transform of Impulse and Step

The impulse $u(k) = \delta_K(k)$ has the following z-transform:

$$u(0) = 1, u(1) = 0, u(2) = 0, \dots \rightarrow U(z) = 1z^0 + 0z^{-1} + 0z^{-2} + \dots \rightarrow U(z) = 1$$

An impulse delayed by one time step $u(k) = \delta_K(k-1)$ has the following z-transform:

$$u(0) = 0, u(1) = 1, u(2) = 0, \dots \rightarrow U(z) = 0z^0 + 1z^{-1} + 0z^{-2} + \dots \rightarrow U(z) = z^{-1}$$

An impulse delayed by d time steps $u(k) = \delta_K(k-d)$ has the following z-transform:

$$u(0) = 0, \dots, u(d-1) = 0, u(d) = 1, u(d+1) = 0, \dots \rightarrow U(z) = z^{-d}$$

The unit step $u(k) = \sigma(k)$ has the following z-transform:

$$u(0) = 1, u(1) = 1, u(2) = 1, \dots \rightarrow U(z) = 1z^0 + 1z^{-1} + 1z^{-2} + \dots \rightarrow U(z) = \frac{1}{1 - z^{-1}}$$

An unit step delayed by d time steps $u(k) = \sigma(k-d)$ has the following z-transform:

$$u(0) = 0, \dots, u(d-1) = 0, u(d) = 1, u(d+1) = 1, \dots \rightarrow U(z) = \frac{z^{-d}}{1 - z^{-1}}$$

The following expressions are identical:

$$\frac{z^{-d}}{1 - z^{-1}} = \frac{z^{-d+1}}{z - 1} = \frac{1}{(1 - z^{-1})z^d} = \frac{1}{(z - 1)z^{d-1}}$$

8.3 Z-Transform

Z-Transform of Geometric Sequences

The geometric sequence $u(k) = a^k$ with any number a commonly occurs because it describes an exponential behavior. This sequence has the following z -transform:

$$u(0) = a^0, u(1) = a^1, u(2) = a^2, u(3) = a^3, \dots \rightarrow U(z) = a^0 z^0 + a^1 z^{-1} + a^2 z^{-2} + a^3 z^{-3} + \dots$$

Further conversions lead to the standard form of a geometric series:

$$U(z) = (az^{-1})^0 + (az^{-1})^1 + (az^{-1})^2 + (az^{-1})^3 + \dots = \sum_{k=0}^{\infty} (az^{-1})^k$$

This *infinite* geometric series can be expressed simply by:

$$U(z) = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}$$

This allows to formulate infinite series as one simple expression. The way back can be carried out by long division.

Long Division:

$$z : (z - a) = 1 + az^{-1} + a^2 z^{-2} + \dots$$

$$\begin{array}{r} z - a \\ \underline{z - a} \\ a \\ \underline{a - a^2 z^{-1}} \\ a^2 z^{-1} \\ \underline{a^2 z^{-1} - a^3 z^{-2}} \\ a^3 z^{-2} \end{array}$$

8.3 Z-Transform

Important Properties of the z-Transform

For limit considerations the cases $t \rightarrow 0$ ($k \rightarrow 0$) or $t \rightarrow \infty$ ($k \rightarrow \infty$) are evaluated. In the frequency range (s or z) this means:

$$t \rightarrow 0: s \rightarrow \infty$$

$$t \rightarrow \infty: s \rightarrow 0$$

$$k \rightarrow 0: z \rightarrow \infty$$

$$k \rightarrow \infty: z \rightarrow 1$$

Start Value

The start value of a sequence can be calculated from its z-transform by:

$$u(k = 0) = \lim_{z \rightarrow \infty} U(z)$$

End Value

The end value (if it exists!) of a sequence can be calculated from its z-transform by:

$$u(k \rightarrow \infty) = \lim_{z \rightarrow 1} (z - 1)U(z)$$

$$z = e^{sT_0}$$

$$u(k) \longleftrightarrow U(z)$$

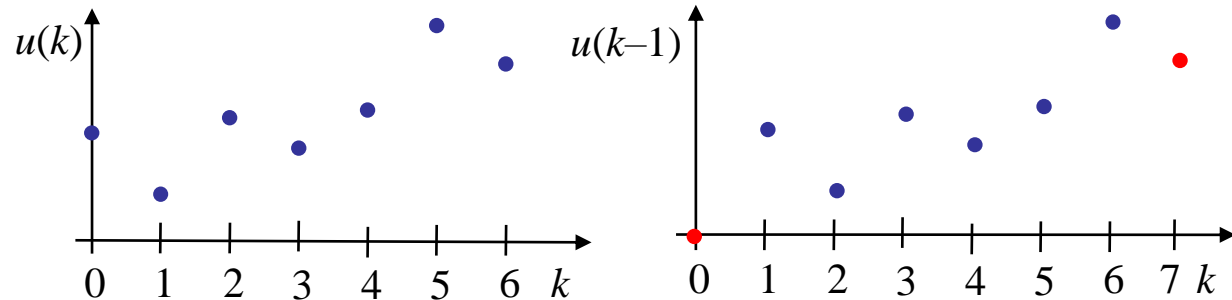
$$U(z) = \mathcal{Z}\{u(k)\}$$

8.3 Z-Transform

Backward Shift (To the Right)

A dead time $T_t = dT_0$ is equivalent to a backward shift (shift to the right) by d samples. This operation corresponds to the Laplace transform e^{-sT_t} . In the z -domain this means:

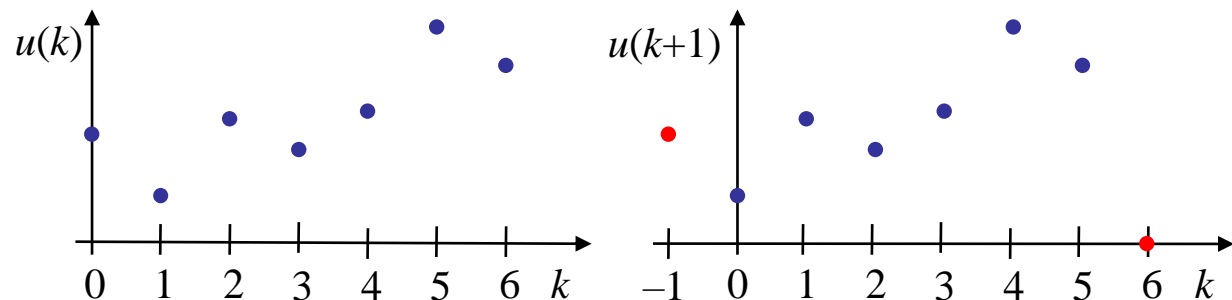
$$u(k-d) \quad \circ \rightarrow \bullet \quad z^{-d}U(z)$$



Forward Shift (To the Left)

A prediction of time $T_p = dT_0$ is equivalent to a forward shift (shift to the left) by d samples. This operation corresponds to the Laplace transform e^{sT_p} . In the z -domain this means:

$$u(k+d) \quad \circ \rightarrow \bullet \quad z^dU(z)$$

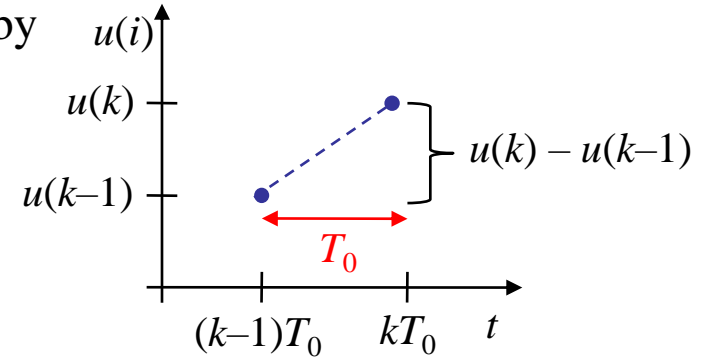


8.3 Z-Transform

Difference / Differentiation

The difference of two subsequently sampled values divided by the sampling time (that passed between their measurement) is called the *difference of first order* and corresponds approximately to a differentiation. In the s -domain it is realized by a multiplication with s . In the z -domain this is given by:

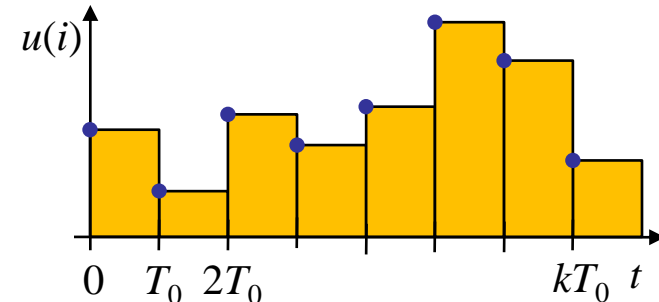
$$u(k) - u(k-1) \quad \circ \bullet \quad \frac{z-1}{z} U(z)$$



Summation / Integration

The sum of all sampled values starting from time 0 multiplied by the sampling time is equal to the lower sum approximation of the area below the samples. That approximately equals the *integration*. In the s -domain this is realized by a division by s . In the z -domain this corresponds to:

$$\sum_{i=0}^k u(i) \quad \circ \bullet \quad \frac{z}{z-1} U(z)$$



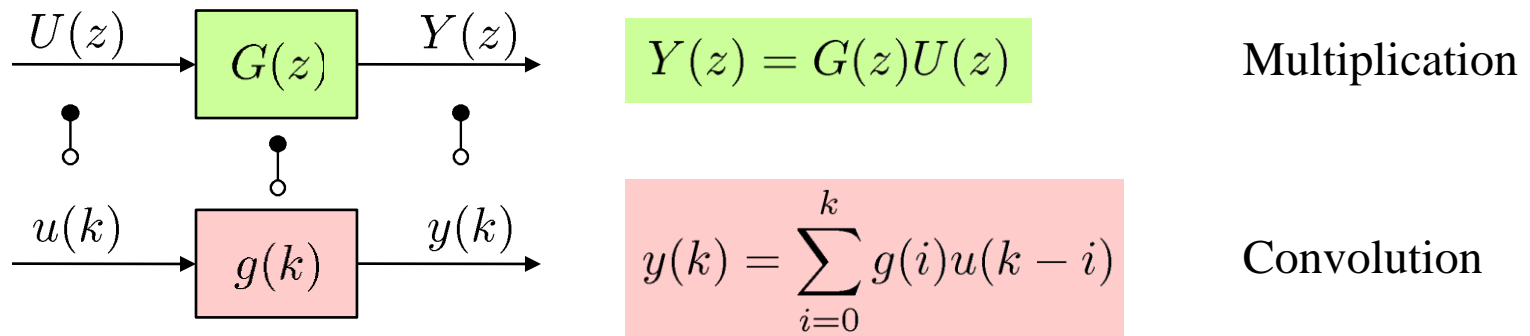
8.4 Transfer Functions

Transfer Function and Impulse Response

The same relationship exists for discrete-time systems between a transfer function in the z -domain and the impulse response sequence as for continuous-time systems between a transfer function in the s -domain and the impulse response function:

$$G(z) \longleftrightarrow g(k)$$

In $G(z)$ as in $g(k)$ all properties of a linear dynamic system are contained. For calculation of the system output over time only the system input over time and either $G(z)$ or $g(k)$ are required.



The multiplication in the z -domain corresponds to the convolution sum in the discrete time domain as the convolution integral in the continuous time domain.

8.4 Transfer Functions

Transfer Function and Impulse Response

We choose a Kronecker-delta impulse as input $u(k) = \delta_K(k)$ or $U(z) = 1$, respectively. This yields the impulse response as output:

$$y(k) = \sum_{i=0}^k g(i)u(k-i) = \sum_{i=0}^k g(i)\delta_K(k-i) = g(k)$$

or

$$Y(z) = G(z)U(z) = G(z) \cdot 1 = G(z)$$

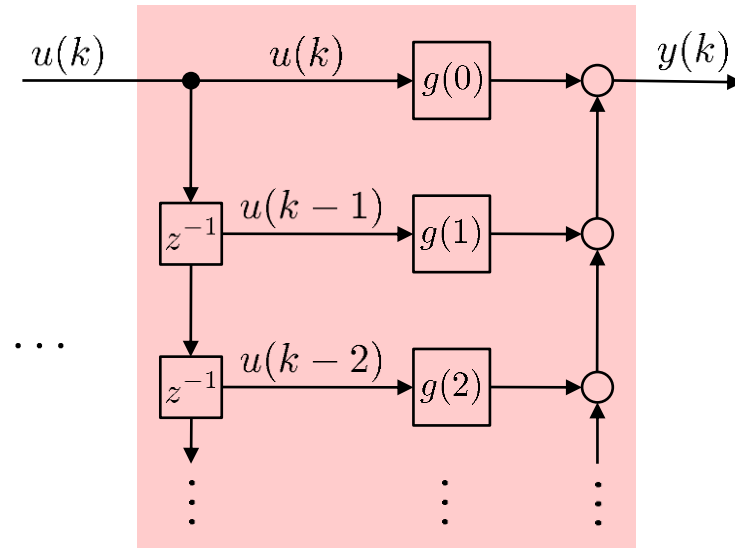
For a general impulse response sequence

$$g(k) = g(0)\delta_K(k) + g(1)\delta_K(k-1) + g(2)\delta_K(k-2) + \dots$$

the corresponding transfer function is:

$$G(z) = g(0)z^0 + g(1)z^{-1} + g(2)z^{-2} + \dots$$

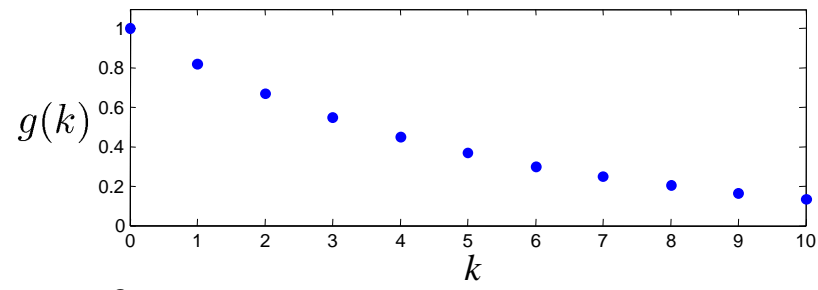
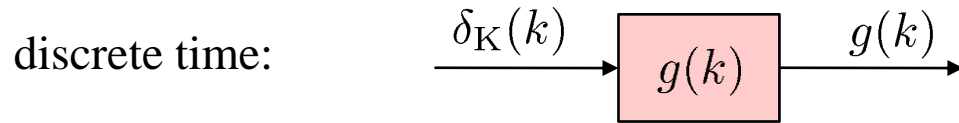
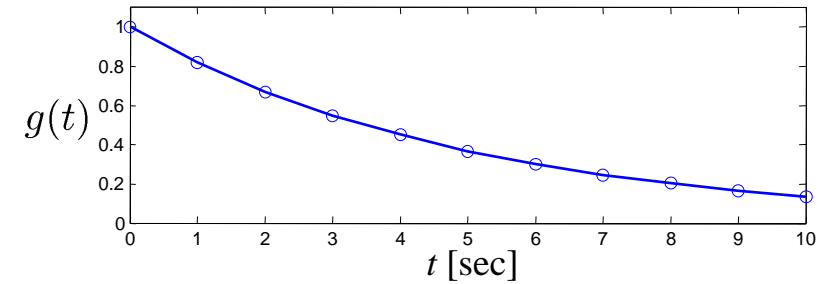
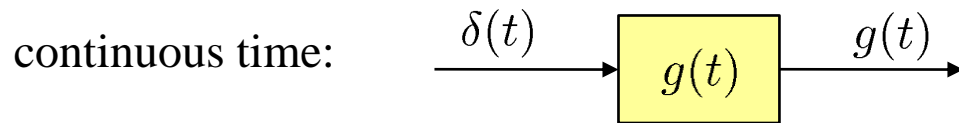
If the impulse response sequence $g(k)$ is of finite length the same is true for the number of terms in $G(z)$. If $g(k)$ is of infinite length, however, the same is also true for $G(z)$ and an easier-to-handle alternative has to be found to avoid an infinite sum.



8.4 Transfer Functions

Example: Transformation Via Impulse Response Invariance

A common method for the transformation from the continuous to the discrete world is to demand identical impulse responses. This is popular for digital filter design. We demand that the discrete impulse response sequence is identical to the sampled continuous impulse response function.



For a first order system this requires a impulse response of:

$$G(s) = \frac{K}{1 + Ts} = \frac{K/T}{s + 1/T} \rightarrow g(t) = \frac{K}{T} e^{-t/T}$$

8.4 Transfer Functions

For $K = 5$ and $T = 5$ sec this results in: $g(t) = e^{-t/5}$. If the sampling time is chosen to $T_0 = 1$ sec then the demand for an impulse response invariance yields:

$$g(t = kT_0) = e^{-kT_0/5} = e^{-k/5} = \left(e^{-1/5}\right)^k = 0.82^k = g(k)$$

Note that this is a *geometric* sequence!

This can also be written with the help of delayed delta impulses:

$$g(k) = 0.82^0 \delta_K(k) + 0.82^1 \delta_K(k - 1) + 0.82^2 \delta_K(k - 2) + 0.82^3 \delta_K(k - 3) + \dots$$

We can easily obtain the corresponding transfer function in the z -domain:

$$G(z) = 0.82^0 z^0 + 0.82^1 z^{-1} + 0.82^2 z^{-2} + 0.82^3 z^{-3} + \dots = \sum_{k=0}^{\infty} 0.82^k z^{-k} = \sum_{k=0}^{\infty} (0.82z^{-1})^k$$

Because this infinite series is difficult to handle we compute the explicit sum with the formula for infinite geometric series with $x = 0.82z^{-1}$:

$$G(z) = \frac{1}{1 - 0.82z^{-1}} = \frac{z}{z - 0.82} \quad \text{Gain: } G(z = 1) = \frac{1}{1 - 0.82} = 5.56 \neq 5$$

Therefore this $G(z)$ corresponds to the $G(s)$ in the sense of *impulse response invariance*.

8.4 Transfer Functions

Example: Transformation Via Step Response Invariance

Another popular method for transformation from continuous to discrete time is the step response invariance. It yields a different result than impulse response invariance. The denominators (and thus poles) are identical but the numerators (and thus zeros) and the gains are different:

$$G(z) = \frac{0.9z^{-1}}{1 - 0.82z^{-1}} = \frac{0.9}{z - 0.82} \quad \text{Gain: } G(z = 1) = \frac{0.9}{1 - 0.82} = 5$$

The choice of the criterion distinguishes all type of such transformations. An invariance of the *impulse responses* accounts for all frequencies in the same way because all frequencies are weighted equally (constant spectrum of an impulse). Therefore it is commonly applied for filter design.

An invariance of the *step response*, however, weights lower frequencies stronger and is the appropriate choice for control applications where the manipulated variable typically is of stepwise character. It also ensures a correct transformation of the *gain*.

8.4 Transfer Functions

Transfer Function → Difference Equation

Consider a general transfer function of numerator degree m and denominator degree n :

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_0 + b_1 z^{-1} + \dots + b_m z^{-m}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}}$$

The coefficient a_0 can be set to 1 through cancelation. This yields the following difference equation in the time-domain:

$$(1 + a_1 z^{-1} + \dots + a_n z^{-n}) Y(z) = (b_0 + b_1 z^{-1} + \dots + b_m z^{-m}) U(z)$$

$$y(k) + a_1 y(k-1) + \dots + a_n y(k-n) = b_0 u(k) + b_1 u(k-1) + \dots + b_m u(k-m)$$

A dead time of $T_t = dT_0$ causes a backward shift by d steps:

$$G(z) = \frac{b_0 + b_1 z^{-1} + \dots + b_m z^{-m}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}} z^{-d} = \frac{b_0 z^{-d} + b_1 z^{-1-d} + \dots + b_m z^{-m-d}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}}$$

$$\rightarrow y(k) + a_1 y(k-1) + \dots + a_n y(k-n) = b_0 u(k-d) + b_1 u(k-1-d) + \dots + b_m u(k-m-d)$$

In contrast to the s -domain, a dead time in the z -domain still keeps the transfer function of rational type (numerator / denominator)!

8.4 Transfer Functions

If the transfer function is written in form of positive powers in z first can be converted in a form with negative powers, i.e., z^{-1} , and afterwards it can be transformed into the time domain.

$$\frac{b'_m z^m + \dots + b'_1 z^1 + b'_0}{a'_n z^n + \dots + a'_1 z^1 + a'_0}$$

WARNING: These are different n and m values compared to the previous slide.

$$\begin{aligned} G(z) &= \frac{\frac{b'_m}{a'_n} z^{m-n} + \frac{b'_{m-1}}{a'_n} z^{m-n-1} + \dots + \frac{b'_0}{a'_n} z^{-n}}{1 + \frac{a'_{n-1}}{a'_n} z^{-1} + \dots + \frac{a'_0}{a'_n} z^{-n}} \\ &= \frac{b_{m-n} z^{m-n} + b_{n-m+1} z^{m-n-1} + \dots + b_n z^{-n}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}} \end{aligned}$$

For $n = m$ this transfer function is identical to the one on the previous slide. For $n > m$ a dead time can be factored out in the numerator:

$$G(z) = \frac{b_{m-n} + b_{n-m+1} z^{-1} + \dots + b_n z^{-m}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}} z^{m-n} = \frac{\tilde{b}_0 + \tilde{b}_1 z^{-1} + \dots + \tilde{b}_m z^{-m}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}} z^{-d}$$

with $d = n - m$. The case $m > n$ does not occur (negative dead time \rightarrow **non-causal**)!

8.4 Transfer Functions

Causality and Properness

A transfer function of the form

$$G(z) = \frac{b_0 + b_1 z^{-1} + \dots + b_m z^{-m}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}}$$

has numerator degree m and denominator degree n which are positive integers. $G(z)$ is causal.

A transfer function of the form

$$G(z) = \frac{b'_m z^m + \dots + b'_1 z^1 + b'_0}{a'_n z^n + \dots + a'_1 z^1 + a'_0}$$

requires: **denominator degree \geq numerator degree** or $n \geq m$. If this requirement is met then **$G(z)$ is causal**. However, if $m > n$, then $G(z)$ is non-causal negative dead times arise, i.e., values in the future have to be predicted.

The condition denominator degree \geq numerator degree is known from the s -domain. There it is a condition for **properness** or **realizability**, i.e., avoiding pure differentiators! For time-discrete systems such limitations do not exist. Every causal system can be realized.

8.4 Transfer Functions

Proper / Strictly Proper

For continuous-time systems the difference between a proper and strictly proper system can be directly seen in the transfer function.

- *Proper*: numerator degree \leq denominator degree: $m \leq n$
- *Strictly proper*: numerator degree $<$ denominator degree: $m < n$

In discrete time a system is *proper* but not *strictly proper* (= “sprungfähig”)

- for transfer functions in z -form (only positive powers of z):

numerator degree $m =$ denominator degree n

- for transfer functions in z^{-1} -form (only negative powers of z):

$b_0 \neq 0$

Only if b_0 exists the input $u(k)$ directly influences the output $y(k)$. If $b_0 = 0$ then a change in the input is delayed by one or more steps until $u(k-1)$ or later until it affects the output $y(k)$.

Terminology: A system follows the difference equation: $y(k) = b_1 u(k-1) + a_1 y(k-1)$

This can be interpreted either a dead time of 1 or as a not strictly proper system:

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_1 z^{-1}}{1 - a_1 z^{-1}} = \frac{\tilde{b}_0}{1 - a_1 z^{-1}} z^{-1} \quad b_1 = \tilde{b}_0$$

8.4 Transfer Functions

Difference Equation → Transfer Function

In order to transform a difference equation into the z -domain, first the equation is rewritten such that $y(k)$ is the newest output value. Then the transformation into the z -domain requires only operators with negative powers like z^{-i} :

Example: $2y(k - 1) + 4y(k) + 3y(k + 3) - u(k) = -u(k - 1)$

1.) New starting time step: $y(k + 3)$

2.) Time transformation such that this value is mapped to $y(k)$: $k := k - 3$

$$\rightarrow 2y(k - 4) + 4y(k - 3) + 3y(k) - u(k - 3) = -u(k - 4)$$

3.) Transformation into the z -domain, separation of $Y(z)$ and $U(z)$, division to obtain transfer function:

$$2z^{-4}Y(z) + 4z^{-3}Y(z) + 3Y(z) - z^{-3}U(z) = -z^{-4}U(z)$$

$$(2z^{-4} + 4z^{-3} + 3)Y(z) = (z^{-3} - z^{-4})U(z)$$

$$\frac{Y(z)}{U(z)} = G(z) = \frac{z^{-3} - z^{-4}}{3 + 4z^{-3} + 2z^{-4}} = \frac{\frac{1}{3}(z^{-3} - z^{-4})}{1 + \frac{4}{3}z^{-3} + \frac{2}{3}z^{-4}} = \frac{\frac{1}{3}(1 - z^{-1})}{1 + \frac{4}{3}z^{-3} + \frac{2}{3}z^{-4}} z^{-3}$$

8.4 Transfer Functions

IIR (*Infinite Impulse Response*)

All impulse response functions $g(t)$ in continuous time are of infinite length. Typically they decay to zero with exponential behavior. By sampling a sequence $g(k)$ of infinite length results. Such systems are named IIR (*infinite impulse response*).

IIR systems have a transfer function with non-trivial denominator, i.e., the denominator is more complex than z^n . This yields at least two different delayed versions of $y(k)$ in the corresponding difference equation. A consequence is that this difference equation can only be calculated *recursively*!

Examples:

$$G(z) = \frac{1}{1 - 0.9z^{-1}}$$

$$G(z) = \frac{z^{-1}}{1 - 0.8z^{-1}}$$

$$G(z) = \frac{z}{1 - 0.7z^{-1}} \quad \text{non-causal!}$$

$$G(z) = \frac{0.4 + 0.5z^{-1} + 0.6z^{-2} + 0.7z^{-3} + 0.8z^{-4}}{1 - 0.9z^{-1}}$$

$$G(z) = \frac{0.4 + 0.5z^{-1} + 0.6z^{-2} + 0.7z^{-3} + 0.8z^{-4}}{(1 - 0.8z^{-1})^2(2 - z^{-1} + 0.3z^{-2} + 0.5z^{-3})}$$

$$G(z) = \frac{z^2 + 0.7z + 0.4}{z^3 + 0.8z + 0.2}$$

$$G(z) = \frac{z^3 + 0.8z + 0.2}{z^2 + 0.7z + 0.4} \quad \text{non-causal!}$$

8.4 Transfer Functions

FIR (*Finite Impulse Response*)

Systems with impulse sequences $g(k)$ of finite length are called FIR systems (*finite impulse response*). They only exist in discrete time! They have no (exact) equivalent in continuous time. However, if the length of an FIR system is allowed to be very long it might be possible to *approximate* a stable IIR system by a long FIR system. Marginally stable or unstable IIR systems, in principle, cannot be approximated by an FIR system because their impulse response does not converge to 0.

FIR systems have a transfer function without denominator or with a denominator of type z^m . A consequence is only one y -term in the difference equation (*feedforward*).

Examples:
$$G(z) = \frac{z^2 - z + 0.25}{z - 0.5} z^{-2} = \frac{(z - 0.5)^2}{z - 0.5} z^{-2} = (z - 0.5) z^{-2} = z^{-1} - 0.5 z^{-2}$$

$$G(z) = 1 - z^{-1} \qquad G(z) = \frac{z^3 + 4z^2 + 3z + 1}{z^5} = z^{-2} + 4z^{-3} + 3z^{-4} + z^{-5}$$

$$G(z) = \sum_{i=0}^{10} b_i z^{-i} \qquad G(z) = \frac{z^3 + 4z^2 + 3z + 1}{z^2} = z + 4 + 3z^{-1} + z^{-2} \quad \text{non-causal!}$$

$$G(z) = z^{-2} \qquad G(z) = z^3 \quad \text{non-causal!}$$

8.4 Transfer Functions

Pole-Zero-Form of a Transfer Function

Up to here we have considered transfer functions in explicit polynomial form. However, a factorized form is often useful because the poles and zeros directly appear in the denominator and numerator. It is simpler to write it in positive powers of z :

$$G(z) = \frac{b_m z^m + \dots + b_1 z^{-1} + b_0}{a_n z^n + \dots + a_1 z^{-1} + a_0} = k \frac{(z - n_1) \cdot (z - n_2) \cdot \dots \cdot (z - n_m)}{(z - p_1) \cdot (z - p_2) \cdot \dots \cdot (z - p_n)}$$

The gain of $G(z)$ can be calculated according to the final value limit theorem of the z -transform by letting $z = 1$:

$$\text{Gain: } K = \frac{b_m + \dots + b_1 + b_0}{a_n + \dots + a_1 + a_0} = k \frac{\prod_{i=1}^m (1 - n_i)}{\prod_{i=1}^n (1 - p_i)}$$

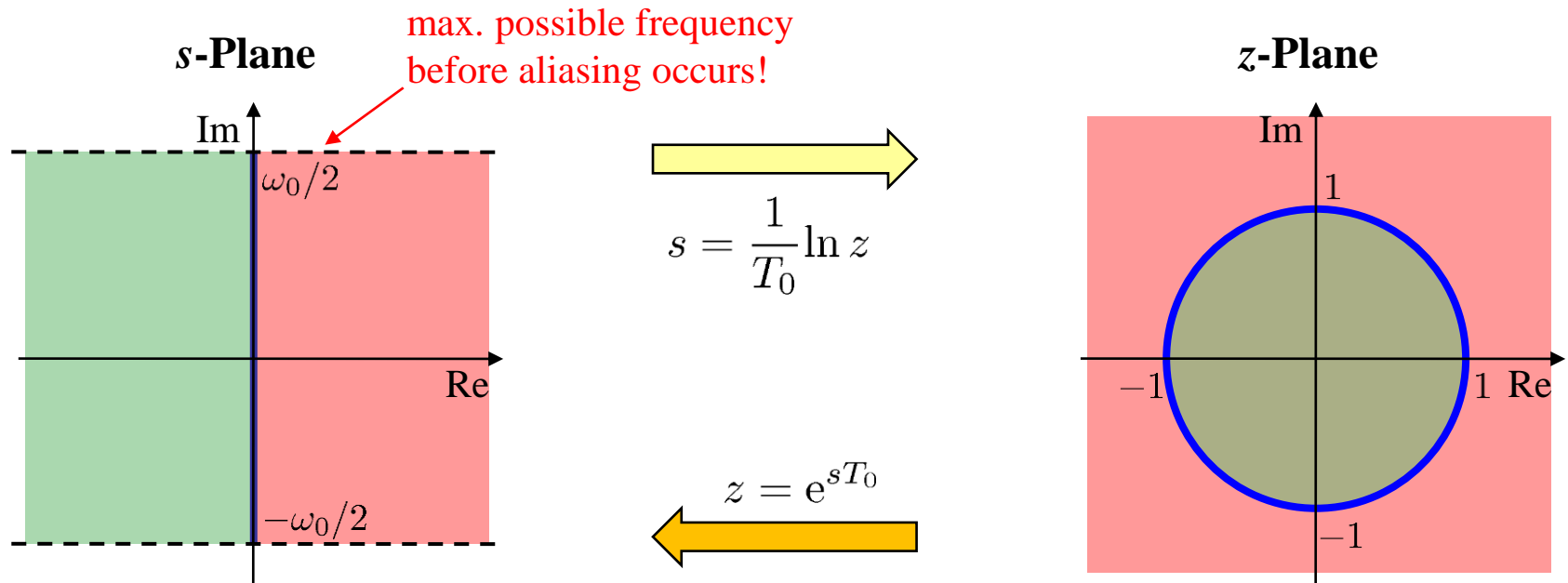
The poles p_i and zeros n_i can be transformed into the s -domain via $s = \frac{1}{T_0} \ln z$ and can be interpreted accordingly.

Immediately conditions for stability and phase minimality for poles and zeros result in the z -domain.

8.4 Transfer Functions

Relation Between s -Plane and z -Plane

- The stability region “left half s -plane” is mapped to the inner region inside the unit circle in the z -plane.
- The imaginary axis of the s -plane is mapped to the unit circle in the z -plane.
- The unstable region “right half s -plane” is mapped to the outer region around the unit circle in the z -plane.



8.4 Transfer Functions

Stability

- A transfer function in the z -domain is **stable** if all poles are *inside* the unit circle.
- If one or more poles are *on* the unit circle (no multiple poles!) and all other poles are inside the unit circle, the system is **marginally stable**.
- If at least one pole exists *outside* the unit circle or a multiple pole is on the unit circle, then the system is **unstable**.
- The stability properties of a transfer function in the s -domain keep valid for transformation in the z -domain because the poles transform according to $z = e^{sT_0}$.

Phase Minimality

- A system has minimum phase if it has only stable and marginally stable poles *and* zeros.

The location of the zeros typically changes during the transformation from the s -domain into the z -domain. Therefore the property “minimum phase” generally is not preserved during the transformation.

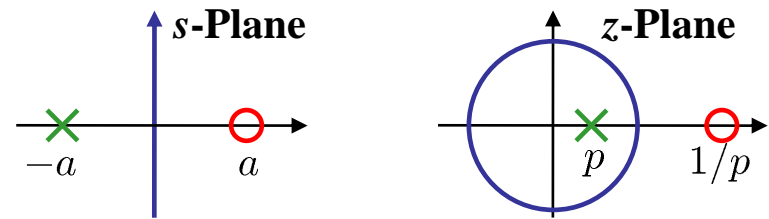
8.4 Transfer Functions

Example: All-Pass in z-Domain

An all-pass is characterized by an amplitude response equal to 1 for all frequencies. Because poles and zeros have the same absolute values, just opposite signs, they cancel in the magnitude. Of course the phase is affected. A simple first order all-pass in the s -domain is:

$$G(s) = \frac{-Ts + 1}{Ts + 1} = \frac{-s + a}{s + a} \quad \text{with } a = \frac{1}{T} > 0$$

Pole:	$s = -a$ (stable)
Zero:	$s = a$ (unstable)



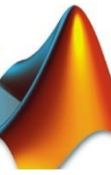
The corresponding all-pass in the z -domain has a stable pole and the inverse zero mirrored at the unit circle. It is not the direct transformation from s to z !

$$G(z) = \frac{pz - 1}{z - p} = \frac{p - z^{-1}}{1 - pz^{-1}} \quad \text{with } |p| < 1$$

Pole:	$z = p$ (stable)
Zero:	$z = 1/p$ (unstable)

The amplitude response is given by $z = e^{i\omega T_0}$:

$$|G(i\omega)| = \frac{|pe^{i\omega T_0} - 1|}{|e^{i\omega T_0} - p|} = \frac{\sqrt{(p \cos \omega T_0 - 1)^2 + p^2 \sin^2 \omega T_0}}{\sqrt{(\cos \omega T_0 - p)^2 + \sin^2 \omega T_0}} = \frac{\sqrt{p^2 - 2p \cos \omega T_0 + 1}}{\sqrt{1 - 2p \cos \omega T_0 + p^2}} = 1$$



Change of Sampling Rate:

`decimate(x,r);1`

*% Reduces the sampling rate of signals x
% by a factor of r with help of a low-pass
% filter.*

`upsample(x,n);1`

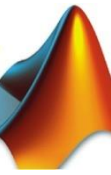
*% Increases the sampling rate by a factor of n,
% by inserting zeros in between the sample
% points*

*% E.g.: x = [1 2 3];
% y = upsample(x,3);
% y = [1 0 0 2 0 0 3 0 0]*

`downsample(x,n);1`

*% Reduction of sampling rate. Only every n-th
% sample is carried over.*

*% E.g.: x = [1 2 3 4 5 6 7 8 9 10];
% y = downsample(x,3);
% y = [1 4 7 10]*



`resample(x,p,q);1` *% Changes the sampling rate of signal vector x*
% by the rational factor p/q

Impulse Response and Step Response:

`impulse;2` *% Calculates the impulse response of a linear*
% system

`step;2` *% Calculates the step response of a linear*
% system

Partial Fraction Expansion:

`[r,p,k] = residuez(b,a);1` *% Performs a partial fraction expansion*
% with the ratio of numerator b(z)
% and denominator a(z).
% The inverse operation is also
% possible.

¹ : *Signal Processing Toolbox*

² : *Control System Toolbox*

9. Transformation into the Frequency Domain

Contents of Chapter 9

9. Transformation of Signals in the Frequency Domain

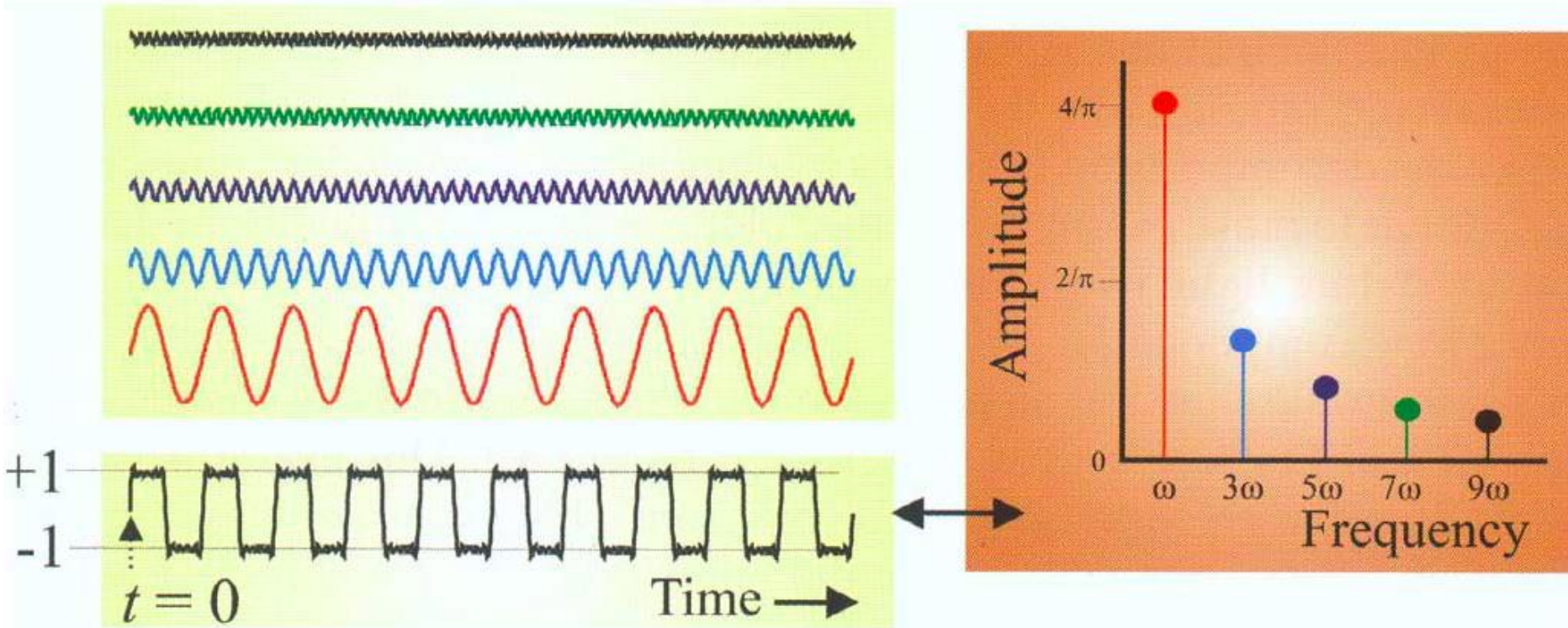
- 9.1 Discrete Fourier Transform (DFT)
- 9.2 Extension: Fast Fourier Transform (FFT)
- 9.3 Frequency Analysis Via DFT
- 9.4 Leakage Effect and Windowing
- 9.5 Non-Stationary Signals und Short-Time-DFT
- 9.6 Outlook: Time-Frequency-Analysis
- 9.7 Outlook: Parametric Frequency Analysis

Joseph Fourier, 1768-1830
(www.wikipedia.org) C



9.1 Discrete Fourier Transform (DFT)

Fourier Series

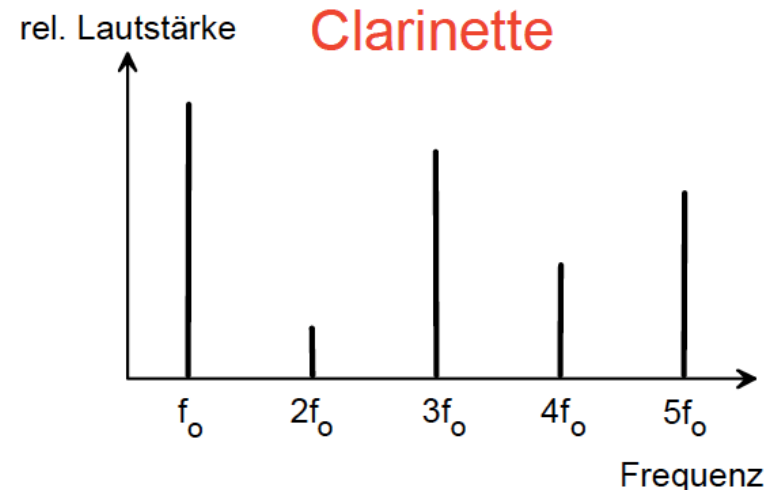
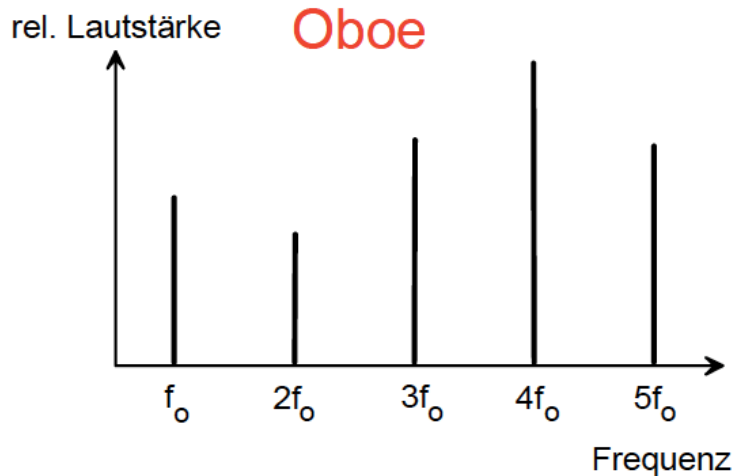
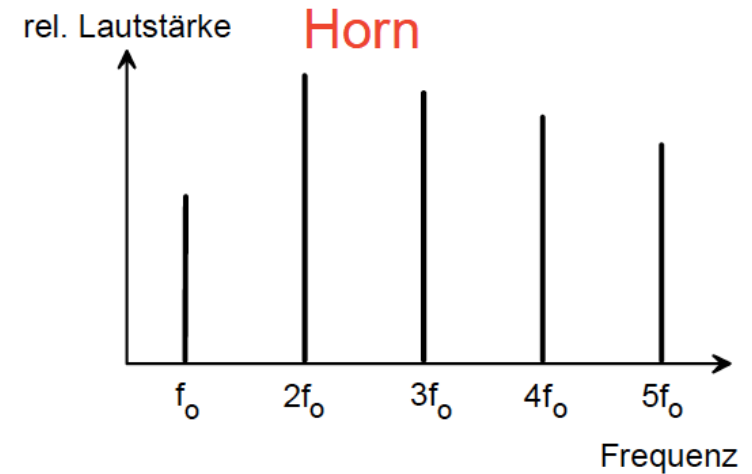
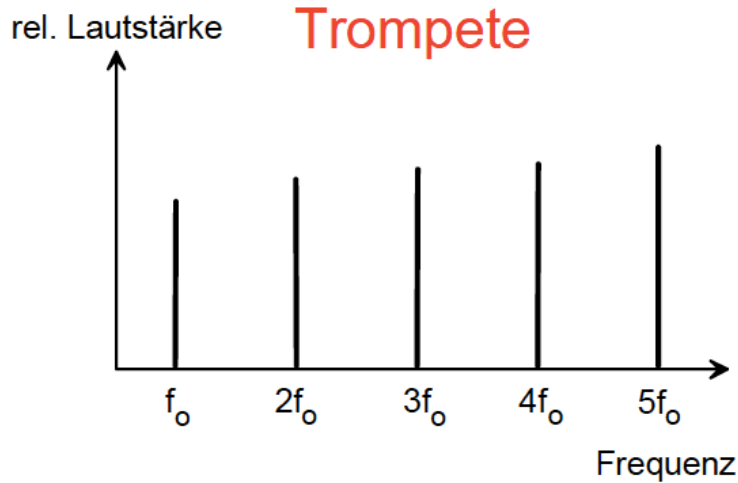


Source: ftp://ftp.ifn-magdeburg.de/pub/MBLehre/sv06_130509-ftp.pdf

9.1 Discrete Fourier Transform (DFT)

Source: http://eitidaten.fh-pforzheim.de/daten/mitarbeiter/blankenbach/vorlesungen/mathe_2/Fourier_Trafo_kurz_Folien.pdf

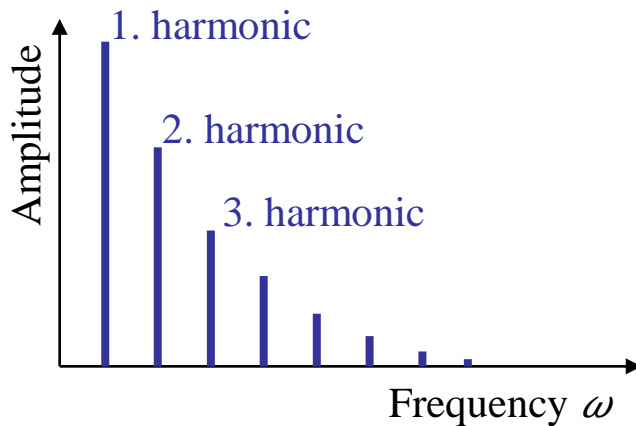
Standard concert pitch A4: $f_0 = 440$ Hz on different music instruments



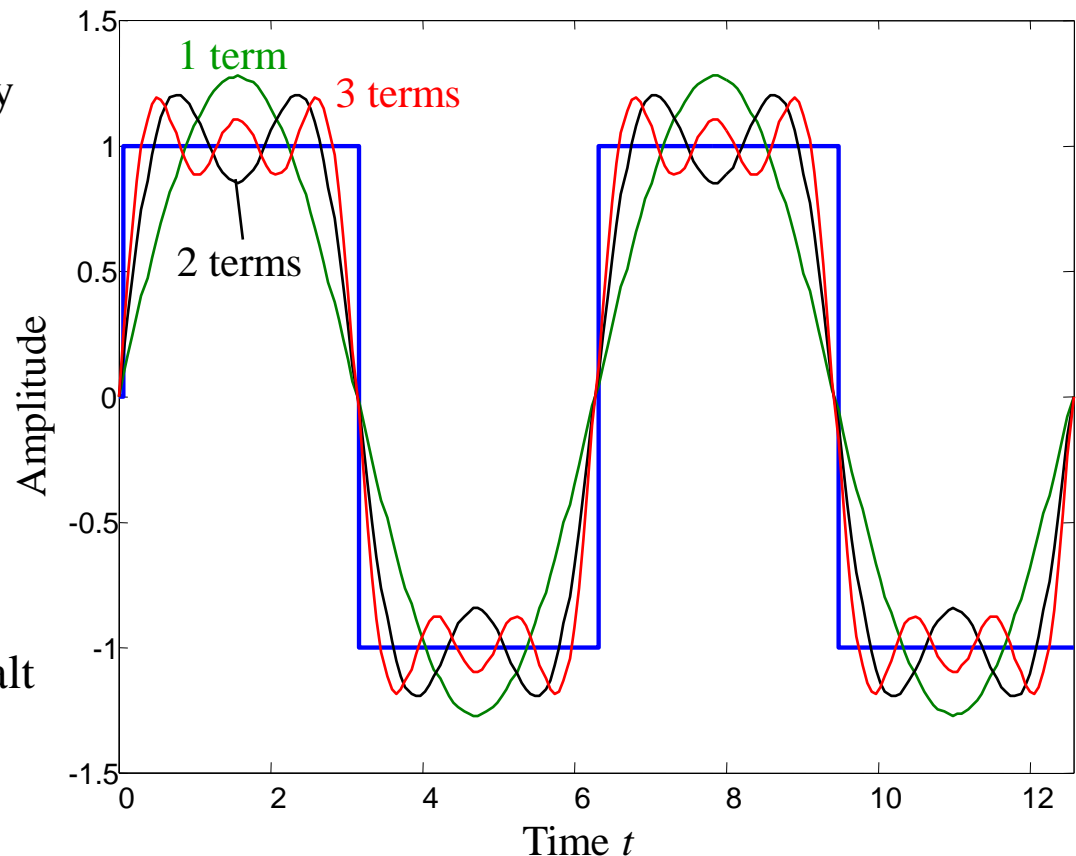
9.1 Discrete Fourier Transform (DFT)

Fourier Series

- Decomposition of a *periodic* signal in its frequency components.
- Signal can be decomposed into an infinite sum of sine and cosine terms.
- Amplitude for each frequency indicates how strong this frequency is contained in the signal.



- If *non-periodic* signals shall be dealt with: period length $\rightarrow \infty$, basic oscillation $\rightarrow 0$.

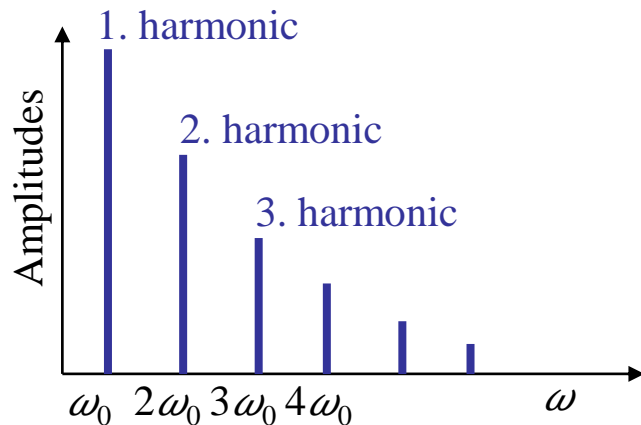


9.1 Discrete Fourier Transform (DFT)

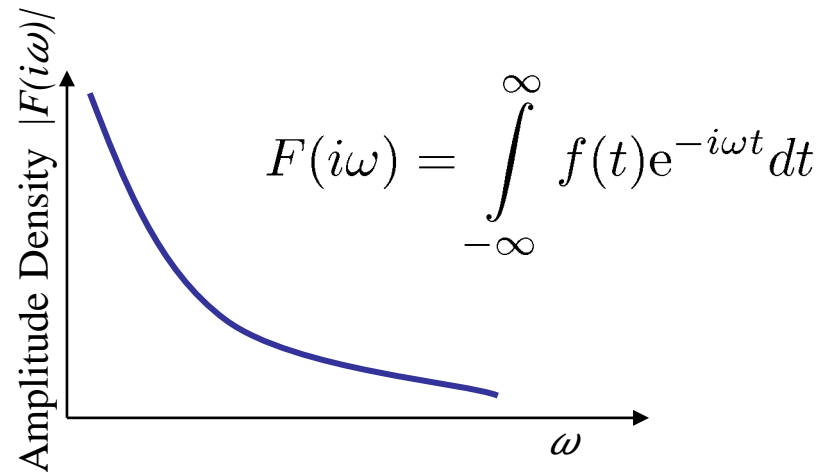
Fourier Transform

- Extension of the Fourier series for non-periodic signals
- Period length $T \rightarrow \infty$, basic oscillation $\omega \rightarrow 0$.
- The spectrum is not composed of discrete frequencies $n \cdot \omega_0$ (i.e., multiples of the basic oscillation). Rather it consists of arbitrary many frequencies (i.e., a real number) – the so-called **amplitude density spectrum** (similar for the phase).

Fourier Series



Fourier Transform



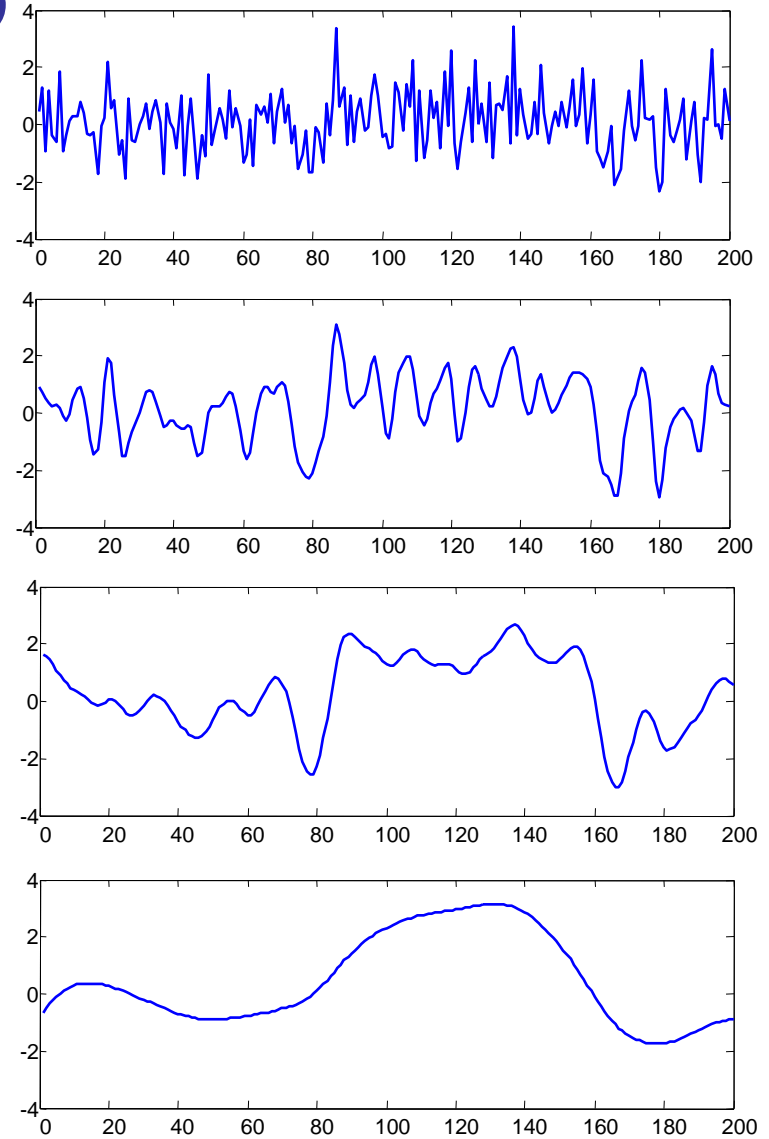
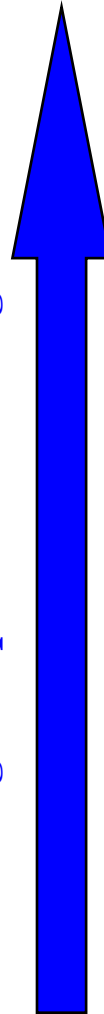
9.1 Discrete Fourier Transform (DFT)

Signals contain many different frequencies.

A transformation from the *time domain* to the *frequency domain* allows to examine how strong which frequencies are contained in the signal.

This is a powerful tool for the analysis further processing of signals.

Share of high frequencies in the signal



Zeit t

9.1 Discrete Fourier Transform (DFT)

Fourier Transform

- Time continuous
- Frequency continuous

$$X(i\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t} dt$$

Remember::

$$\omega = 2\pi f$$

$$f_0 = \frac{1}{T_0}$$

sampling frequency \nearrow
 sampling time \nwarrow

Time-Discrete Fourier Transform

- Time discrete: $t = kT_0$
- Frequency continuous

$$X(i\omega) = \sum_{k=-\infty}^{\infty} x(kT_0)e^{-i\omega kT_0} \rightarrow X(i\Omega) = \sum_{k=-\infty}^{\infty} x(k)e^{-i\Omega k}$$

$$x(k) = x(kT_0)$$

$$\Omega = \omega T_0 = 2\pi f T_0 = 0 \dots 2\pi$$

$$f = 0 \dots f_0$$

Discrete Fourier Transform

- Time discrete N samples: $t = 0, T_0, 2T_0, \dots, (N - 1)T_0$
- Frequency discrete in N samples:

$$\omega_n = 0, \frac{1}{N}\omega_0, \frac{2}{N}\omega_0, \dots, \frac{N-1}{N}\omega_0 \text{ or}$$

$$\Omega_n = 0, \frac{1}{N}2\pi, \frac{2}{N}2\pi, \dots, \frac{N-1}{N}2\pi$$

Für $n = 0, 1, 2, \dots, N - 1$:

$$X(n) = X(i\omega_n) = X(i\Omega_n) = \sum_{k=0}^{N-1} x(k)e^{-i2\pi nk/N}$$

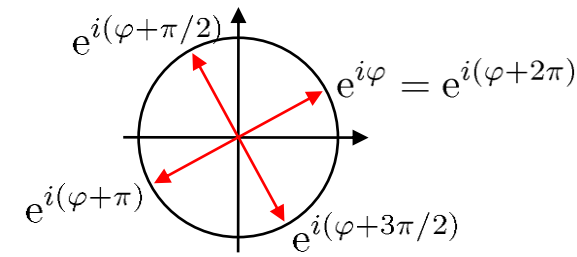
9.1 Discrete Fourier Transform (DFT)

Properties

- *Periodicity* in the frequency range (see sampling theorem).
Because the exp-function is periodic with $i2\pi$:

$$e^{i\varphi} = e^{i(\varphi+2\pi)} = e^{i(\varphi+4\pi)} = \dots$$

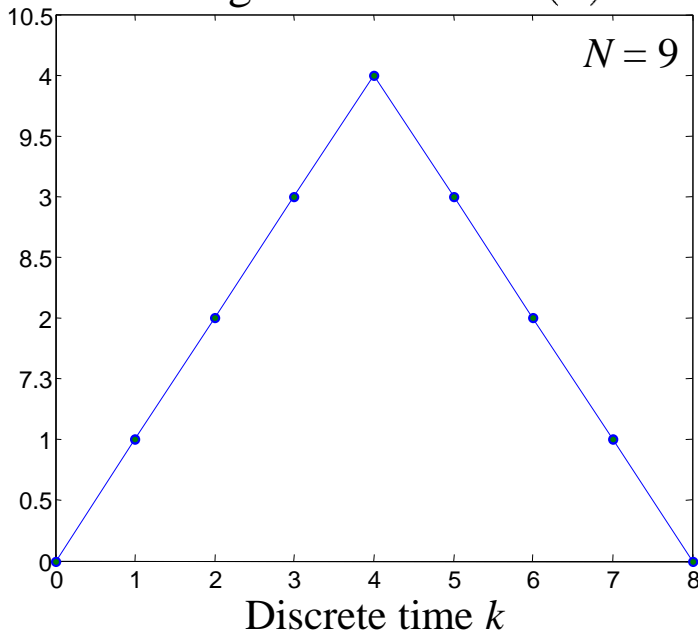
$$\rightarrow X(n) = X(n + N) = X(n + 2N) = \dots$$



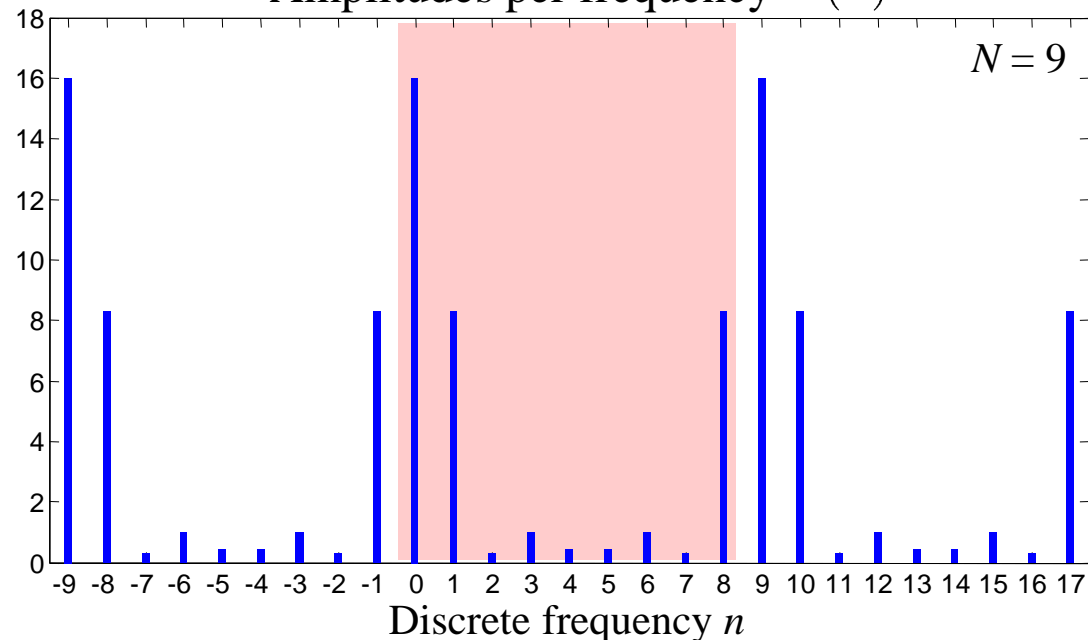
Interpretation of time and frequency axes:

E.g. for $f_0 = 50$ Hz $\rightarrow T_0 = 0.02$ s
 $k = 4 \rightarrow t = 4 \cdot 0.02$ s = 0.08 s
 $n = 4 \rightarrow f = 4/N \cdot 50$ Hz = 28.2 Hz

Signal over time $x(k)$



Amplitudes per frequency $X(n)$



9.1 Discrete Fourier Transform (DFT)

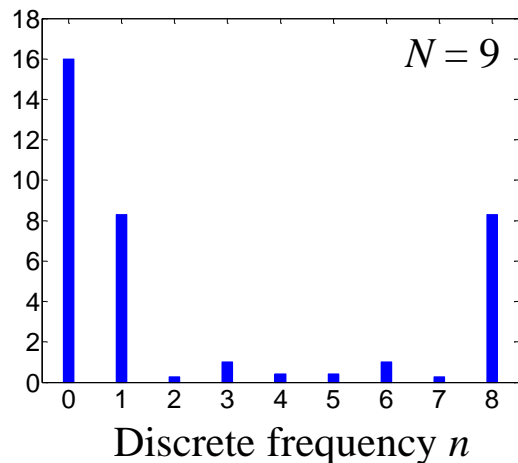
Properties

- *Periodicity* in the time range.

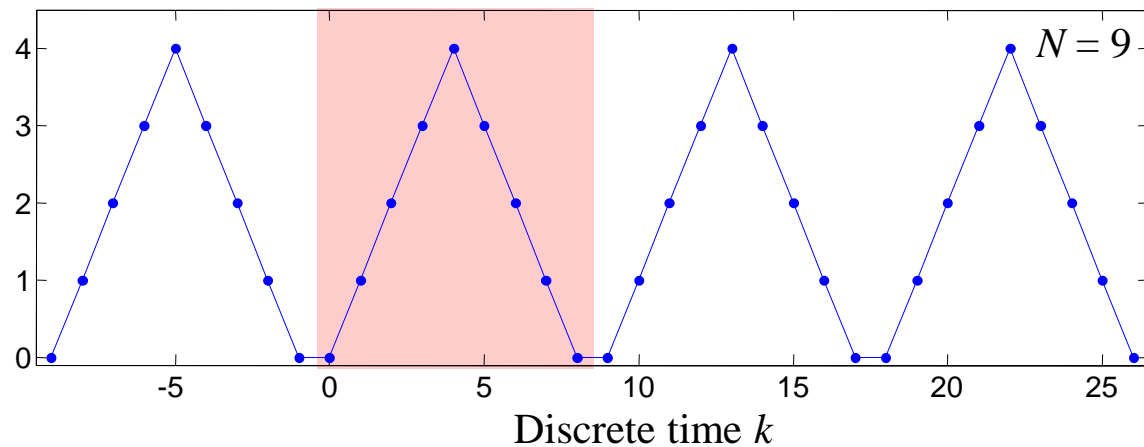
Because the 2π -periodic exp-function also occurs in the backward transformation, in contrast to the continuous-time transform, for the DFT the time signal appears to be periodic. The discretization of the frequency axis causes this effect.

$$\rightarrow x(k) = x(k + N) = x(k + 2N) = \dots$$

Amplitudes per frequency $X(n)$



Signal over time $x(k)$

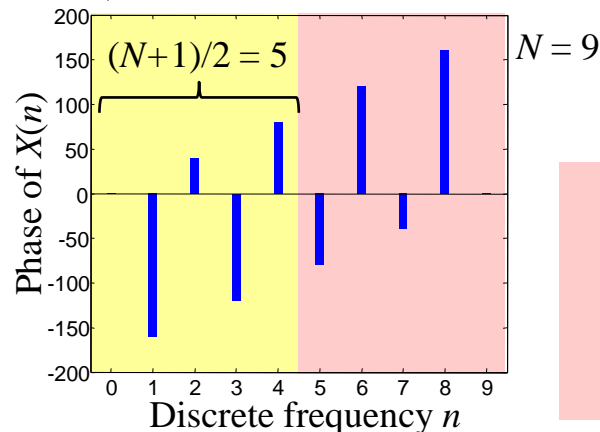
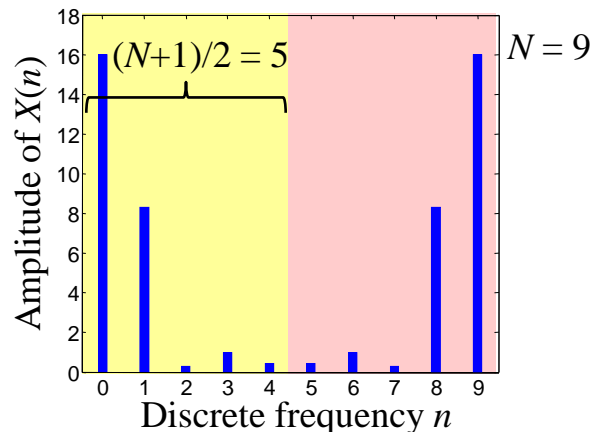


9.1 Discrete Fourier Transform (DFT)

Properties

- If $x(k)$ is real (normal case) then the *amplitude response* is an *even function* and the *phase response* is an *odd function*, i.e., both are determined completely by half of the points; the other half can be generated by mirroring:
 - N is even: $N/2+1$ points are required.
 - N is odd: $(N+1)/2$ points are required.

Reason: The time signal $x(k)$ contains only frequencies up to $f_0/2$ (sampling theorem!) otherwise we would get aliasing. Therefore it only makes sense to display the frequency plot in the range $f = 0 \dots f_0/2$ or $\Omega = 0 \dots \pi$. The part $f = f_0/2 \dots f_0$ or $\Omega = \pi \dots 2\pi$ respectively $f = -f_0/2 \dots 0$ or $\Omega = -\pi \dots 0$ is redundant!



This range contains no new information and can be generated by mirroring. Commonly therefore only the left range is displayed!

9.1 Discrete Fourier Transform (DFT)

$$\text{DFT}\{x(k)\} = X(n)$$

Further properties of the DFT are already known from the continuous Fourier Transform:

- Linearity: $\text{DFT}\{ax_1(k) + bx_2(k)\} = aX_1(n) + bX_2(n)$
- Time shift: $\text{DFT}\{x(k + l)\} = X(n) \cdot e^{-i2\pi nl/N}$
- Frequency shift: $\text{IDFT}\{X(n + l)\} = x(k) \cdot e^{i2\pi kl/N}$
- Convolution: $\text{DFT}\{x_1(k) * x_2(k)\} = X_1(n) \cdot X_2(n)$
- Multiplication: $\text{DFT}\{x_1(k) \cdot x_2(k)\} = X_1(n) * X_2(n)$

Inverse DFT

For completeness, here the formula for the transformation back into the time-domain:

$$\text{IDFT}\{X(n)\} = x(k) = \frac{1}{N} \sum_{n=0}^{N-1} X(n)e^{i2\pi kn/N}$$

9.1 Discrete Fourier Transform (DFT)

Implementation of the DFT

$$\text{DFT}\{x(k)\} = X(n) = \sum_{k=0}^{N-1} x(k)e^{-i2\pi nk/N} = \sum_{k=0}^{N-1} x(k)W_N^{nk}$$

Abbreviation:

with $W_N = e^{-i2\pi/N}$

This can be written for $n = 0, 1, 2, \dots, N-1$ as the following equation system:

$$\begin{pmatrix} X(0) \\ X(1) \\ X(2) \\ \vdots \\ X(N-1) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N^1 & W_N^2 & \cdots & W_N^{N-1} \\ 1 & W_N^2 & W_N^4 & \cdots & W_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)^2} \end{pmatrix} \begin{pmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(N-1) \end{pmatrix}$$

To carry out this matrix-vector multiplication, the following amount of computation is necessary:

- N^2 complex multiplications
- N^2 complex additions

9.2 Extension: Fast Fourier Transform (FFT)

Idea for the Fast Fourier Transform (FFT)

- Efficient implementation of the DFT with identical result.
- Split of an DFT of size N (number of data points) in 2 DFTs of size $N/2$ by a trick.
- Further split of 2 DFTs of size $N/2$ in 4 DFTs of size $N/4$, etc.
- These splits are continued up to $N/2^s = 1$; s represent the number of splits necessary.
- Works only if $N = 2^s$, i.e., a power of 2. If this is *not* the case, the signal $x(k)$ is filled with zeros such that the number of points is equal to 2^s (*zero padding*).

Complexity of the FFT

- Only $N \lg(N)$ complex multiplications and addition are required.
- Example: $N = 1024$
 - computational demand DFT $\sim N^2 \approx 1.000.000$
 - computational demand FFT $\sim N \lg(N) = 1024 \cdot 10 \approx 10.000 \rightarrow$ Factor 100 savings!

Info: $\lg()$ is the logarithm to base 2

$$2^s = x \rightarrow s = \lg(x)$$

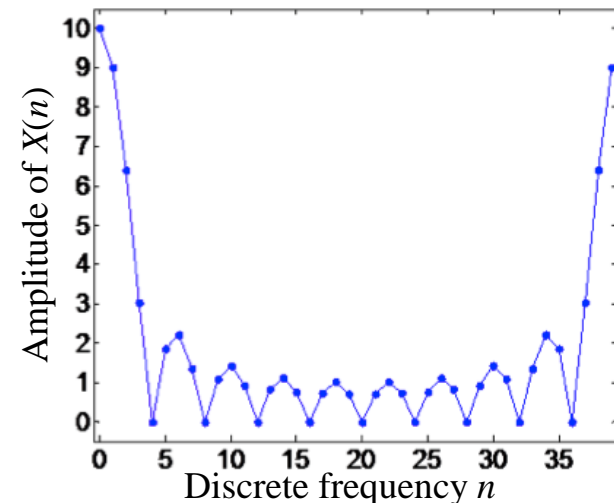
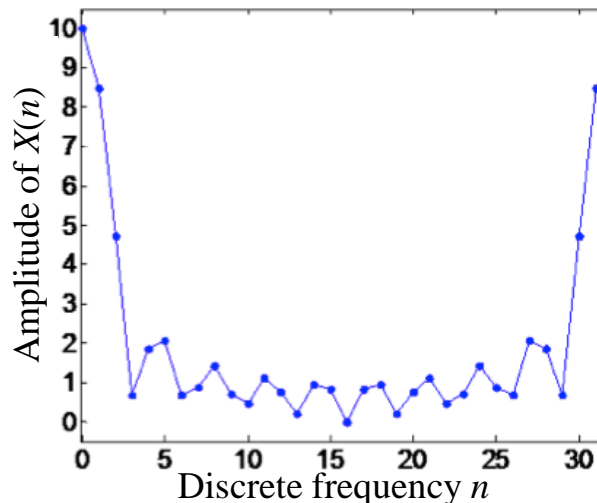
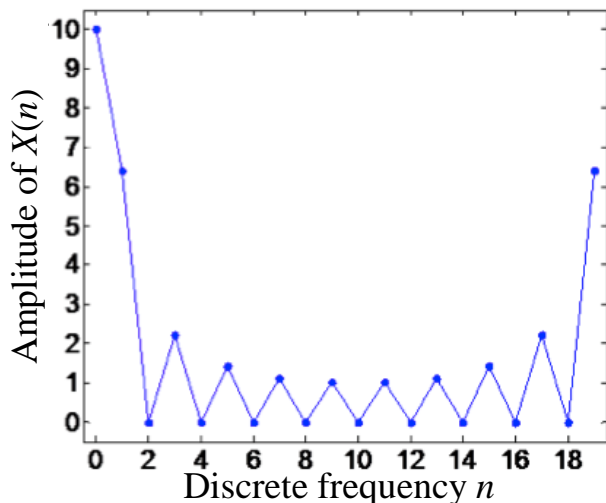
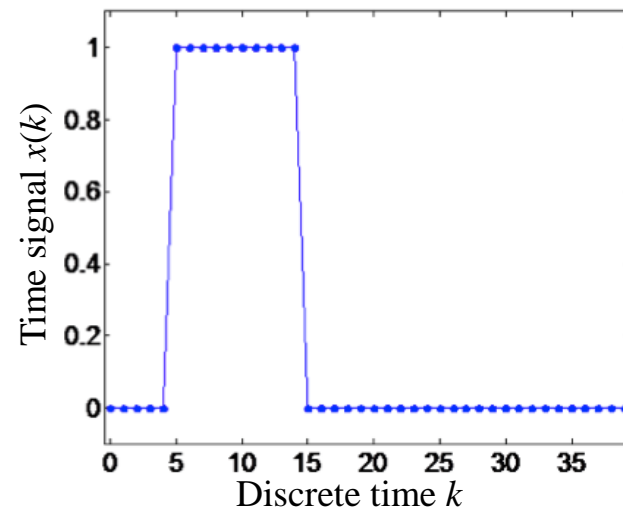
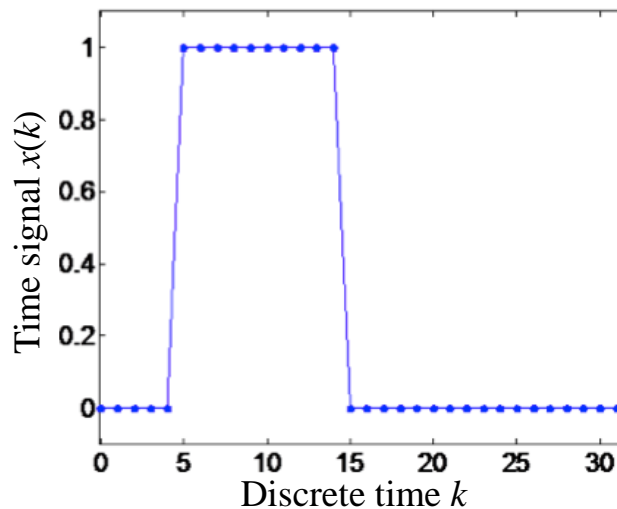
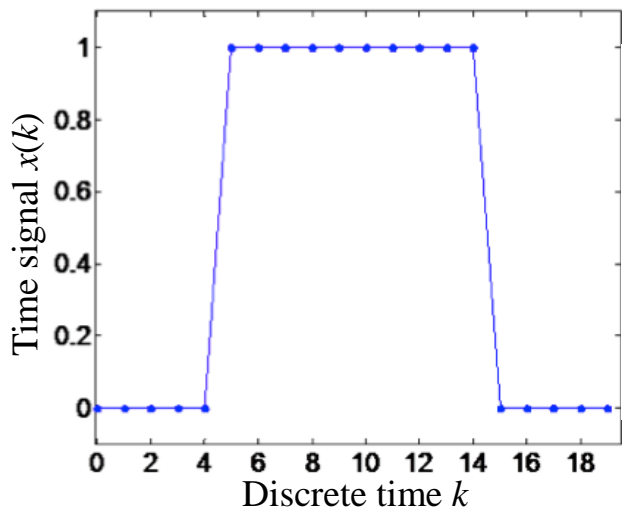
9.3 Frequency Analysis Via DFT

Choice for the amount of data N

$N = 20$

$N = 32$

$N = 40$



9.3 Frequency Analysis Via DFT

Observations:

- All time signals are chosen identically, only the number of zeros filled in vary, so that the total number of points are $N = 20, 32, 40$.
- The resolution in the frequency domain depends on N . The frequency axes are scaled as follows:
 - $N = 20$: $\omega_n = 0, \frac{1}{20}\omega_0, \frac{2}{20}\omega_0, \dots, \frac{19}{20}\omega_0$ $\Delta\omega = \frac{1}{20}\omega_0$
 - $N = 32$: $\omega_n = 0, \frac{1}{32}\omega_0, \frac{2}{32}\omega_0, \dots, \frac{31}{32}\omega_0$ $\Delta\omega = \frac{1}{32}\omega_0$
 - $N = 40$: $\omega_n = 0, \frac{1}{40}\omega_0, \frac{2}{40}\omega_0, \dots, \frac{39}{40}\omega_0$ $\Delta\omega = \frac{1}{40}\omega_0$
- A clever choice for N by zero padding can achieve frequency intervals $\Delta\omega$ of desired size even if the original signal is shorter than N values.
If a certain frequency ω^* is interesting and the amplitude for this frequency is important to know with high accuracy, it should be *exactly* contained in the frequency discretization by an appropriate choice of N (see *picket fence effect*)!

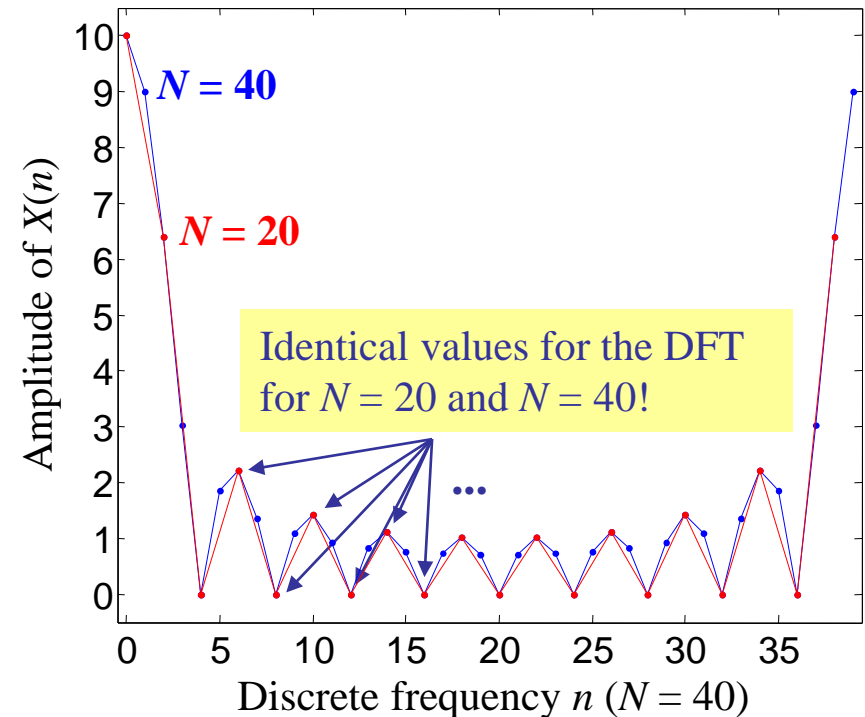
9.3 Frequency Analysis Via DFT

Observation:

- Doubling the number of points $N = 20 \rightarrow 40$ doubles frequency resolution. The DFT for $N = 20$ yields identical values (for every second point) as the DFT for $N = 40$.

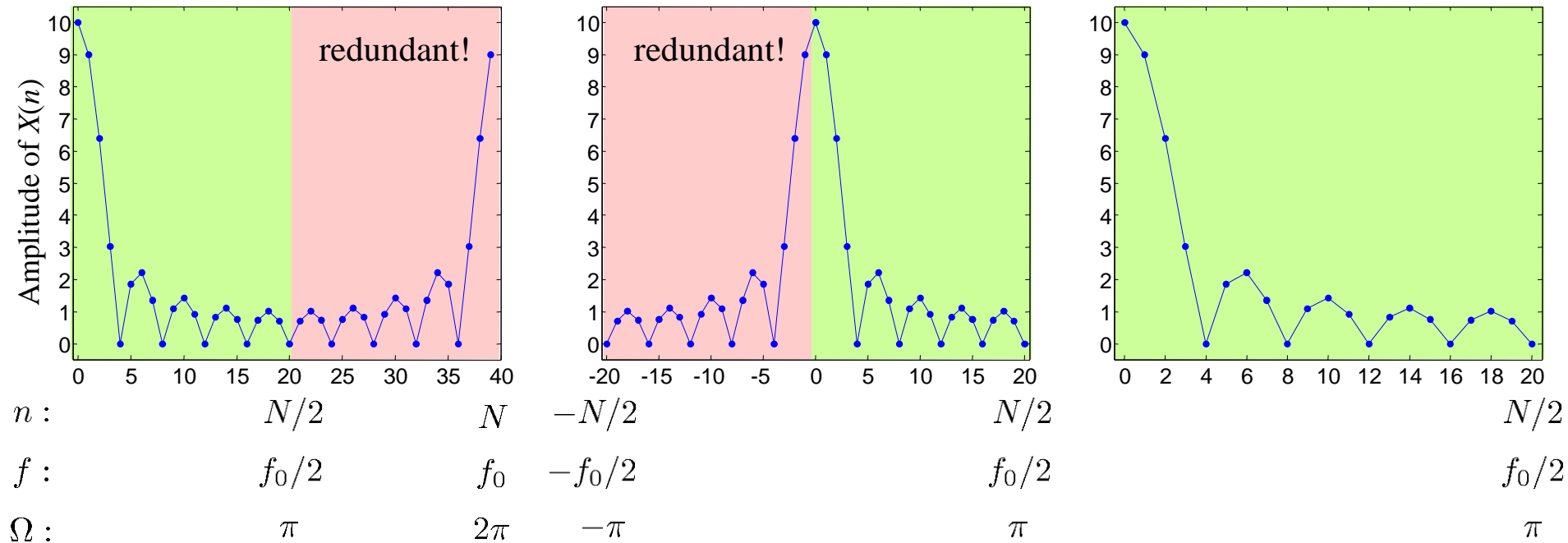
Remark:

- The phase of $X(n)$ sometimes is interesting, as well. We focus on the amplitudes but an analysis of the phase can also be important.
- MATLAB creates the plots shown in these lecture notes. `fft()` yields $X(n)$ in the frequency range 0 to f_0 .
- Commonly the upper half of the spectrum is omitted because it does not carry any additional information. Also a symmetric plot around the origin from $-f_0/2$ to $+f_0/2$ is popular.



9.3 Frequency Analysis Via DFT

Equivalent Types of Plots for the Spectrum

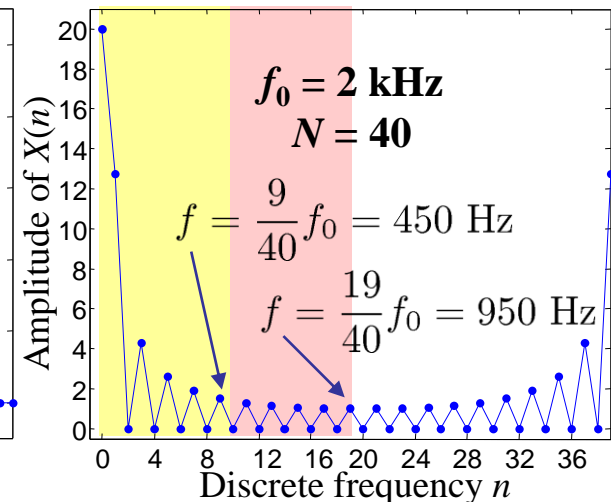
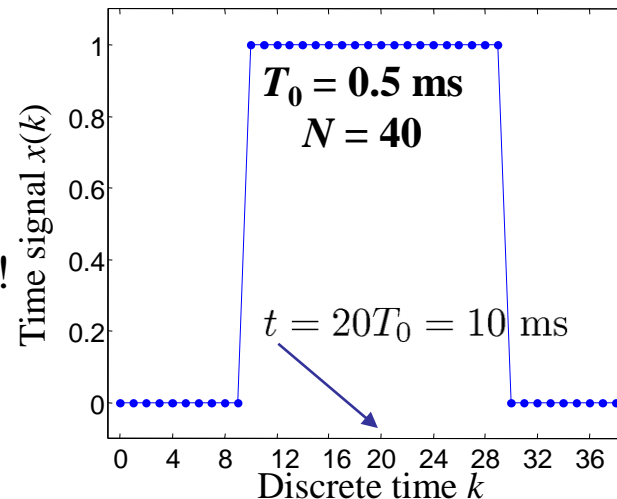
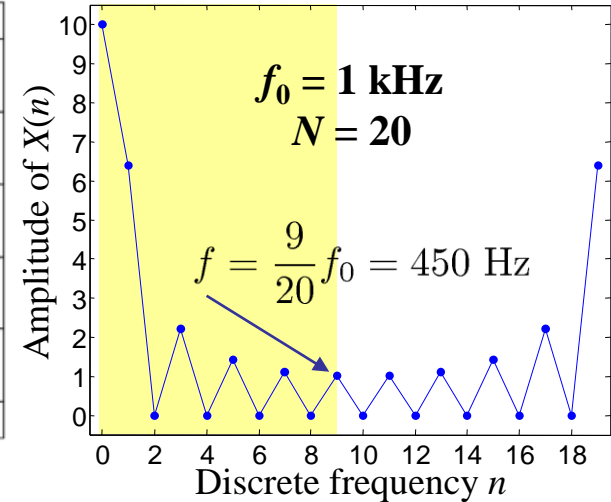
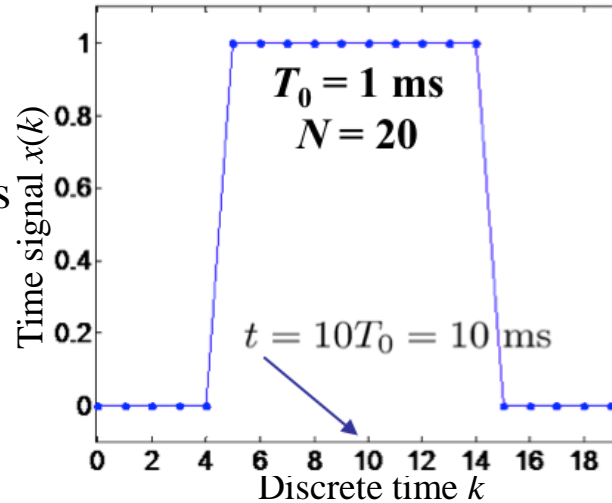


9.3 Frequency Analysis Via DFT

Choice of Sampling Time T_0 / Sampling Frequency f_0

- The faster the signal is sampled, the wider is its frequency range.
- In practice, the amplitudes typically become smaller at higher frequencies.
- As the sampling theorem tells us, the sampling frequency should be chosen such that the highest significant signal frequency lies below $f_0/2$. Otherwise we get aliasing!

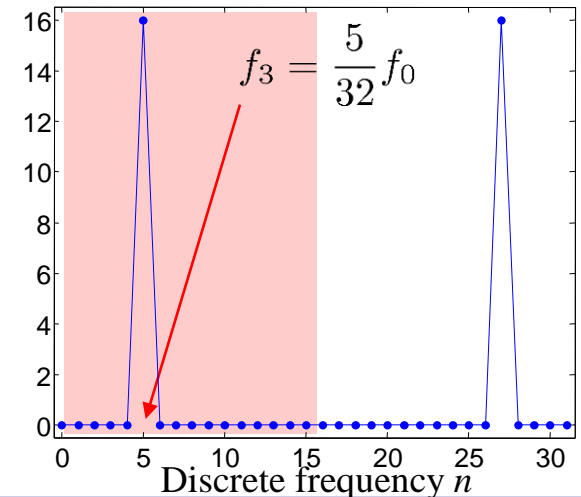
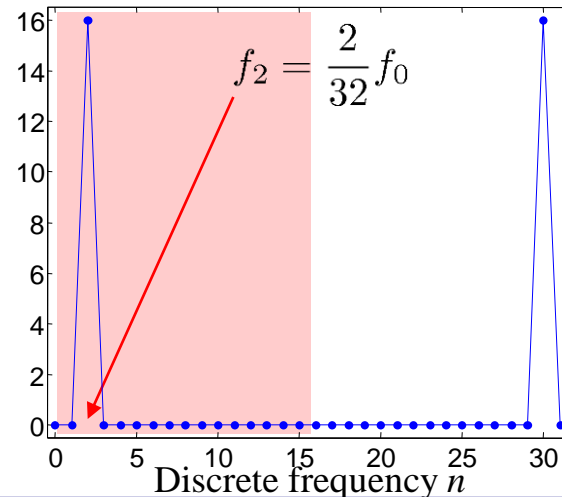
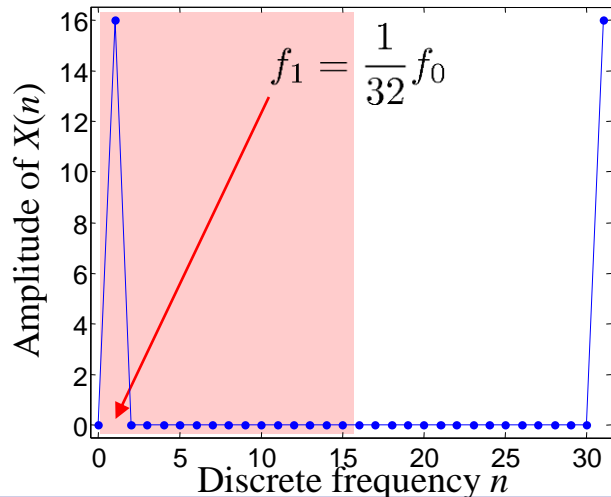
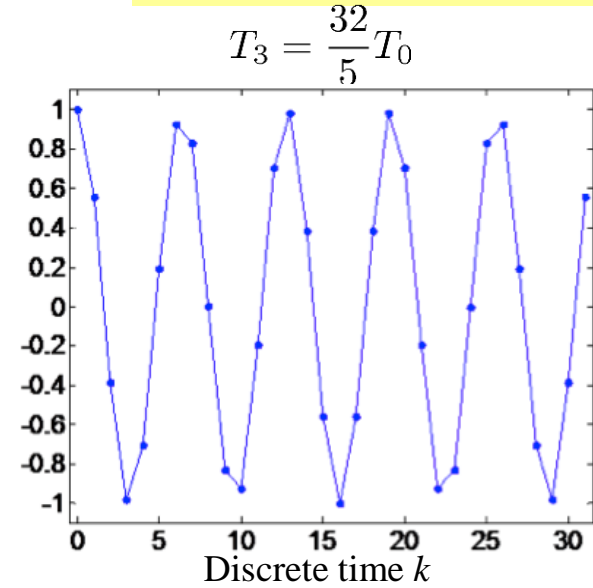
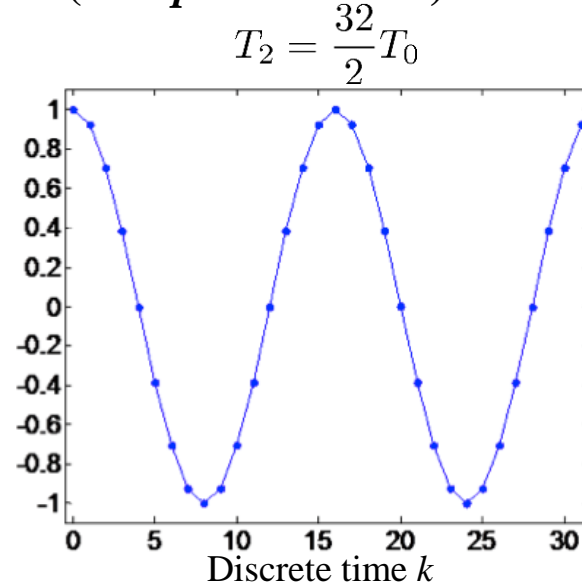
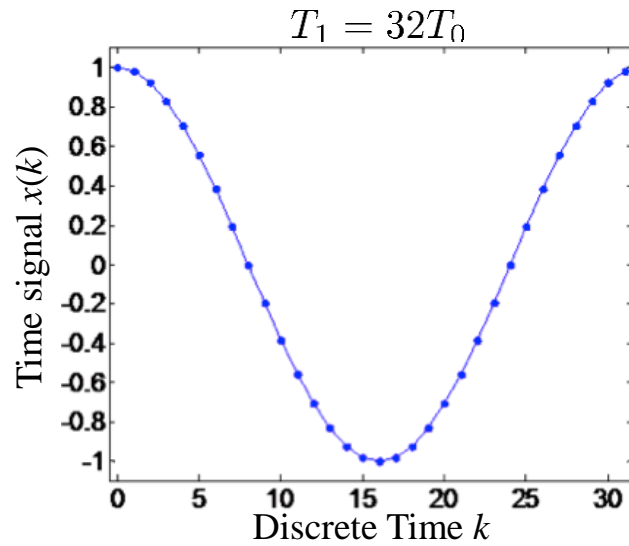
This range is generated by the increase of the sampling frequency.



9.3 Frequency Analysis Via DFT

DFT of Sin- or Cos-Type Signals (*Complete Periods*)

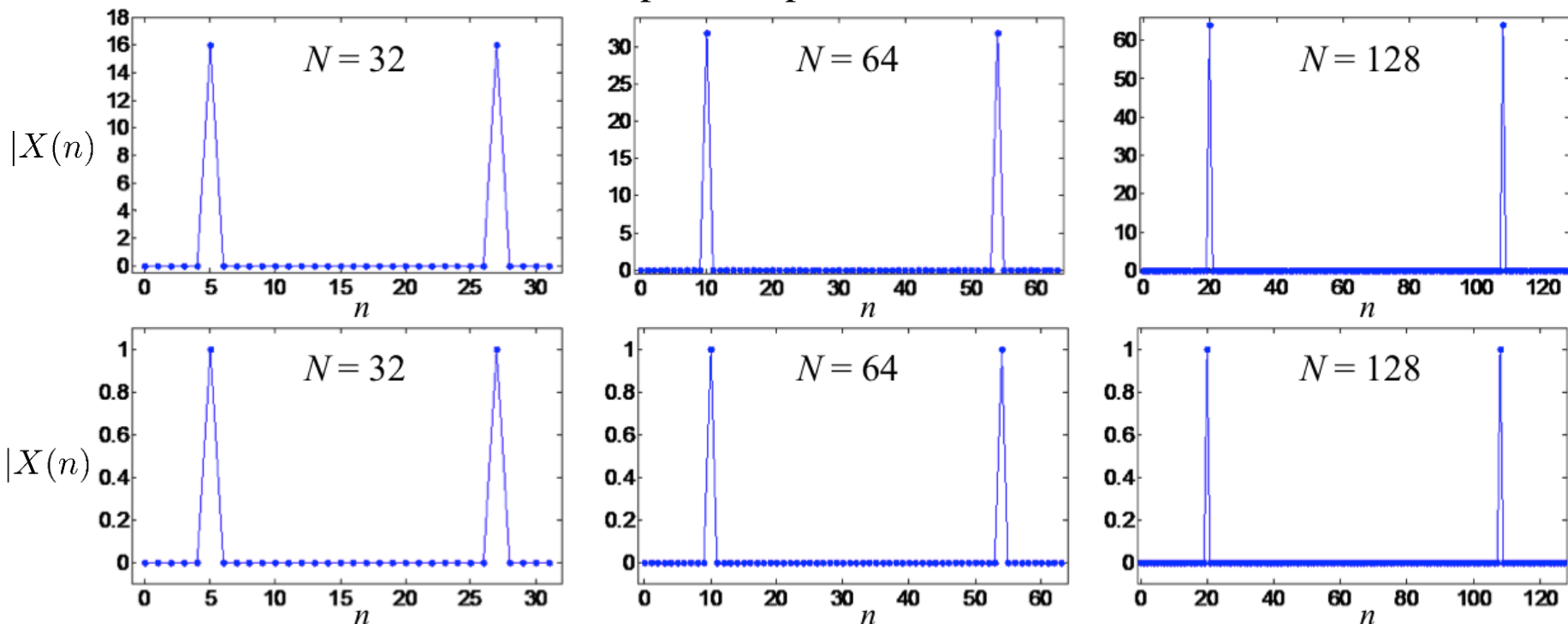
Example:
Cos-type signals with 1, 2, 5 complete periods!
 $N = 32$



9.3 Frequency Analysis Via DFT

Observations:

- The amplitude obtained from the DFT for the signal frequency is $N/2$ if the original signal had amplitude 1. It is clear that this number is proportional to the number of data points N because so many points have to be summed up.
- Thus, the amplitude axes of the frequency response are commonly scaled with a factor $2/N$ to make the axes in the plot independent of N .

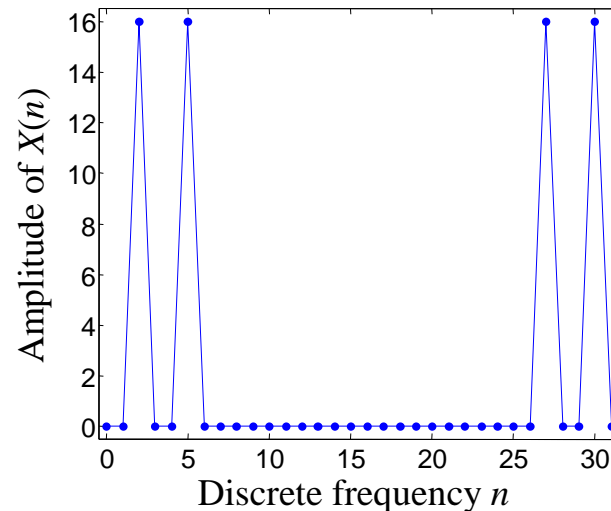
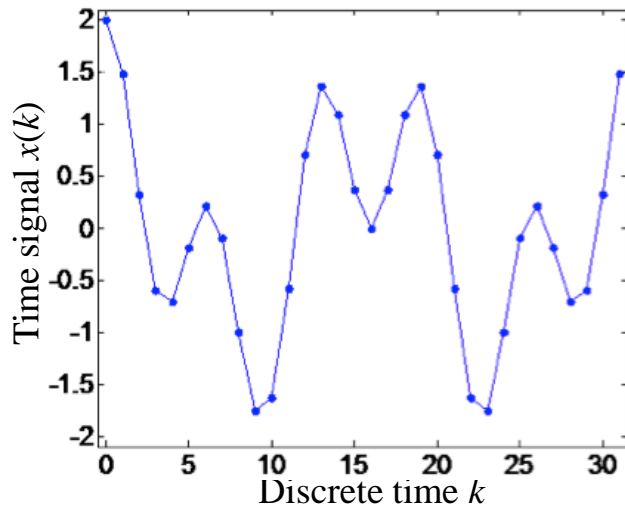


9.3 Frequency Analysis Via DFT

Observations:

- If the length N of the time signal is an *exact multiple* of the period length of an oscillation then the DFT reveals the amplitude of this oscillation exactly in the spectrum:
 - The complete energy is concentrated on one peak (if we have just one oscillation).
 - This peaks lies exactly at the correct frequency.
- Due to the linearity property of the DFT these facts are valid for an additive mixture of oscillations, as well.

Example: Superposition of two oscillations at $f_2 = 2/32 f_0$ and $f_3 = 5/32 f_0$:



9.3 Frequency Analysis Via DFT

Reason for the Exact Frequency Representation

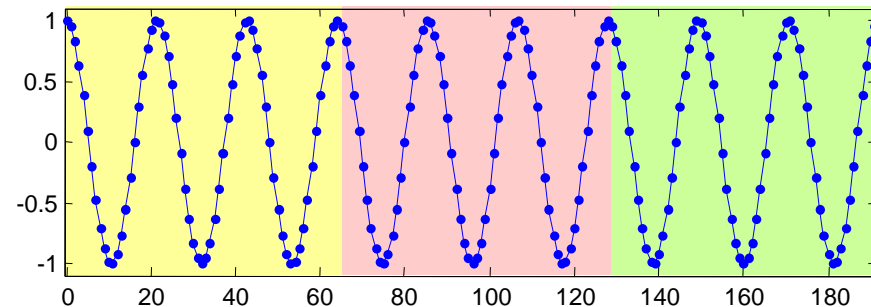
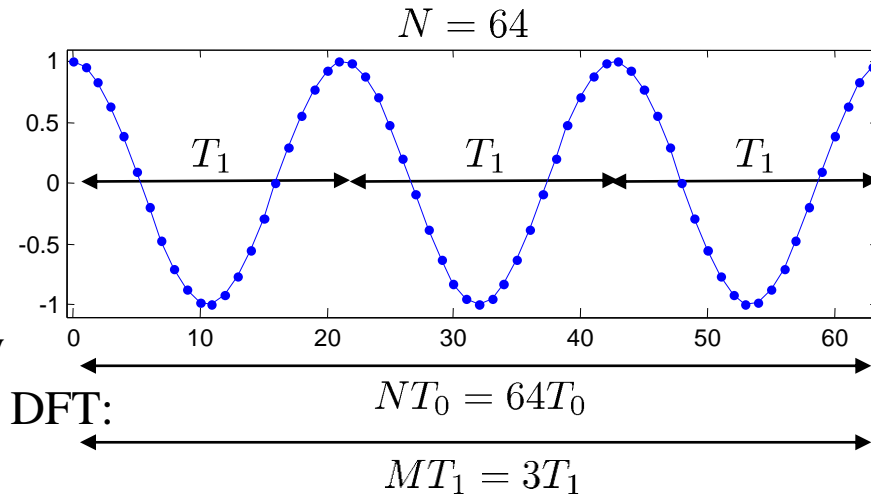
1. If a time signal $x(k)$ with $k = 0, 1, \dots, N-1$ contains exactly M periods of a sin- or cos-signal of duration T_1 this holds:

$$MT_1 = NT_0$$

Thus the frequency f_1 automatically is exactly equal to one of the discrete frequencies of the DFT:

$$T_1 = \frac{N}{M}T_0 \rightarrow f_1 = \frac{M}{N}f_0$$

2. Due to periodicity of the complex exp-function the DFT “thinks” the signal repeats itself infinitely often, i.e., the original signal for $k = 0, 1, \dots, N-1$ is repeated for $k = N, N+1, \dots, 2N-1$ and $k = 2N, 2N+1, \dots, 3N-1$, etc. Because the oscillation are full periods, they fit together exactly at the points $N, 2N$, etc. (continuity).

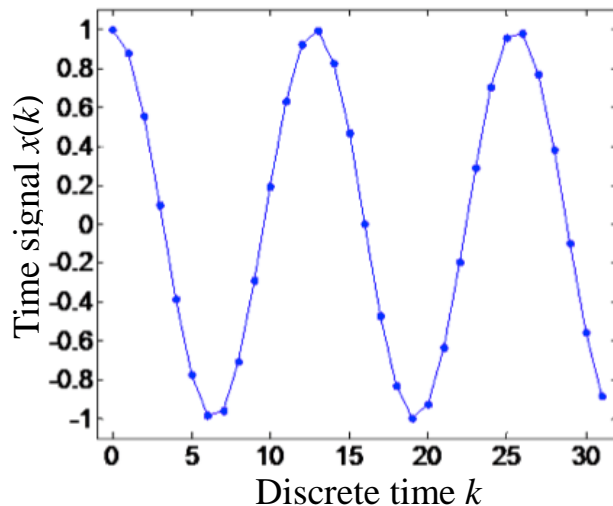


9.3 Frequency Analysis Via DFT

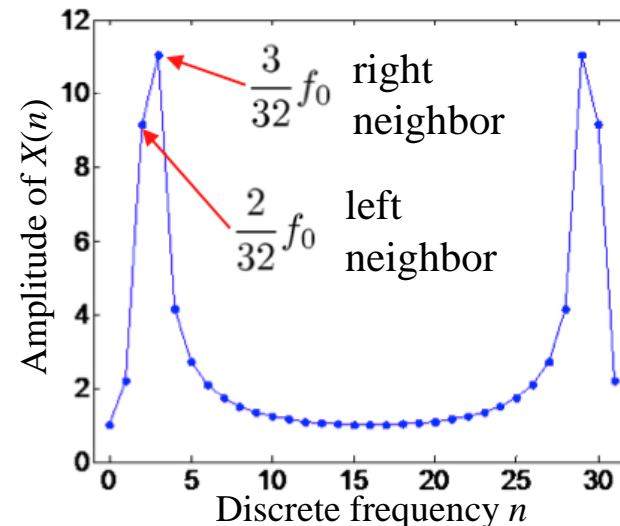
DFT of incomplete sinus-type signals

- Typically it is not possible to choose N such that all included oscillations exhibit an integer multiple of periods. Reasons:
 - The period length of the interesting oscillation is not known.
 - Many oscillations of various period lengths are interesting and it is impossible to find a reasonable value for N fulfills all conditions concurrently.

What happens if an oscillation is not present for an integer number of periods?



$$T_1 = \frac{32}{2.5} T_0$$



Lies between $n = 2$ and $n = 3$.

$$f_1 = \frac{2.5}{32} f_0$$

9.3 Frequency Analysis Via DFT

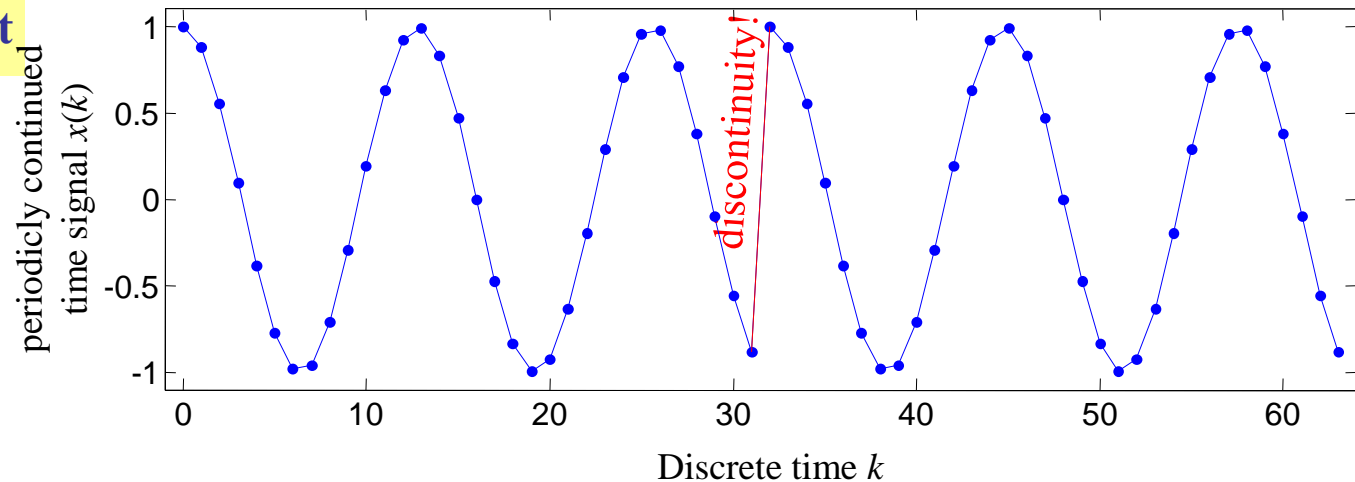
Observations:

- The frequency f_1 of a periodic signal does not exactly exist in the frequency discretization! Therefore the amplitude belonging to $2.5/32 f_0$ splits between $2/32 f_0$ and $3/32 f_0$.

→ Picket Fence Effect

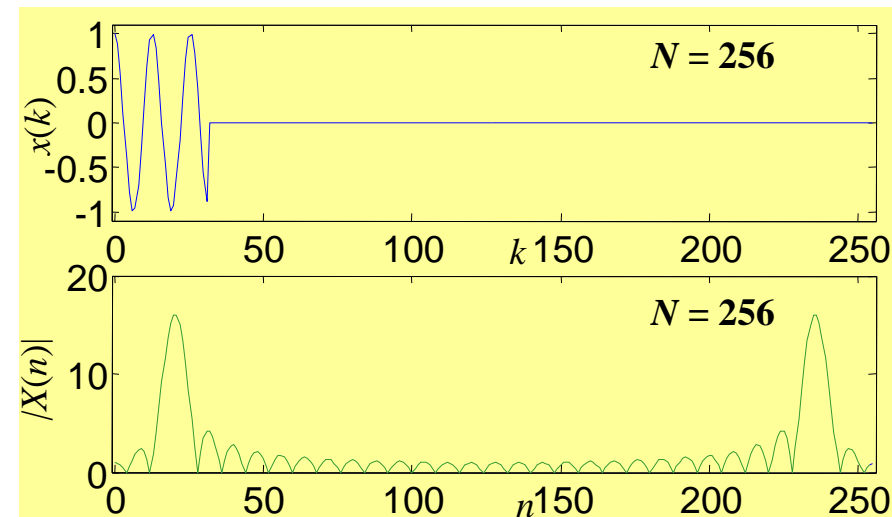
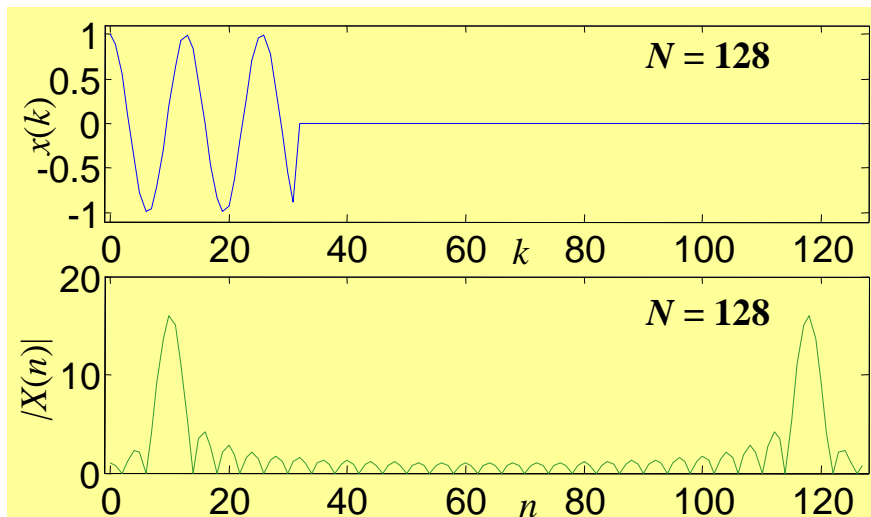
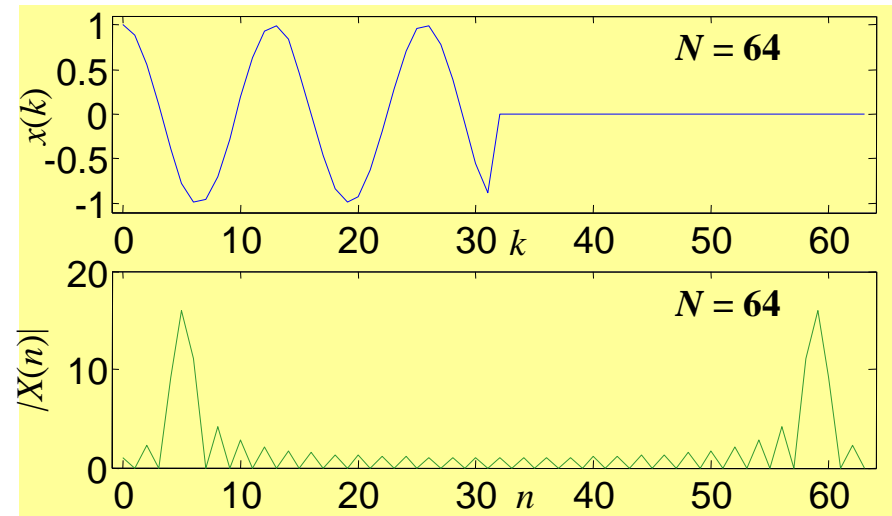
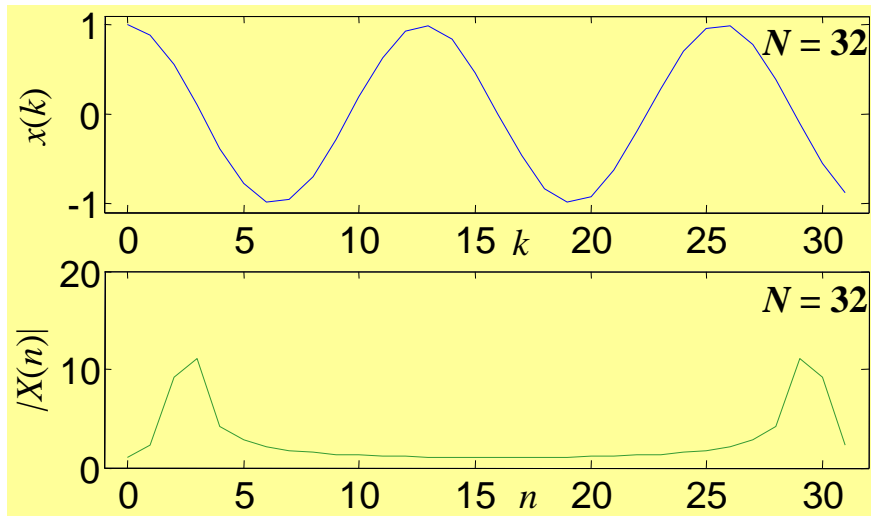
- Additionally the spectrum “smears” (*leaks*) across the whole frequency range. This is a direct consequence of the **discontinuity** of the time signals that induces disturbing “steps” in the (thought) periodic signal.

→ Leakage Effect



9.4 Leakage Effect and Windowing

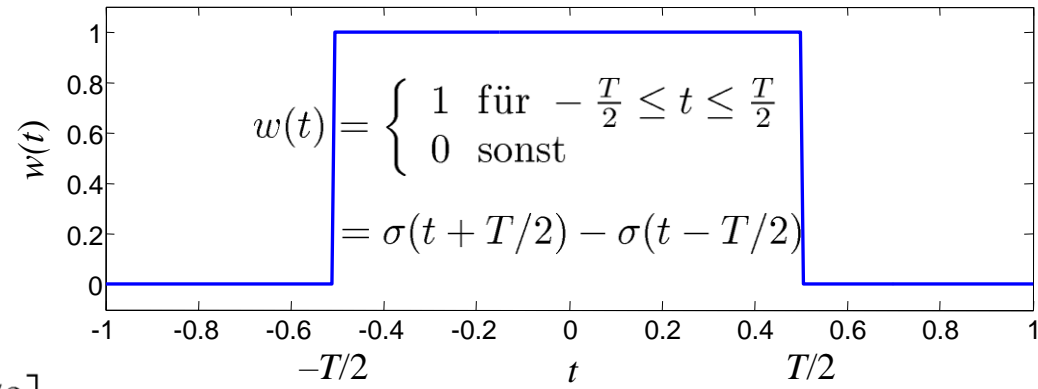
Identical Example for Different Resolutions, i.e., of Different Lengths N



9.4 Leakage Effect and Windowing

Observations:

- For $N \rightarrow \infty$ the DFT result converges to the amplitude and phase response.
- We have to distinguish two negative effects that can occur: (i) The maximum amplitude is split into its neighbors due to discretization (*picket fence effect*) and (ii) the spectrum is smeared across (*leakage effect*).
- A rectangular window has a sinc-function as Fourier Transform. The original signal can be thought of as a multiplication with the **rectangle** or convolution with *sinc()*.

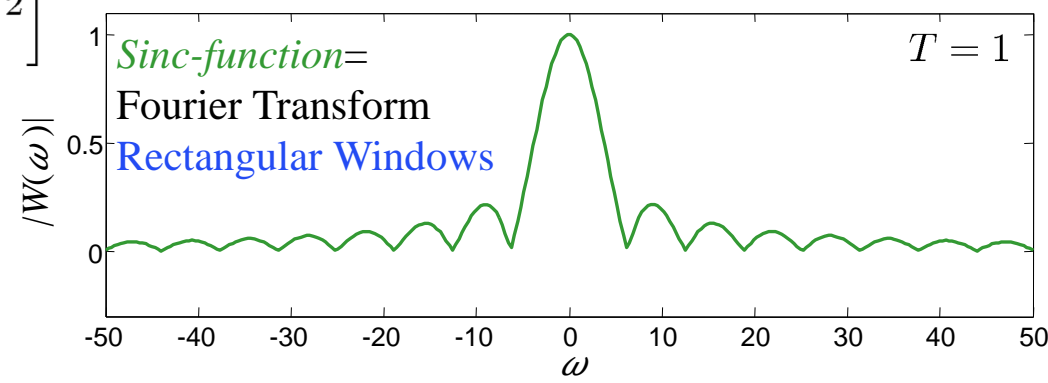


$$W(\omega) = \mathcal{F}\{w(t)\} = \frac{1}{i\omega} \left[e^{i\omega T/2} - e^{-i\omega T/2} \right]$$

$$= \frac{i2\sin(\omega T/2)}{i\omega} = \frac{\sin(\omega T/2)}{\omega/2}$$

$$= \frac{\sin(\omega/2)}{\omega/2} = \frac{\sin(\tilde{\omega})}{\tilde{\omega}} = \text{sinc}(\tilde{\omega})$$

for $T = 1$



9.4 Leakage Effect and Windowing

Explanation of the Leakage Effect

- A band-limited time signal $x(k)$ of length L can be created from a signal of length infinity or large N by multiplication with a rectangular window $w(k)$ of length L :

$$x(k) = w(k) \cdot x_p(k)$$

- This multiplication in the time-domain corresponds to a convolution in the frequency-domain:

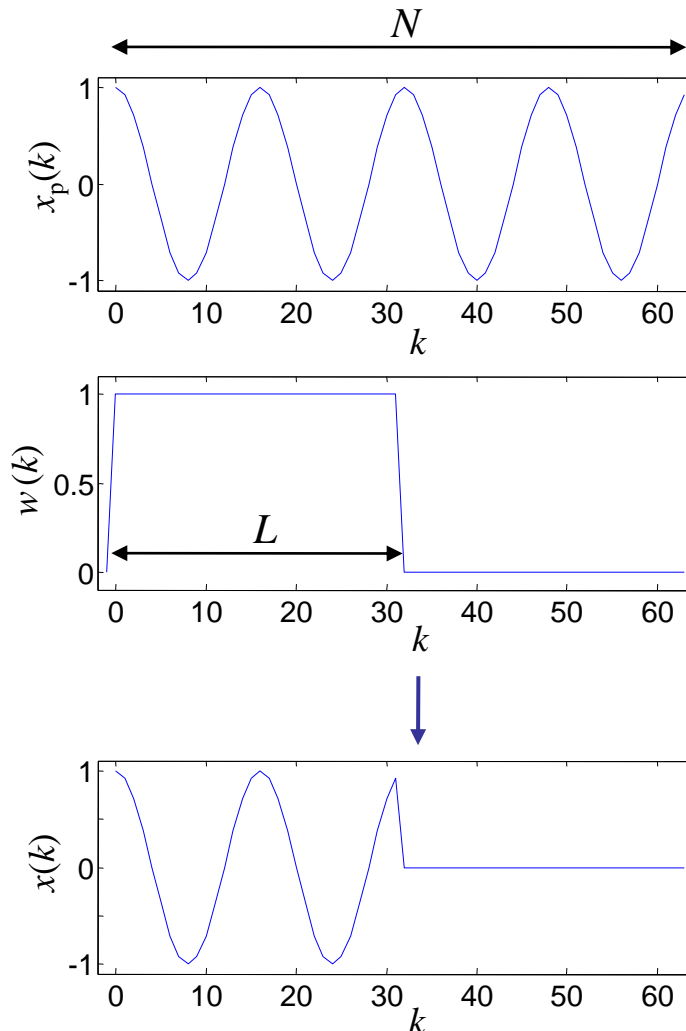
$$X(n) = W(n) * X_p(n)$$

Here $W(n)$ is the Fourier transform and DFT of the rectangular window $w(k)$:

$$W(\Omega) = \frac{\sin(L\Omega/2)}{\sin(\Omega/2)} e^{-i(L-1)\Omega/2}$$

$$W(n) = \frac{\sin(\pi Ln/N)}{\sin(\pi n/N)} e^{-i\pi(L-1)n/N}$$

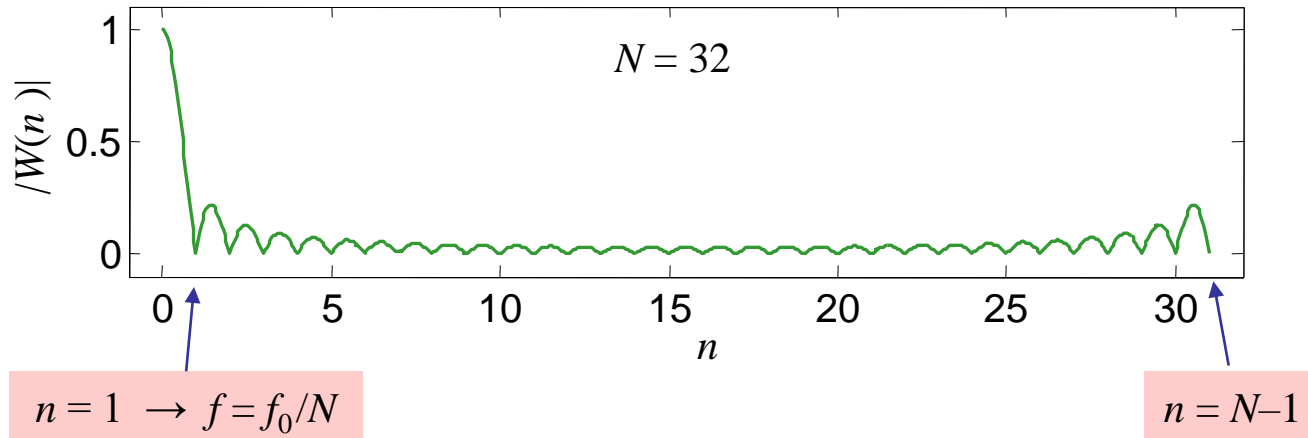
$$L = 32, N = 64 > 32$$



9.4 Leakage Effect and Windowing

The DFT of a rectangular window of length L looks like the sinc-function. In practice, usually $L = N$. Zero-padding is equivalent with $L < N$ since anyway $w(k) = x(k) = 0$ for $k > L$:

$$W(n) = \frac{\sin(\pi n)}{\sin(\pi n/N)} e^{-i\pi(N-1)n/N} \quad \rightarrow \quad |W(n)| = \left| \frac{\sin(\pi n)}{\sin(\pi n/N)} \right|$$



The zeros of the DFT of the rectangular window of length N lie at multiples of f_0/N . If the time signal is an oscillation of frequency Mf_0/N , then the zeros are at integer values of n . This means that in this case a convolution with such a signal is trivial and no leakage effect results.

9.4 Leakage Effect and Windowing

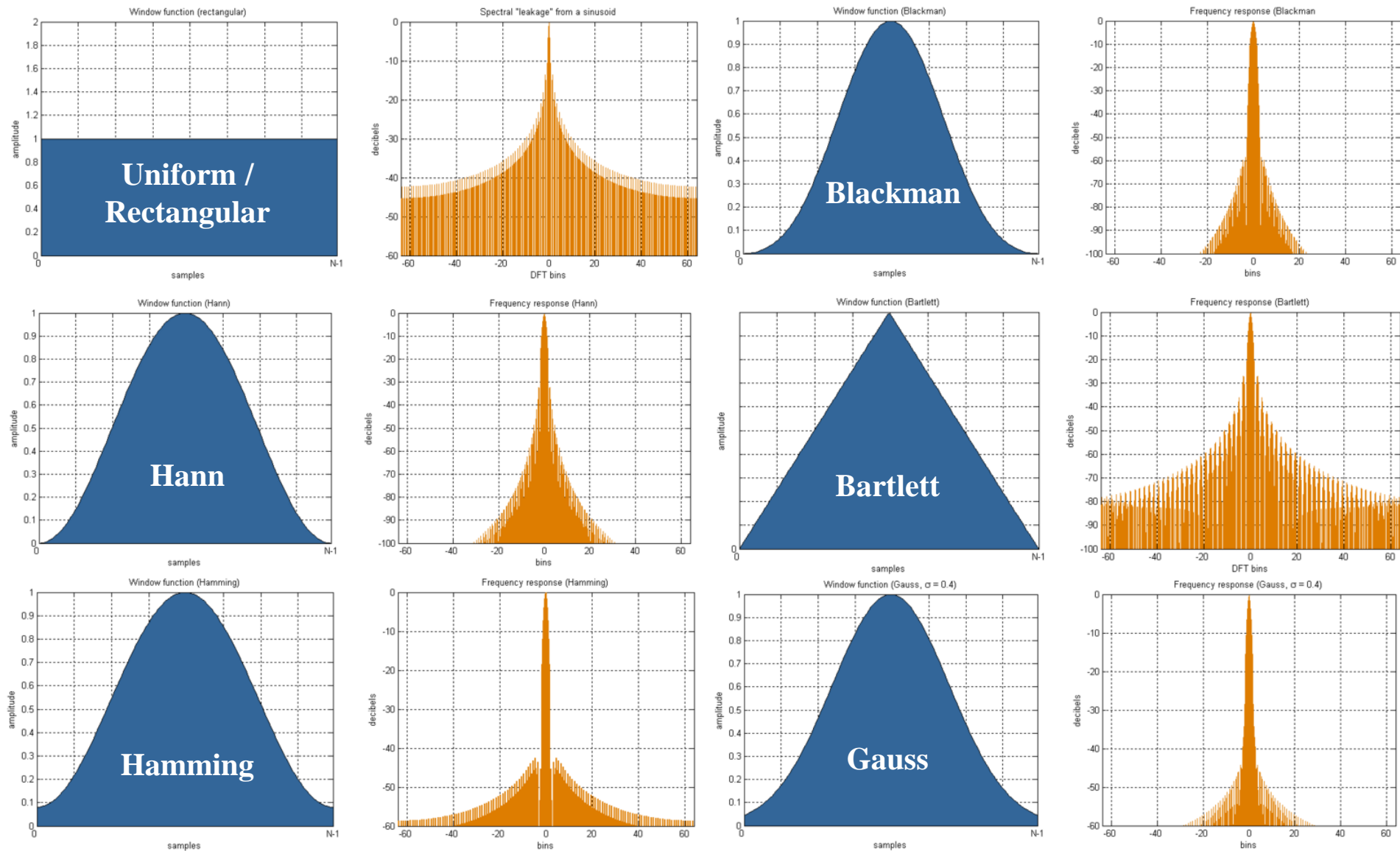
Summary: Rectangular Window

- Every finite real signal of length N thus can be *thought* of being constructed by a multiplication of an infinite length signal with period length N by a rectangular signal of length N .
- The rectangular window leads to discontinuities, i.e., abrupt changes. This means high frequencies are induced.
- The errors caused by windowing with a rectangle or not windowing at all (which is the same thing!) thus are extremely large (picket fence and leakage effects)

Room for Improvement

- A smoother shape of the window would help to induce not so high frequencies.
- Many alternative windows are commonly used, see next slide.
- All these windows are similar. They reduce the leakage effect. However, they necessarily distort the signal by their smooth transition at the beginning and end.

9.4 Leakage Effect and Windowing



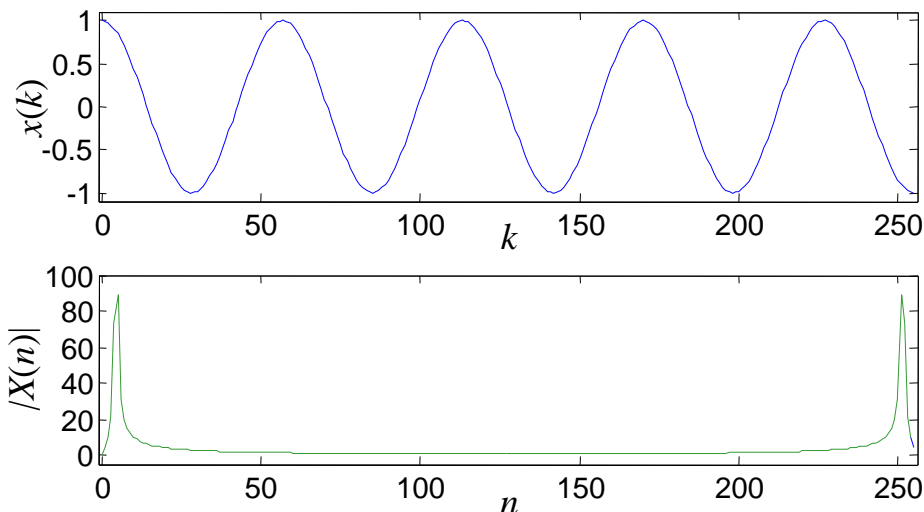
9.4 Leakage Effect and Windowing

Example: Windowing with Uniform/Rectangular and Hann Window

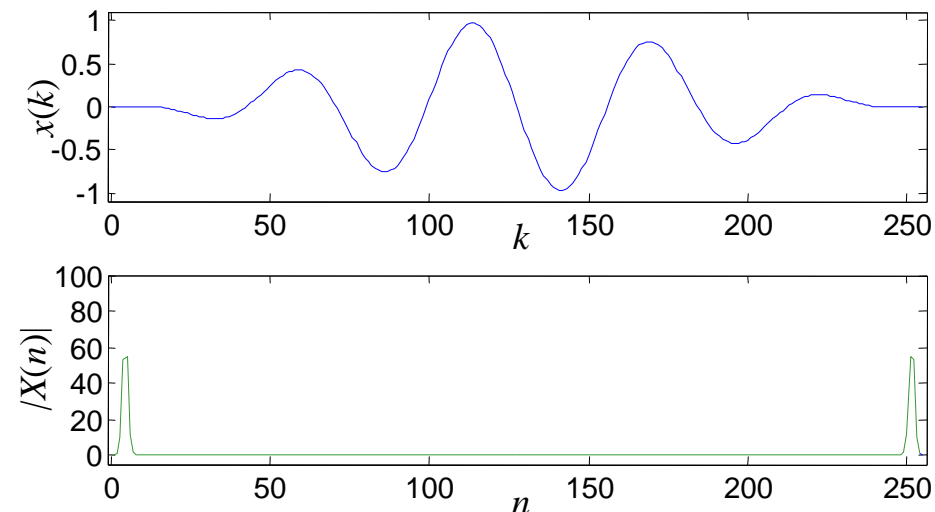
- Using the uniform/rectangular window is like using no window at all.
- The Hann window (and similar alternatives) reduce the leakage effect significantly. By the smoother transitions at the window edges less disturbing high frequencies are induced.
- Hann window of length L (usually $L = N$):

$$w_{\text{Hann}}(k) = 0.5 \left(1 - \cos \frac{2\pi k}{L-1} \right)$$

Uniform/Rectangular $N = 256$

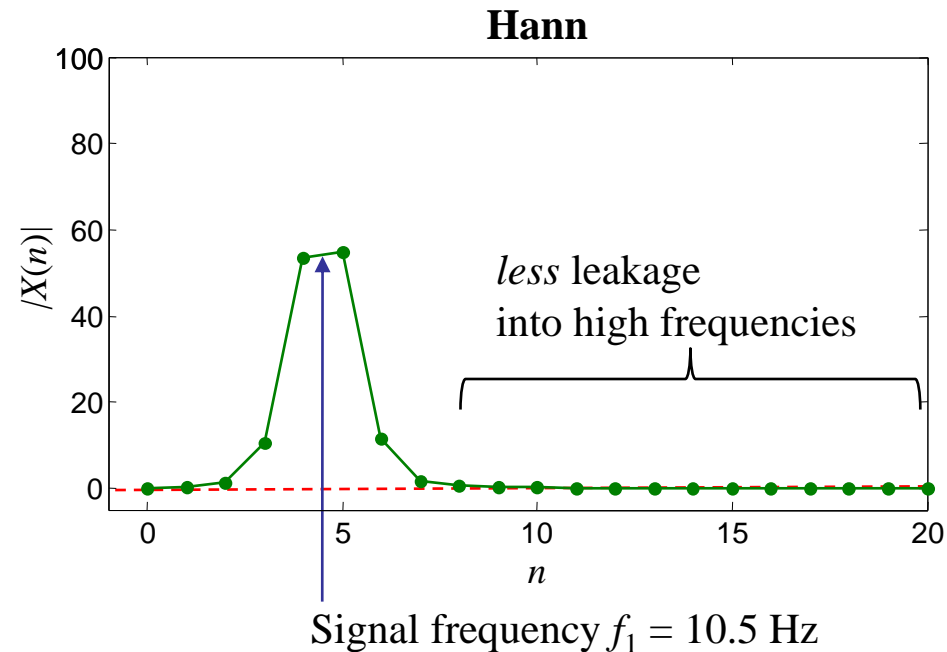
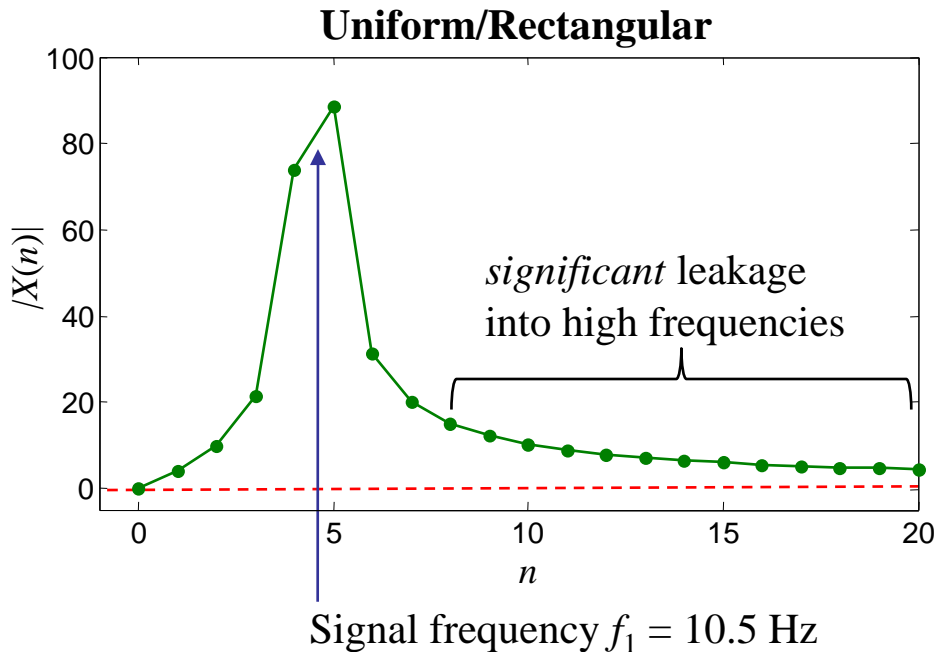


Hann $N = 256$



9.4 Leakage Effect and Windowing

Zoom:



Observations:

- Hann window reduces leakage effect significantly.
- Since the Hann window has a smaller area than the rectangular window signal energy is lost and the amplitudes in the spectrum are smaller. It makes sense to normalize with respect to the window area in order to compensate for this influence.

9.4 Leakage Effect and Windowing

Correction of Signal Damping with Windowing

Windowing distorts the original signal in 2 ways:

- Amplitude: The signal amplitude is reduced
- Energy: The signal energy (effective value RMS, “area under the signal“) is reduced

One of these effects can be corrected by multiplying the DFT with a correction factor (> 1):

Window Type	Correction Amplitude	Correction Energy
Uniform/Rectangular	1	1
Hann	2	1,63
Hamming	1,85	1,59
Blackman	2,8	1,97

Source: <https://community.plm.automation.siemens.com/t5/Testing-Knowledge-Base/Window-Correction-Factors/ta-p/431775>

9.5 Non-Stationary Signals and Short-Term-DFT

Stationary Signals:

- Signals that *do not change* their characteristics / properties over *time*.
- Up to this point we implicitly assumed that all signals are stationary.

Non-stationary Signals:

- Signals that *do change* their characteristics / properties over *time*.
- In practice most signals are non-stationary. However, for a short time interval they can be considered, at least approximately, stationary. Examples:
 - Signals with trends, i.e., with slowly changing mean. This is typical for larger time scales. If we look at stock indices over years (not days!). A varying mean changes the d.c. value of the spectrum for $n = 0$ or $f = 0$ Hz
 - By wear the properties of construction elements change over time. Certain signals of machines (rotation speed, sound, ...) might change their characteristics like the frequency of their peak value.
 - Instead of wear also a failure can be the cause for such changes. However, this happens much faster!

9.5 Non-Stationary Signals and Short-Term-DFT

Problem by Applying a Fourier Transform or DFT to Non-Stationary Signals:

- It is averaged by integration or summation over the complete signal. If the spectrum changes over time its frequency components are weight with their relevance.
- The transform reveals no information about *when* which frequency occurs how strongly in the signal!

Solving this Problem

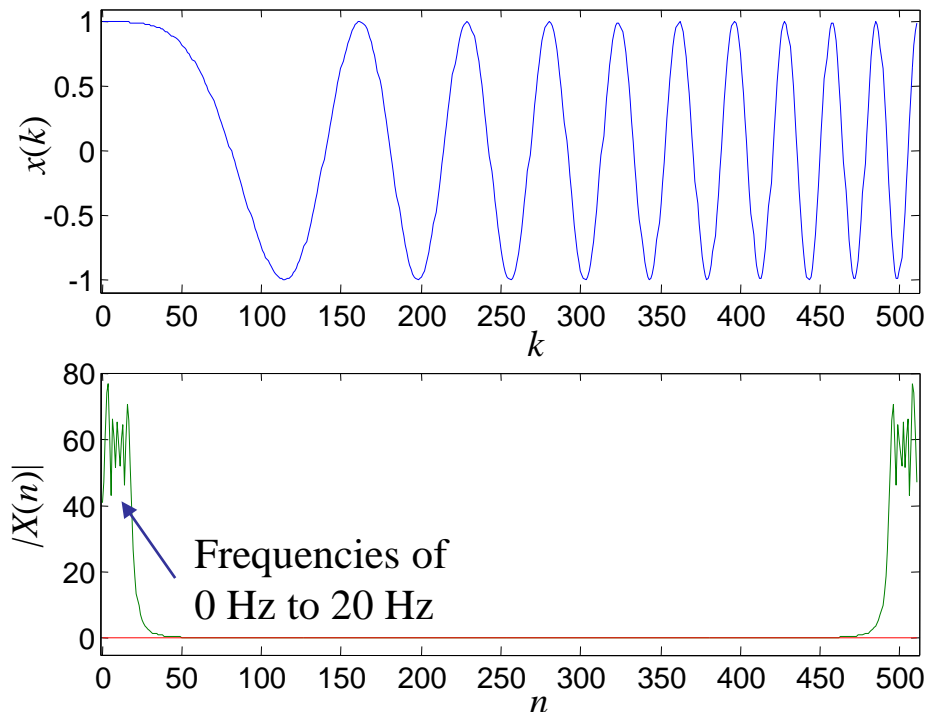
1. Transform only short intervals of the signal into the frequency-domain. Within the short intervals the signal can be assumed to be approximately stationary:
→ Short-time Fourier transform or short-time DFT.
2. Modification of the Fourier Transform such that it does not look for oscillations of infinite length (like the original transform) but rather for wave packages that are active only in certain time intervals:
→ Wavelet transform.

9.5 Non-Stationary Signals and Short-Term-DFT

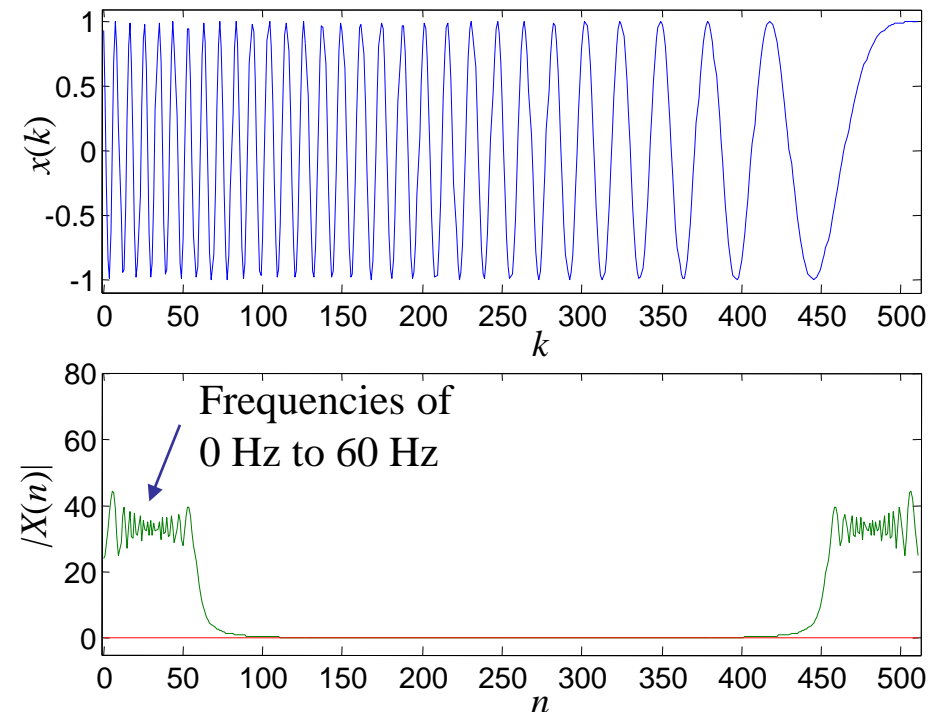
Illustration of the Difficulties by Applying a DFT to Non-Stationary Signals

- Order in which frequencies occur is irrelevant.
- The result (spectrum) is affected by the frequencies according to their time dominance.
- DFT is not meaningful!

Chirp-Signal 0 ... 20 Hz



Chirp-Signal 60 ... 0 Hz



9.5 Non-Stationary Signals and Short-Term-DFT

Short-Time Discrete Fourier Transform (STDFT)

- Windowed DFT
- Width of the window determines the time resolution and also the frequency resolution. The width is a parameter defined by the user. It should be guided by the expected rate of change in the spectrum:
 - Signal changes its frequency properties quickly → narrow window.
 - Signal changes its frequency properties slowly → wide window.
- The DFT does not only depend on the frequency f or n but also on a second variable: the time shift of the window t_0 . It indicates the time t_0 around which the DFT is valid

Windowed Fourier Transform with Window $w(t)$:

$$X_w(f, t_0) = \int_{-\infty}^{\infty} x(t) \cdot w(t - t_0) \cdot e^{-i2\pi ft} dt$$

Windowed DFT with Window $w(k)$:

$$X(n, k_0) = \sum_{k=0}^{N-1} x(k) \cdot w(k - k_0) \cdot e^{-i2\pi nk/N}$$

9.5 Non-Stationary Signals and Short-Term-DFT

Gaussian as Window

- Strongly decreasing form from center towards outer regions.
- Symmetrical.
- Fourier transform of a Gaussian is again a Gaussian, i.e., it is symmetrical in its time-frequency properties.

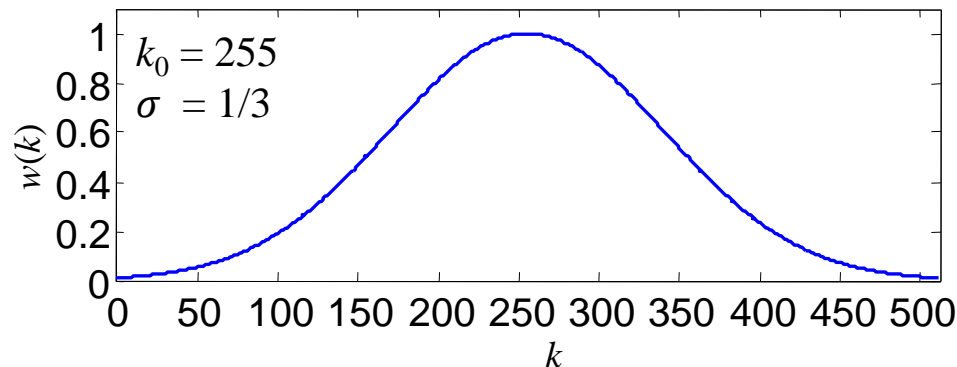
Gauss-Window for Fourier Transform

$$w(t, t_0) = e^{-\frac{1}{2} \left(\frac{t-t_0}{\sigma} \right)^2}$$

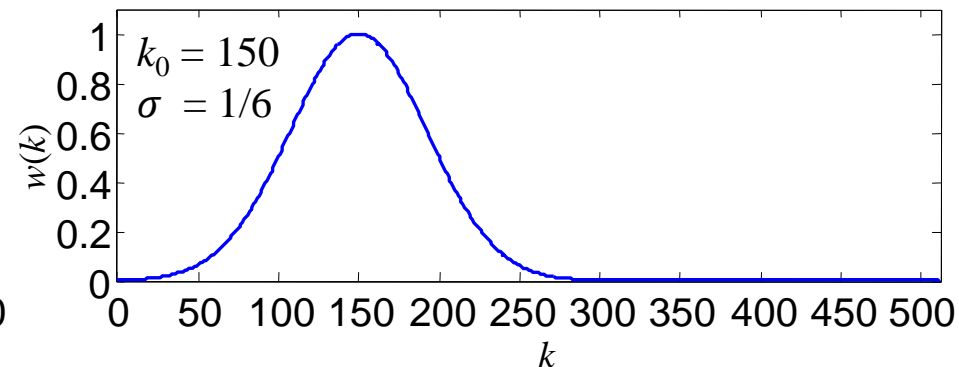
Gauss-Window for DFT

$$w(k, k_0) = e^{-\frac{1}{2} \left(\frac{k-k_0}{\sigma(N-1)/2} \right)^2}$$

Gauss-Window for DFT

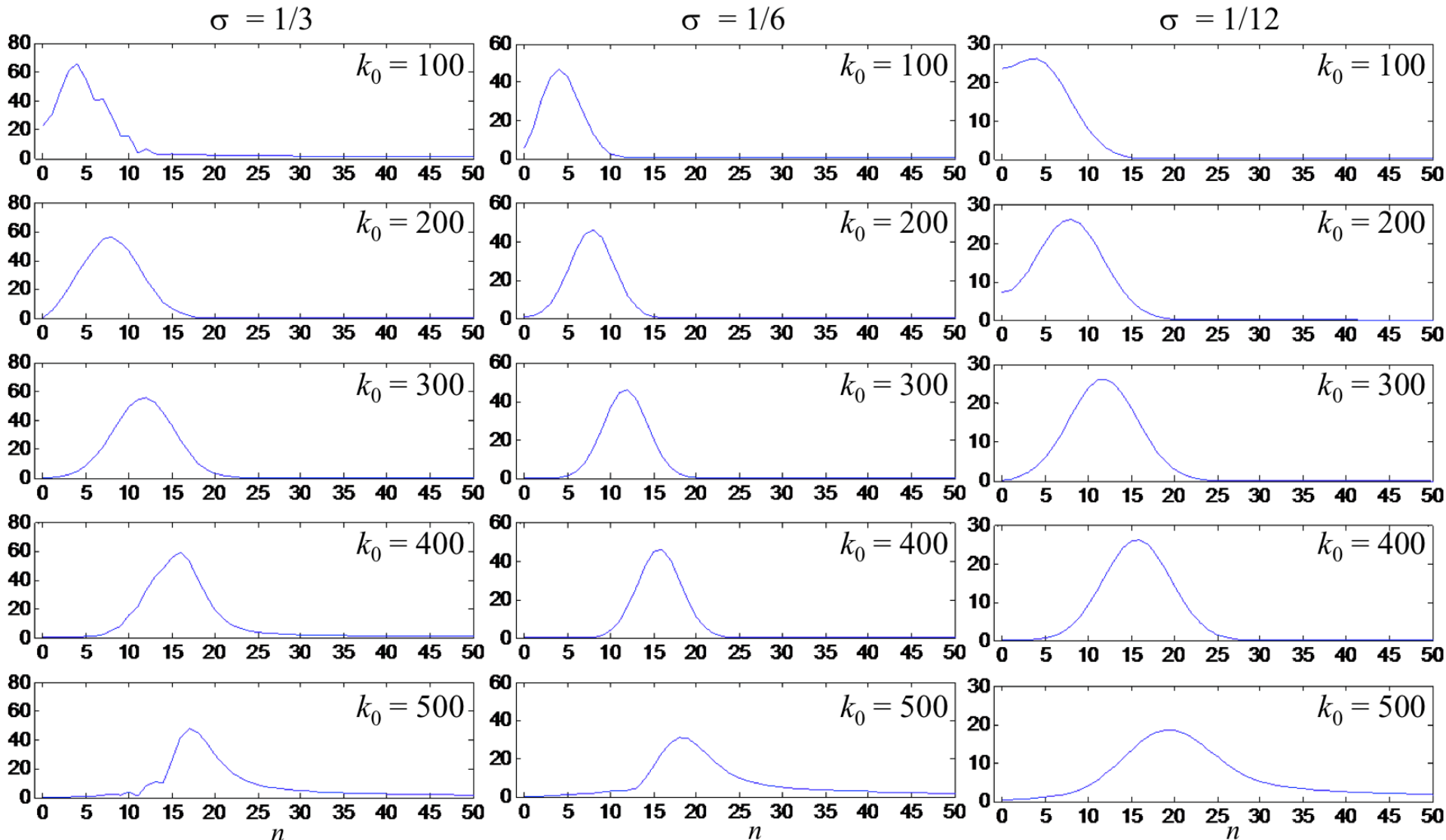


Gauss-Window for DFT



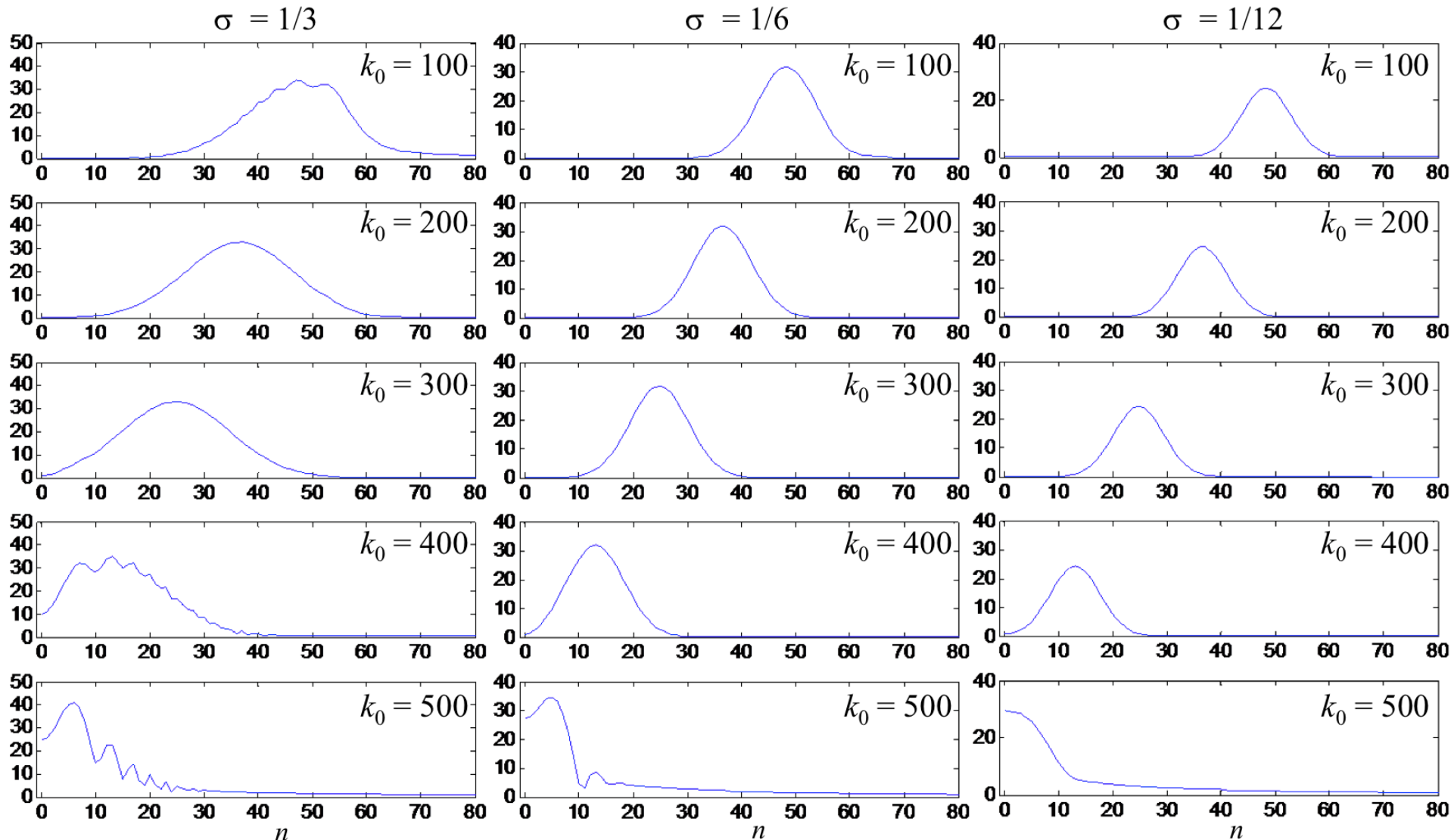
9.5 Non-Stationary Signals and Short-Term-DFT

Short-time DFTs of the 1. chirp-signal with a Gauss-window



9.5 Non-Stationary Signals and Short-Term-DFT

Short-time DFTs of the 2. chirp-signal with a Gauss-window



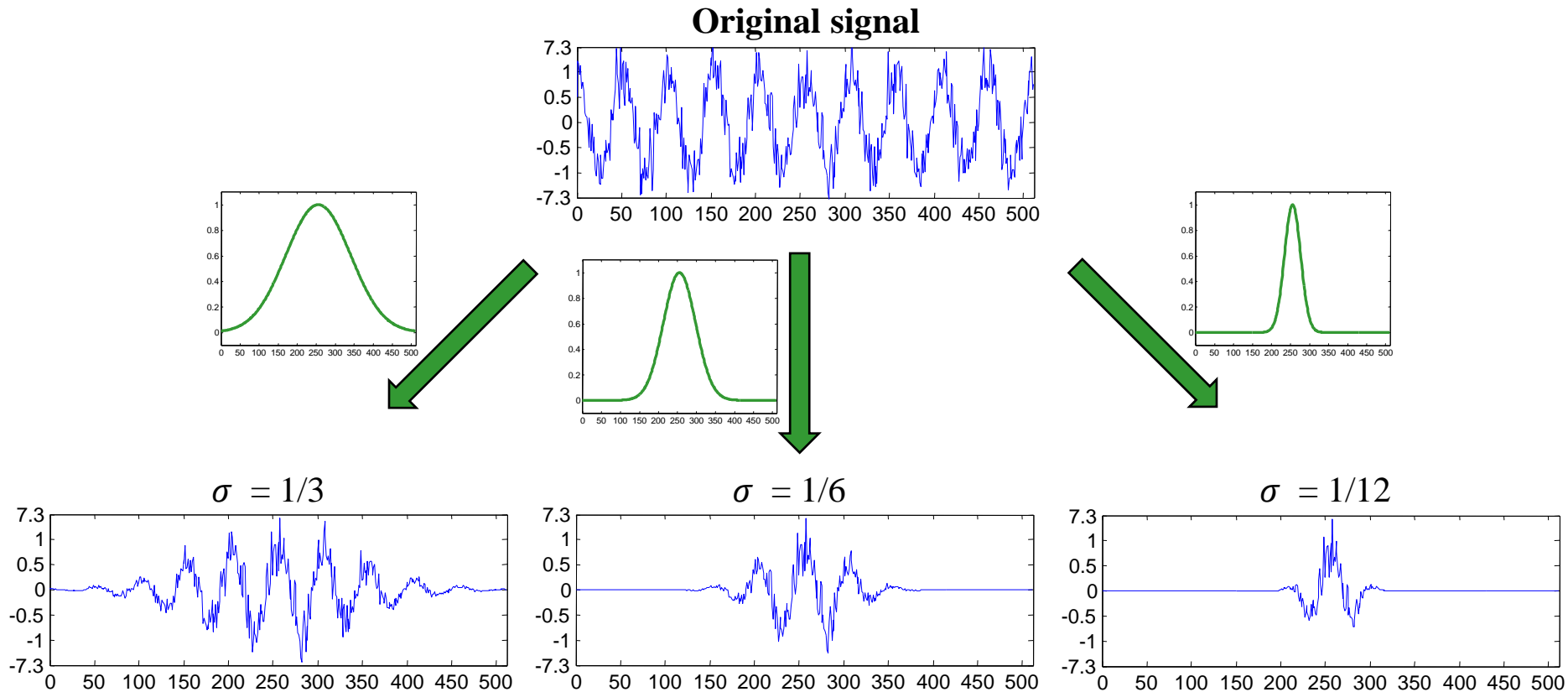
9.5 Non-Stationary Signals and Short-Term-DFT

Observations:

- By shifting the window via k_0 the time range that shall be analyzed can be selected.
- Because the window is not infinitesimally narrow all signal properties inside mix.
- A too wide window with respect to the signal spectrum change rate ($\sigma = 1/3$) yield an unnecessarily large averaging effect over time.
- At $k_0 = 500$ the leakage effect is easy to see. The reason for this is as follows: The Gauss-window is close to the end of the data at 511 and it has significant values where the data stops. This induces similar high frequencies like a uniform/rectangular window.
- For the 2. chirp-signal even the width $\sigma = 1/6$ is a bit too wide. That can be seen in the low quality of the bottom plot. That is because the 2. chirp-signal changes its frequency 3 times as fast as the 1.
- The window should not be chosen too narrow to ensure a certain robustness, see next slides.

9.5 Non-Stationary Signals and Short-Term-DFT

Effect of window width in a short-time DFTs of noisy signals



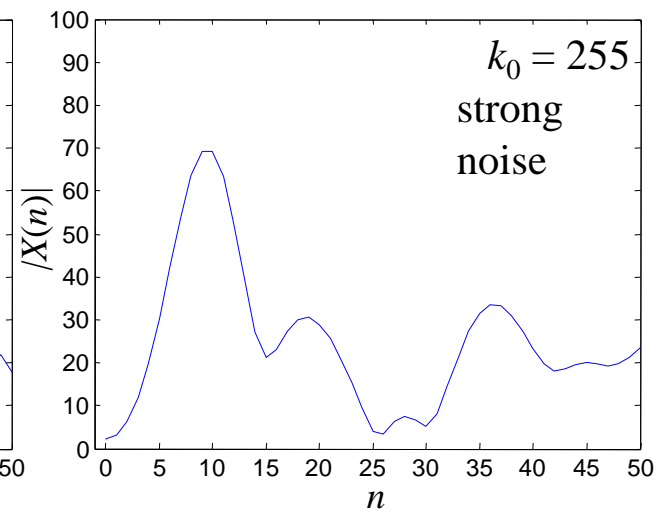
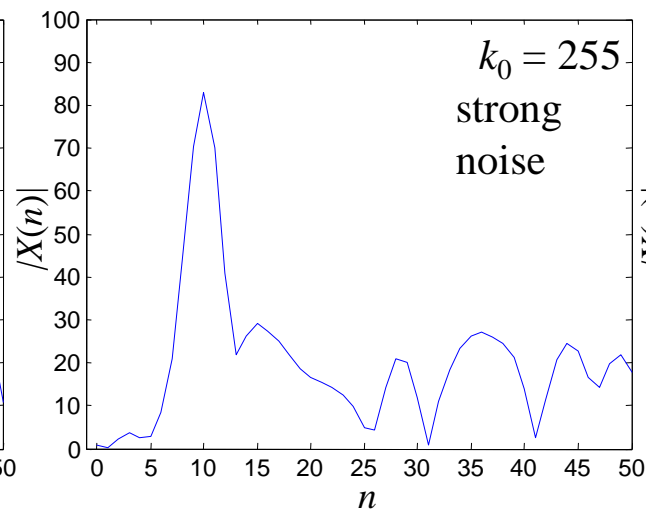
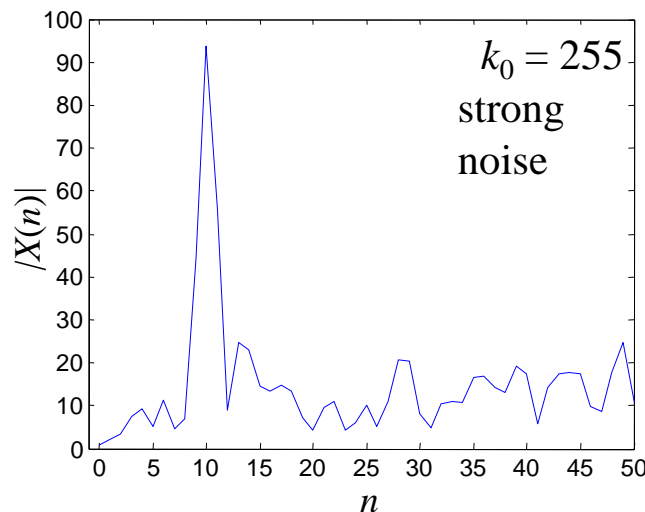
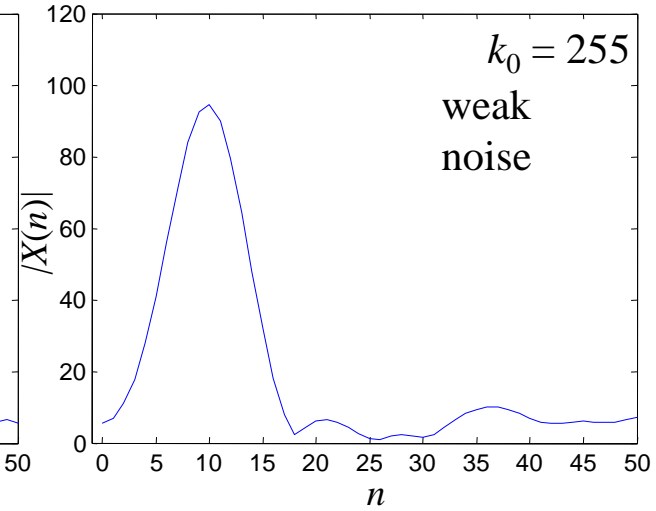
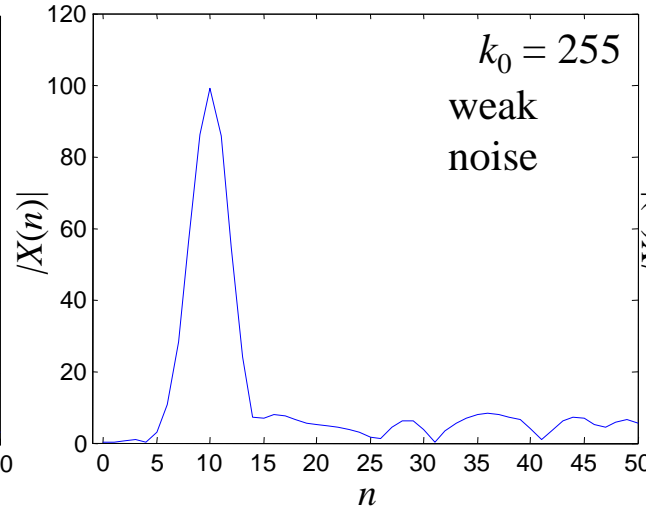
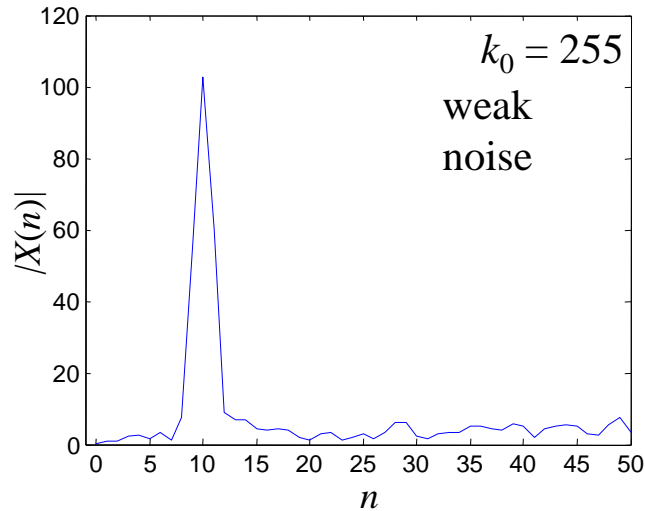
9.5 Non-Stationary Signals and Short-Term-DFT

Short-time DFTs of *noisy* 1. Chirp-signal with a Gauss-window (normalized w.r.t. window area!)

$\sigma = 1/3$

$\sigma = 1/6$

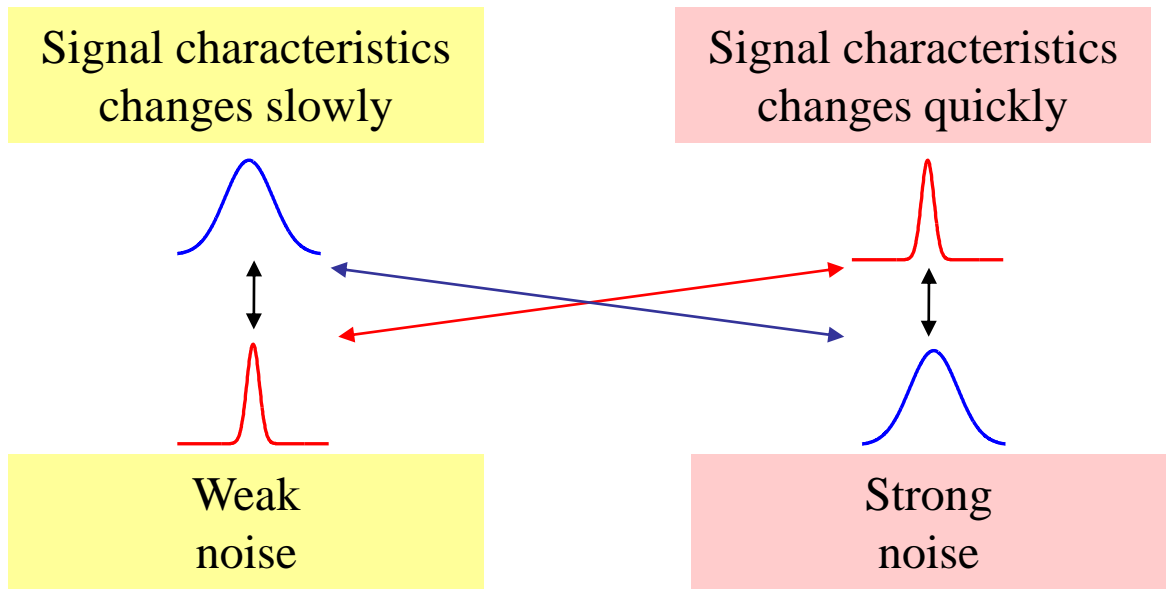
$\sigma = 1/12$



9.5 Non-Stationary Signals and Short-Term-DFT

Observations:

- The window width determines theoretically the maximal possible resolution. The wider a window is the more accurate the frequencies can be determined.
- The window width determines the robustness with respect to the noise in the original signal. The wider a window is the less significant the noise deteriorates the result. Wider windows mean more data is utilized!
- The optimal window width is a compromise between both goals:



Optimal Window Width:

Compromise:

Optimal Window Width:

9.6 Outlook: Time-Frequency-Analysis

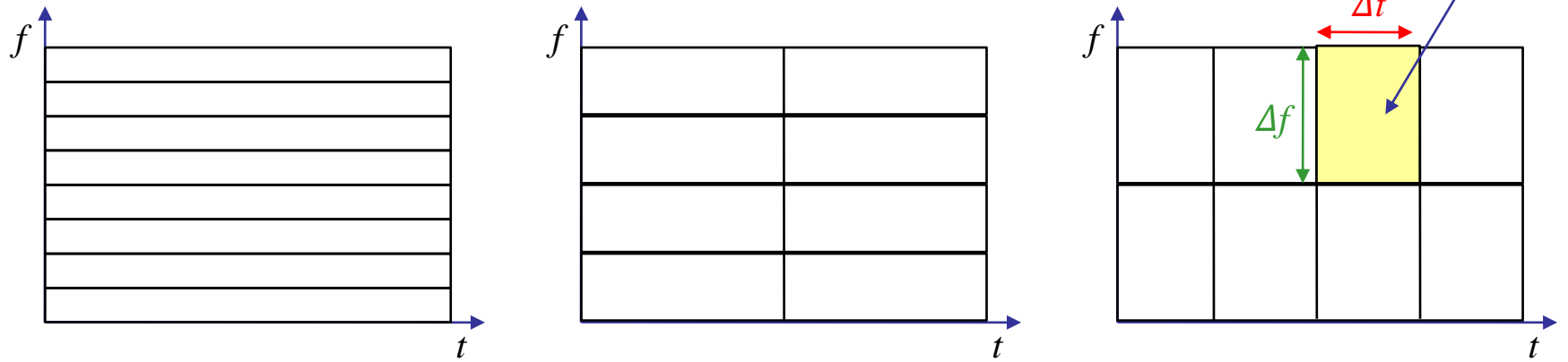
Heisenberg's Uncertainty Principle:
 Energy / Time: $\Delta E \Delta t \geq \text{const.}$
 Impuls w / Position: $\Delta p \Delta x \geq \text{const.}$

Goals of Time-Frequency Analysis

- Good overview on *which frequencies* are present at *what times*.
- Illustration: Strength of frequency component by grey tones:
white = frequency does not exist / *black* = frequency is strongly present
- Best possible resolution in time Δt and frequency (or energy) Δf are coupled by *Heisenberg's uncertainty principle* and thus relate anti-proportionally:

$$\Delta t \sim \frac{1}{\Delta f} \quad \text{or} \quad \Delta t \cdot \Delta f = \text{const.}$$

- With the width of the window in the short-time DFT not only the time resolution Δt but implicitly also the frequency resolution Δf is fixed.



9.6 Outlook: Time-Frequency-Analysis

Example 1: Analysis of a periodic signal with varying frequency over time

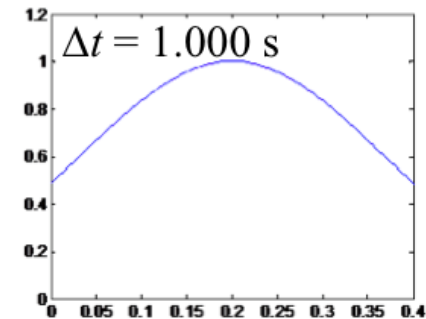
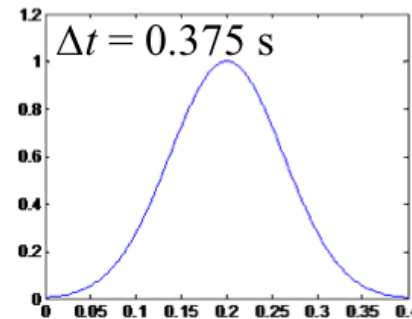
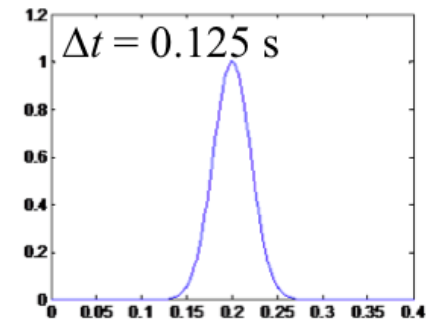
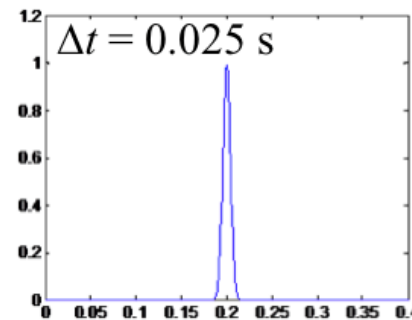
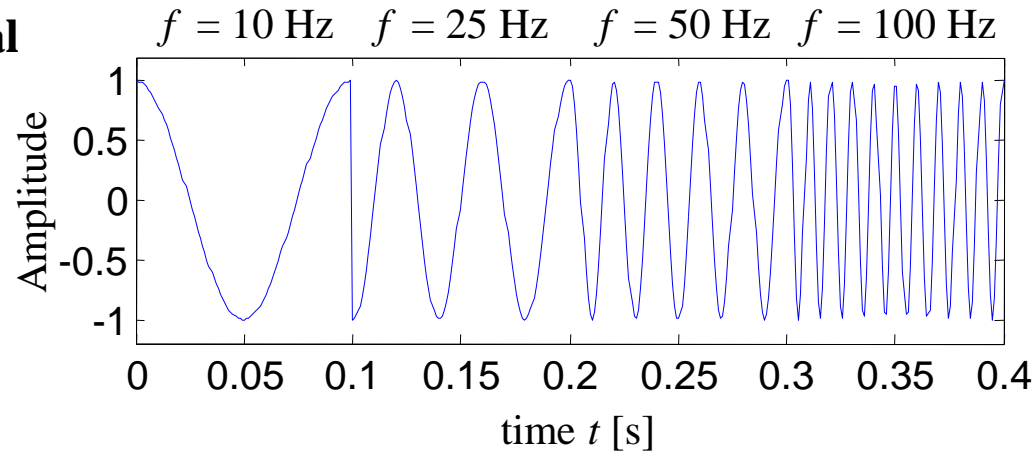
Goal of a short-time DFT:

Frequency analysis of the signal in dependency of time. We want to know when which frequency occurs.

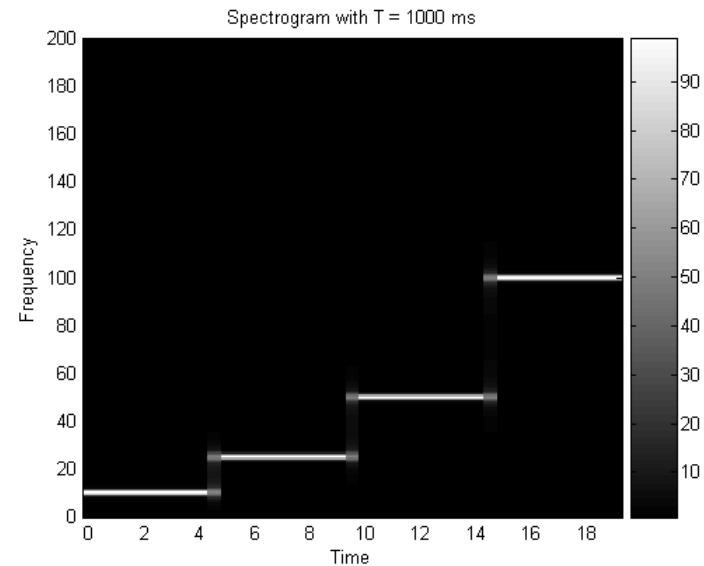
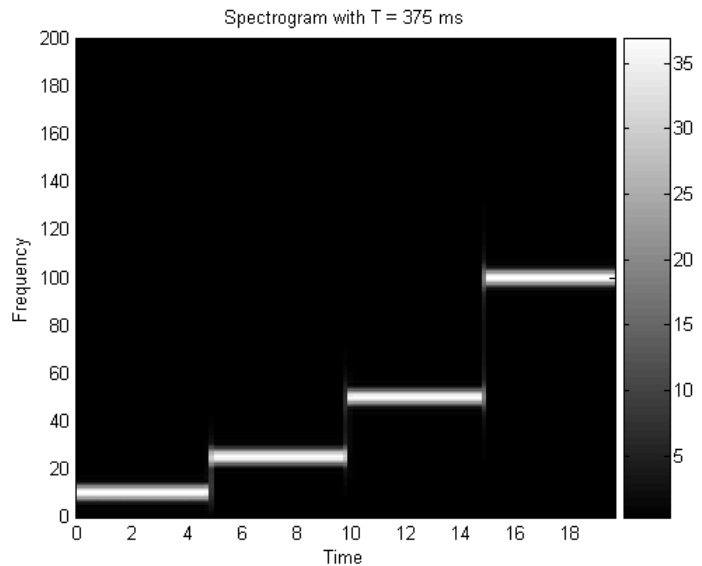
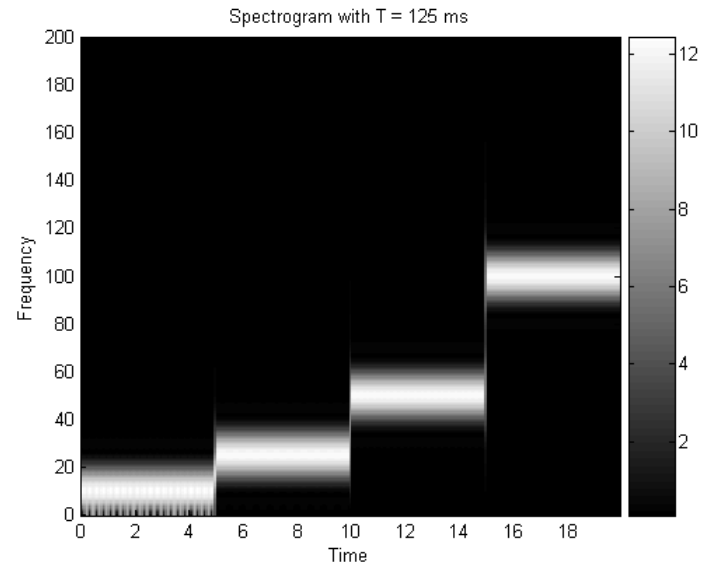
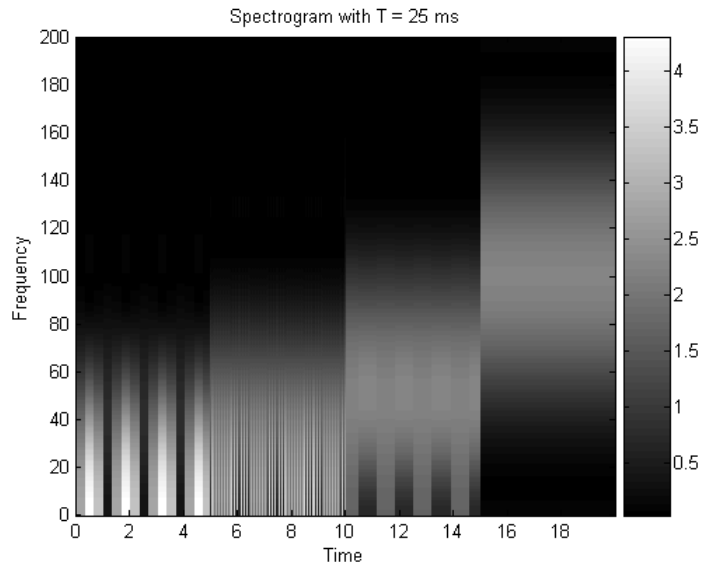
Choice for window width:

- Determines the time resolution Δt .
- Implicitly also determines the frequency resolution Δf because both are anti-proportional:

$$\Delta t \sim \frac{1}{\Delta f}$$



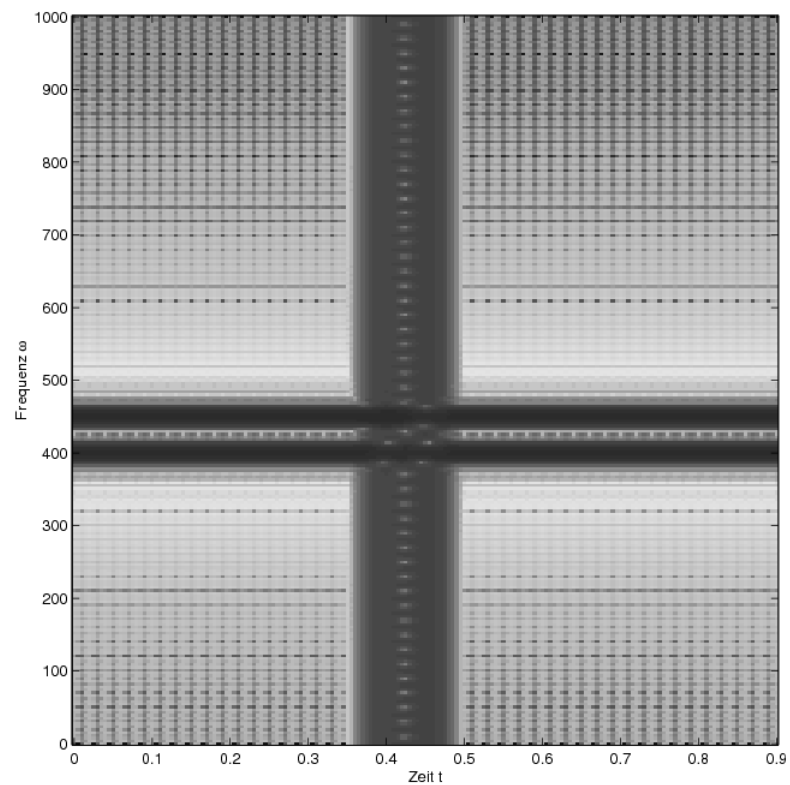
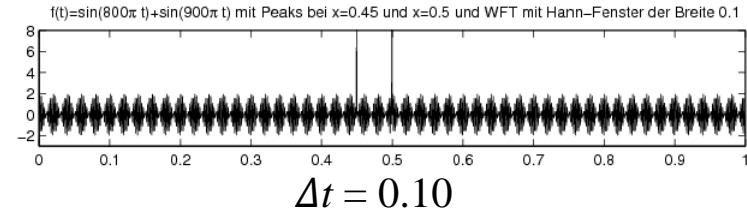
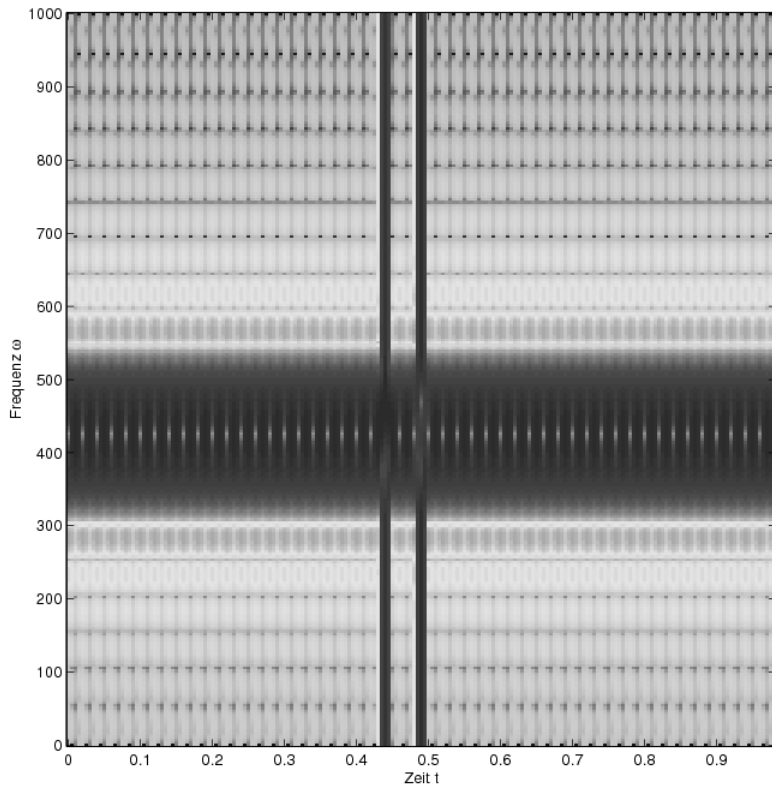
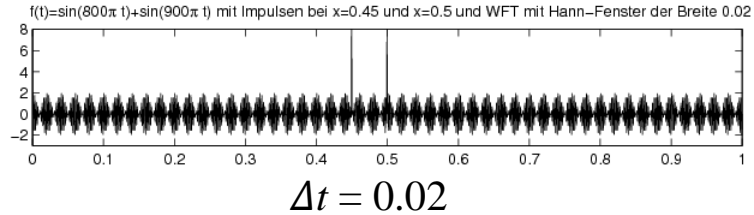
9.6 Outlook: Time-Frequency-Analysis



Quelle:
Wikipedia

9.6 Outlook: Time-Frequency-Analysis

Example 2: Detection and sensitivity with respect to a short peak disturbance



Quelle: Skript „time-frequency-Analyse und Wavelettransformationen“ of M. Clausen und M. Müller, Universität Bonn

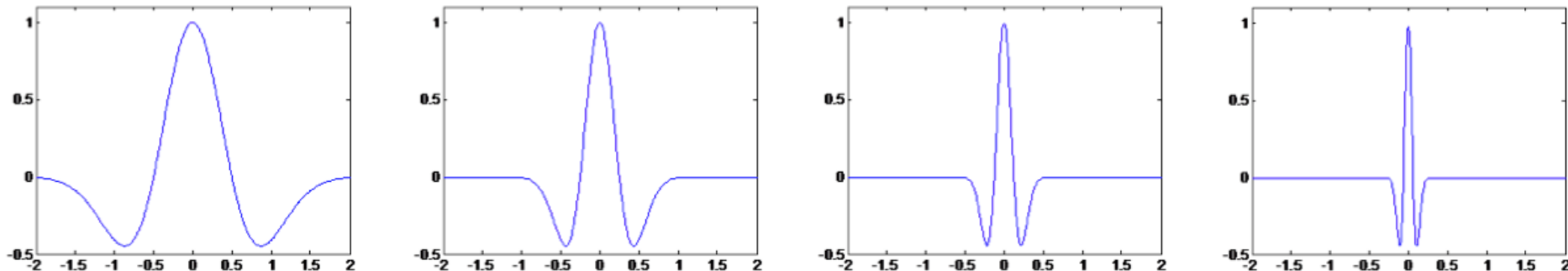
9.6 Outlook: Time-Frequency-Analysis

Fourier Transform

- Looks for periodic signals of infinite length but of all frequencies.
- Ad-hoc fix: focus on a certain time range by windowing.
- Window possesses a certain width → Determination of time and frequency resolution.

Wavelet Transform

- Looks for wave packages of different lengths and frequency.
- Long wave packages are of low frequency → high frequency res. but low time res.
- Short wave packages are of high frequency → low frequency res. but high time res.
- *Idea:* High frequencies commonly occur briefly and thus should be resolved more accurately than low frequencies that typically are present for long time intervals.



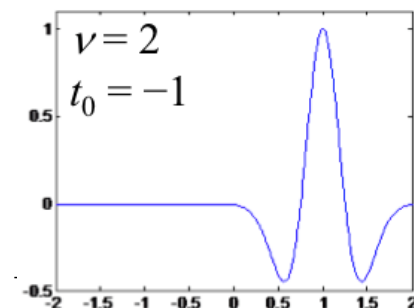
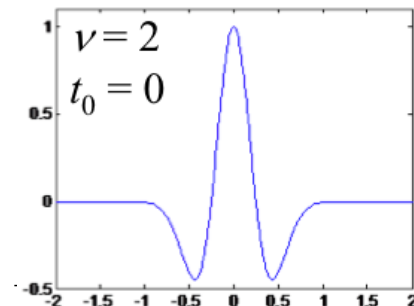
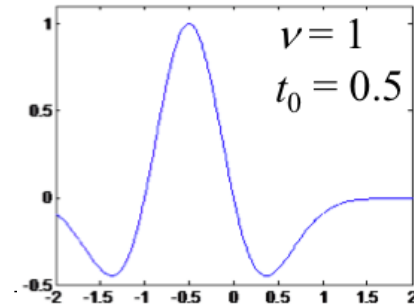
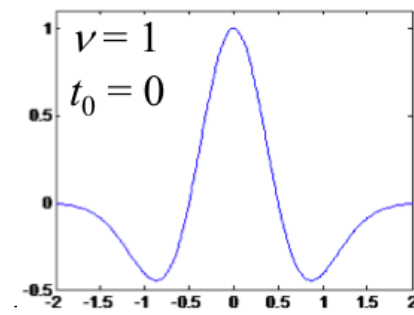
9.6 Outlook: Time-Frequency-Analysis

Construction of Wavelets

- Basis wavelet (*mother wavelet*) as master copy.
- All wavelets are derived from the mother wavelet by time shifts and time scalings (typically with factors 2^{-n}).
- Time shift t_0 for localization of a certain part of the signal.
- Time scaling by a factor ν for a certain frequency component.

Properties of Wavelets

- Through the time shift the signal can be analyzed around $t = t_0$.
- Through the time scaling ν various frequency components can be analyzed.
- In contrast to the Fourier transform where sin-signals of infinite length are analyzed, the length of a wavelet is coupled to its frequency (scaling)!

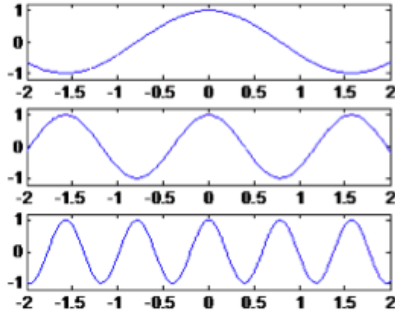


9.6 Outlook: Time-Frequency-Analysis

can be compared to a frequency f

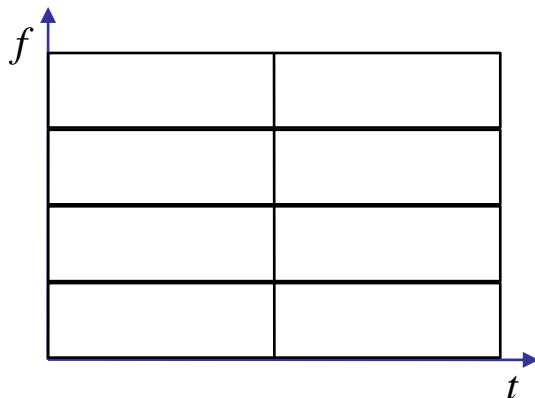
Fourier Transform

$$X(f) = \int_{-\infty}^{\infty} x(t) \cdot e^{-i2\pi ft} dt$$



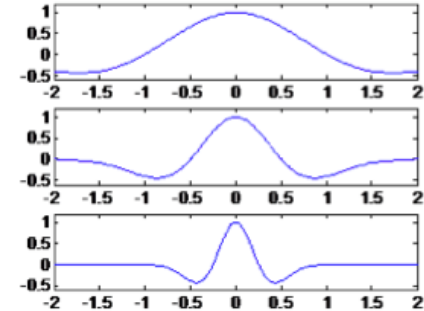
Windowed Fourier Transform

$$X_w(f, t_0) = \int_{-\infty}^{\infty} x(t) \cdot w(t - t_0) \cdot e^{-i2\pi ft} dt$$

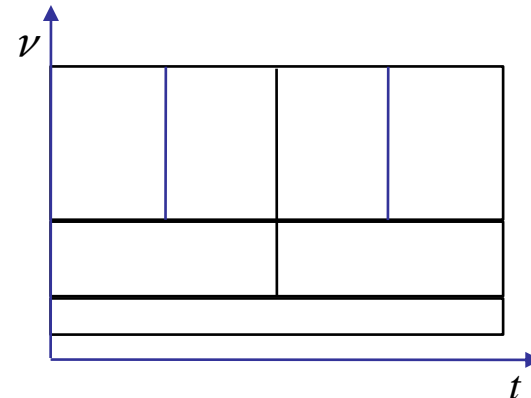


Wavelet Transform

$$X(\nu, t_0) = \int_{-\infty}^{\infty} x(t) \cdot \Psi(\nu(t - t_0)) dt$$



Windowing is not necessary since wavelets are local themselves!



9.7 Outlook: Parametric Frequency Analysis

Parametric Methods

- A large number of data samples N is modeled by estimating a *small* number of n parameters ($n \ll N$).
- These parameters typically results from structural considerations and not only as a means for model accuracy.
- These parameters are physically or with other first principles interpretable or easy to convert in interpretable quantities.
- Examples: IIR or transfer function models, AR or ARMA models, ...

Non-parametric Methods

- A large number of data samples N is described with a *large* number of n parameters. Often $n = N$, i.e., no averaging or noise suppression in the statistical sense takes place.
- The parameters themselves and their number has no direct physical motivation. It just reflects such issues as accuracy, resolution, variance, etc.
- The parameters have no direct physical meaning or interpretation.
- Examples: FIR models (= impulse response models), DFT, ...

9.7 Outlook: Parametric Frequency Analysis

Idea of a Parametric Frequency Analysis

- Signal model: Impulse response of a parametric transfer function.
- Estimation of the parameters of this transfer function.

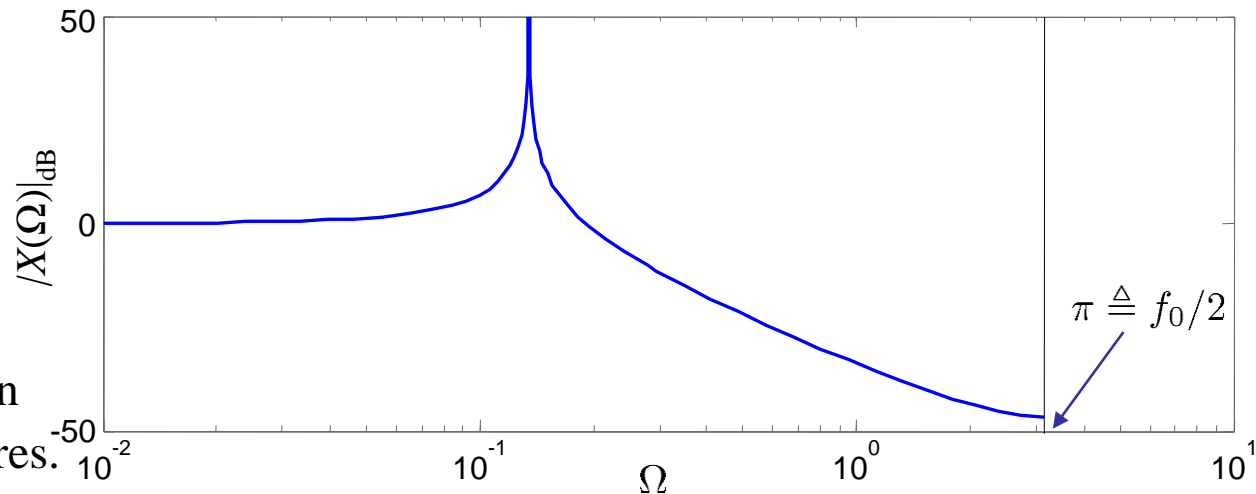
Example: Autoregressive model of 2. order (AR(2)):

$$G(z) = \frac{X(z)}{U(z)} = \frac{(1 + a_1 + a_0)z^2}{z^2 + a_1z + a_0} = \frac{1 + a_1 + a_0}{1 + a_1z^{-1} + a_0z^{-2}}$$

Gain = 1

$$\rightarrow x(k) = (1 + a_1 + a_0)u(k) - a_1x(k - 1) - a_0x(k - 2)$$

- Modeling of *one* damped oscillation.
- Pole locations determine frequency and damping.
- 2 parameters are required for each oscillation and can be estimated by least squares.



9.7 Outlook: Parametric Frequency Analysis

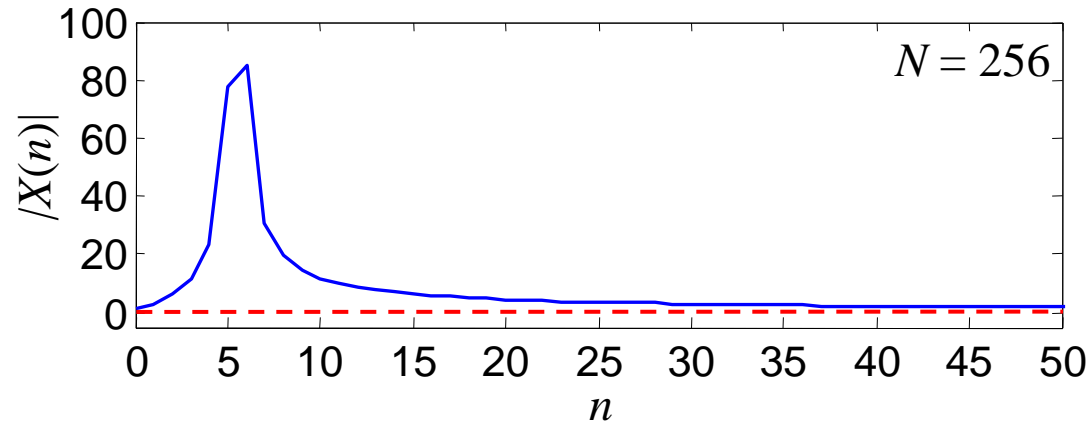
time signal:

$$x(k) = \cos(2\pi f_1 k / 256)$$

$$f_1 = 5.5 \text{ Hz}$$

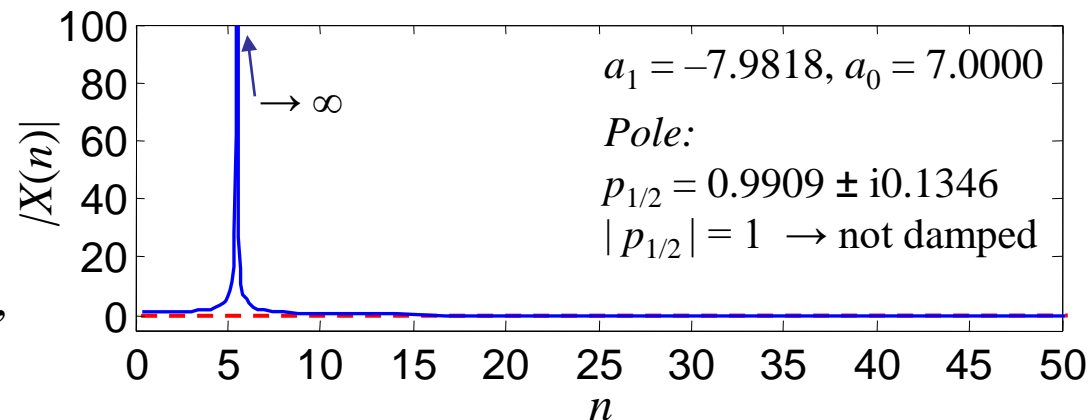
DFT

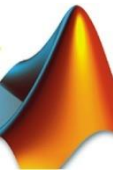
- From 256 samples of the time signal 256 frequency values are computed. → High sensitivity with respect to noise!
- Leakage and picket fence effect distort the spectrum from a peak at $f_1 = 5.5 \text{ Hz}$ to a broader bump.



Parametric AR(2) Estimation

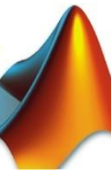
- From 256 samples of the time signal 2 parameters of the AR model are estimated. → Very insensitive with respect to noise!
- An exact frequency (a real number, not discretized!) is computed.



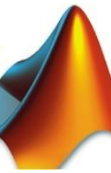


Differences between DFT and Z-Transform

	DTF	Z-Transform
Operates with	Numbers	Variables (symbolic)
Time	Discrete: $0 \dots T_{\max}$	Discrete: $0 \dots \infty$
Frequency	Discrete	Continuous
Use	Signals	Signals & Systems



```
Y = fft(X) ;           % Discrete Fourier Transform (1-D) .  
                    % The algorithm uses an FFT.  
  
y = ifft(X) ;        % Inverse Discrete Fourier Transform  
  
A = dftmtx(n) ;1    % Matrix of the discrete Fourier Transform  
                    % (DFT) . The matrix product with a vector  
                    % calculated the DFT of this vector.  
  
spectrum;1          % Different methods for estimating the  
                    % spectrum (see MATLAB help)  
  
window;1           % Function to perform windowing of signals  
                    % (e.g. gausswin, hamming, etc.)  
  
S = spectrogram(x) ;1 % Calculates the short-time Fourier transform  
                    % (STFT) of a signal.
```



```
Pxx = pcov(x,p);1      % Calculates the spectral density function  
                        % of the vector x by means of the  
                        % covariance method  
                        % p is the order of the predictor (AR).
```

¹ : *Signal Processing Toolbox*

10. Filters

Contents of Chapter 10

4. Filter

10.1 Requirements

10.2 FIR and IIR Filters

10.3 Design of FIR Filters

- Window Method
- Optimization Method (Parks-McClellan)

10.4 Design of IIR Filters

- Method of Bilineare Transformation
- Overview of Analog Filter Typs

10.5 Implementation of Filters

10.6 Nonlinear Filters

10.7 Non-Causal Filters

10.8 Adaptive Filters

10.1 Requirements

What is a Filter?

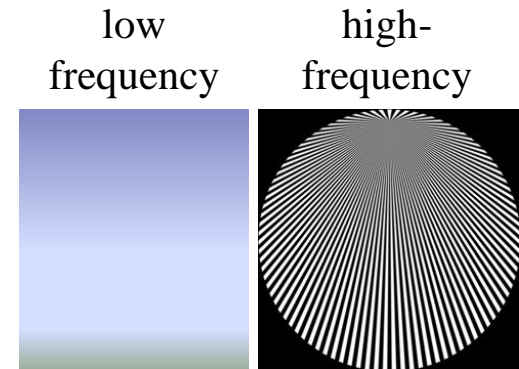
A filter is a system that modifies certain properties (characteristics) of a signals, e.g. it suppresses or enhances. Typical filters are *dynamic systems* and frequency selective, i.e., they block certain frequencies or frequency ranges or let them pass.

Digital Filter

We focus to *digital filters*, i.e., filters that are discrete in time and can be described by difference equations. They can be implemented directly in digital electronic circuits (hardware) but usually are implemented by programs on a computer (software).

Time ↔ Frequency

Usually we consider signals as functions of continuous or discrete time t or k : $x(t)$ or $x(k)$. In a lot of applications, however, the signals rather depend on other variables like location. This is the case for the vast field of image processing. “Frequency” then means the inverse of space (like normally frequency is the inverse of time).

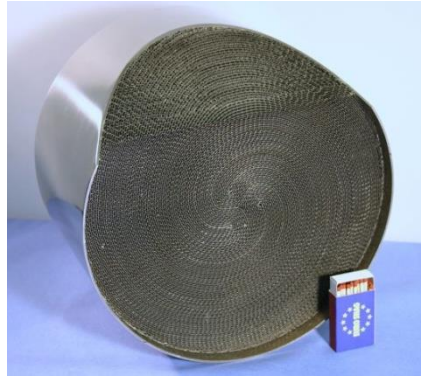


10.1 Requirements

Different Filters



*Optical filter:
Lets only certain colors pass!*



*Soot filter:
Lets only small particles pass!*



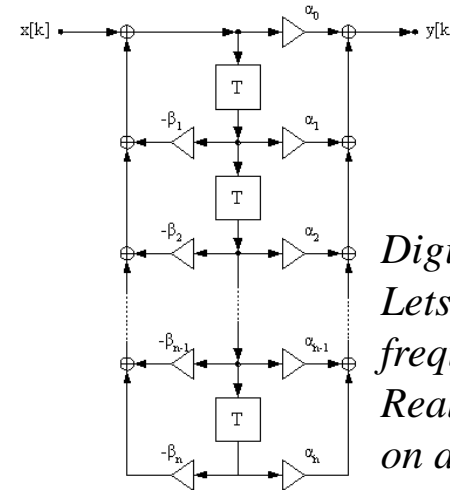
*Air filter:
Lets only small particles pass!*



*Coffee filter:
Lets only liquids pass!*



*Analog electronic filter:
Lets certain frequencies pass!
Realized as R-L-C-circuit.*



*Digital filter:
Lets certain frequencies pass!
Realized in software on a computer.*

10.1 Requirements

Three Steps for *Deriving* a Filter

1. Specification: What should the filter do and under what restrictions?
2. Design: Which filter fulfills these specifications?
3. Implementation (realization): How is the filter build (in hardware) or programmed (in software)?

Four Steps for *Designing* a Filter

- a) Choice of a system class: E.g. linear, stable, causal, time-invariant, dynamic systems.
- b) Choice of a filter structure:
e.g. FIR (*finite impulse response*)
or IIR (*infinite impulse response*).
- c) Determination of the filter order.
- d) Determination of the filter parameters.

10.1 Requirements

Signal-to-Noise Ratio

abbreviated SNR is the ratio between the averaged power of a signal (meaningful information) and the averaged power of noise (disturbance)

$$\text{SNR} = \frac{P_{\text{Signal}}}{P_{\text{Noise}}}$$

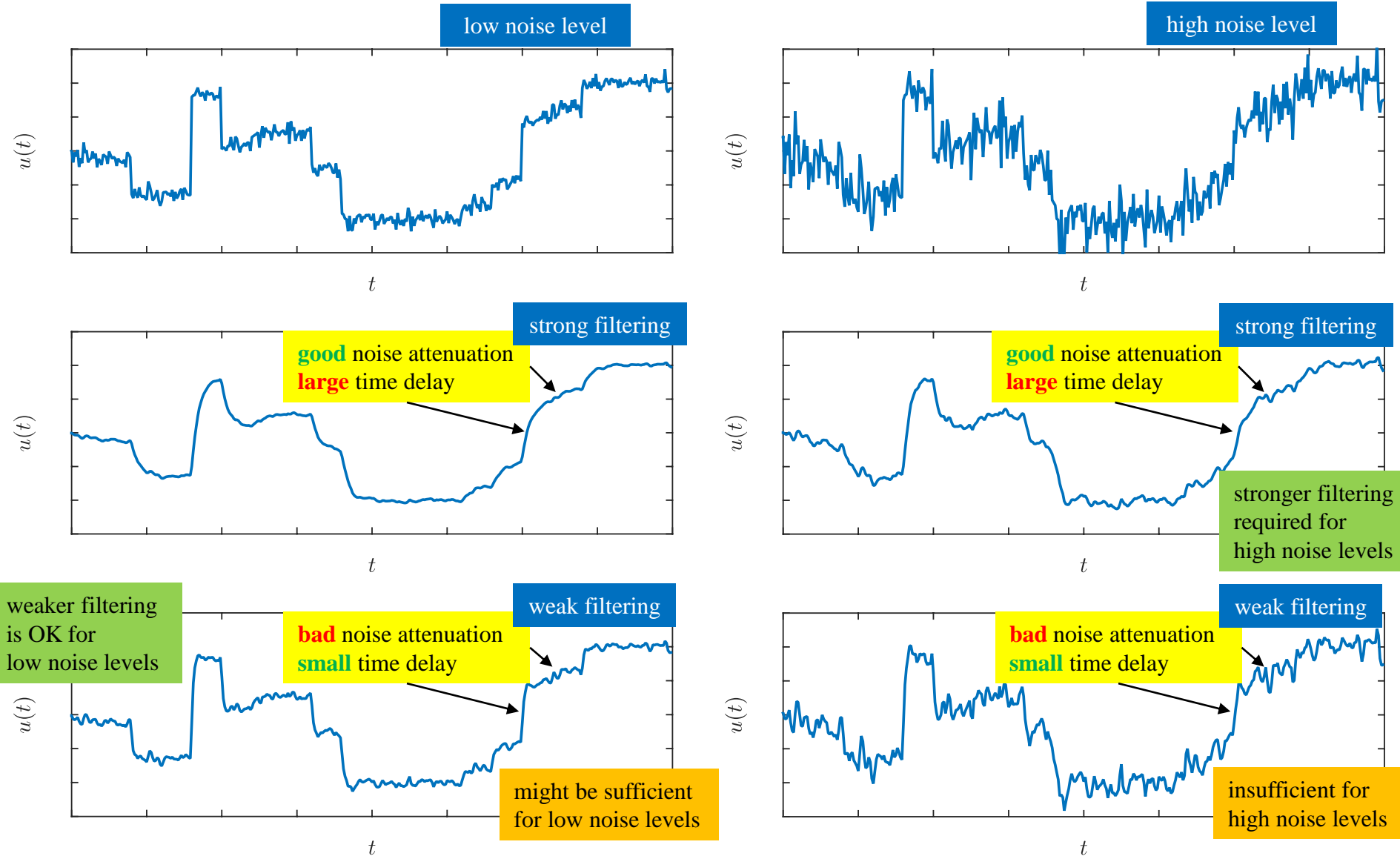
Often it is given in logarithmic scale, in decibel:

$$\text{SNR} = 10 \lg \left(\frac{P_{\text{Signal}}}{P_{\text{Noise}}} \right) \text{ dB}$$

Since it relates *powers* (\sim squares of amplitudes) the 3 dB corner frequency marks the 1/2 drop-off, not the $1/\sqrt{2}$ drop-off as it is known from the magnitude bode diagram used in control theory which shows *amplitudes*!

Task of a filter is to improve (i.e., increase) the SNR. This is typically possible because signal and noise are dominant in different frequency ranges. The corner frequency of the filter represents the boundary between signal frequency range and the noise frequency range.

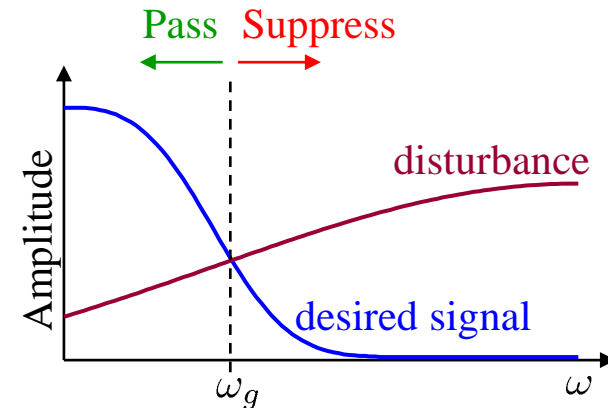
10.1 Requirements



10.1 Requirements

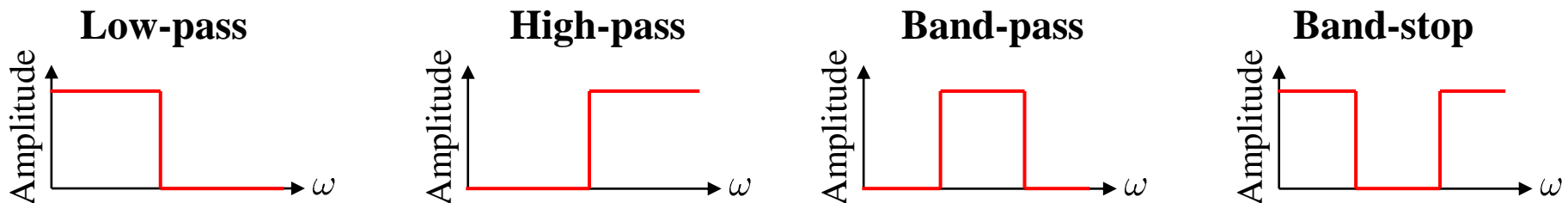
Limit (Cut-off) Frequency

A frequency selective filter can only be useful if the **desired signal** and the **disturbance** are in different frequency ranges. Then it is possible to place the limit (cut-off) frequency ω_g in such a way that a significant part of the desired signal can pass while a significant part of the disturbance cannot.



Filter Types

If, as in the above example, the desired signal lies mostly in the low-frequency range while the disturbance lies mostly in the high frequency range, a **low-pass filter** can improve the signal quality a lot. A low-pass filter lets all low frequency components pass but suppresses all high frequency components. That is the most common used filter type. In many applications, however, the desired signal and disturbance are in other frequency ranges.



10.1 Requirements

Application Examples for Different Filter Characteristics

- Low-pass: Suppression of high frequency noise to improve the quality of the signal.
- High-pass: Suppression of a slowly changing signal change like offsets (frequency 0) or trends or drifts.
- Band-pass: Extraction of a frequency band. Typical for radio or TV receivers. The signal is modulated on a high frequency carrier that it needs to be extracted from before further processing.
- Band-stop: Suppression of certain (typically narrow) frequency ranges. Commonly applied to actuation signals in the aerospace industry to avoid damages due to an excitation of resonances (weakly damped modes). Also called a notch filter.

Ideal Filter

- Perfect output of the signal in the *passband*, i.e., $|G(i\omega)| = 1 = 0 \text{ dB}$.
- Perfect suppression of the signal in the *stopband*, i.e., $|G(i\omega)| = 0$.
- Infinitely steep transition from passband to stopband, i.e., steepness = ∞ .
- No phase shift (no delay) of the signal, i.e., $\angle G(i\omega) = 0$.

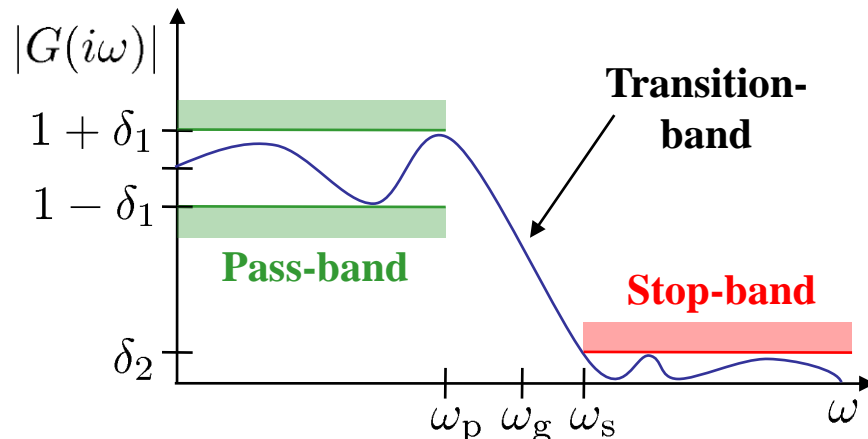
10.1 Requirements

Real Filters

In the real world, these properties cannot be exactly realized. The demands for an ideal filter can never be achieved. Thus we ease the requirements and accept tolerances.

Specification of Real Lowpass Filters

- Pass-band: Gain close to 1, between $1 - \delta_1$ and $1 + \delta_1$.
- Stop-band: Gain smaller than δ_2 .
- Pass-band: $\omega < \omega_p$, transition-band: $\omega_p < \omega < \omega_s$, stop-band: $\omega > \omega_s$.
- No requirements on the phase. Sometimes linear phase is demanded, see later.



Remarks:

- The closer ω_p and ω_s lie together and the smaller δ_1 and δ_2 are chosen, the more extreme are the requirements.
- More extreme requirements necessarily lead to more complex filters.

10.1 Requirements

Restriction to Filters With the Following Properties (see Chapter 8)

- Stable
- Linear (for nonlinear filters see Section 10.6)
- Causal (for non-causal filters, see Section 10.7)
- Time-invariant (for time-variant filters, see Section 10.8)

Furthermore it is sometimes *desirable*, particularly in communications:

- linear in its phase

This means that *every* oscillation is identically shifted by the filter in phase. This is independent of the oscillation frequency. This property is important in acoustic environments (audio components) because the ears are very sensitive the frequency-dependent phase differences. If the linear phase property is not at least approximately fulfilled this means low and high frequency sounds arrive at the ear at different times! This would disturb any acoustic sensation. In control systems with linear phase have a different name: they are called

- systems with a pure *dead time* with no other phase delay.

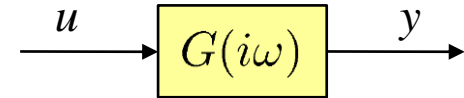
10.1 Requirements

Property: Linear Phase

Mathematically “linear phase” more precisely means the phase is a linear function of the frequency:

$$G(i\omega) = |G(i\omega)|e^{-i\omega T_t}$$

with a real T_t



A filter with such a transfer function has an output $y(t)$ to an input oscillation $u(t)$ with an amplitude A_1 , frequency ω_1 and phase φ_1 after transients are decayed of:

$$u(t) = A_1 \sin(\omega_1 t + \varphi_1) \quad \rightarrow \quad y(t) = A_1 |G(i\omega_1)| \sin(\omega_1 t + \varphi_1 - \omega_1 T_t)$$

Because the phase shift is linear in the frequency this can be written as:

Amplitude gain

Phase shift

$$y(t) = A_1 |G(i\omega_1)| \sin(\omega_1(t - T_t) + \varphi_1)$$

The phase φ_1 of the input signal $u(t)$ is not changed by the filter.

And this is the case independent of the frequency of the signal ω_1 .

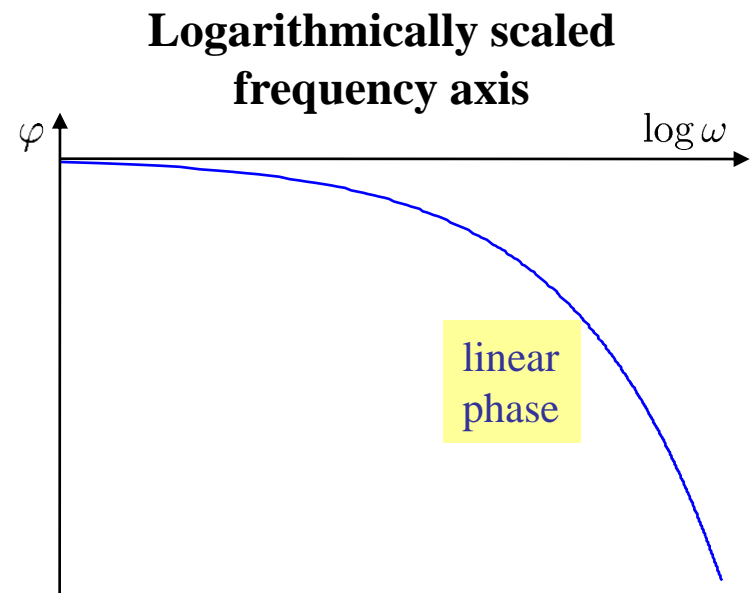
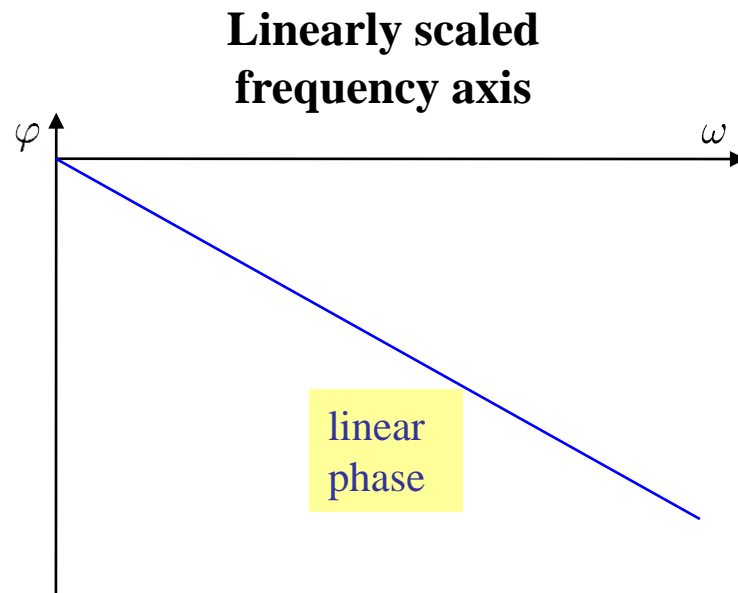
Time shift
(dead time)

The dead time T_t is commonly also called *group propagation delay*: $\tau_g = -\frac{d\varphi}{d\omega}$

10.1 Requirements

Property: Linear Phase

- Can *exactly* only be achieved by FIR filters.
- For IIR filters the phase can only be *approximately* linear in a certain frequency band.
- Especially in the audio and communications fields linear phase is a commonly requested property of big importance.



10.1 Requirements

Property: Zero Phase

A particularly simple special case of a filter with linear phase is a filter with zero phase, i.e., with a phase response = 0 for all frequencies. This is the case for transfer functions that are purely real and non-negative. Such a transfer function $F(z)$ can be constructed from an arbitrary transfer function $G(z)$ with arbitrary phase as follows:

$$F(z) = G(z)G(z^{-1})$$

This leads to a purely real frequency response:

$$F(e^{i\omega T_0}) = G(e^{i\omega T_0})G(e^{-i\omega T_0}) = |G(e^{i\omega T_0})|^2$$

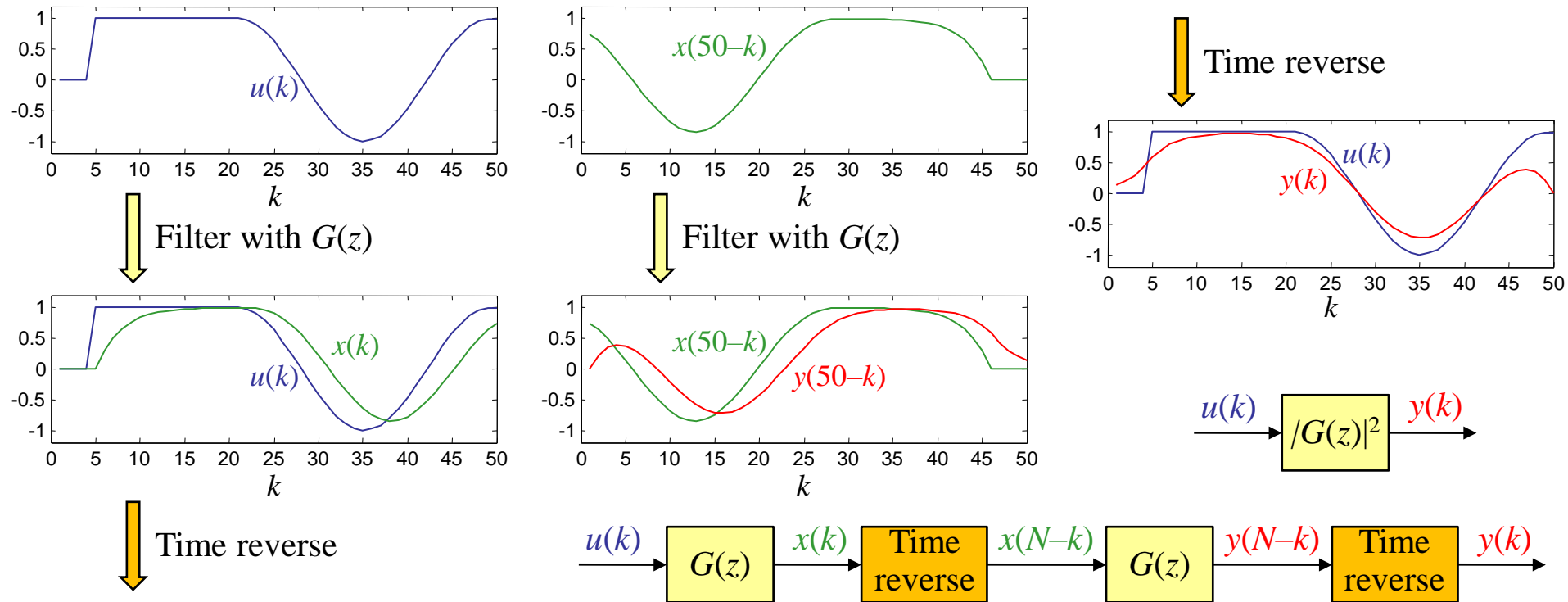
This means: $F(z)$ has for every zero z_n a mirrored zero at $z_n^{-1} = 1/z_n$ and for every pole z_p a mirrored pole at $z_p^{-1} = 1/z_p$. If z_n and z_p are inside the unit circle (stable!) then $1/z_p$ and $1/z_p$ automatically are outside the unit circle (unstable!). Consequently, zero phase filters have the following properties:

- FIR: non-causal.
- IIR: unstable and non-causal.

10.1 Requirements

Implementation of Zero Phase Filters

Because they are necessarily non-causal, zero phase filters can only be implemented offline. A simple possibility is to filter the data with a causal $G(z)$ and subsequently filter the outcome in backward direction again with $G(z)$. The phase delay induced by the first filter exactly will be reversed by the second backward filtering process.



10.2 FIR and IIR Filters

Equivalent descriptions of a linear dynamic systems in discrete time:

Difference equation of order n

$$y(k) = \sum_{i=0}^m b_i u(k-i) - \sum_{i=1}^n a_i y(k-i)$$

→ Can be implemented directly ($m \leq n$).

Usually $m = n$. If $m < n$ we can assume $m = n$ with $b_i = 0$ for $i > m$.

Properties

- Order n is small: e.g. $n = 2, 3, 4, \dots$
- Feedforward: $b_i u(k-i)$
- Feedback: $a_i y(k-i)$
- Infinite impulse response (IIR)

Impulse response

$$y(k) = \sum_{i=0}^{\infty} g_i u(k-i)$$

→ Cannot be implemented directly!

Approximation with $m+1$ terms:

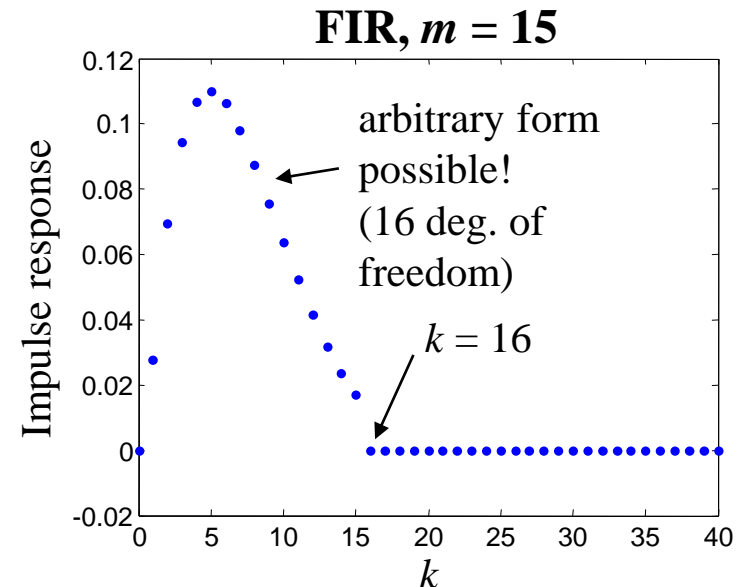
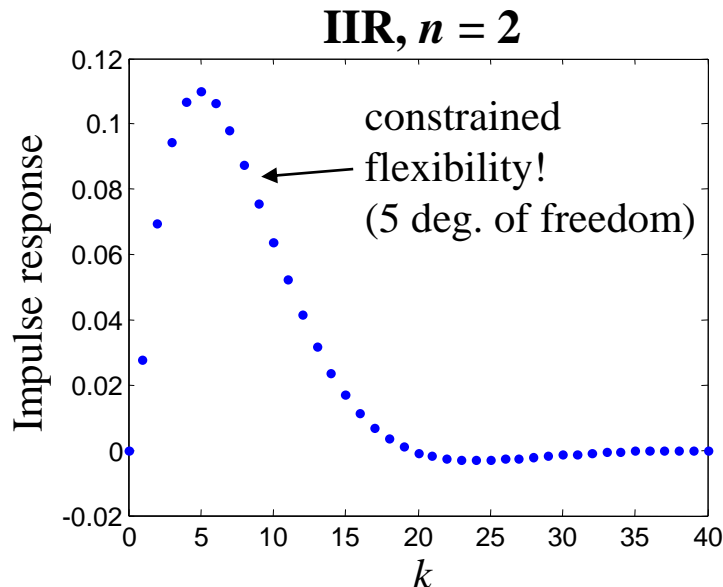
$$y(k) = \sum_{i=0}^m b_i u(k-i)$$

- Order m is large: $m = 10, 20, 30, \dots$
- Feedforward: $b_i u(k-i)$
- No feedback!
- Finite impulse response (FIR)

10.2 FIR and IIR Filters

Difference in the Order

- IIR filters usually have significantly fewer parameters (a_i & b_i) than FIR filters (b_i).
- IIR filters need fewer memory for storage of previous data.
- IIR impulse response usually asymptotically exponentially decays towards zero. FIR impulse response is exactly equal to zero after time steps $k > m$.
- IIR filters can be unstable. FIR filters are inherently stable (no feedback).
- IIR filters have an analog correspondence. FIR filters exist only in the digital world.



10.2 FIR and IIR Filters

Transfer Functions

IIR Filter

$$G(z) = \frac{b_0 + b_1 z^{-1} + \dots + b_m z^{-m}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}}$$

$$G(z) = \frac{b_0 z^n + b_1 z^{n-1} + \dots + b_m z^{n-m}}{z^n + a_1 z^{n-1} + \dots + a_n}$$

- m zeros at arbitrary locations
- n Pole at arbitrary locations
- $b_0 = 0$ for strictly proper systems
- Complex relationship between parameters and impulse response
- Not well suited for adaptation
 - feedback structure
 - stability problems

FIR Filter

$$G(z) = b_0 + b_1 z^{-1} + \dots + b_m z^{-m}$$

$$G(z) = \frac{b_0 z^m + b_1 z^{m-1} + \dots + b_m}{z^m}$$

- m zeros at arbitrary locations
- m poles at 0
- $b_0 = 0$ for strictly proper systems
- $b_i = g_i$ are the first $m+1$ steps of the impulse response, all subsequent ones = 0
- Well suited for adaptation
 - feedforward structure
 - inherent stability

10.2 FIR and IIR Filters

Example:

A system with impulse response $g(k) = a^k$ can be exactly realized by an **IIR filter of 1. order**:

$$\begin{aligned} G_{\text{IIR}}(z) &= 1 + az^{-1} + a^2z^{-2} + a^3z^{-3} + \dots \\ &= \sum_{k=0}^{\infty} (az^{-1})^k \end{aligned}$$

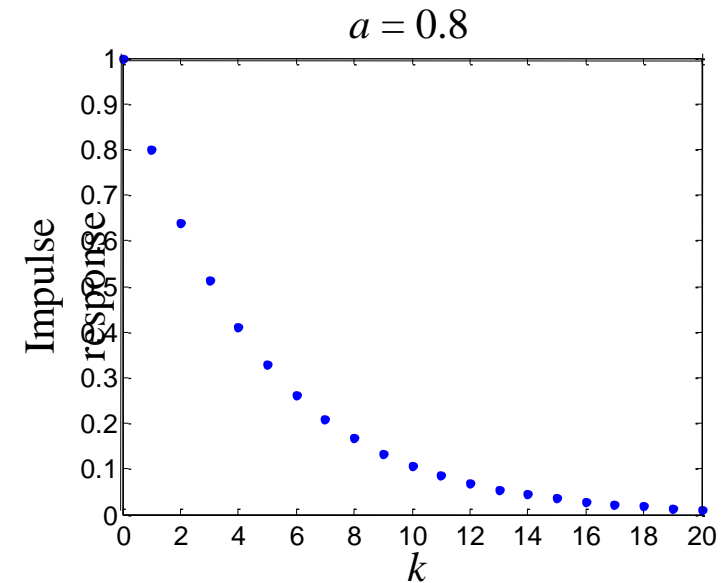
This infinite geometric series can exactly be written as:

$$G_{\text{IIR}}(z) = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}$$

The gain of this IIR filter is:

$$K_{\text{IIR}} = G_{\text{IIR}}(z = 1) = \frac{1}{1 - a}$$

The marginally stable case (integral behavior) is achieved for $a = 1$.
Then the gain cannot be calculated anymore.



10.2 FIR and IIR Filters

An FIR filter can approximately represent every stable impulse response. For the first $m+1$ terms an **FIR of m . order** can exactly describe every sequence:

$$G_{\text{FIR}}(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \dots + b_{m-1} z^{m-1} + b_m z^{-m} = \sum_{k=0}^m b_k z^{-k}$$

A natural choice for the filter parameters would be:

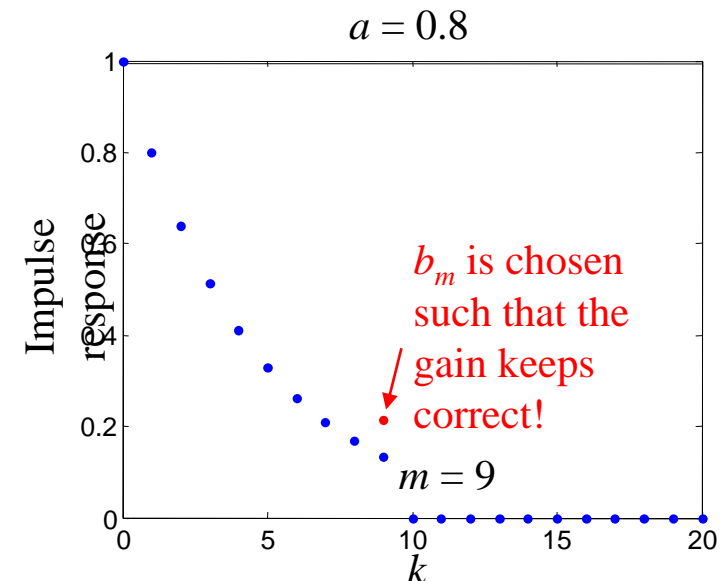
$$b_k = g(k) = a^k \quad \text{for } k = 0, 1, \dots, m$$

However, this would yield a wrong (too small) gain because all summands for $k > m$ are simply missing.

Alternatively the last parameter (summand) can be adjusted in order to make the gain correct, i.e., for $\omega \rightarrow 0 / z \rightarrow 1$:

$$b_m = \frac{1}{1-a} - b_0 - b_1 - \dots - b_{m-1}$$

This is a reasonable approach for low-pass filters. For high-pass filter an alternative could be to require identical gains for $\omega \rightarrow \infty / z \rightarrow \infty$.



10.2 FIR and IIR Filters

Properties of FIR filters:

- stable,
- can realize linear phase,
- very flexible because many degrees of freedom (parameters) → frequency response can be shaped as desired,
- only forward path → simple to implement,
- easy to adapt.

Properties of IIR filters:

- can become unstable,
- no linear phase possible,
- with the help of a few parameters significant effects can be realized,
- high steepness even for low orders,
- feedback path → more complex to implement,
- complex to adapt (stability problems, not linear in its parameters).

10.3 Design of FIR Filters

General Remarks About FIR Filter Design

The design of digital filters commonly is based on the mature field of design of analog filters. Because FIR filters only exist in the digital world, no analog correspondence is available. New design method must be developed. The three standard approaches are:

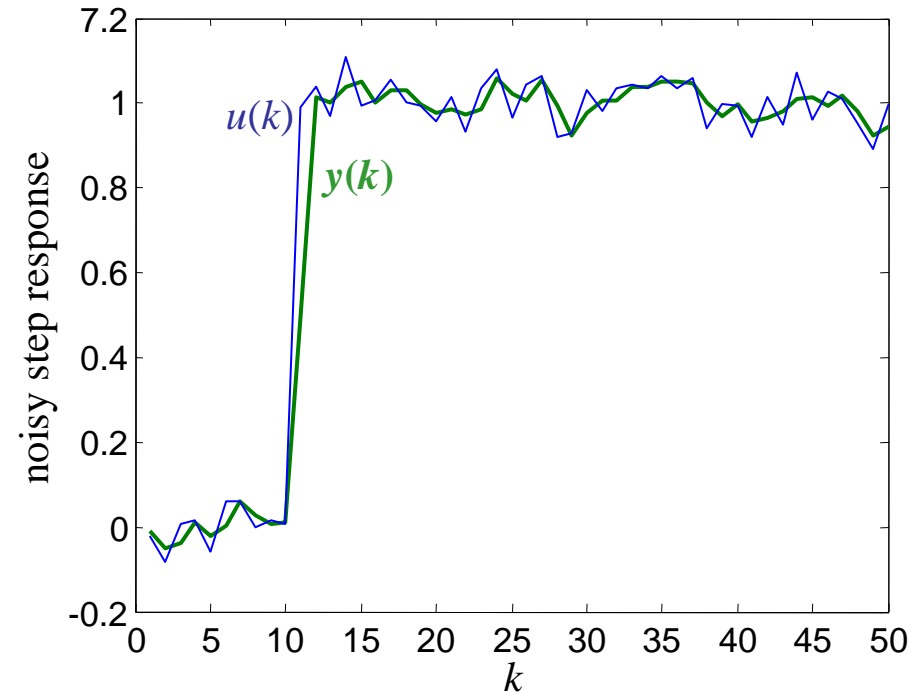
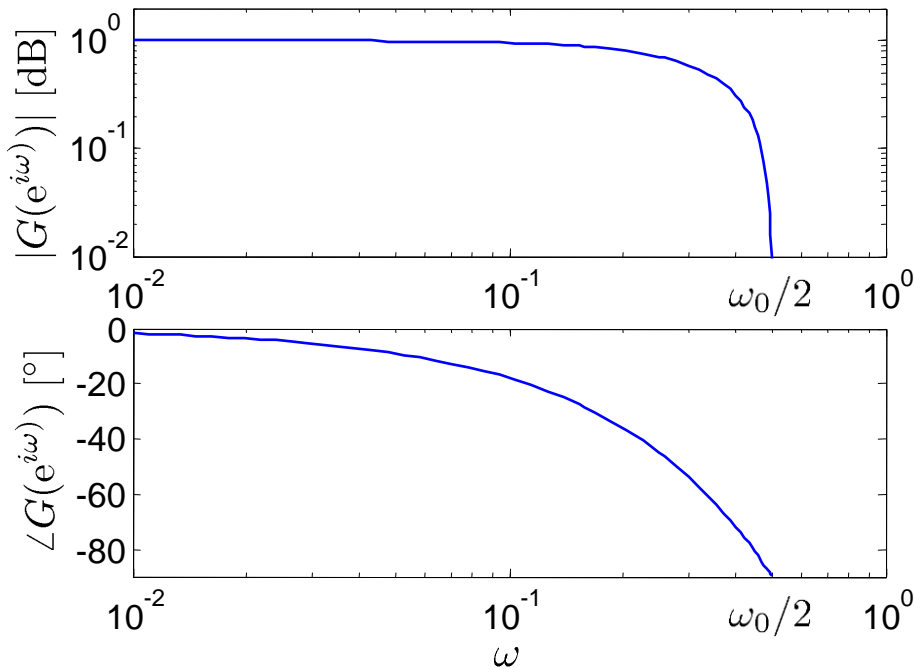
- *Window method*: A simple approach that can be pursued by hand. The desired amplitude response is established. Then it is transformed by the inverse Fourier transform to the impulse response in the time domain. Since the impulse response is usually of infinite length the filter order m must be reduced/cropped to a realizable number. This causes an approximation error and thus the method is not very accurate.
- *Frequency sampling method*: This is a very universal approach and also possible for recursive filters. The desired frequency response is sampled and transformed with the inverse DFT to the impulse response.
- *Optimal filter method*: With support of a software tool this is the most powerful and flexible approach. A *minmax optimization* problem is solved via a *Chebyshev* approximation that minimizes the maximal deviation between the frequency response of the filter and the desired frequency response. This is carried out with the algorithm proposed by *Parks-McClellan* and implemented in the MATLAB signal processing toolbox.

10.3 Design of FIR Filters

Example: A simple FIR filter of 1. order

For FIR filters the output is calculated as a weighted average of the current and previous inputs (*moving average, MA*). A simple low-pass filter can look like:

$$y(k) = \frac{u(k) + u(k - 1)}{2} = 0.5u(k) + 0.5u(k - 1)$$



10.3 Design of FIR Filters

Remarks on FIR Filters With Linear Phase

FIR filters have to fulfill certain condition in order to have a linear (or affine) phase:

- Linear phase, i.e., $\varphi(\omega) = \alpha\omega$: symmetrical impulse response.
- Affine phase, i.e., $\varphi(\omega) = \alpha\omega + \beta$: centrosymmetrical impulse response.

Remember:

Addition of two *conjugate complex* numbers:

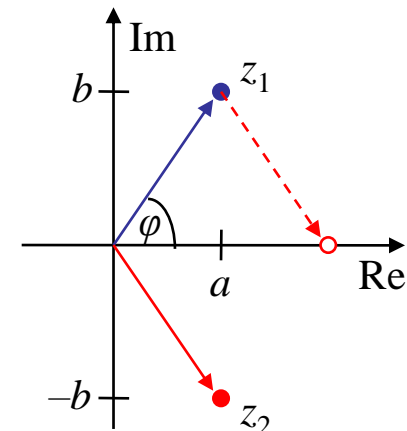
$$z_1 + z_2 = a + ib + (a - ib) = 2a \rightarrow \text{purely real!}$$

Same numbers written in absolute value and phase form:

$$\begin{aligned} z_1 + z_2 &= ce^{i\varphi} + ce^{-i\varphi} = c(e^{i\varphi} + e^{-i\varphi}) \rightarrow \text{purely real!} \\ &= c(\cos\varphi + i\sin\varphi + \cos\varphi - i\sin\varphi) = 2c\cos\varphi \end{aligned}$$

→ **Sum of two *conjugate complex* numbers is purely real!**

$$\begin{aligned} z_1 &= a + ib = ce^{i\varphi} \\ c &= \sqrt{a^2 + b^2} \\ \varphi &= \arctan \frac{b}{a} \end{aligned}$$



10.3 Design of FIR Filters

Example: Symmetrical FIR Filter of Length $L = 9$ (Order $m = 8$)

Transfer function of the filter:

$$G(z) = g(0) + g(1)z^{-1} + g(2)z^{-2} + g(3)z^{-3} + g(4)z^{-4} + g(5)z^{-5} + g(6)z^{-6} + g(7)z^{-7} + g(8)z^{-8}$$

Because of symmetry we have:

$$g(0) = g(8), \quad g(1) = g(7), \quad g(2) = g(6), \quad g(3) = g(5)$$

$$G(z) = g(0) [1 + z^{-8}] + g(1) [z^{-1} + z^{-7}] + g(2) [z^{-2} + z^{-6}] + g(3) [z^{-3} + z^{-5}] + g(4)z^{-4}$$

Factoring z^{-4} out yields:

$$G(z) = z^{-4} \{g(0) [z^4 + z^{-4}] + g(1) [z^3 + z^{-3}] + g(2) [z^2 + z^{-2}] + g(3) [z^1 + z^{-1}] + g(4)\}$$

The frequency response is obtained for $z = e^{i\omega T_0}$. Expression of the following form

$$z^n + z^{-n} \Big|_{z=e^{i\omega T_0}} = e^{in\omega T_0} + e^{-in\omega T_0}$$

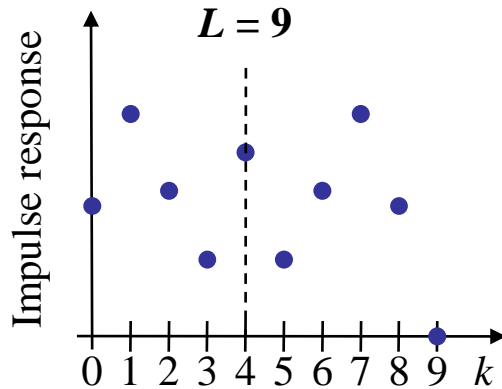
are purely real and therefore have phase = 0. Thus the phase of this filter finally is:

$$z^{-4} \Big|_{z=e^{i\omega T_0}} = e^{-i4\omega T_0} \quad \rightarrow \quad \varphi(\omega) = -4T_0 \omega = \alpha \omega \quad (+ \pi \text{ if the sign of “\{...\}” is negative!})$$

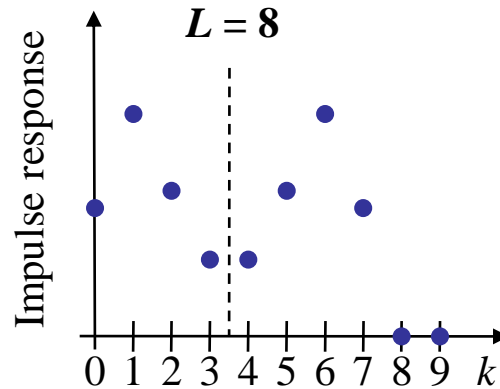
10.3 Design of FIR Filters

The 4 Types of FIR Filters With Linear Phase

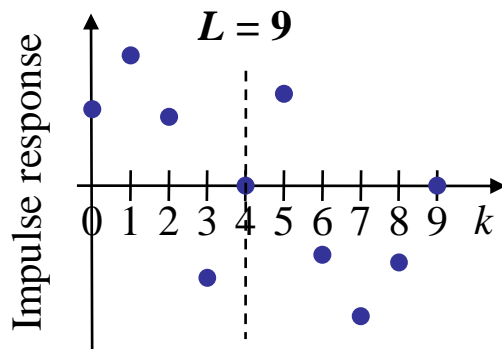
Type 1: Symmetry odd length



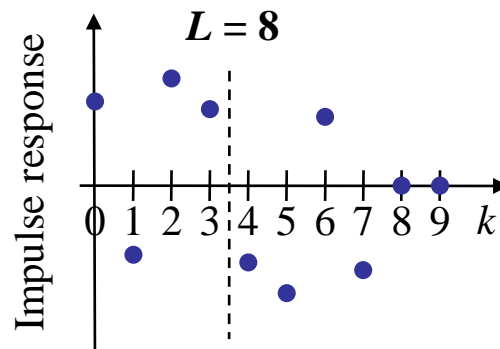
Type 2: Symmetry even length



Type 3: Centrosymmetry odd length



Type 4: Centrosymmetry even length



Symmetry:

- $g(k) = g(L - 1 - k)$
- Phase is linear, i.e.
 $\varphi(\omega) = \alpha \omega$

Centrosymmetry:

- $g(k) = -g(L - 1 - k)$
- Phase is affine, i.e.
 $\varphi(\omega) = \alpha \omega + \beta$
 $\beta = \pi/2$

10.3 Design of FIR Filters

Window Method

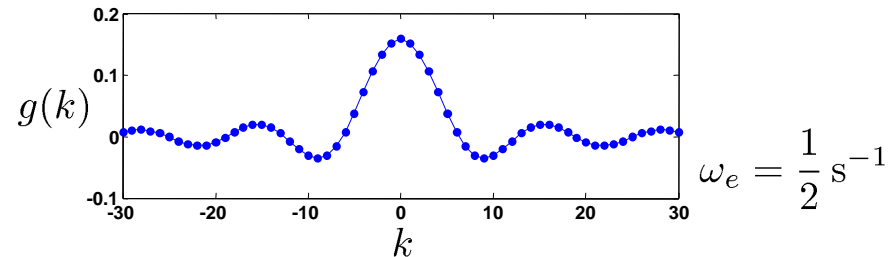
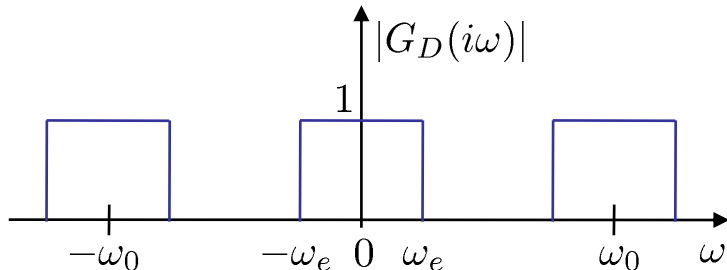
The idea behind this approach is to design a filter that has a desired frequency response $G_D(i\omega)$ ($D = \text{desired}$). Subsequently the impulse response $g(k)$ can be calculated via the inverse Fourier transform as follows:

$$g(k) = \frac{1}{2\pi} \int_{-\omega_0/2}^{\omega_0/2} G_D(i\omega) e^{i\omega T_0 k} d\omega$$

This impulse response typically is non-causal and of infinite length. We have to shift it and crop it at a certain finite order m to make the FIR filter realizable.

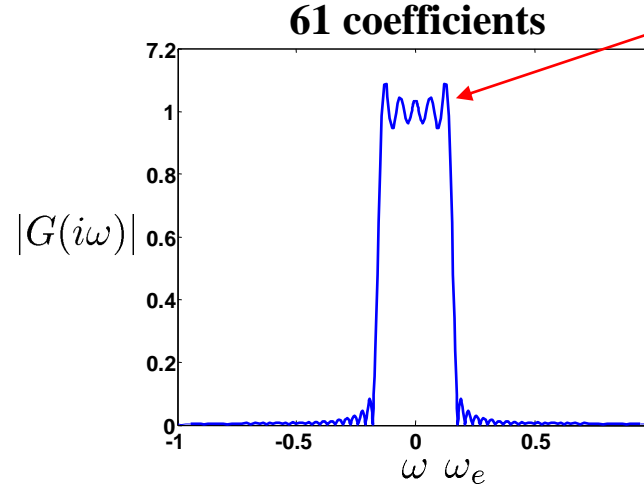
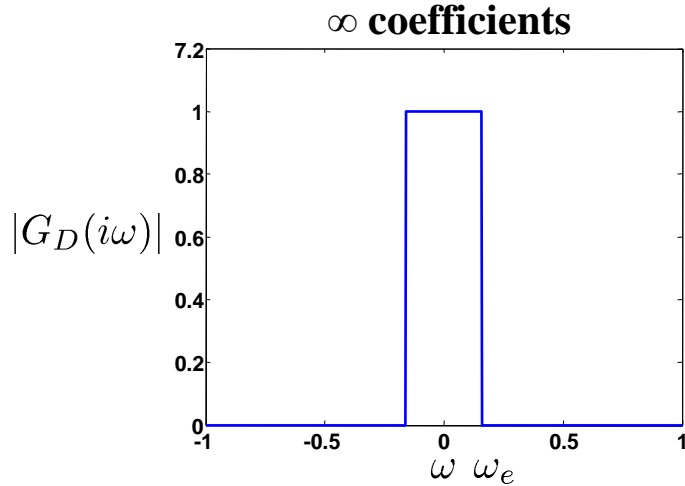
Example: Low-pass with cut-off frequency ω_e (sampling rate $T_0 = 1$ s)

$$g(k) = \frac{1}{2\pi} \int_{-\omega_0/2}^{\omega_0/2} G_D(i\omega) e^{i\omega T_0 k} d\omega = \frac{1}{2\pi} \int_{-\omega_e}^{\omega_e} 1 \cdot e^{i\omega T_0 k} d\omega = \frac{\omega_e \sin \omega_e k}{\pi \omega_e k} = \frac{\sin \omega_e k}{\pi k}$$

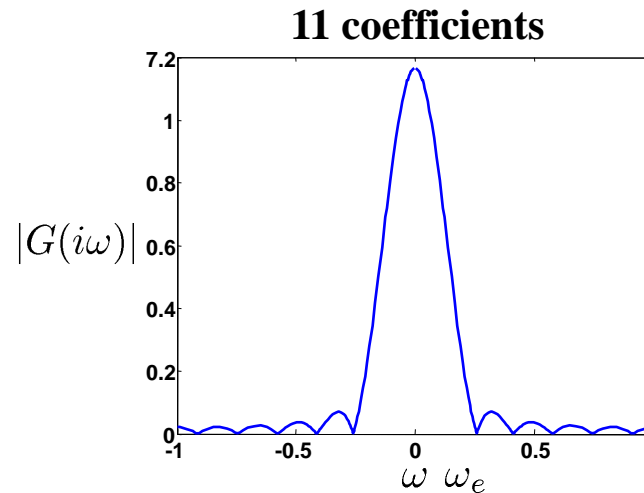
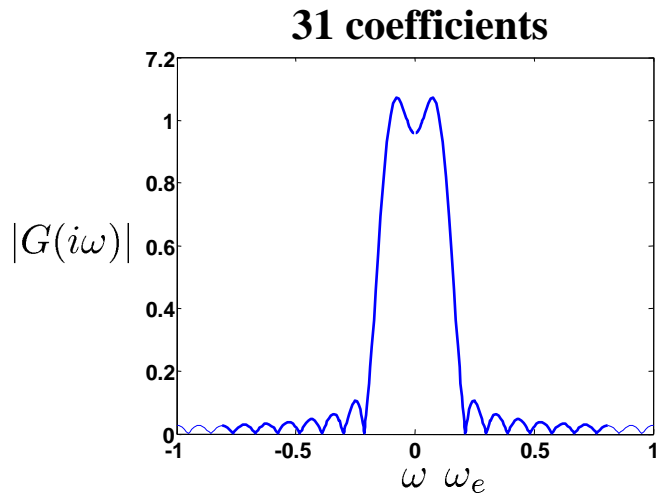


10.3 Design of FIR Filters

Approximation Error Through Cropping the Impulse Response



Unwanted behavior!
Gibbs Phenomenon



10.3 Design of FIR Filters

Consequences From Unwanted Behavior of FIR-Filters

The *ripple* in the amplitude response of the FIR filter can easily be explained. The cropping of the impulse response $g(k)$ is identical to a windowing with a uniform/rectangular window $w(k)$. In the frequency domain this corresponds to a convolution with the Fourier transform of the rectangular window $W(i\omega)$, the sinc-function:

$$g(k) \cdot w(k) \quad \circ \rightarrow \bullet \quad G(i\omega) * W(i\omega)$$

This explains the ripples. Unfortunately they do not become smaller if more coefficients are spent to describe the impulse response more accurately. This is the so-called **Gibbs phenomenon** (see math, “Fourier series”).

In order to reduce this undesirable effect, the impulse response is multiplied with a smoother window like in the DFT context. Such a window can reduce high frequencies by letting the impulse response slowly decay towards zero. For FIR filter design the so-called *Kaiser* window is commonly applied.

10.3 Design of FIR Filters

Optimal Filter Design Method

With most optimization methods the quadratic error $|E(i\omega)|^2$ between the desired filter characteristics $H_D(i\omega)$ and the real filter characteristics $H(i\omega)$ is minimized:

$$E(i\omega) = H_D(i\omega) - H(i\omega) \quad \int_{-\omega_0/2}^{\omega_0/2} |E(i\omega)|^2 d\omega \rightarrow \min$$

However, the algorithm according to Parks and McClellan minimizes the maximal (not squared) error because it has yield more reliable results:

$$\max\{|E(i\omega)|\} \rightarrow \min$$

The minimization of the maximal absolute value ensures that the ripples are *equally* distributed over all frequencies which led to the name *Equiripple* filter. The criterion is also important in many other approaches to robust optimization and control.

Because the absolute value of the error is magnitudes larger in the pass-band than in the stop-band, it is important to multiply the errors with a normalization weight that guarantees no frequency ranges are preferred:

$$\max\{|W(i\omega)E(i\omega)|\} \rightarrow \min$$

10.3 Design of FIR Filters

To achieve a filter with equally large (small) ripples in the pass- and stop-band the following frequency weight must be chosen for a low-pass filter:

$$W(i\omega) = \begin{cases} 1 & \text{für } 0 \leq \omega \leq \omega_p \\ \delta_1/\delta_2 & \text{für } \omega_s \leq \omega \leq \omega_0/2 \end{cases} \quad \text{or} \quad W(i\omega) = \begin{cases} \delta_2/\delta_1 & \text{für } 0 \leq \omega \leq \omega_p \\ 1 & \text{für } \omega_s \leq \omega \leq \omega_0/2 \end{cases}$$

MATLAB offers the Parks/McClellan minimax algorithm and least-squares optimization tools:

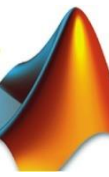


The default mode of operation of `firls` and `firpm` is to design type I or type II linear phase filters, depending on whether the order you desire is even or odd, respectively. A lowpass example with approximate amplitude 1 from 0 to 0.4 Hz, and approximate amplitude 0 from 0.5 to 1.0 Hz is

```
n = 20; % Filter order
f = [0 0.4 0.5 1]; % Frequency band edges
a = [1 1 0 0]; % Desired amplitudes
b = firpm(n,f,a); % Parks-McClellan FIR Design
```

From 0.4 to 0.5 Hz, `firpm` performs no error minimization; this is a transition band or "don't care" region. A transition band minimizes the error more in the bands that you do care about, at the expense of a slower transition rate. In this way, these types of filters have an inherent trade-off similar to FIR design by windowing. To compare least squares to equiripple filter design, use `firls` to create a similar filter.

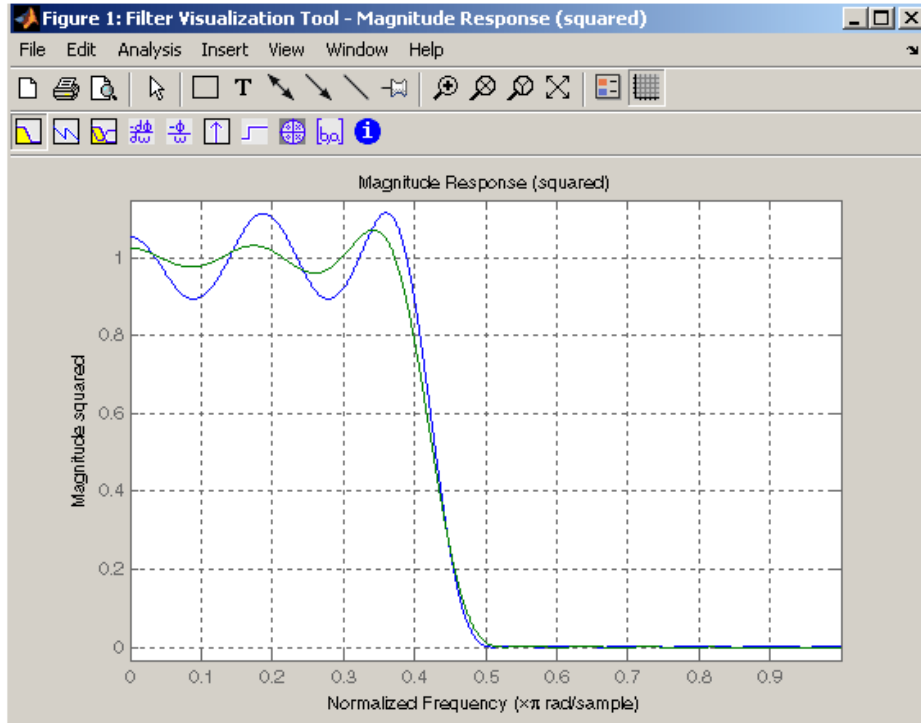
10.3 Design of FIR Filters



Type `bb = fir1s(n,f,a)`;

and compare their frequency responses using FVTool: `fvtool(b,1,bb,1)`

Note that the y-axis shown in the figure below is in Magnitude Squared. You can set this by right-clicking on the axis label and selecting Magnitude Squared from the menu.



The filter designed with `firpm` exhibits equiripple behavior. Also note that the `fir1s` filter has a better response over most of the passband and stopband, but at the band edges ($f = 0.4$ and $f = 0.5$), the response is further away from the ideal than the `firpm` filter. This shows that the `firpm` filter's maximum error over the passband and stopband is smaller and, in fact, it is the smallest possible for this band edge configuration and filter length.

10.3 Design of FIR Filters

Removing Periodic Signals of Known Frequency (High-pass Approach)

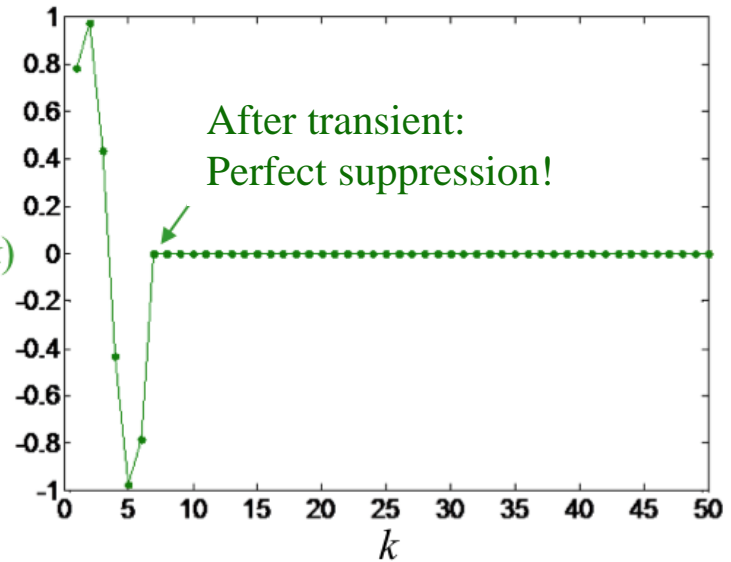
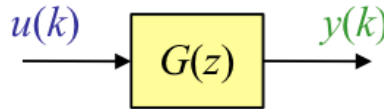
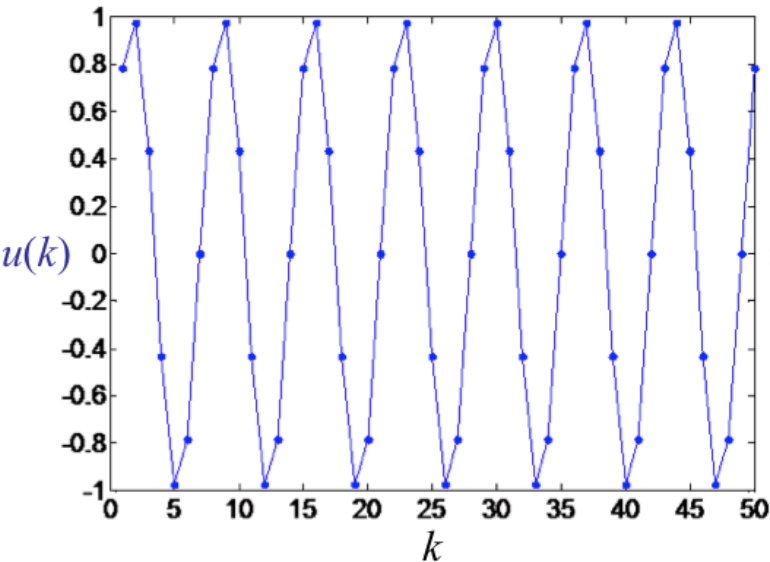
With an FIR filter a arbitrary periodic signal of known frequency can be removed perfectly. Typical applications:

- Carrier frequency of a radio signal
- Hum (50 Hz and multiples as upper harmonics)

The following FIR filter removes all periodic signals with period length $T_p = m \cdot T_0$ or frequency $\omega_p = \omega_0/m$, respectively:

$$G(z) = 1 - z^{-m}$$

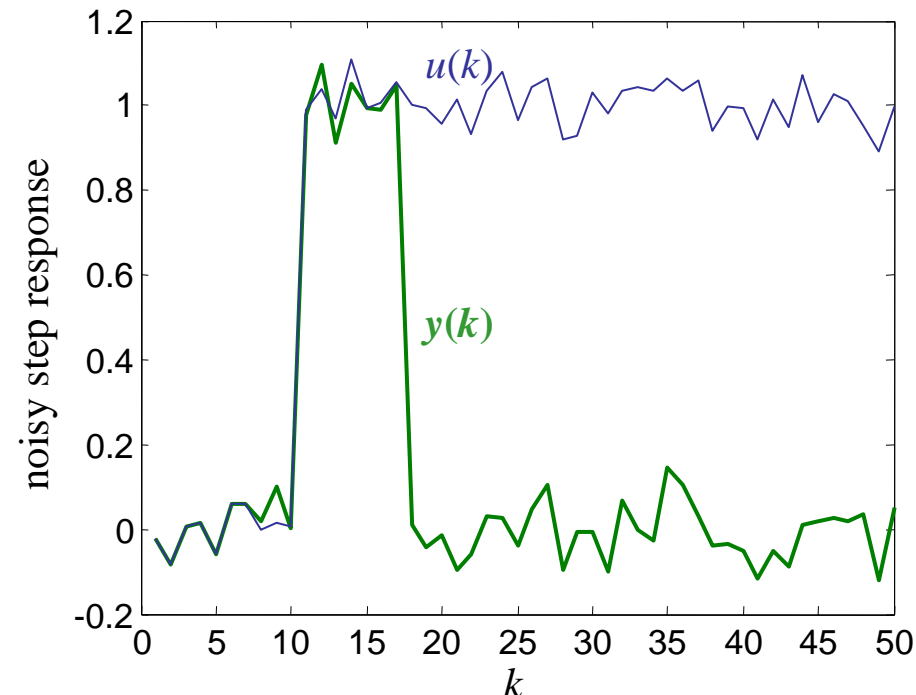
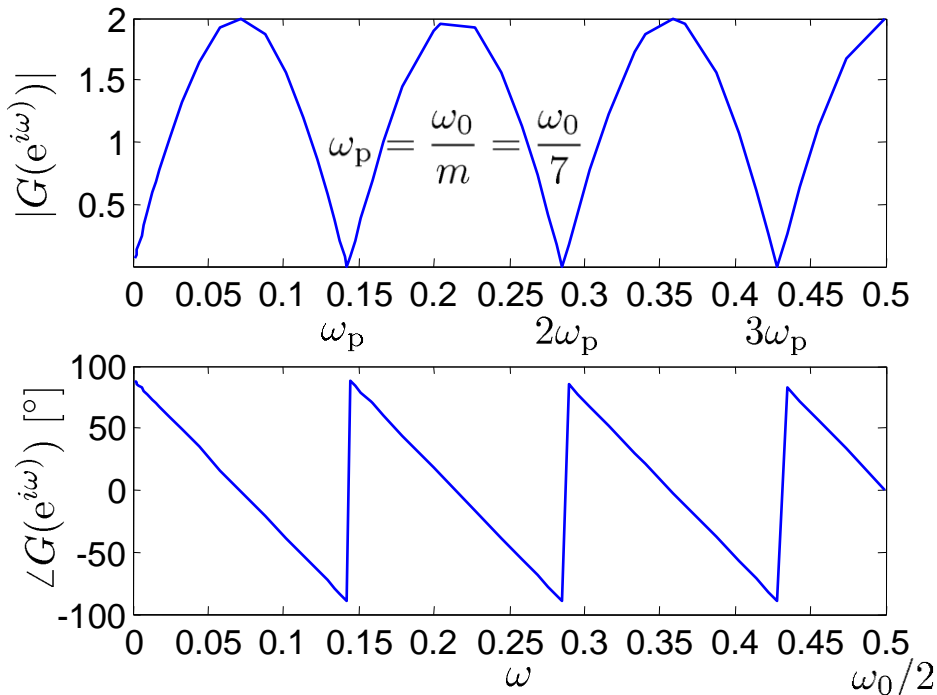
$m = 7$



10.3 Design of FIR Filters

The special properties of such a filter are:

- Independence of the shape of the signal (depends only on the period length).
- Removes all multiples of ω_p perfectly.
- Perfect damping with $-\infty$ dB (infinite steepness!).
- High-pass! Removes all low frequencies (and d.c. values) as well.

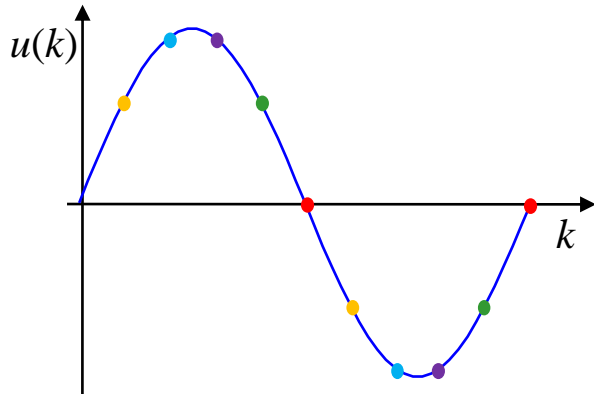


10.3 Design of FIR Filters

Removing Periodic Signals of Known Frequency (Low-pass Approach)

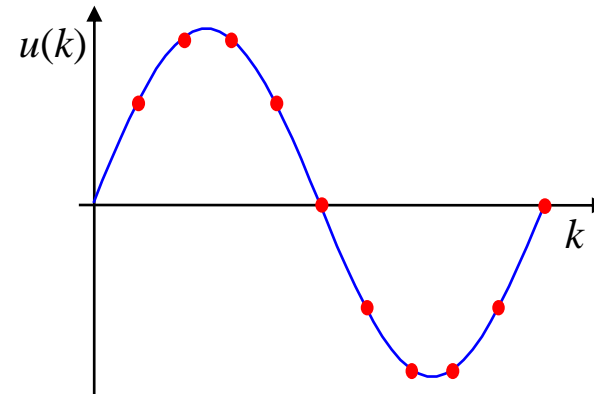
If the same task as before is requested with a low-pass filter instead of a high-pass the following two possibilities with gain = 1 suggest itself:

$$G_1(z) = \frac{1}{2} (1 + z^{-m/2})$$



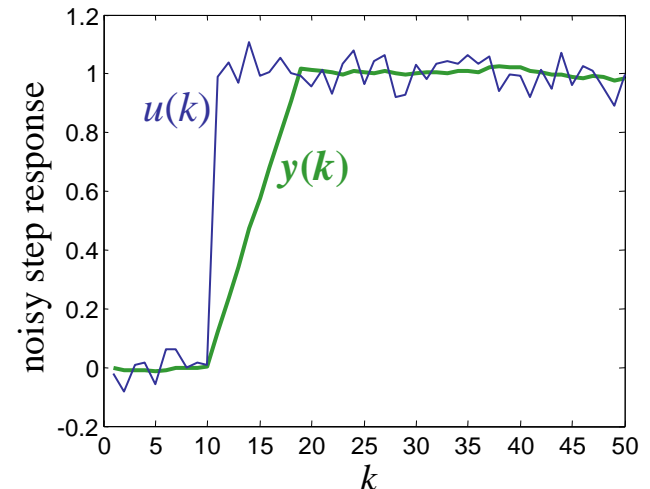
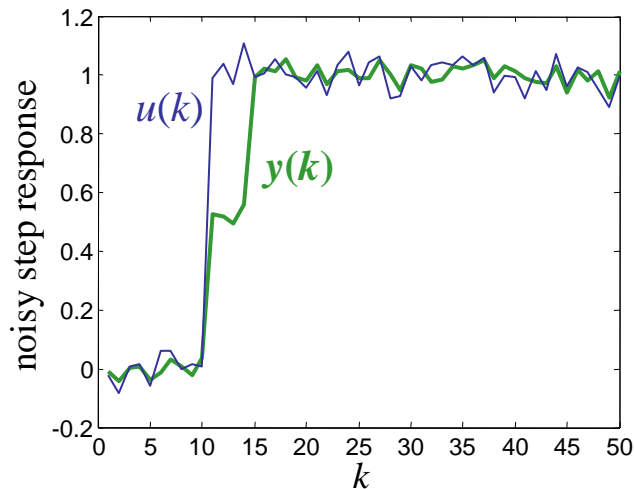
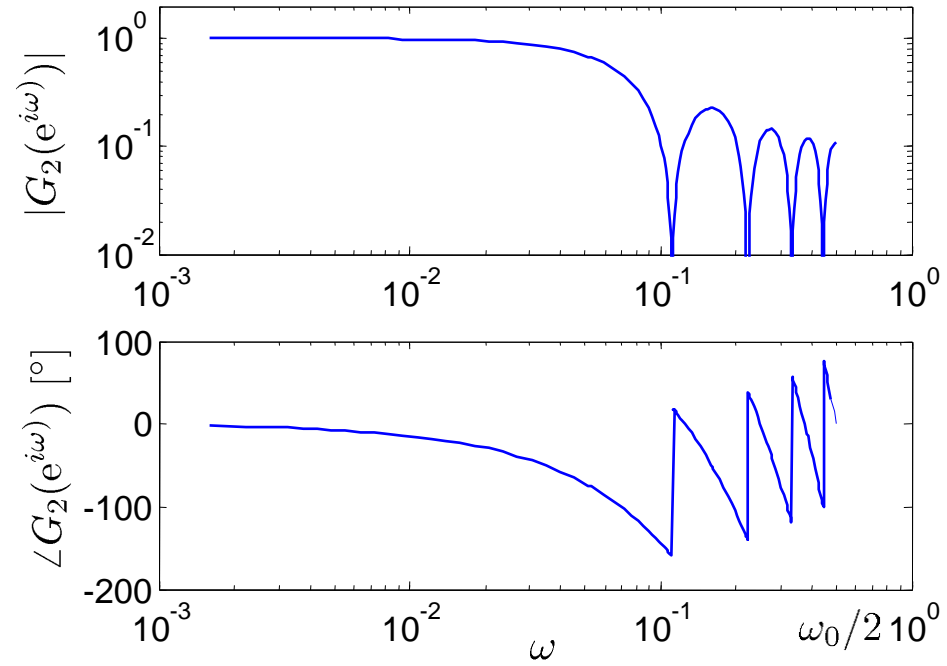
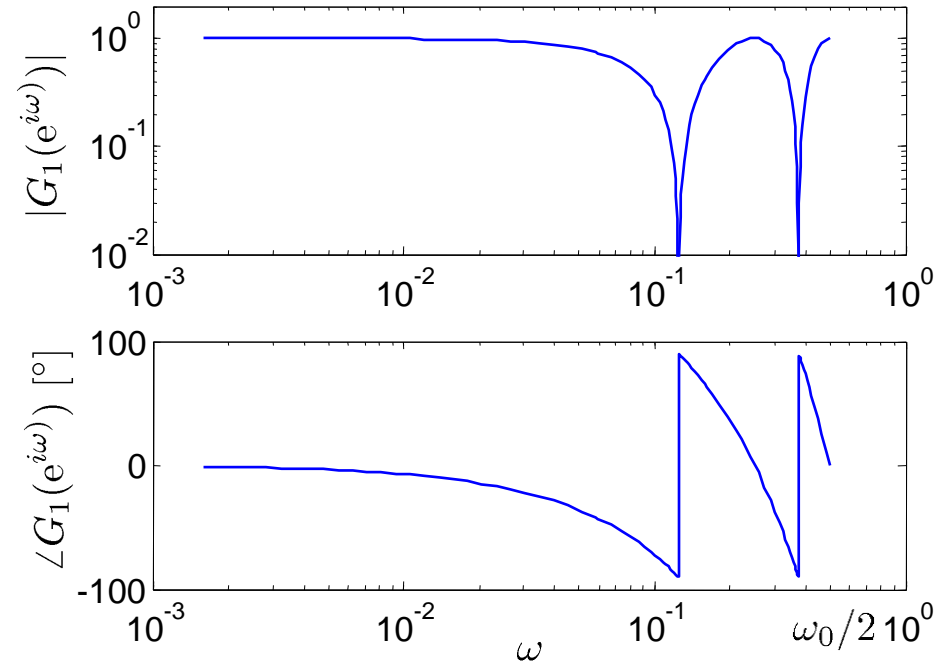
- Positive and negative half waves have to be symmetrical in order to cancel each other.
- m has to be even.
- Little distortion for other frequencies.
- Removes only multiples of $2\omega_p$.

$$G_2(z) = \frac{1}{m+1} (1 + z^{-1} + z^{-2} + \dots + z^{-m})$$



- Positive and negative half waves must accumulated to zero.
- Strong averaging (low-pass effect).
- Removes only multiples of ω_p .

10.3 Design of FIR Filters



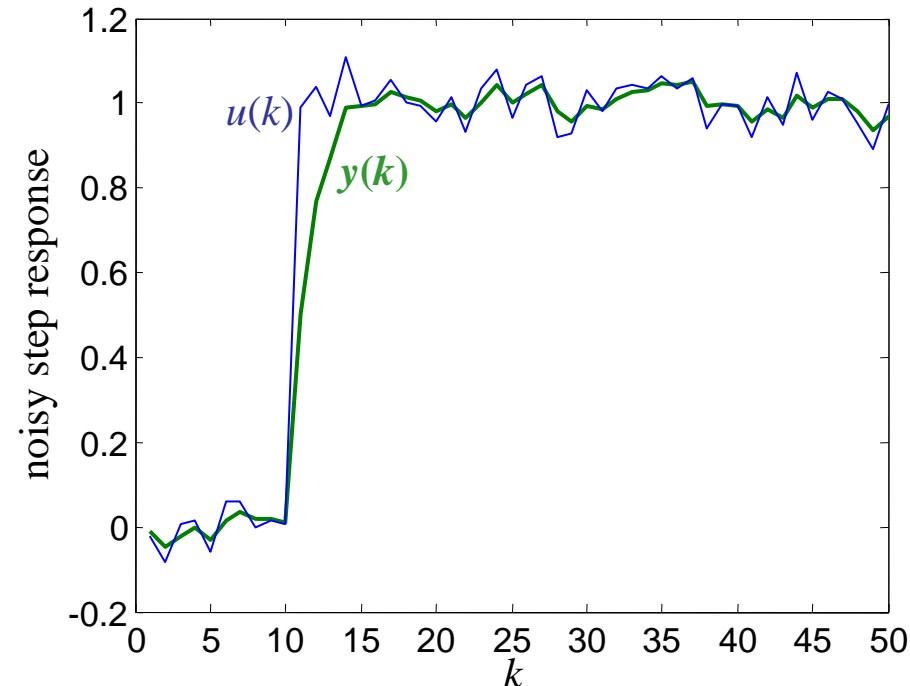
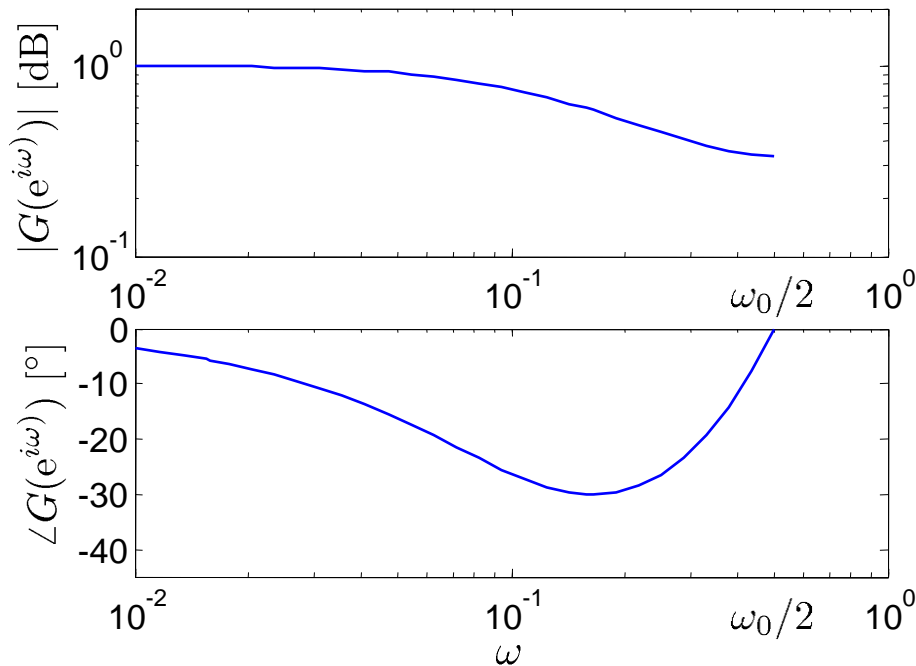
10.4 Design of IIR Filters

Example: A Simple IIR Filter of 1. Order

For IIR filters the output is a weighted average of the current and previous inputs (*moving average, MA*) and previous outputs (*autoregressive, AR*) \rightarrow ARMA. The most simple first order IIR filter is a PT₁-system, i.e.:

$$\begin{aligned} y(k) &= 0.5u(k) + 0.5y(k - 1) \\ &= 0.5u(k) + 0.25u(k - 1) + 0.125u(k - 2) + \dots \end{aligned}$$

In comparison to FIR filters, here implicitly infinitely old inputs $u(k-i)$ influence the output!



10.4 Design of IIR Filters

Transformation from Analog in Digital

Typically IIR filters are designed with traditional methods in the analog world. In a second step they are transformed from the analog in the digital world. For this transformation various approaches are common, dependent on the application area:

- Impulse invariance method: Demand identical impulse response in the analog and digital.
- Bilinear transformation (also called: *Tustin* formula): The s -variable in the analog frequency domain is *approximated* by a rational fractional function in z such that a numerator / denominator expression in the s -domain becomes a numerator / denominator expression in the z -domain (and vice versa).

Furthermore there exist other method that are more popular in digital control than in filter design:

- Identical time signals with zero or first order hold.
- Identical poles and zeros.

In the following, we focus on the **bilinear transformation** approach.

10.4 Design of IIR Filters

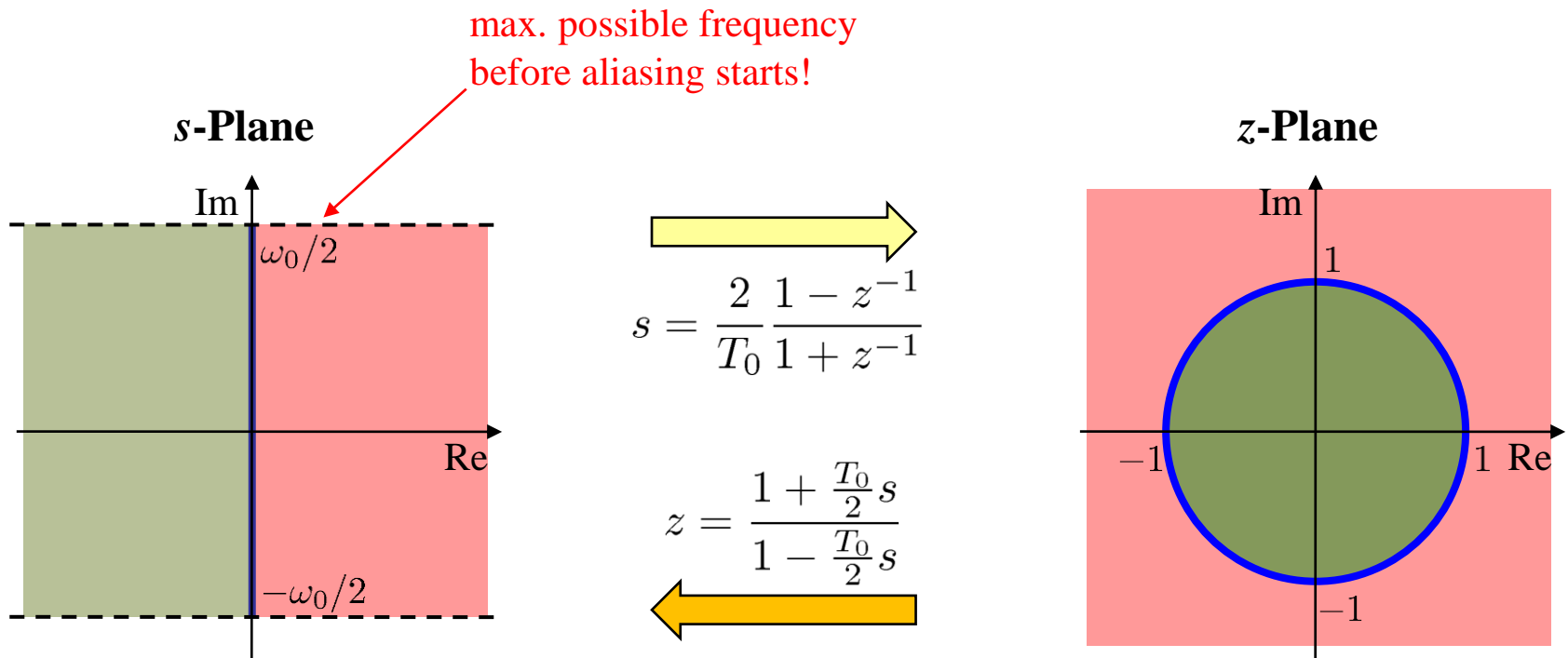
Bilinear Transformation (Tustin Formula)

The exact transformation between s and z is nonlinear and would destroy the fractional rational function form. Linear system theory would not apply anymore.

$$z = e^{sT_0}$$

$$s = \frac{1}{T_0} \ln z$$

Via the bilinear transformation this form is preserved. The stability properties stay identical, as well.



10.4 Design of IIR Filters

Comparison: Frequency Response in the Analog and Digital World

A transfer function $G_a(s)$ in the s -domain can approximately be transformed by the bilinear transformation into the z -domain:

$$G_a(s) = G_d \left(\frac{2}{T_0} \frac{1 - z^{-1}}{1 + z^{-1}} \right)$$

The *frequency response* in the analog can be obtained by $s = i\omega_a$ and correspondingly by going through the unit circle in the z -domain $z = e^{i\omega_d T_0}$. Since the bilinear transformation is just an *approximation*, the analog frequency ω_a differs from the digital frequency ω_d :

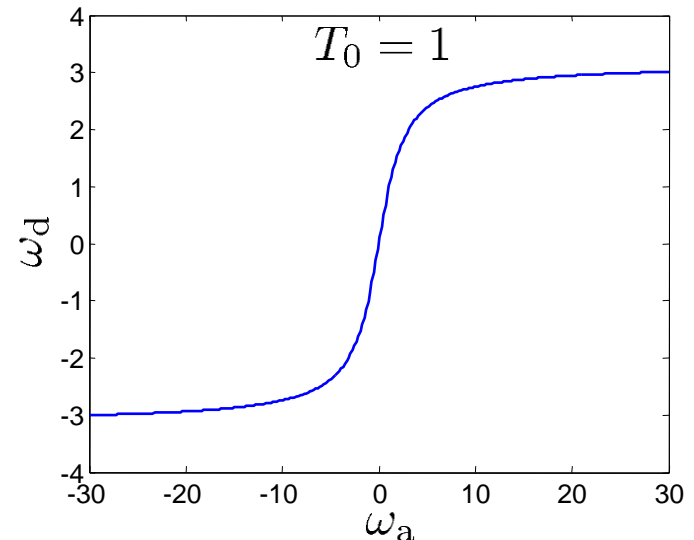
$$i\omega_a = \frac{2}{T_0} \frac{1 - e^{-i\omega_d T_0}}{1 + e^{-i\omega_d T_0}} = i \frac{2}{T_0} \tan \frac{\omega_d T_0}{2}$$

$$\omega_a = \frac{2}{T_0} \tan \frac{\omega_d T_0}{2}$$

$$\omega_d = \frac{2}{T_0} \arctan \frac{\omega_a T_0}{2}$$

The upper bound for the digital frequency is given by the half sampling frequency according to Shannon:

$$\omega_{d,\max} = \frac{\pi}{T_0} = \pi f_0 = \frac{\omega_0}{2}$$



10.4 Design of IIR Filters

Bilinear Transformation (Tustin Formula) = Trapezoidal Rule for Integration

In discrete time a continuous integration can be approximated in different ways. More accurate than calculating the lower or upper sum (see next slide) is the trapezoidal rule:

$$y(k) = y(k - 1) + T_0 \frac{u(k) + u(k - 1)}{2}$$

width
medium height

In the z-domain this results in:

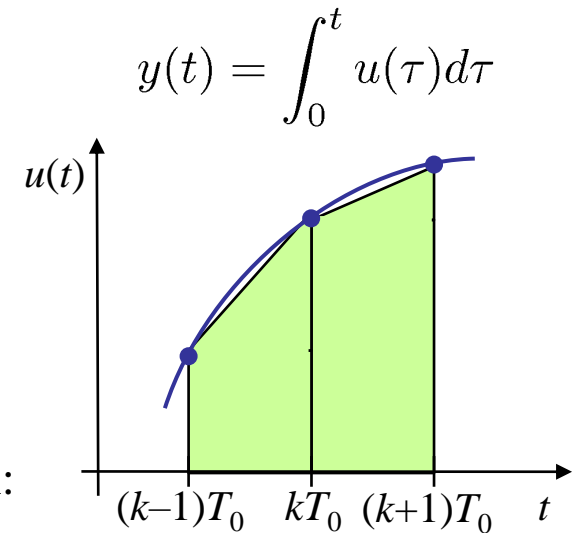
$$Y(z) = \frac{T_0}{2} \frac{1 + z^{-1}}{1 - z^{-1}} U(z)$$

This formula shall correspond to an integration in the s-domain:

$$Y(s) = \frac{1}{s} U(s)$$

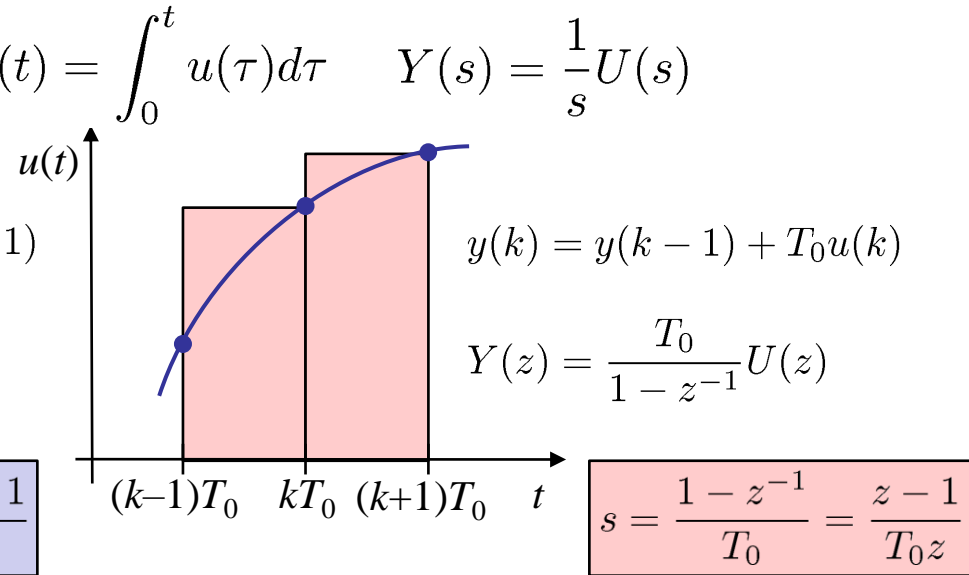
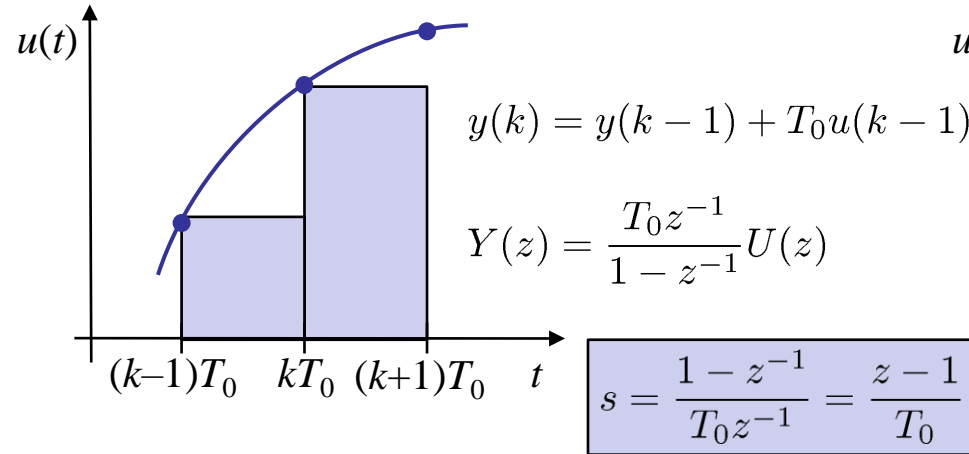
This exactly yields the bilinear transformation:

$$\frac{1}{s} = \frac{T_0}{2} \frac{1 + z^{-1}}{1 - z^{-1}} \rightarrow s = \frac{2}{T_0} \frac{1 - z^{-1}}{1 + z^{-1}}$$

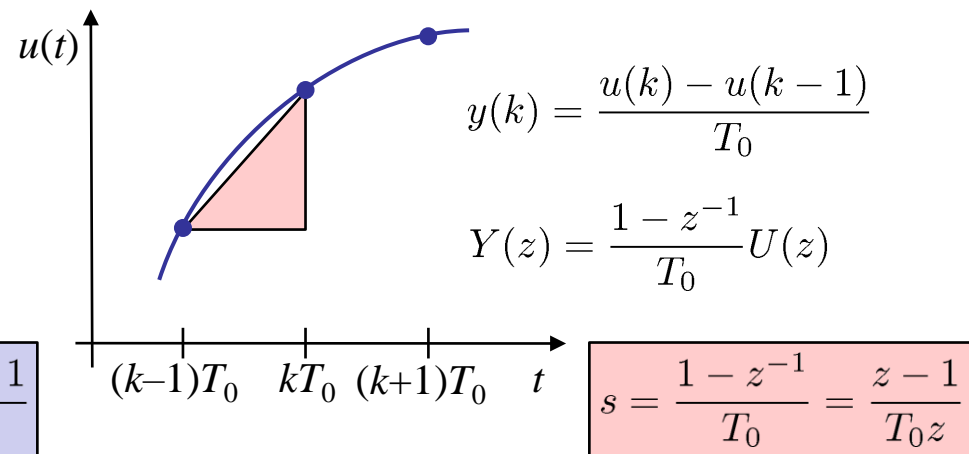
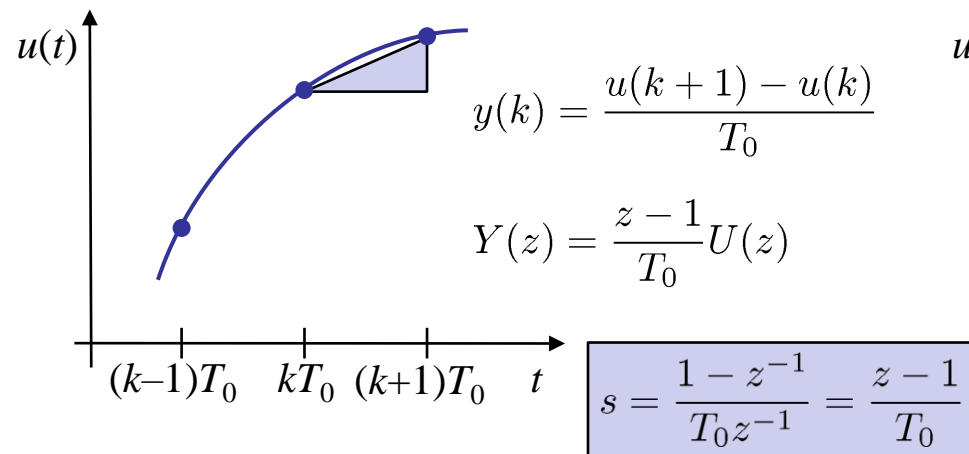


10.4 Design of IIR Filters

Integration with Lower and Upper Sum $y(t) = \int_0^t u(\tau)d\tau$ $Y(s) = \frac{1}{s}U(s)$



Differentiation with Forward and Backward Differences $y(t) = \frac{d}{dt}u(t)$ $Y(s) = sU(s)$



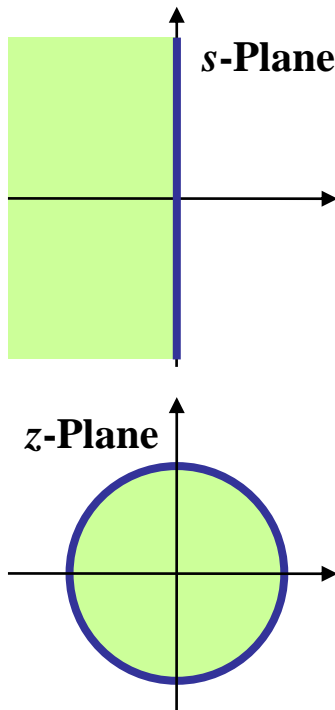
10.4 Design of IIR Filters

Comparison

Bilinear Transformation (Trapezoidal Integration)

$$s = \frac{2}{T_0} \frac{1 - z^{-1}}{1 + z^{-1}}$$

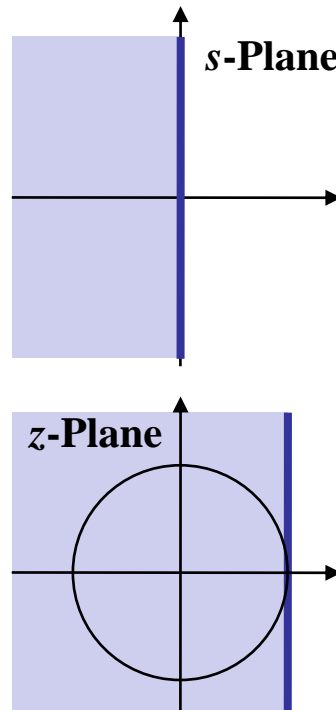
$$z = \frac{1 + \frac{T_0}{2}s}{1 - \frac{T_0}{2}s}$$



Forward Differences (Lower Sum Integration)

$$s = \frac{1 - z^{-1}}{T_0 z^{-1}}$$

$$z = 1 + T_0 s$$



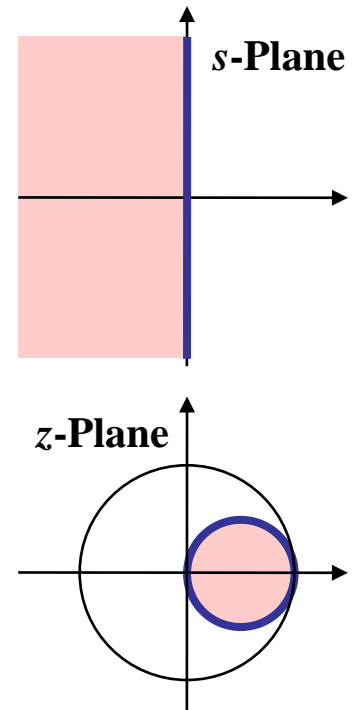
Stability area is mapped too large



Backward Differences (Upper Sum Integration)

$$s = \frac{1 - z^{-1}}{T_0}$$

$$z = \frac{1}{1 - T_0 s}$$



small

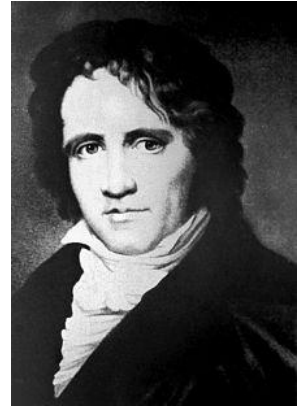


10.4 Design of IIR Filters

Procedure for Filter Design Via Bilinear Transformation

1. Specification is either directly made in the analog world or it is transformed from the digital in the analog world.
2. Filter design in the analog world.
3. Transformation of the final analog filter in the digital world.

Friedrich Bessel, 1784-1846
(www.wikipedia.org)



The Following IIR Filters are Common:

- Bessel filter: Approximately linear phase in the pass-band
- Butterworth filter: Monotone amplitude response
- Chebyshev filter type 1: Ripple in the pass-band
- Chebyshev filter type 2: Ripple in the stop-band
- Cauer filter (elliptic filter): Ripple in the pass- and stop-band

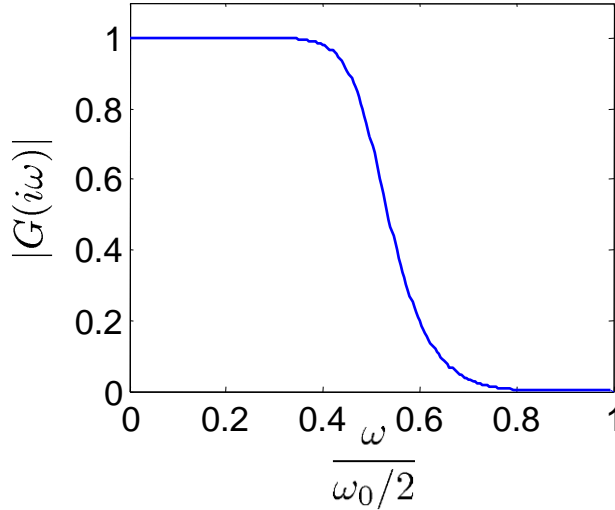
For the steepness of filters of identical orders (i.e., comparable complexity):

Bessel < Butterworth < Chebyshev < Cauer

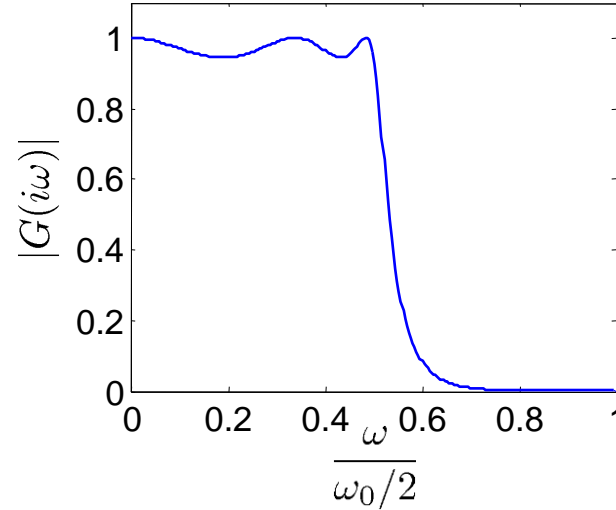
10.4 Design of IIR Filters

Overview on Analog Filters

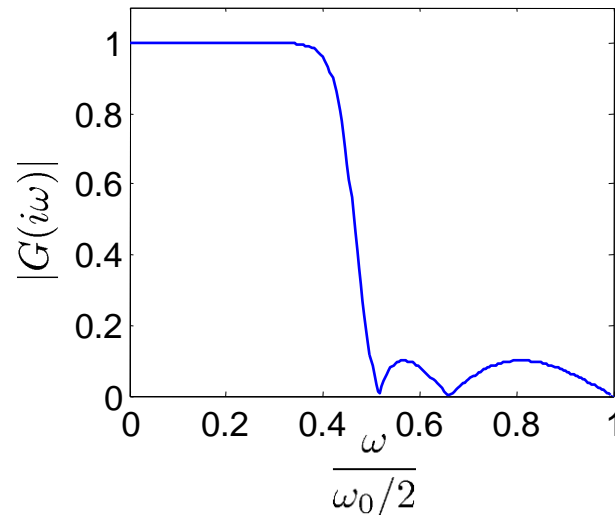
Butterworth



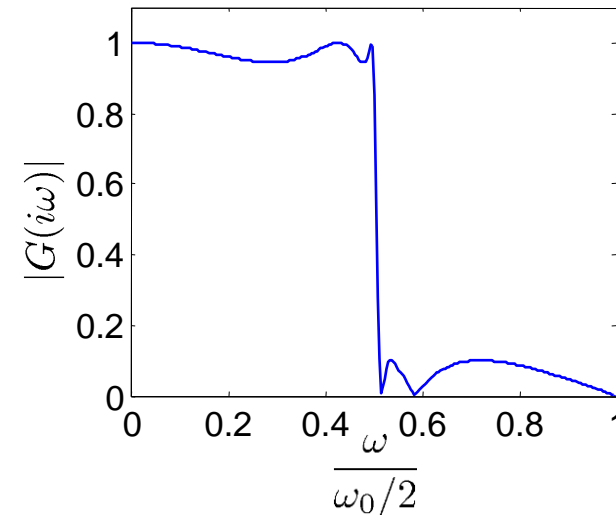
**Chebyshev
Type I**



**Chebyshev
Type II**



**Cauer
(elliptic)**



10.4 Design of IIR Filters

Butterworth Filter

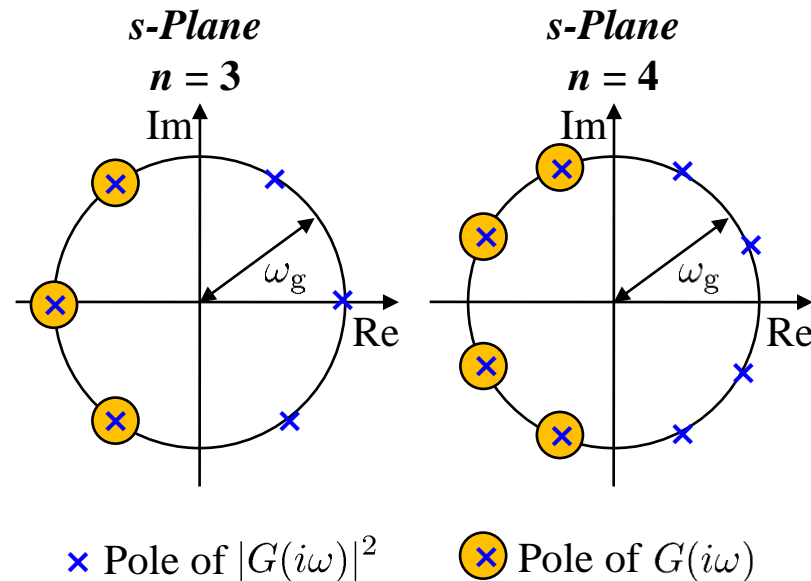
- Design with focus on maximal flatness of the amplitude response close to the limit frequency ω_g .
- Monotone amplitude response, i.e., no ripples.
- Fast drop-off in the amplitude response at the limit frequency.
- Strong overshoot of the impulse response.
- Relative low steepness with $20 \cdot n$ dB / decade (n = filter order).

Amplitude Response:

$$|G(i\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_g}\right)^{2n}}$$

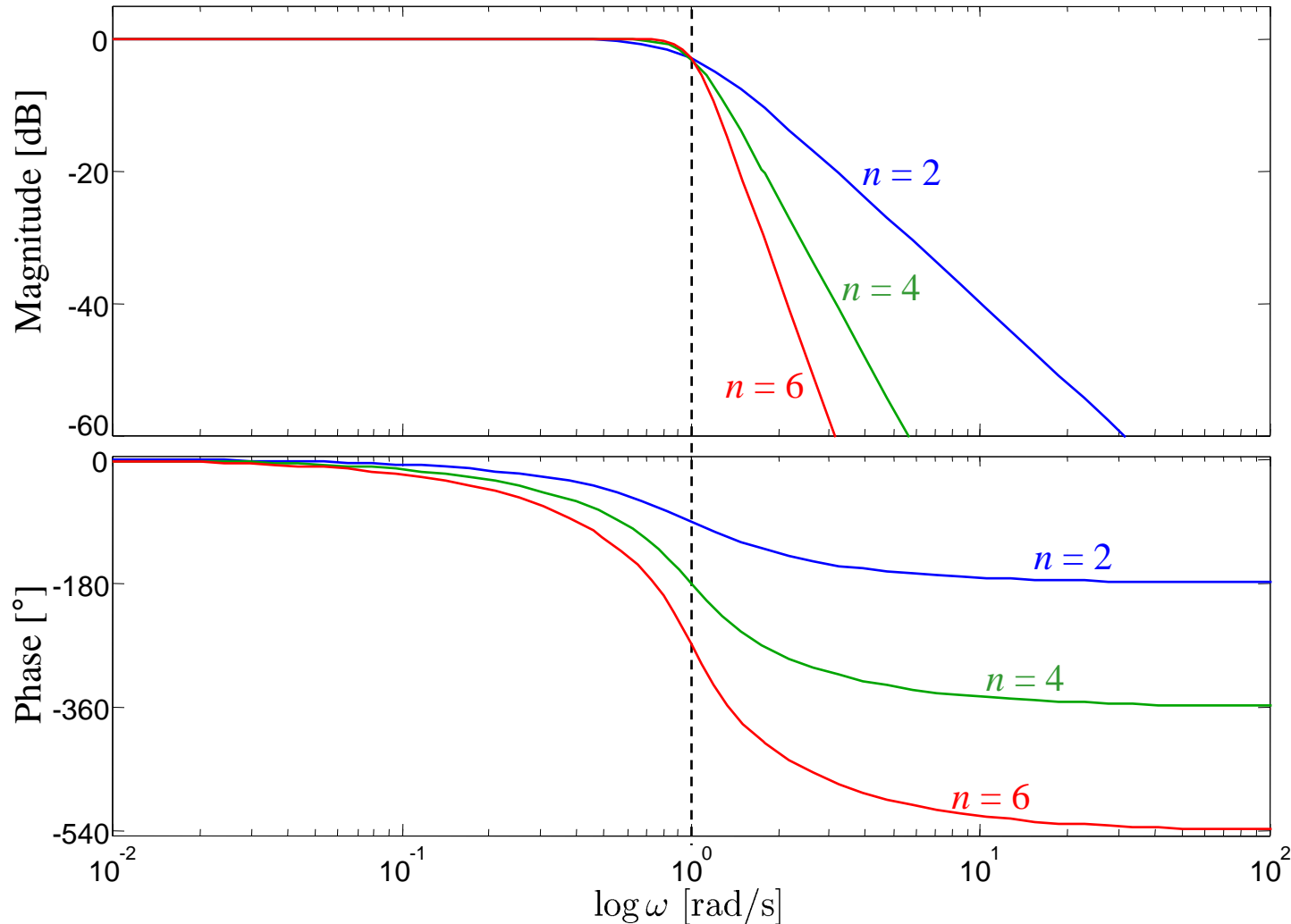
$$G(i\omega) = \prod_{i=1}^n \frac{s_i}{s_i - s}$$

where s_i are the n stable poles of the $2n$ -root of $-\omega_g$.

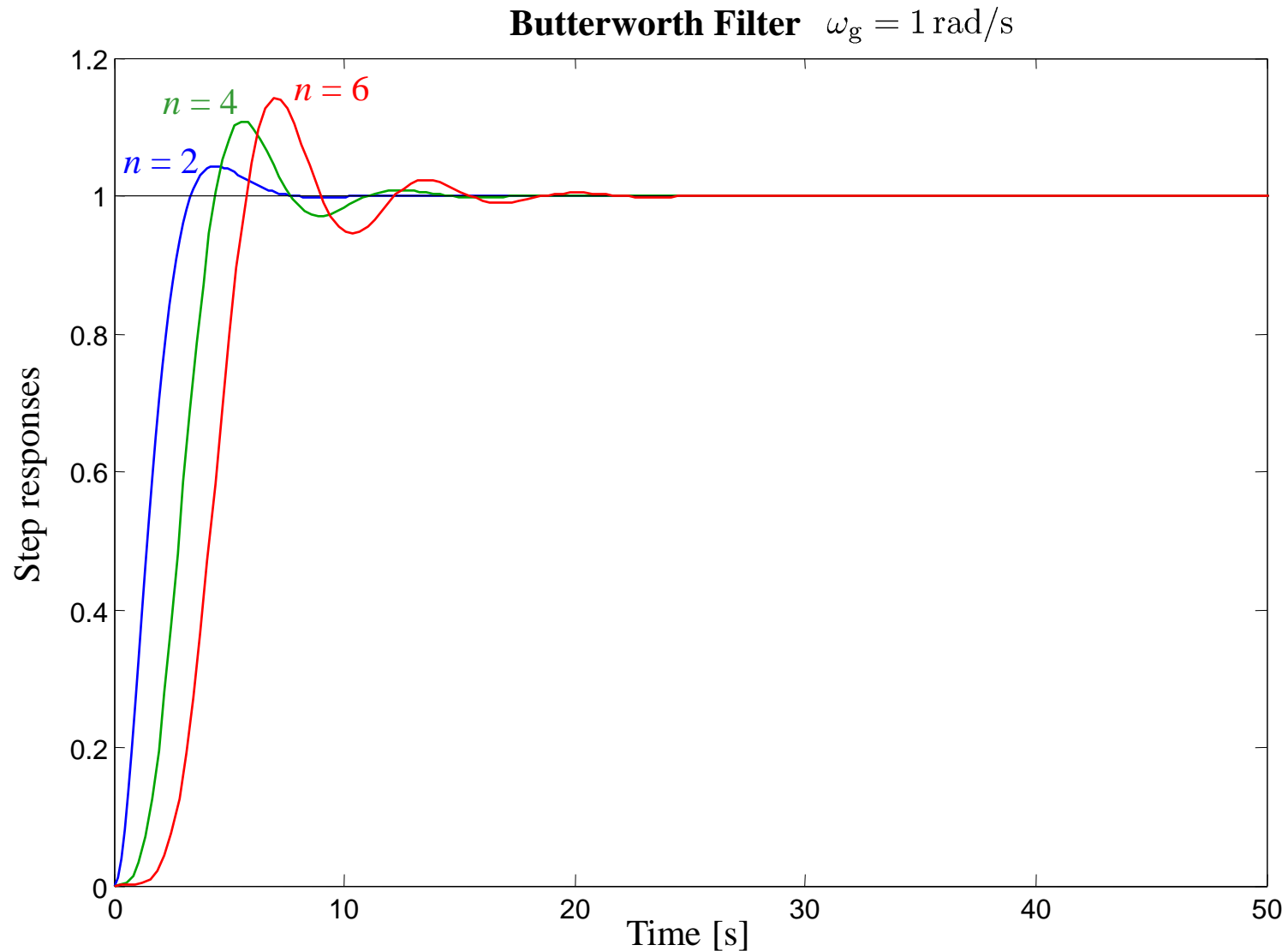


10.4 Design of IIR Filters

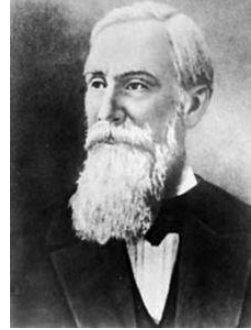
Butterworth-Filter $\omega_g = 1 \text{ rad/s}$



10.4 Design of IIR Filters



10.4 Design of IIR Filters



Chebyshev Filter

- Steeper than Butterworth filter.
- Ripples in pass-band (type I) *or* stop-band (type II) in the amplitude response. Acceptance of ripple drawback for benefits in steepness.
- Step response oscillates more than for Butterworth filter.
- Transposes into Butterworth filter if the allowed ripple factor $\varepsilon \rightarrow 0!$
- Design parameters: limit frequency ω_g , order n , allowed ripple factor ε .

$$|G(i\omega)|_I^2 = \frac{1}{1 + \varepsilon^2 T_n^2\left(\frac{\omega}{\omega_g}\right)}$$

ε : ripple factor

Because the Chebyshev polynomial changes in the pass-band between 0 and 1 a lower limit on the gain is given by:

$$1 - \delta_1 = \frac{1}{\sqrt{1 + \varepsilon^2}}$$

Chebyshev Polynomial of Order n :

$$T_n(x) = \begin{cases} \cos(n \arccos x) & \text{für } |x| \leq 1 \\ \cosh(n \operatorname{arccosh} x) & \text{für } |x| > 1 \end{cases}$$

$$T_0(x) = 1$$

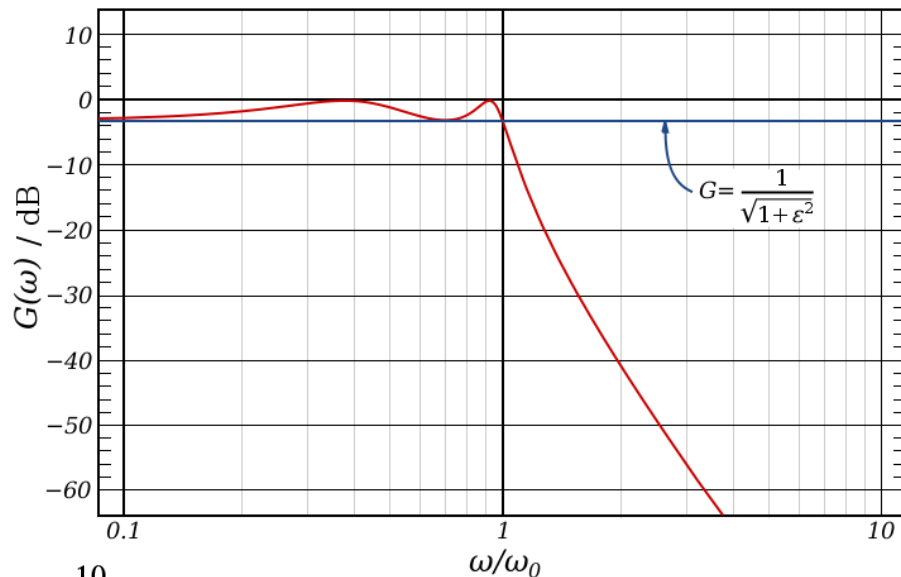
$$T_1(x) = x$$

$$T_2(x) = 2x^2 - 1$$

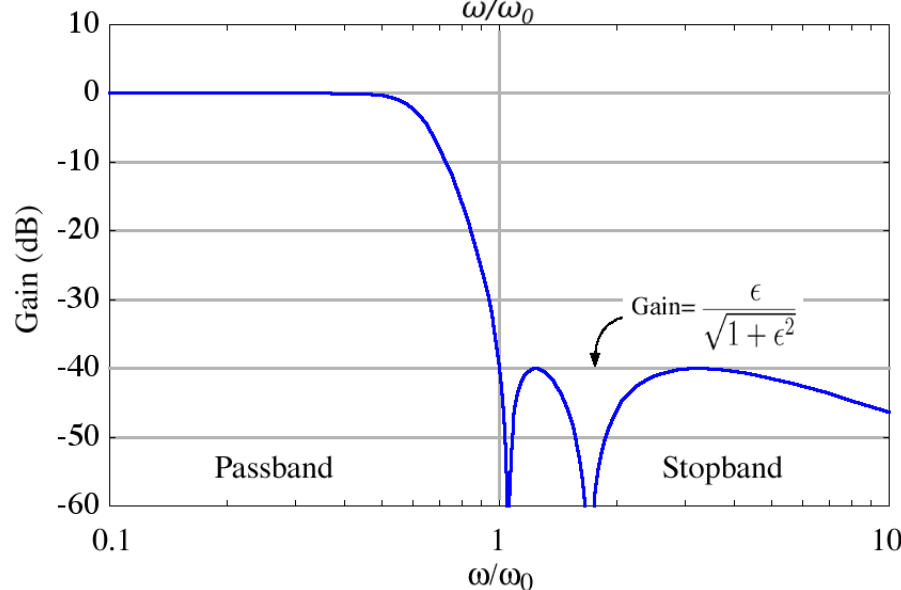
$$T_3(x) = 4x^3 - 3x$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

10.4 Design of IIR Filters



The frequency response of a fourth-order type I Chebyshev low-pass filter with $\epsilon = 1$

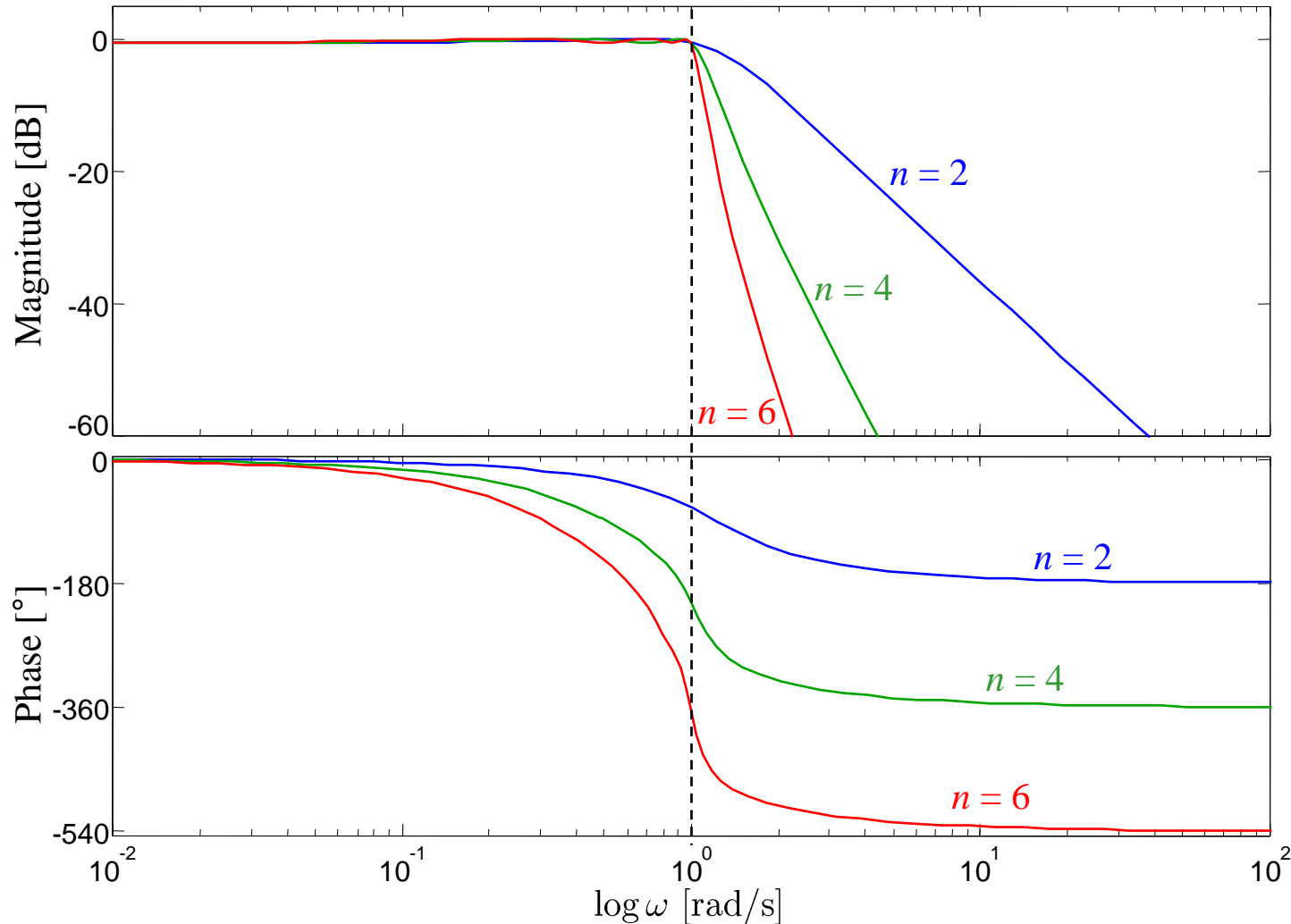


The frequency response of a fifth-order type II Chebyshev low-pass filter with $\epsilon = 0.01$

Source: https://en.wikipedia.org/wiki/Chebyshev_filter

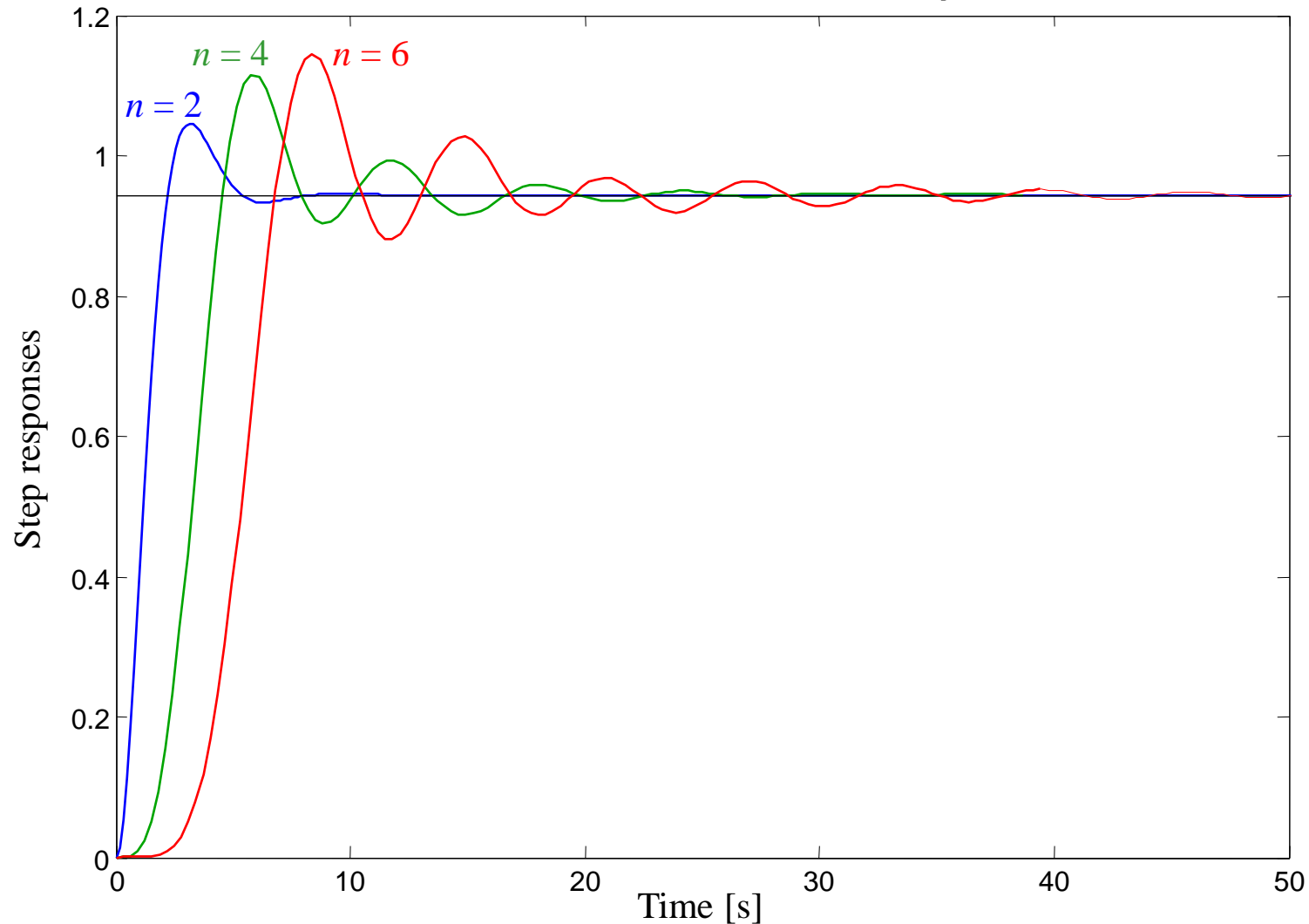
10.4 Design of IIR Filters

Chebyshev Filter Type I $\omega_g = 1 \text{ rad/s}$



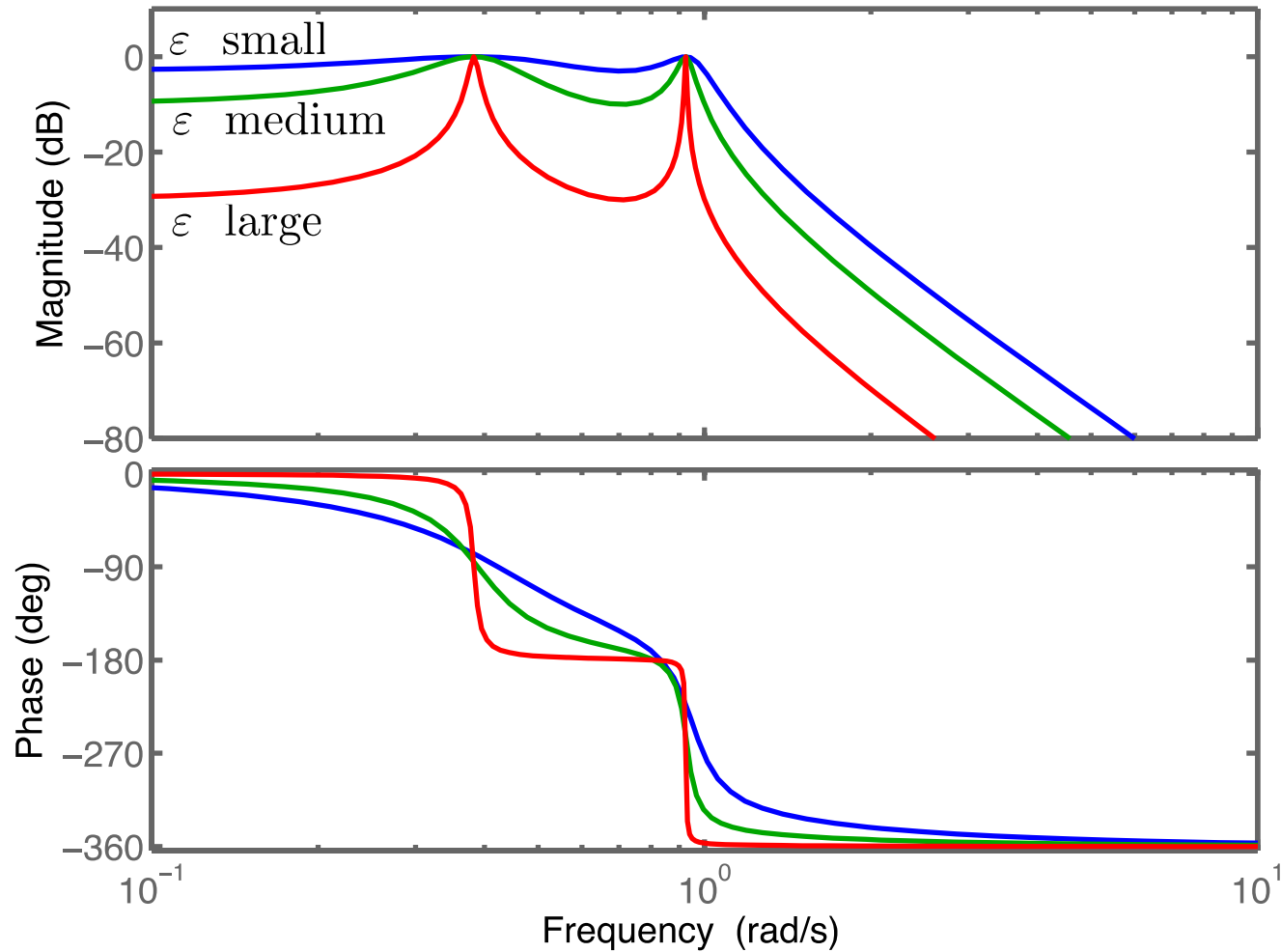
10.4 Design of IIR Filters

Chebyshev Filter Type I $\omega_g = 1 \text{ rad/s}$



10.4 Design of IIR Filters

Chebyshev Filter Type I ($n = 4$) $\omega_g = 1$ rad/s



10.4 Design of IIR Filters

Cauer Filter (Elliptic Filter)

- Steeper than Chebyshev filter, even the steepest possible (for linear filters).
- Ripples in the pass-band *and* stop-band in the amplitude response.
- Step response oscillates stronger than for the Chebyshev filter.
- Transposes into Chebyshev filter type I if the steepness factor $\xi \rightarrow \infty!$
- Design parameters: limit frequency ω_g , order n , ripple ε and steepness ξ .

$$|G(i\omega)|^2 = \frac{1}{1 + \varepsilon^2 R_n^2 \left(\frac{\omega}{\omega_g}, \xi \right)}$$

ε : Ripple for pass-band

ξ : Steepness (selectivity factor)

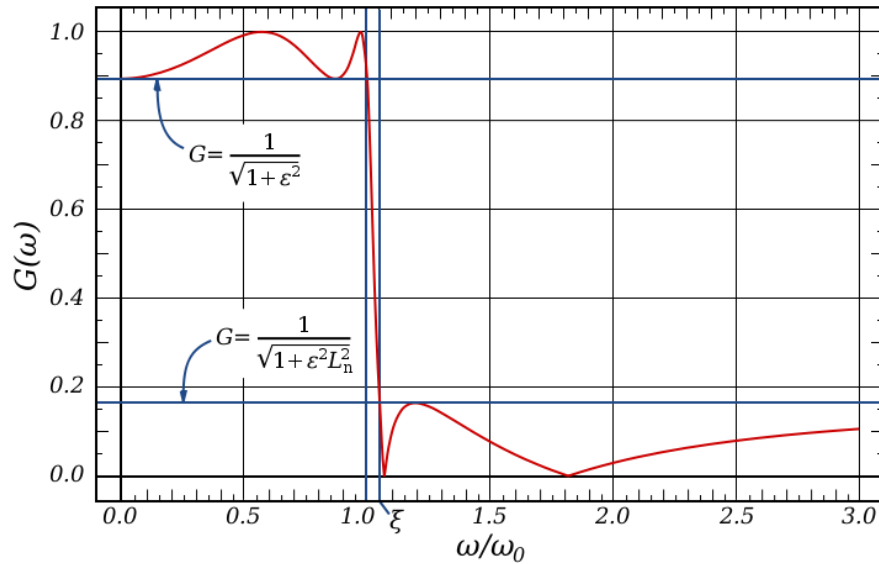
x_{ni} : zeros ← calculated according to a complex formula in dependence on ξ .
 x_{pi} : poles ←

Maximal steepness at $x = 1$.

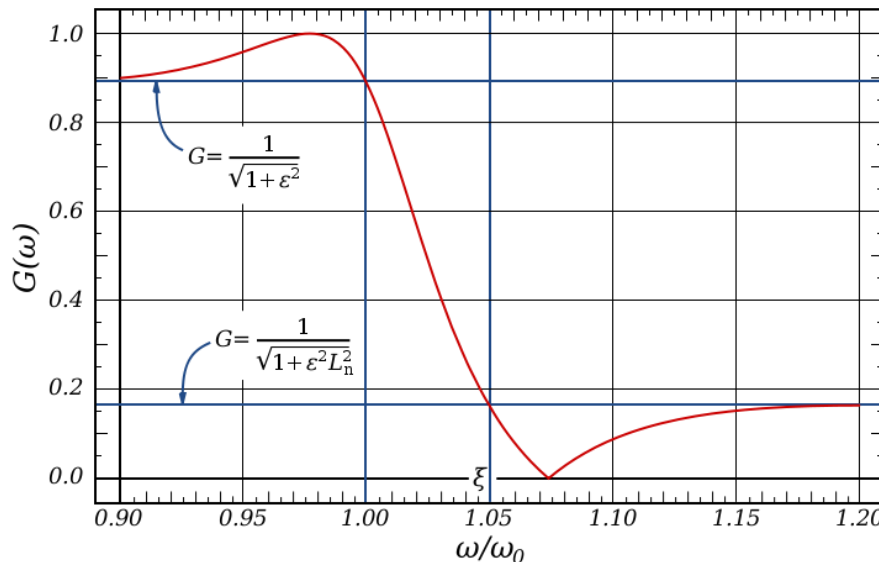
Elliptic functions of order n :

$$R_n(x, \xi) = \begin{cases} r_0 \frac{\prod_{i=0}^n (x - x_{ni})}{\prod_{i=0}^{n-1} (x - x_{pi})} & \text{for even } n \\ r_0 x \frac{\prod_{i=0}^{n-1} (x - x_{ni})}{\prod_{i=0}^{n-1} (x - x_{pi})} & \text{for odd } n \end{cases}$$

10.4 Design of IIR Filters



The frequency response of a fourth-order elliptic low-pass filter with $\epsilon=0.5$ and $\xi=1.05$. Also shown are the minimum gain in the passband and the maximum gain in the stopband, and the transition region between normalized frequency 1 and ξ

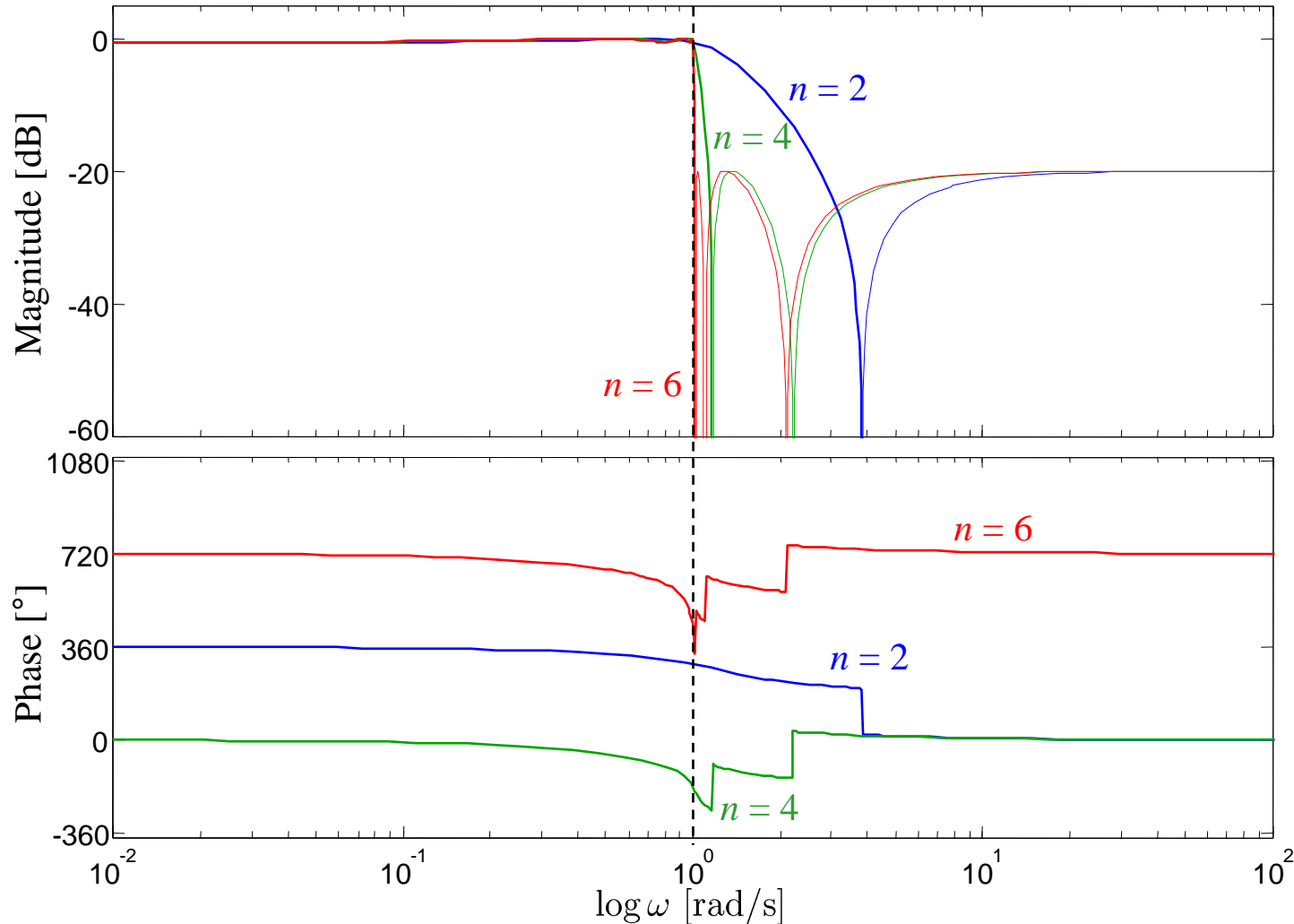


A closeup of the transition region of the above plot.

Source: https://en.wikipedia.org/wiki/Elliptic_filter

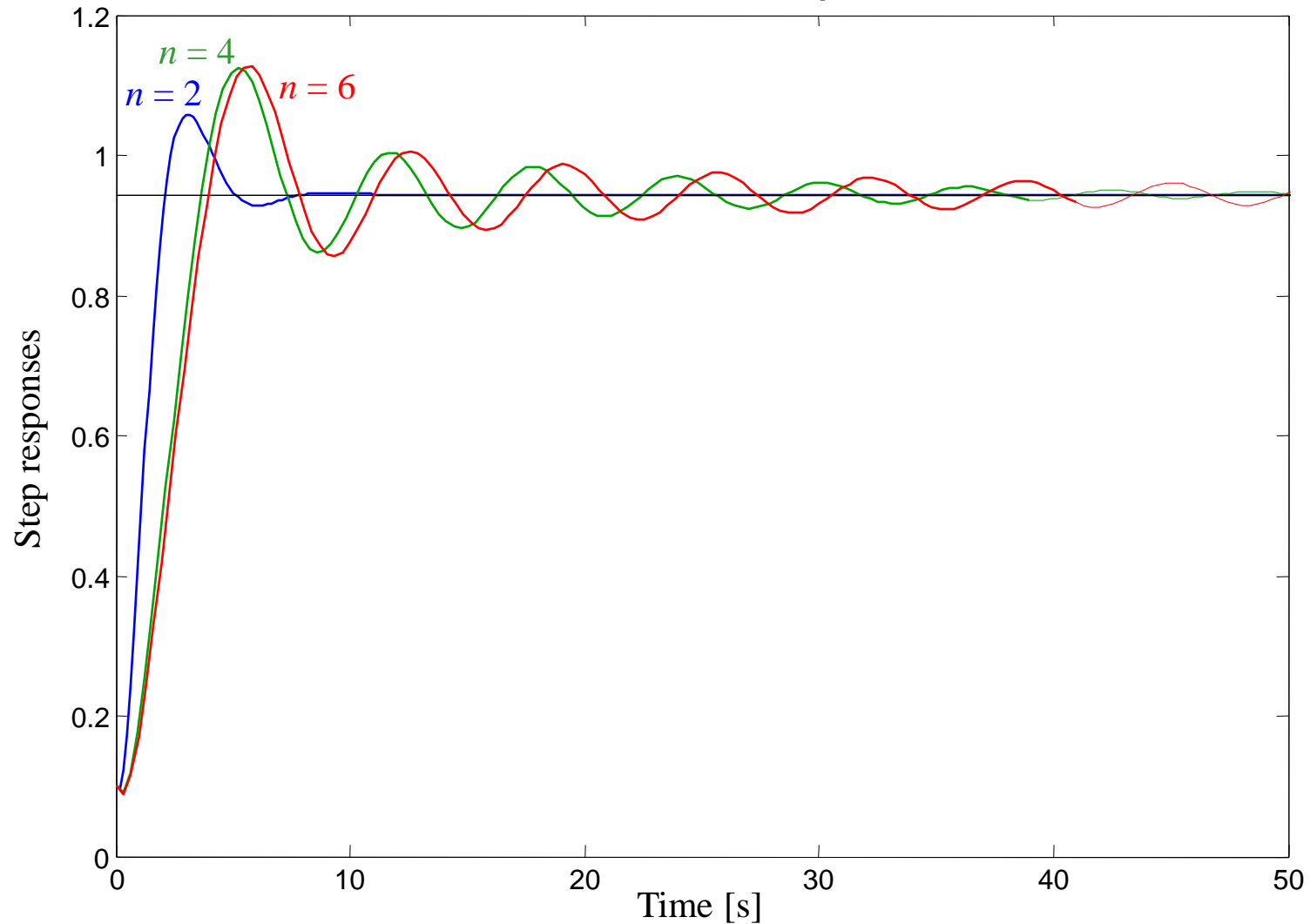
10.4 Design of IIR Filters

Cauer Filter $\omega_g = 1 \text{ rad/s}$



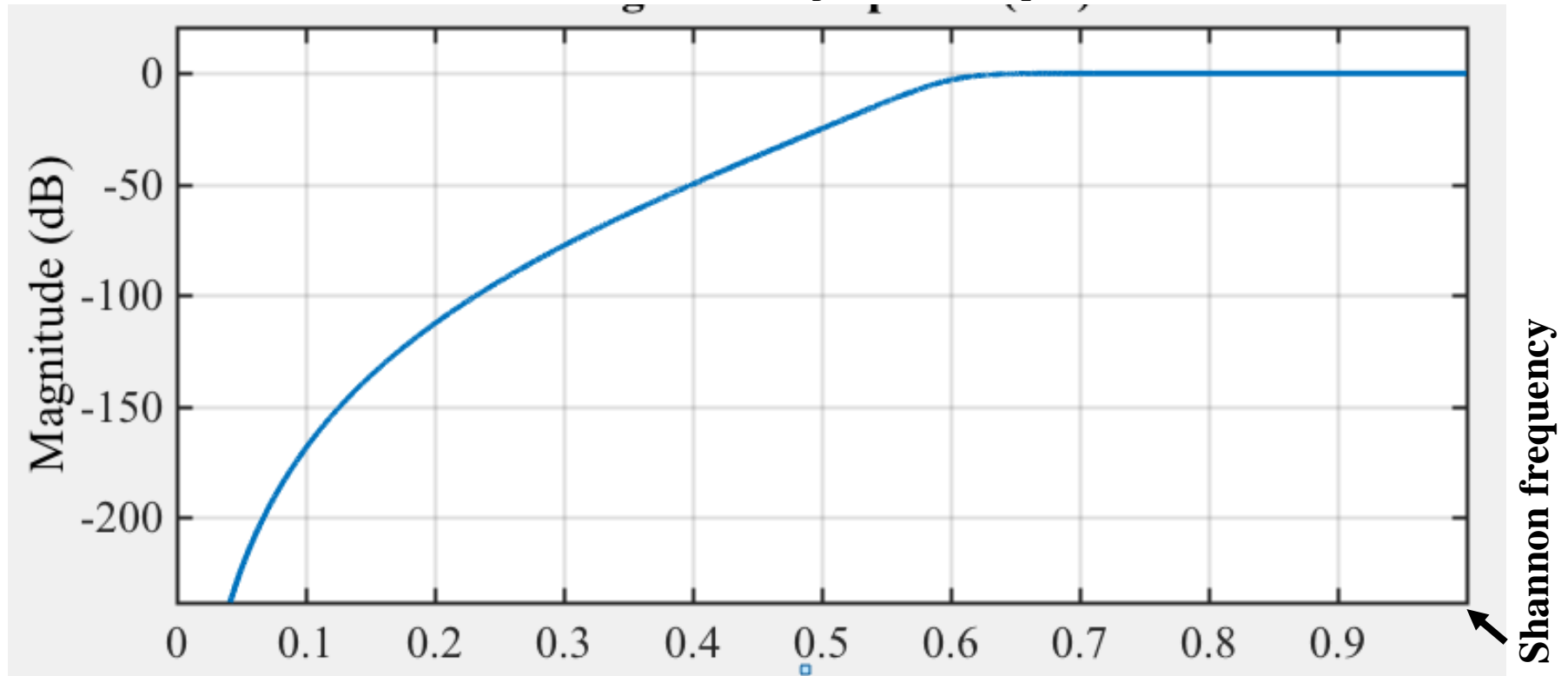
10.4 Design of IIR Filters

Cauer Filter $\omega_g = 1 \text{ rad/s}$



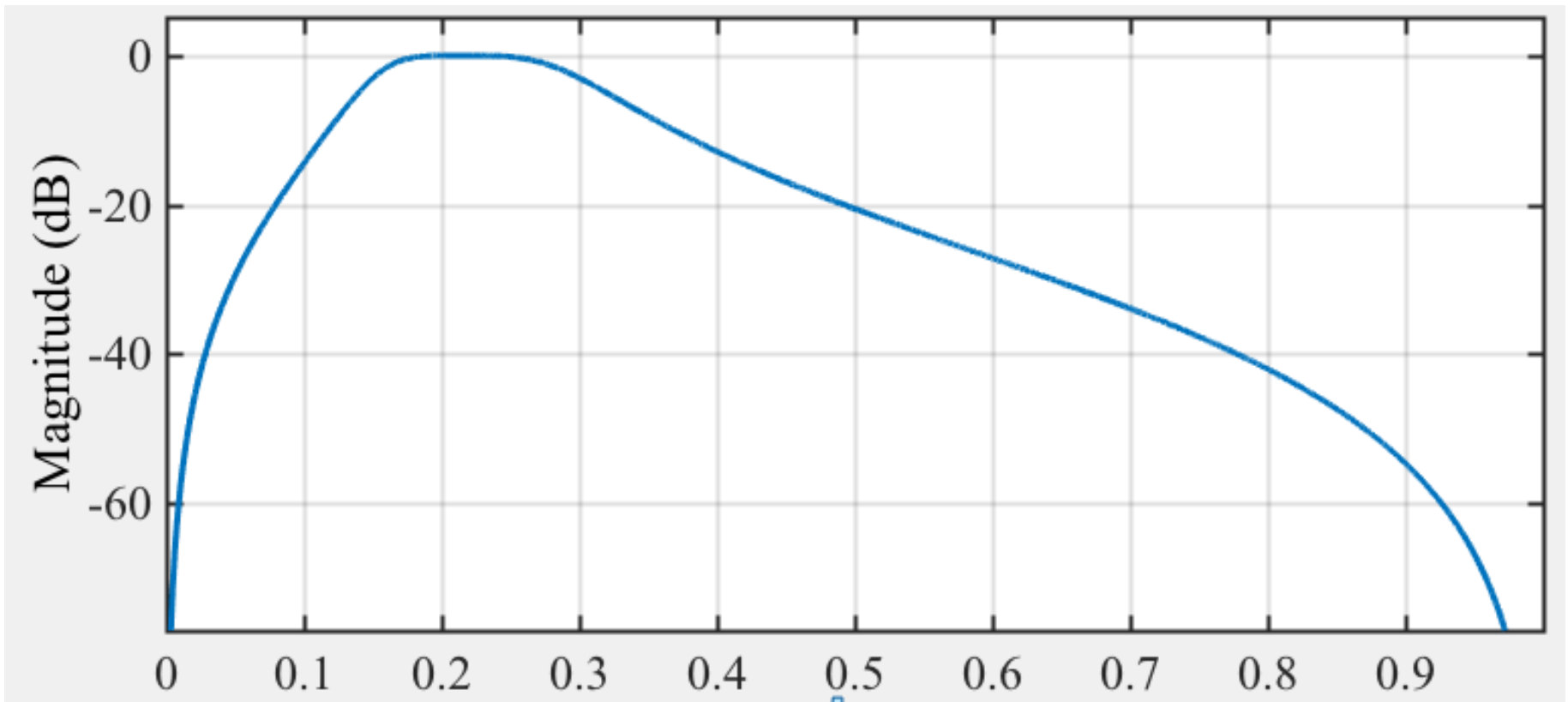
4.4 Entwurf von IIR-Filtern

```
% For data sampled at 1000 Hz, design a 9th-order highpass
% Butterworth filter with cutoff frequency of 300Hz.
Wn = 300/500; % Normalized cutoff frequency
[z,p,k] = butter(9,Wn,'high'); % Butterworth filter
[sos] = zp2sos(z,p,k); % Convert to SOS form
h = fvtool(sos); % Plot magnitude response
```



4.4 Entwurf von IIR-Filtern

```
% Design a 4th-order butterworth band-pass filter which passes
% frequencies between 0.15 and 0.3.
[b,a]=butter(2,[.15,.3]);           % Bandpass digital filter design
h = fvtool(b,a);                   % Visualize filter
```



4.4 Entwurf von IIR-Filtern

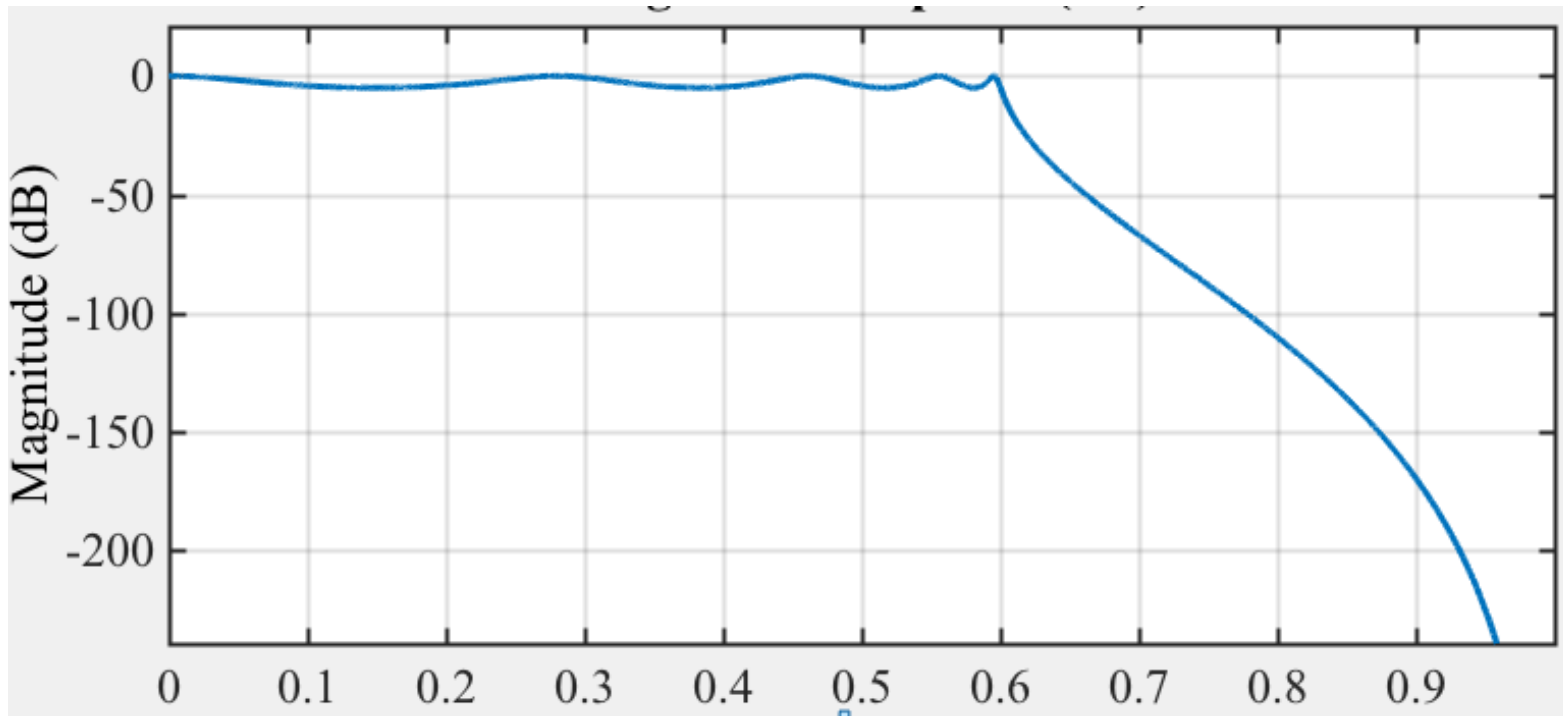
```
% For data sampled at 1000 Hz, design a 9th-order lowpass Chebyshev  
% Type I filter with 5 dB of ripple in the passband, and a passband  
% edge frequency of 300Hz.
```

```
Wn = 300/500; % Normalized passband edge frequency
```

```
[z,p,k] = cheby1(9,5,Wn);
```

```
[sos] = zp2sos(z,p,k); % Convert to SOS form
```

```
h = fvtool(sos) % Plot magnitude response
```

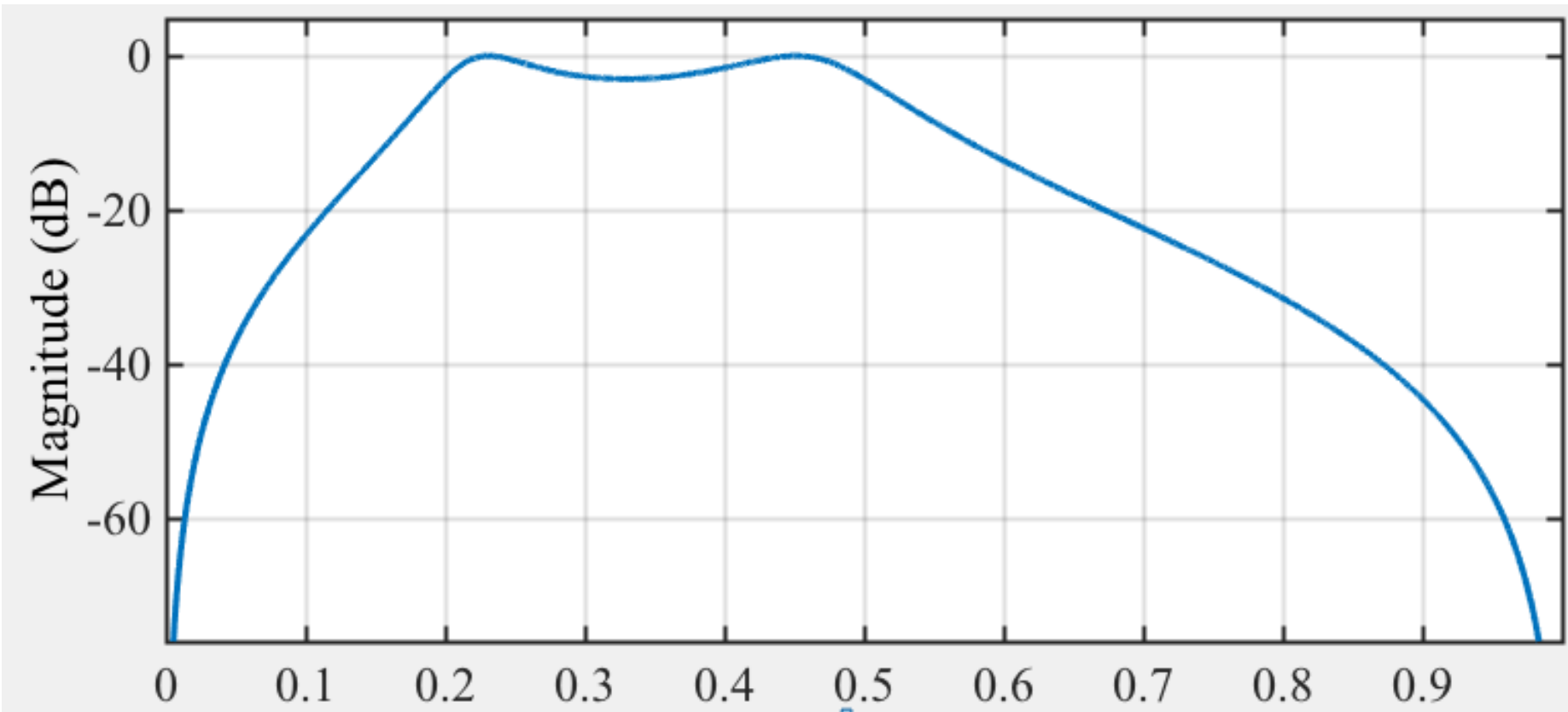


4.4 Entwurf von IIR-Filtern

```
% Design a 2nd-order Chebyshev Type I band-pass filter which passes  
% frequencies between 0.2 and 0.5 with 3 dB of ripple in the  
% passband.
```

```
[b,a]=cheby1(2,3,[.2,.5]);           % Bandpass digital filter design
```

```
h = fvtool(b,a);                     % Visualize filter
```



4.4 Entwurf von IIR-Filtern

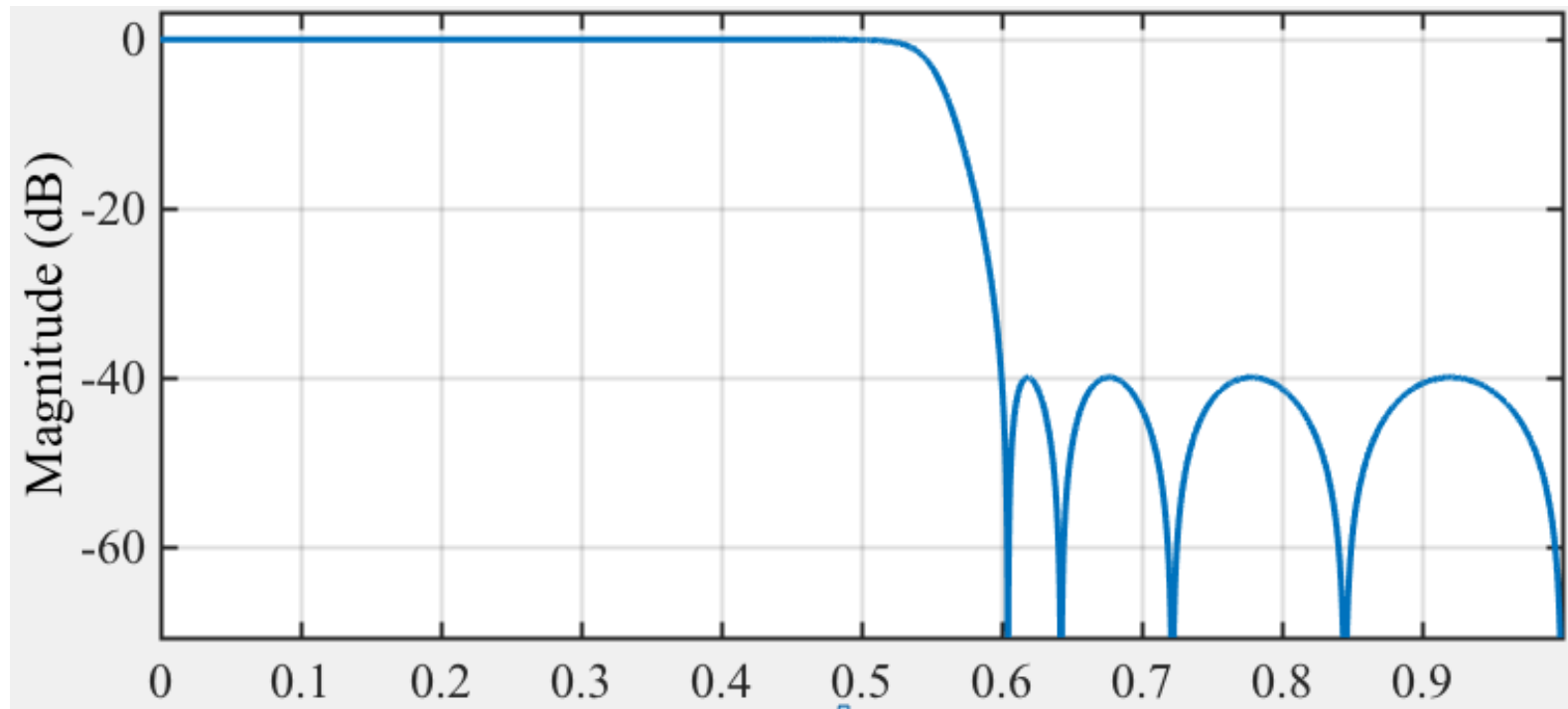
```
% For data sampled at 1000 Hz, design a ninth-order lowpass  
% Chebyshev Type II filter with stopband attenuation 40 dB down from  
% the passband and a stopband edge frequency of 300Hz.
```

```
Wn = 300/500; % Normalized stopband edge frequency
```

```
[z,p,k] = cheby2(9,40,Wn);
```

```
[sos] = zp2sos(z,p,k); % Convert to SOS form
```

```
h = fvtool(sos) % Plot magnitude response
```

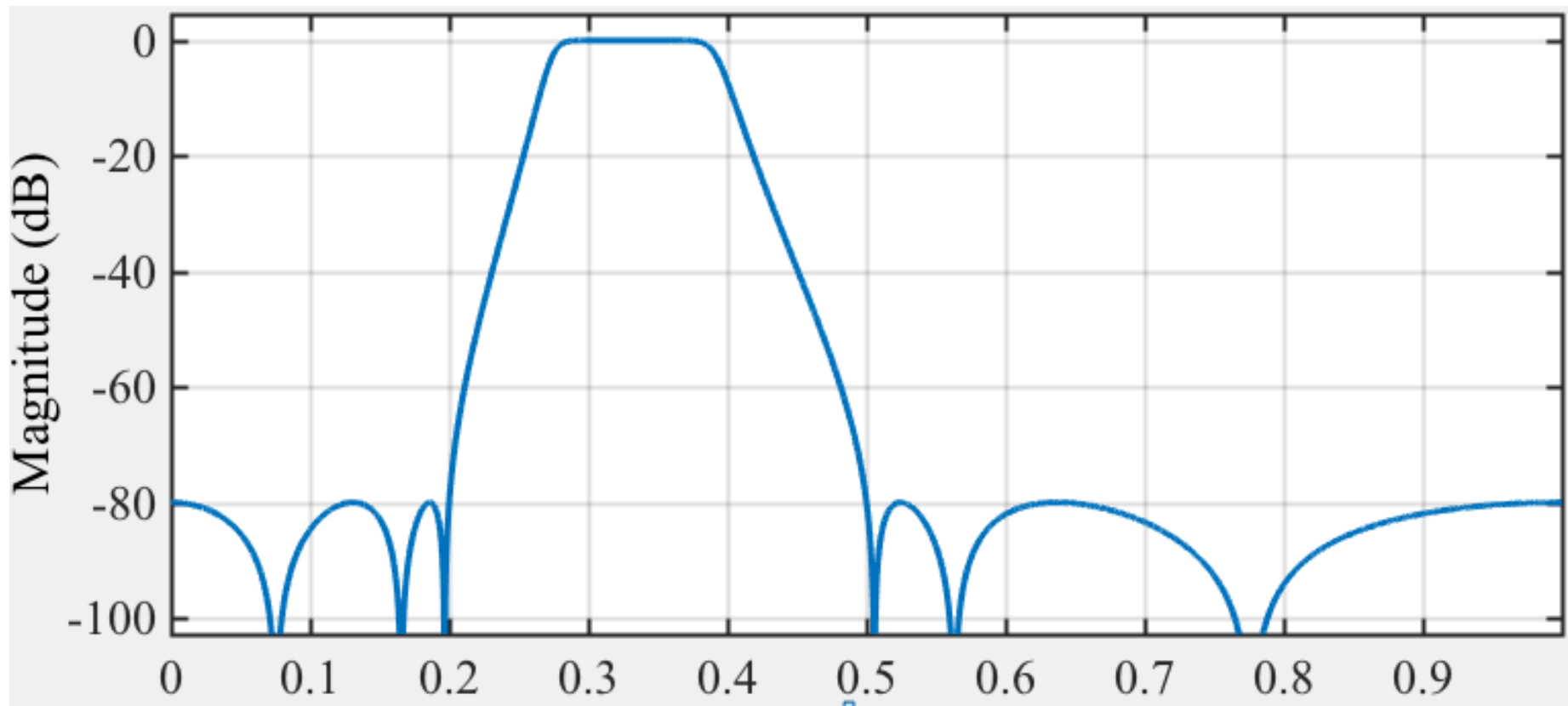


4.4 Entwurf von IIR-Filtern

```
% Design a 6th-order Chebyshev Type II band-pass filter which passes  
% frequencies between 0.2 and 0.5 and with stopband attenuation 80 dB  
% down from the passband.
```

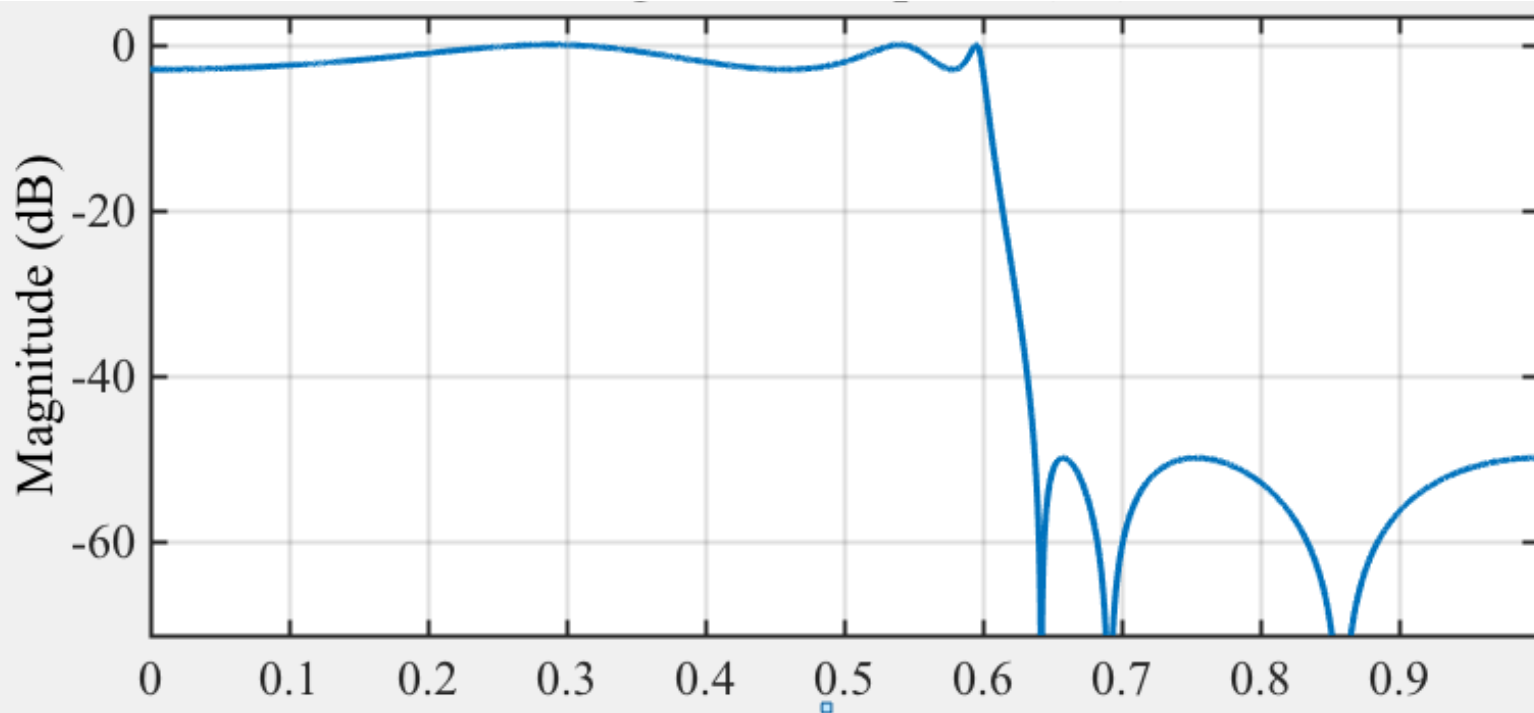
```
[b,a]=cheby2(6,80,[.2,.5]); % Bandpass digital filter design
```

```
h = fvtool(b,a); % Visualize filter
```



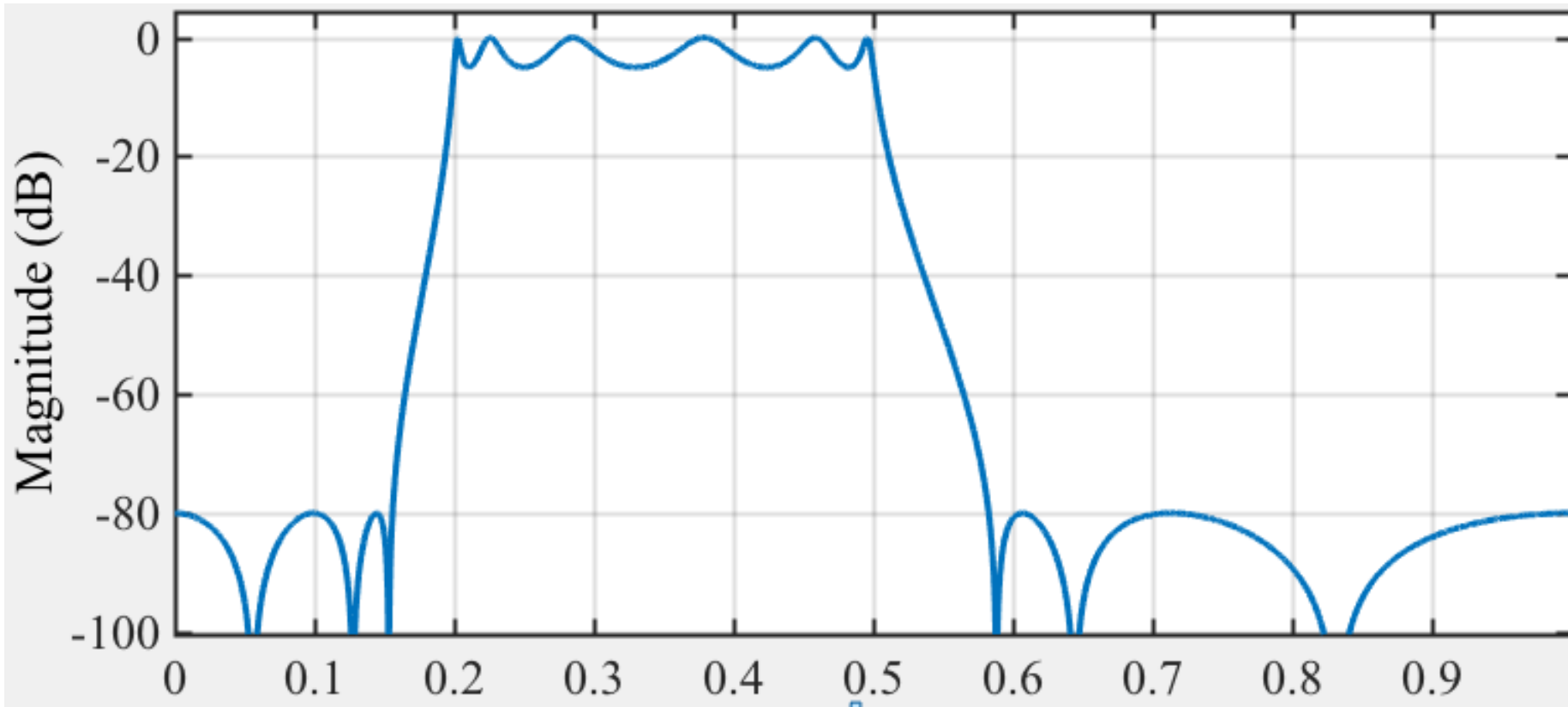
4.4 Entwurf von IIR-Filtern

```
% For data sampled at 1000 Hz, design a sixth-order lowpass
% elliptic filter with a passband edge frequency of 300Hz, 3 dB of
% ripple in the passband, and 50 dB of attenuation in the stopband.
Wn = 300/500;           % Normalized passband edge frequency
[z,p,k] = ellip(6,3,50,Wn);
[sos] = zp2sos(z,p,k);  % Convert to SOS form
h = fvtool(sos)         % Plot magnitude response
```



4.4 Entwurf von IIR-Filtern

```
% Design a 6th-order Elliptic band-pass filter which passes
% frequencies between 0.2 and 0.5, and with 5 dB of ripple in the
% passband, and 80 dB of attenuation in the stopband
[b,a]=ellip(6,5,80,[.2,.5]);    % Bandpass digital filter design
h = fvtool(b,a);              % Visualize filter
```



10.4 Design of IIR Filters

Normalization and Transformation

Up to here we have focused on **low-pass filters**. But with the help of simple transformations this knowledge can be carried over to all kind of filters.

Starting point is the design of a low-pass filter with normalized limit frequency $\omega_g = 1$ rad/s. All other filters can be easily derived from this:

Low-pass with limit frequency ω_g :
$$s \rightarrow \frac{s}{\omega_g}$$

High-pass with limit frequency ω_g :
$$s \rightarrow \frac{\omega_g}{s}$$

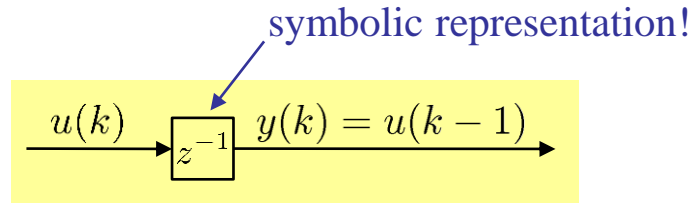
Band-pass with limit frequencies ω_{g1} and ω_{g2} :
$$s \rightarrow \frac{s^2 + \omega_{g1}\omega_{g2}}{s(\omega_{g2} - \omega_{g1})}$$

Band-stop with limit frequencies ω_{g1} and ω_{g2} :
$$s \rightarrow \frac{s(\omega_{g2} - \omega_{g1})}{s^2 + \omega_{g1}\omega_{g2}}$$

10.5 Implementation of Filters

Block Diagram of Digital Filters

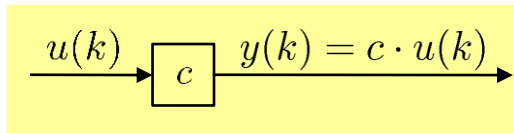
Delay of one sampling time step:



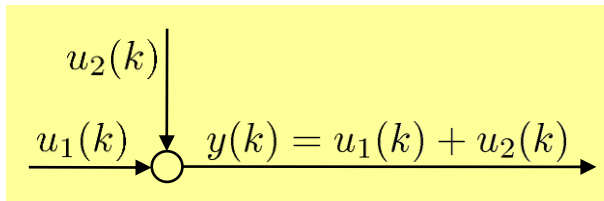
WARNING: Formally such a block diagram is wrong because it mixes time and frequency domain. However, such a sloppy representation is commonly found and easy to read. More strictly the following time delay is meant:

$$Y(z) = z^{-1}U(z) \rightarrow y(k) = u(k - 1)$$

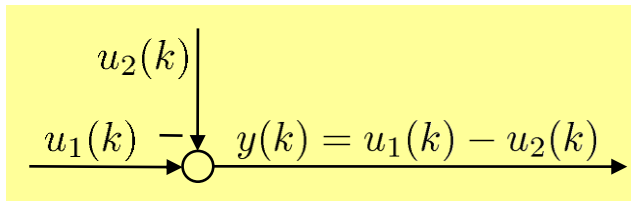
Multiplication with a factor:



Addition:



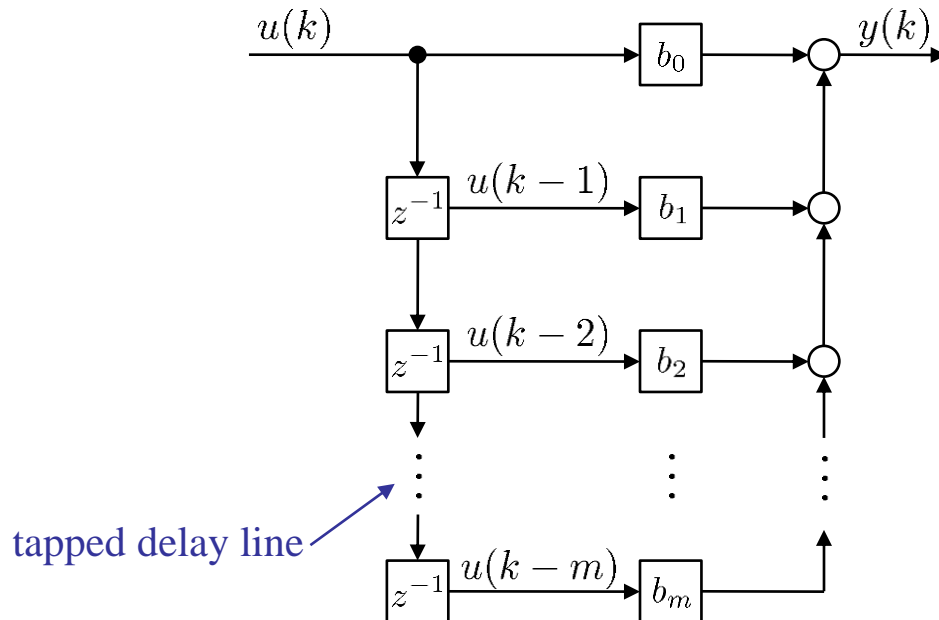
Subtraction:



10.5 Implementation of Filters

FIR Filter

- m memory elements
- $m+1$ multiplications and m additions
- No feedback
- For symmetrical filter with $b_0 = b_m, b_1 = b_{m-1}, \dots$ or $b_0 = -b_m, b_1 = -b_{m-1}, \dots$, half of the multiplications can be save by first adding $u(k)$ with $u(k-m)$ and $u(k-1)$ with $u(k-m-1)$, etc.



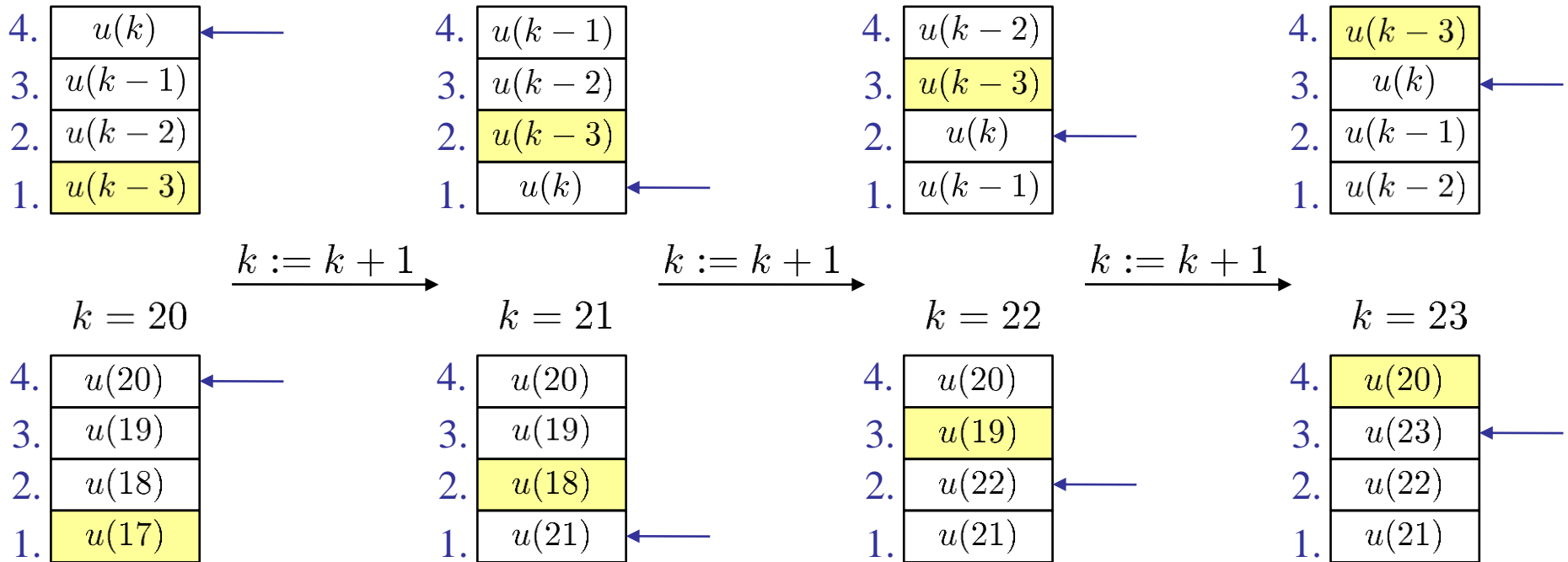
10.5 Implementation of Filters

Efficient Realization of a Tapped Delay Line in Software

Example for $m = 3$:

The pointer \leftarrow moves up one memory block in each time step. When it moves out at the top it jumps back to the bottom. This can be implemented with the modulo operator:

$adr := (adr + 1) \bmod m$. In each time step only *one* memory block has to be overwritten instead of moving all of them one step further!



10.5 Implementation of Filters

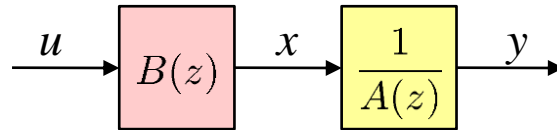
IIR Filter

An IIR filter of order n can be written as

$$G_{\text{IIR}}(z) = \frac{B(z)}{A(z)} = \frac{b_0 + b_1 z^{-1} + \dots + b_n z^{-n}}{1 + a_1 z^{-1} + \dots + a_n z^{-n}}$$

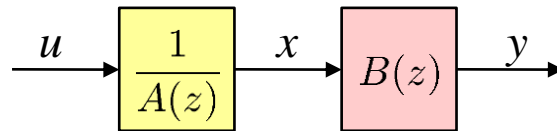
If the order of the numerator is smaller than the order of the denominator ($m < n$) then simply the lacking $b_i = 0$ for $i > m$. This transfer function can be split into two part in two ways:

Direct Form I:



$$G_{\text{IIR}}(z) = B(z) \cdot \frac{1}{A(z)} = (b_0 + b_1 z^{-1} + \dots + b_n z^{-n}) \cdot \frac{1}{1 + a_1 z^{-1} + \dots + a_n z^{-n}}$$

Direct Form II:

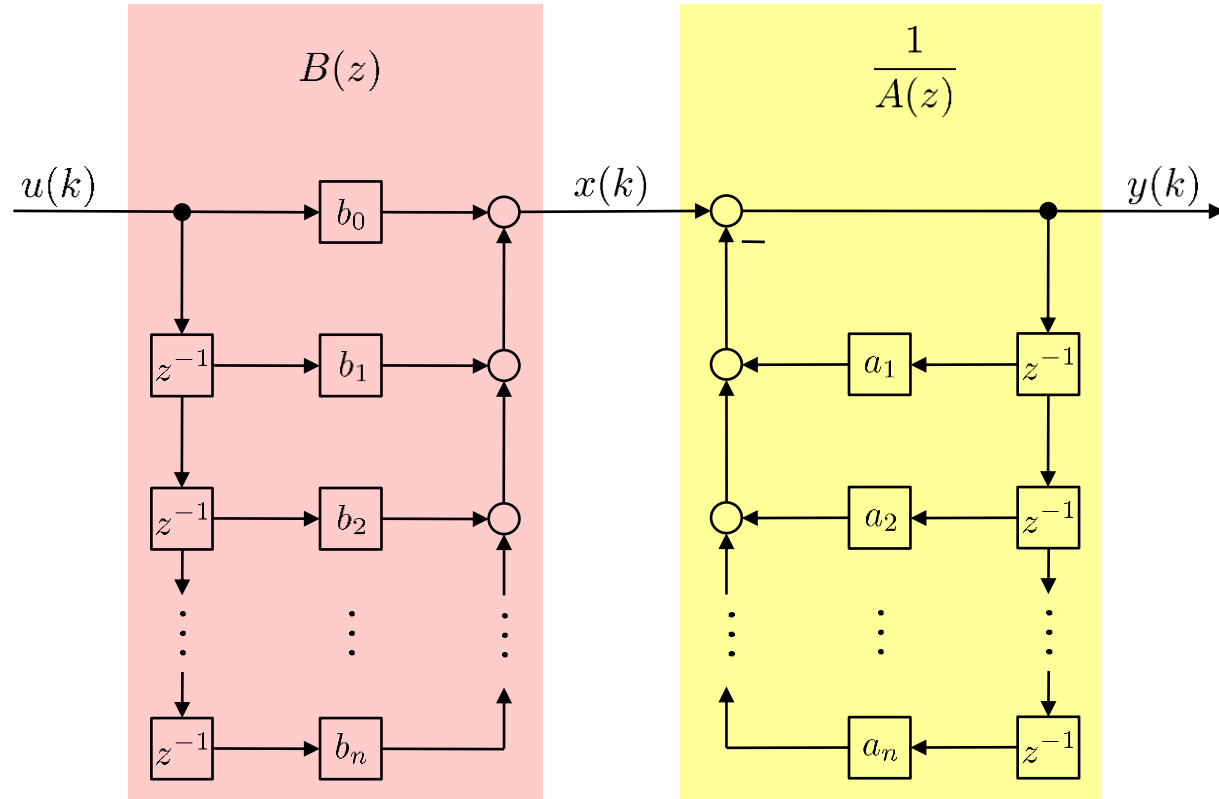


$$G_{\text{IIR}}(z) = \frac{1}{A(z)} \cdot B(z) = \frac{1}{1 + a_1 z^{-1} + \dots + a_n z^{-n}} \cdot (b_0 + b_1 z^{-1} + \dots + b_n z^{-n})$$

10.5 Implementation of Filters

IIR Filter in Direct Form I

- $2n$ memory blocks
- $2n+1$ multiplications and $2n$ additions
- n feedback paths



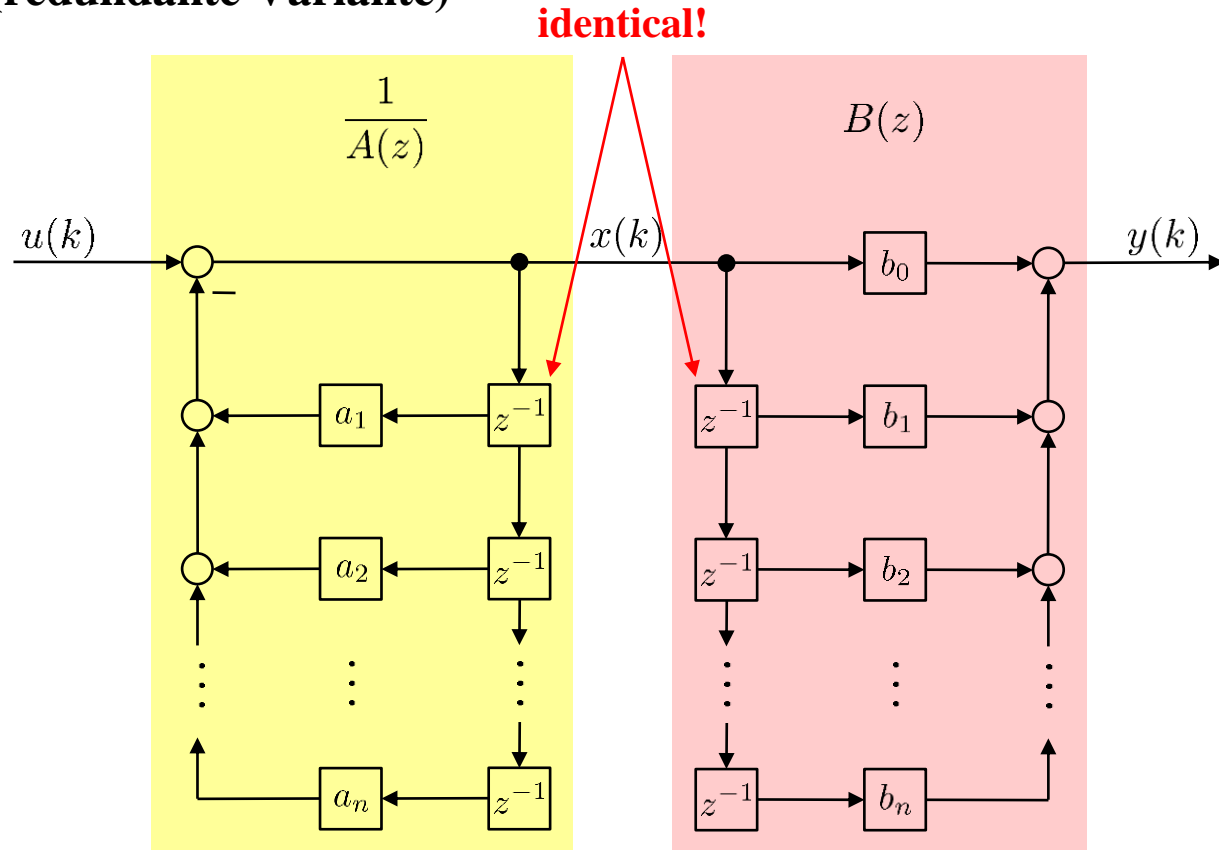
$$x(k) = b_0u(k) + b_1u(k - 1) + b_2u(k - 2) + \dots + b_nu(k - n)$$

$$y(k) = x(k) - a_1y(k - 1) - a_2y(k - 2) - \dots - a_ny(k - n)$$

10.5 Implementation of Filters

IIR Filter in Direct Form II (redundante Variante)

- $2n$ memory blocks
- $2n+1$ multiplications and $2n$ additions
- n feedback paths



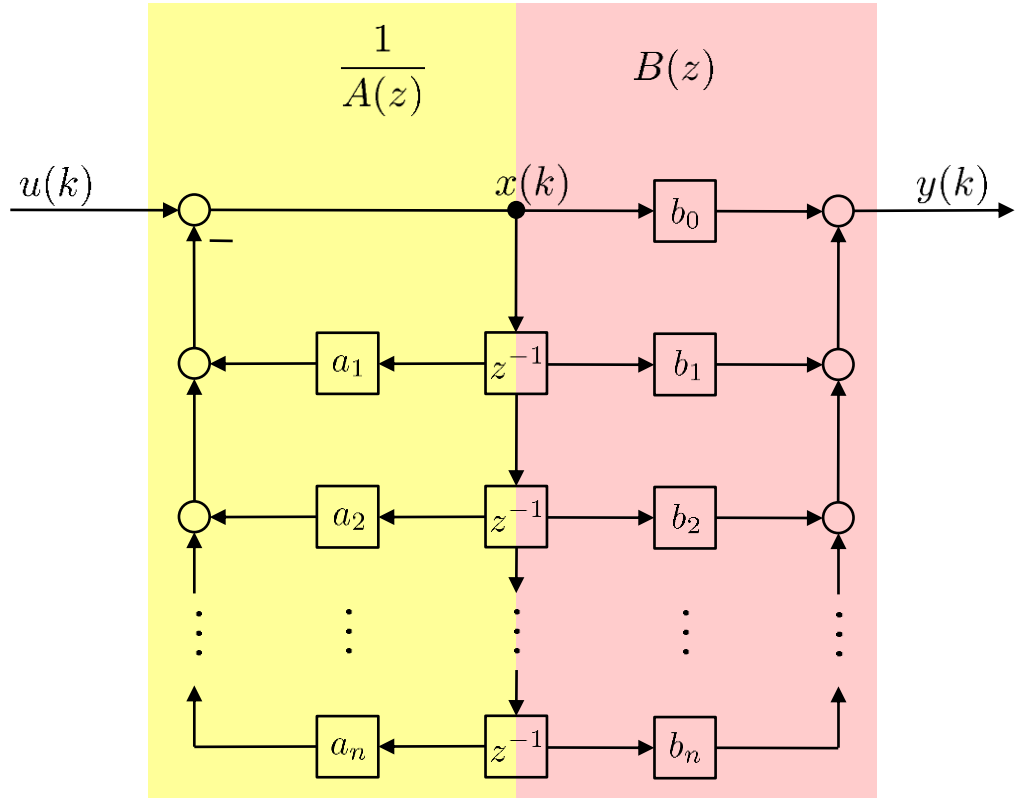
$$x(k) = u(k) - a_1x(k-1) - a_2x(k-2) - \dots - a_nx(k-n)$$

$$y(k) = b_0x(k) + b_1x(k-1) + b_2x(k-2) + \dots + b_nx(k-n)$$

10.5 Implementation of Filters

IIR Filter in Direct Form II (Non-Redundant Variant)

- n memory blocks
- n memory blocks correspond to the n states of the filter (see state space in control)
- $2n+1$ multiplications and $2n$ additions
- n feedback paths



$$x(k) = u(k) - a_1x(k-1) - a_2x(k-2) - \dots - a_nx(k-n)$$

$$y(k) = b_0x(k) + b_1x(k-1) + b_2x(k-2) + \dots + b_nx(k-n)$$

10.5 Implementation of Filters

Cascade Form

Consists of a series circuit of IIR filters of 2. order in direct form II:

$$G(z) = \prod_{i=1}^l \frac{b_{i0} + b_{i1}z^{-1} + b_{i2}z^{-2}}{1 + a_{i1}z^{-1} + a_{i2}z^{-2}}$$

In this product each factor represents a second order system with two conjugate complex or two real poles. For an even order n of the complete filter $l = n/2$.

For an odd n we have $l = (n+1)/2$ and $b_{l2} = a_{l2} = 0$.

Parallel Form

Consists of a parallel circuit of filters derived from a partial fraction expansion:

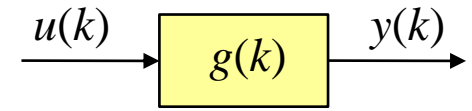
$$G(z) = \sum_{i=1}^{l_1} c_i z^{-i} + \sum_{i=1}^{l_2} \frac{d_i}{1 + a_i z^{-1}} + \sum_{i=1}^{l_3} \frac{g_i(1 + e_i z^{-1})}{(1 + f_i z^{-1})(1 + f_i^* z^{-1})}$$

This means that filters with poles at 0, with real poles at $-a_i$ and with conjugate complex pole pairs at $-f_i$ and $-f_i^*$ are run in parallel..

Ladder Form and Lattice Form

Representations in form of continued fractions or lattice structures are sophisticated filter forms that possess advantages with respect to robustness against round-off errors.

10.6 Non-Causal Filters

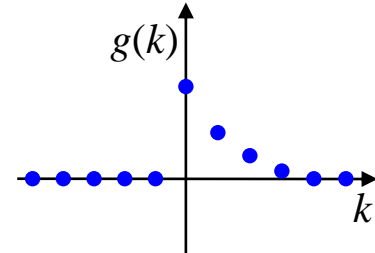


Causal Filters

For a *causal* filter its output $y(k)$ depends only on the *current* and *previous* input $u(k-i)$ with $i \geq 0$. This automatically means that the impulse response is equal to zero for negative times:

$$y(k) = \sum_{i=-\infty}^{\infty} g(i)u(k-i) = \sum_{i=0}^{\infty} g(i)u(k-i)$$

since $g(i) = 0$ for $i < 0$,
otherwise the future inputs would influence the now: $u(k-i)$

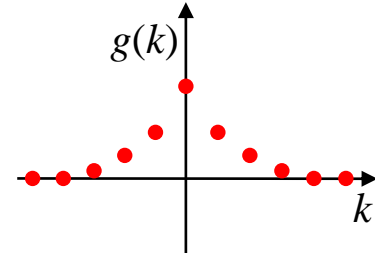


Non-Causal Filters

For a *non-causal* filter its output $y(k)$ also depend on the *future* input $u(k-i)$ with $i < 0$. This automatically means that the impulse response is *not* equal to zero for negative times:

$$y(k) = \sum_{i=-\infty}^{\infty} g(i)u(k-i) \neq \sum_{i=0}^{\infty} g(i)u(k-i)$$

since $g(i) \neq 0$ for $i < 0$, because future inputs are relevant: $u(k-i)$

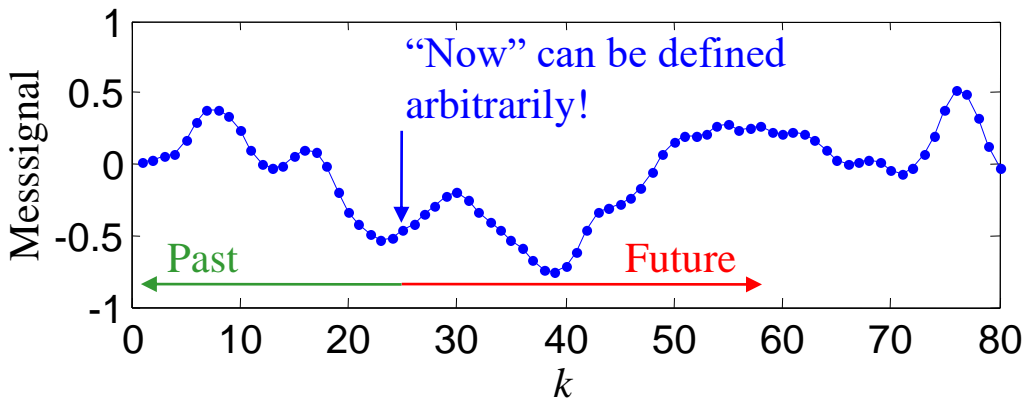


commonly symmetrical, but this is not necessary

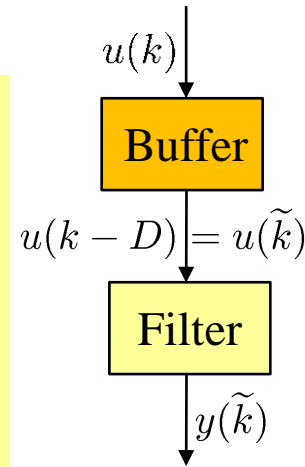
10.6 Non-Causal Filters

How the *Future* is Known to Calculate a Non-Causal Filter?

- *Offline data processing:* The data set is available from start to end in the computer. Then the “now” can be arbitrarily chosen by the user.
- *Buffers in online data processing:* Data is stored in a buffer for a couple of sampling time steps, say D steps, before being processed further. The whole processing is therefore delayed by D steps. Relative to this delayed “now” there exist the possibility to look D steps into the future up to $g(-D)$. It is important to note that in order to look D steps into the future with a non-causal filter, we have to buffer D steps of the signal, thus introducing a dead time of D steps.



Signal processing is based on buffered signals that are D steps back with respect to real time $\tilde{k} = k - D$ and thus as many steps can be predicted :

$$\begin{aligned} \tilde{k} + 1 &= k - D + 1 \\ &\vdots \\ \tilde{k} + D &= k \end{aligned}$$


10.6 Non-Causal Filters

Advantages of Non-Causal Filters

- An impulse or step response that is symmetric to $k = 0$ has a real frequency response, i.e., no phase delay (see green dashed filter response)!

Symmetry implies: $G(z) = \dots + b_2z^2 + b_1z^1 + b_0 + b_1z^{-1} + b_2z^{-2} \dots$

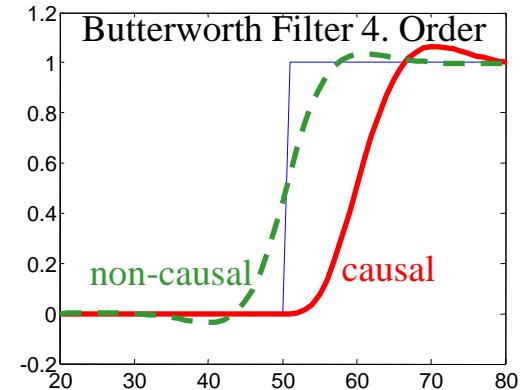
Rearrange:

$$G(z) = b_0 + b_1 \underbrace{(z^1 + z^{-1})}_{\substack{\text{conj. compl.} \\ \text{pair} \rightarrow \text{real}}} + b_2 \underbrace{(z^2 + z^{-2})}_{\substack{\text{conj. compl.} \\ \text{pair} \rightarrow \text{real}}} + \dots \quad \leftarrow \text{real for } z = e^{i\omega T_0}$$

- By forward and backward filtering of the data (which is possible only offline) every phase delay introduced by the forward filtering is exactly compensated again by the backward filtering. This fact is independent on the nature of the filter and thus is true for every type (FIR, IIR, nonlinear).

However, it is filtered twice. This means we effectively have the amplitude response of $|G(i\omega)|^2$.

- Because a non-causal filter can “react” to a step input before it actually happens, such a filter is much faster!



10.6 Non-Causal Filters

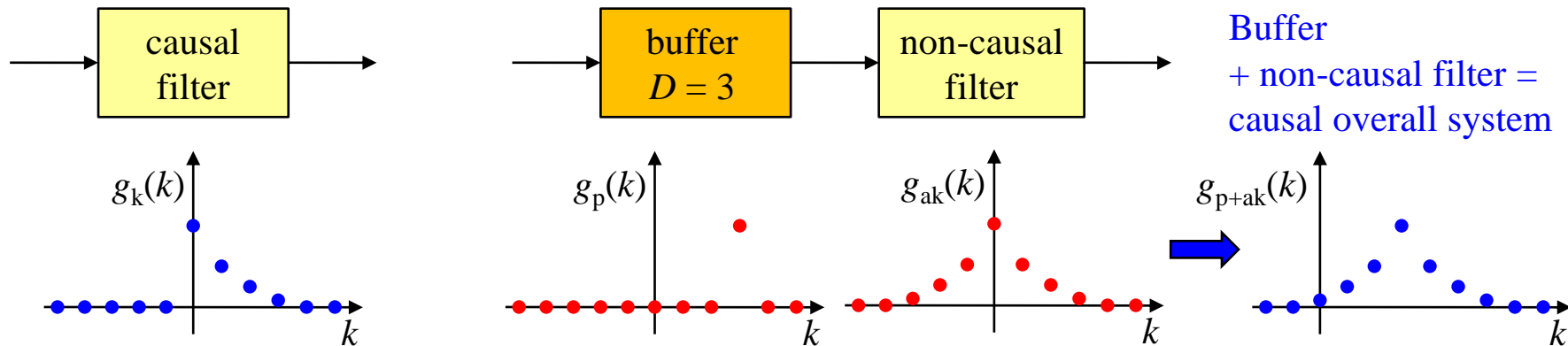
Drawbacks of Non-Causal Filters

- Can hardly be applied for applications with strict real-time requirements such as feedback control because any delays deteriorate the performance significantly.

In communication systems, however, delays introduced by buffers usually are

- irrelevant/unimportant since communication is unidirectional (radio, TV),
- negligible when communication is bidirectional (telephone) because signal run times introduce the major part of the delay anyway.

In feedback control a buffer would introduce an additional dead time. This has severe consequences for the control quality (reduced phase margin, danger of instability). These drawbacks are typically more important than the achievable improvements in signal quality.

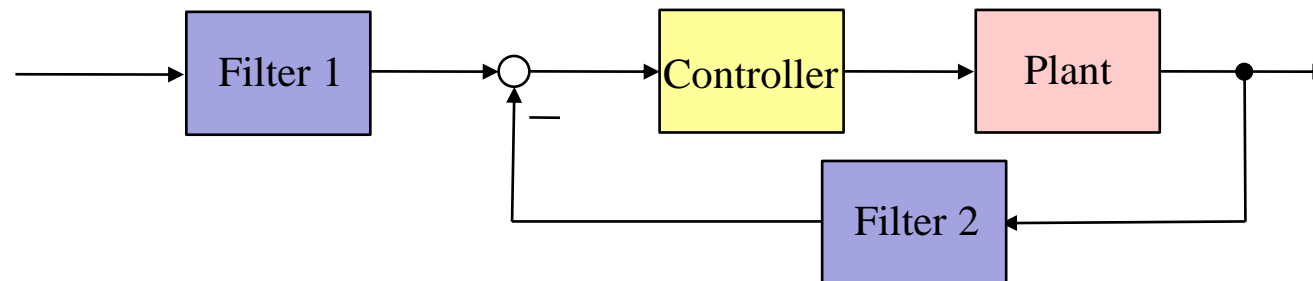


10.6 Non-Causal Filters

Non-Causal Filters in Feedback Control

Feedback control gives nice examples for non-causal filters:

1. Reference input filter: Commonly the future course of the reference value is known a priori. The non-causal filters can easily be exploited to utilize this knowledge.
2. Feedback filter: The comparison between desired and control value requires the control value as fast as possible. A non-causal filter with buffer would introduce a dead time which deteriorates the control performance because it causes phase lag. There non-causal filter would be counterproductive. A “truly” non-causal filter cannot be employed because the future control variable is unknown.



10.7 Nonlinear Filters

Nonlinear filters are seldom applied due to the additional complexity in their handling and design. In the field of image processing they are however, more common. Most frequently simple nonlinear operators like max-, min- or other order/sorting-operators can be found.

Median Filter

Probably the most important and frequently used nonlinear filter is the median filter. It is helpful in eliminating **outliers**. In contrast to the arithmetic average, the median gives the number which is right in the middle of a sorted sequence, i.e., half of the numbers are larger, half of the numbers are smaller.

Example:

Sequence: 4, 7, 20, 21, 30 \rightarrow median = 20, arithmetic average = 16.4

Sequence: 4, 7, 20, 21, 1000 \rightarrow median = 20, arithmetic average = 210.4

The median is commonly used to eliminate outliers e.g. in statistics where the arithmetic average does not represent the “typical” case like study program duration, house prices, etc.

10.7 Nonlinear Filters

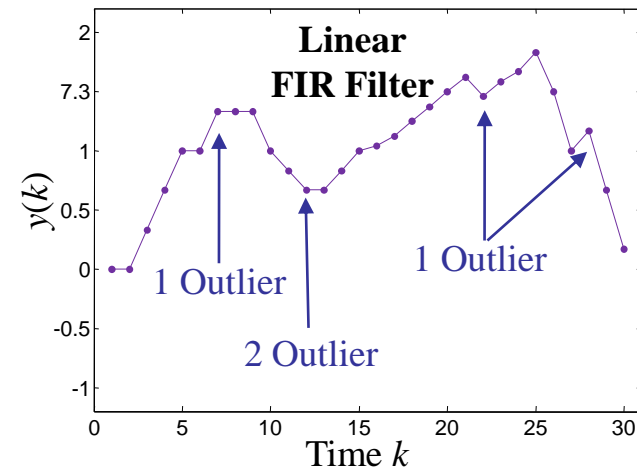
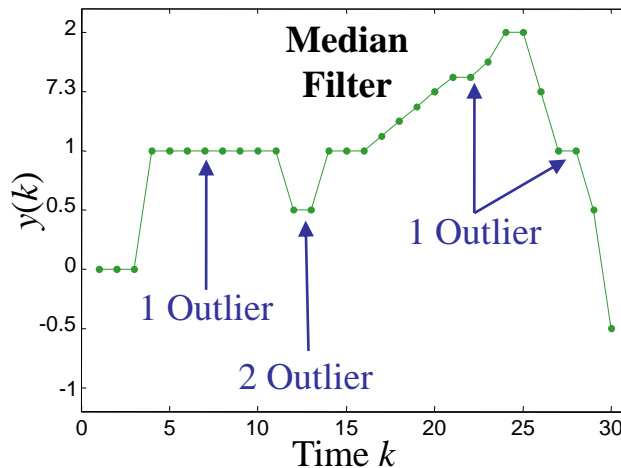
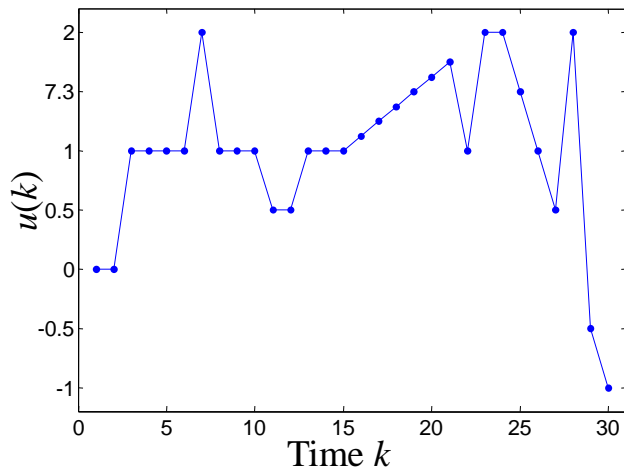
Median Filter for Elimination of Outliers

A median filter of n . order has an output $y(k)$ that is calculated as the median of the last n data samples $u(k), u(k-1), \dots, u(k-n+1)$. With a median filter of n . order from n subsequent data samples $(n-1)/2$ outliers in series can be filtered out and removed without distorting the signal very much.

Example: Median filter of 3. order versus linear average FIR filter

Median filter: $y(k) = \text{median} \{u(k), u(k-1), u(k-2)\}$

Linear FIR filter: $y(k) = \frac{1}{3}u(k) + \frac{1}{3}u(k-1) + \frac{1}{3}u(k-2)$



10.8 Outlook: Adaptive Filters

What is an Adaptive Filter?

An adaptive filter has no fixed parameters but they change over time in order to meet changing requirements. The time-varying parameters are typically changed according to some adaptation law in order to improve the performance of the filter. Typical applications are:

- *Online system identification:* A time-variant process shall be identified (modeled by measurement data). Because the process behavior changes over time the filter has to track these changes.
- *Channel equalization:* A signal is distorted from sender to receiver by the dynamic channel in between (obstacles, reflections, ...). This distortion must be compensated (canceled) at the receiver to improve the quality. E.g. built in cell phones!
- *Echo compensation:* To avoid (or weaken) acoustic feedback distortions, adaptive filters are applied to eliminate the part from the sound signal back from the speaker to the microphone.
- *Active noise suppression:* An adaptive filter can model a measured disturbance in order to actively compensate it by adding it to the signal with 180° phase shift (destructive interference).

10.8 Outlook: Adaptive Filters

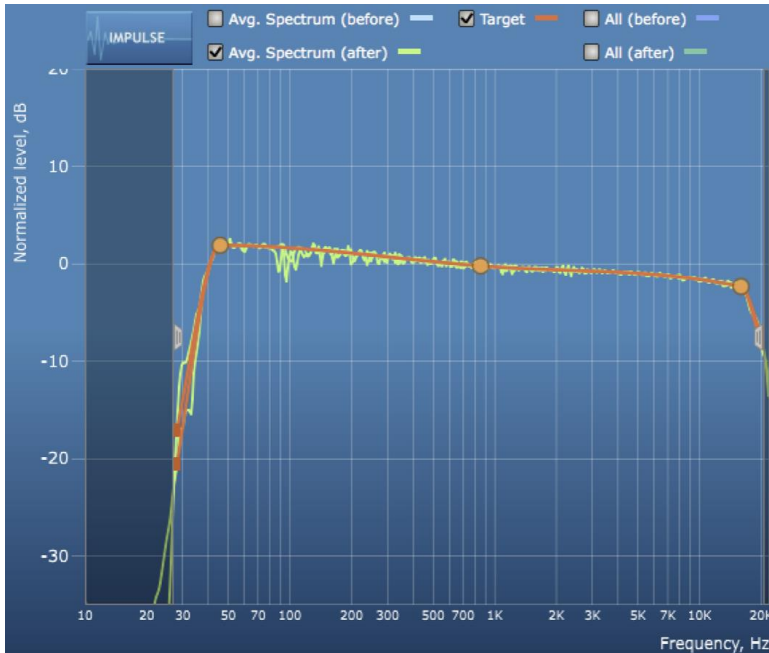
Automatic Room Acoustic Correction

Room amplitude response for left and right channel:

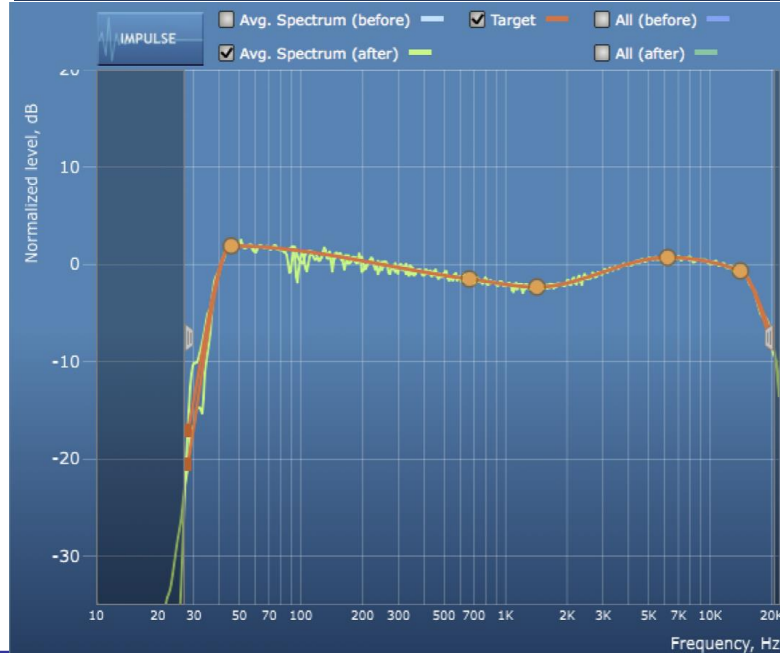
- white: neural amplifier (before correction)
- orange: desired characteristics
- green: optimized amplifier (after correction)

[mactechnews]

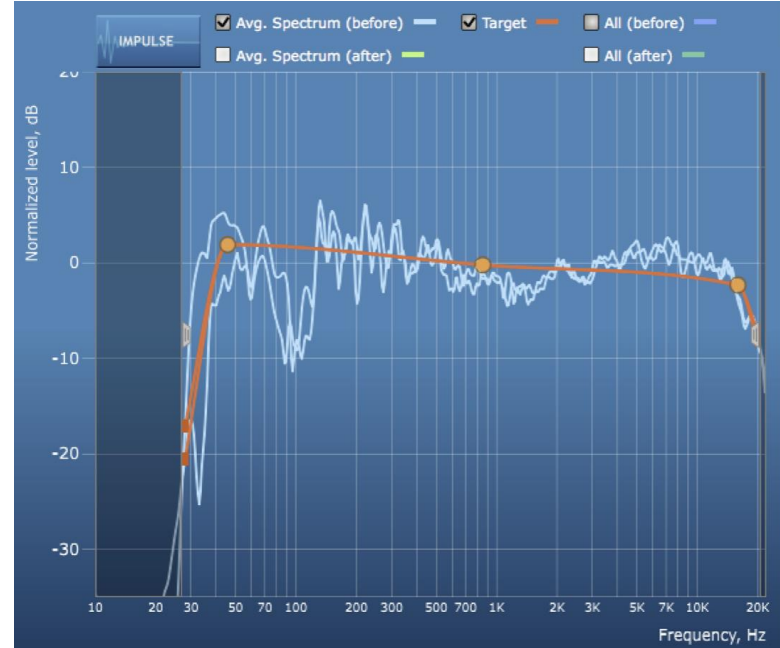
Amplitude Response: Optimal
all frequencies equally considered



Amplitude Response: Optimal
low and high frequencies emphasized



Amplitude Response: Original



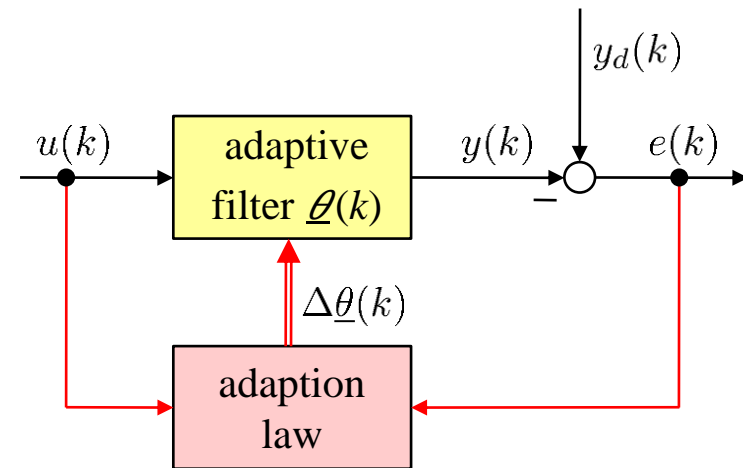
10.8 Outlook: Adaptive Filters

Principles of an Adaptive Filter

- Comparison between desired filter output $y_d(k)$ and actual filter output $y(k)$.
- Calculation of the error $e(k)$.
- In the adaptation law the change of the filter parameters is computed from the error. This usually is done by an **update** of the filter parameters according to:

$$\underline{\theta}(k + 1) = \underline{\theta}(k) + \Delta\underline{\theta}(k)$$

- Different adaptation laws distinguish each other by different calculations of this parameter update $\Delta\underline{\theta}(k)$. The following goal are pursued and for each application a suited compromise must be sought:
 - convergence speed
 - tracking speed
 - computational demand in each update step
 - numerical robustness (round-off errors!)
- Typical adaptation laws are:
 - least mean squares (LMS): gradient method
 - recursive least squares (RLS): Newton's method

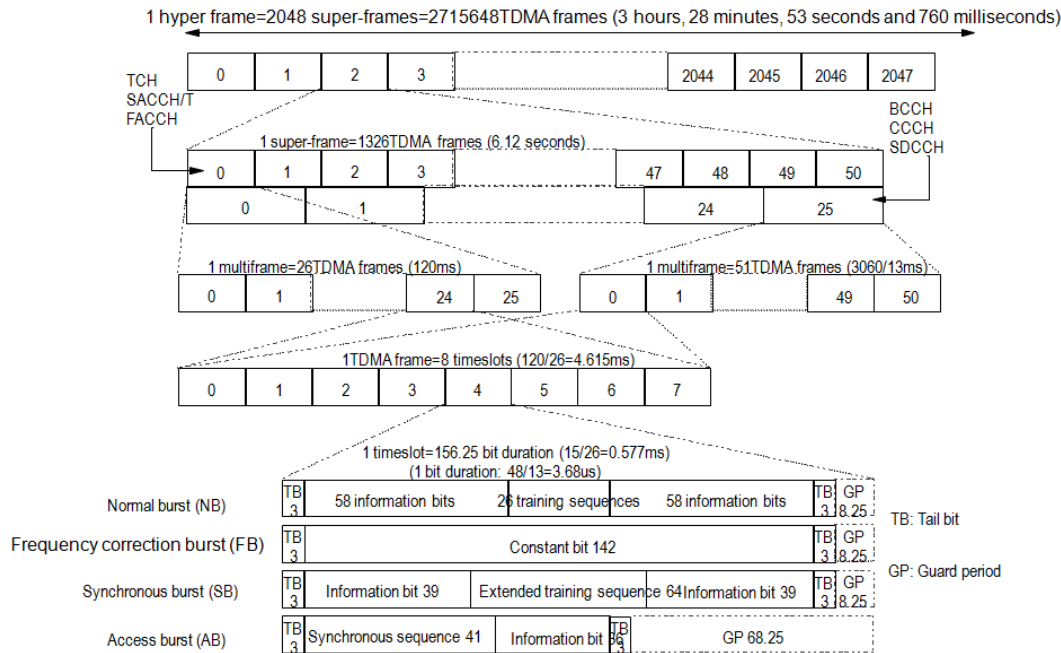


10.8 Outlook: Adaptive Filters

GSM: Mobile Communication

- Data send in packages of 148 bits each.
- Hereof 26 bits represent a reference signal for training of the adaptive filter in cell phone.
- This leads to an overhead of approx. 17%.
- One data package is send and received every 0.577 ms.

Frame



10.8 Outlook: Adaptive Filters

Online Adaptation

The gradient method tries to minimize the quadratic error $e^2(k)$ by changing the parameter vector in direction opposite to the steepest ascent (gradient) by a step proportional to the step size or length η :

$$\Delta \underline{\theta}(k) = -\eta \frac{de^2(k)}{d\underline{\theta}(k)}$$

Commonly adaptive filters are of FIR type, i.e.:

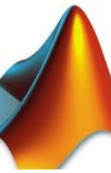
$$y(k) = \theta_1 u(k-1) + \theta_2 u(k-2) + \dots + \theta_m u(k-m)$$

$$\text{with } \Delta \underline{\theta} = \begin{pmatrix} \Delta \theta_1 \\ \Delta \theta_2 \\ \vdots \\ \Delta \theta_m \end{pmatrix}$$

Thus the parameter update becomes (Remember: $e(k) = y_d(k) - y(k)$):

$$\Delta \underline{\theta}(k) = -2\eta e(k) \frac{de(k)}{d\underline{\theta}(k)} = 2\eta e(k) \frac{dy(k)}{d\underline{\theta}(k)} = \eta' e(k) \begin{pmatrix} u(k-1) \\ u(k-2) \\ \vdots \\ u(k-m) \end{pmatrix}$$

This means the update is proportional to the (new) step size η' , to the error $e(k)$ and to the “excitation” (regressor) of the corresponding parameter θ_i by $u(k-i)$.



Discrete-time transfer function:

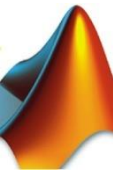
```
sys = filt(num,den);2 % Assigning a discrete-time transfer function
```

FIR filter:

```
fir1;1 % FIR filter using the window method  
firls;1 % FIR-Filter using least squares optimization  
firpm;1 % FIR-Filter using Parks-McClellan optimization
```

IIR filter:

```
besself;1 % Bessel filter  
butter;1 % Butterworth filter  
cheby1;1 % Chebyshev filter type 1  
cheby2;1 % Chebyshev filter type 2  
ellip;1 % Cauer filter (elliptic filter)
```

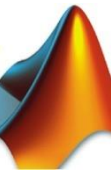


```
y = filter(b,a,X); % Digital IIR filter (direct form II)
y = filtfilt(b,a,X);1 % Corresponding non-causal filter
% with forward and backward path
% WARNING: The amplitude response has
% the squared (twice) effect

[b,a] = yulewalk(n,f,m);1 % Digital, recursive IIR filter.
% Uses least squares to model the
% frequency response

H = dfilt.structure(in1,...);1 % Yields discrete-time filter according
% to the method 'structure', see
% MATLAB help

[b,a] = prony(h,n,m);1 % Filter design in the time-domain
% according to the "Prony" method
```



Filter-Parameter-Identifikation:

```
[b,a] = invfreqz(h,w,n,m);1 % Identifies a discrete-time amplitude  
% and phase response (continuous-time:  
% "invfreqs")
```

¹ : *Signal Processing Toolbox*

² : *Control System Toolbox*

11. Selected Methods in Signal Processing

Contents Chapter 11

6. Selected Methods in Signal Processing

6.1 Principal Component Analysis (PCA)

6.2 Clustering

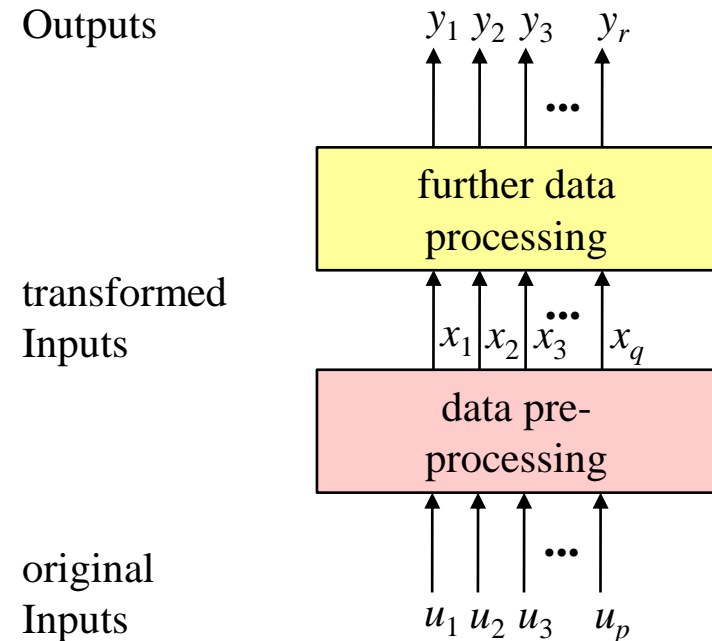
11.1 Principal Component Analysis

Data Preprocessing

Complex tasks in signal processing often are partitioned into two or more steps that each can be handled simpler individually. Typically, a early (first) steps is called signal *preprocessing*. Dependent on the specific task, signal preprocessing can be:

- Filtering, smoothing, interpolation
- Transformation of data into a new coordinate system
- Dimension reduction, data compression
- Transformation into the frequency domain
- Feature extraction
- Nonlinearity transform

Some of the most common an important data preprocessing approaches will be discussed in the following.



11.1 Principal Component Analysis

Supervised versus Unsupervised Learning

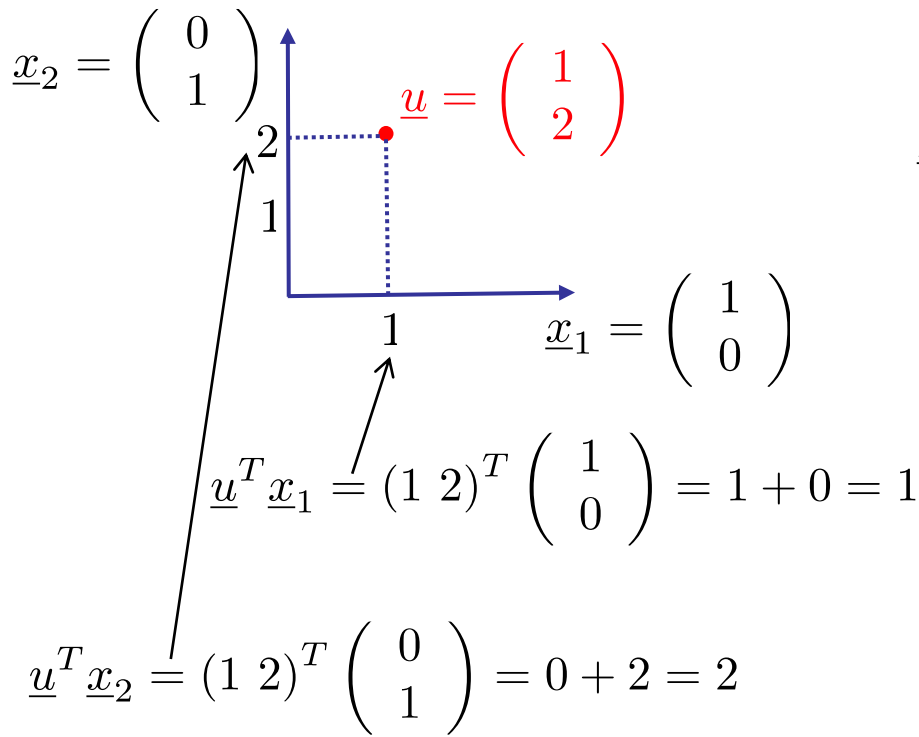
Two approaches to learning can be distinguished:

- *Supervised learning*: The desired output y is known and is compared with the result of the used method \hat{y} . A loss function to measure the quality of the method that depends on y and \hat{y} is calculated and often optimized. Frequently the mean squared error (MSE) is used for that purpose.
- *Unsupervised Learning*: The desired output y is unknown or at least not used. Rather in interim goal is defined which can be calculated solely on the input data $\{u_i(k)\}$, $i = 1, 2, \dots, p$ and $k = 1, 2, \dots, N$. Frequently the distribution of data in the input space plays an important role.

Unsupervised learning is much simpler to realize than supervised learning. The interim goal is easier to achieve than the final one. However, the risk exists that the interim goal is not as helpful as assumed. Therefore the success of unsupervised learning is not always guaranteed. The methods presented here are unsupervised and require little computational effort.

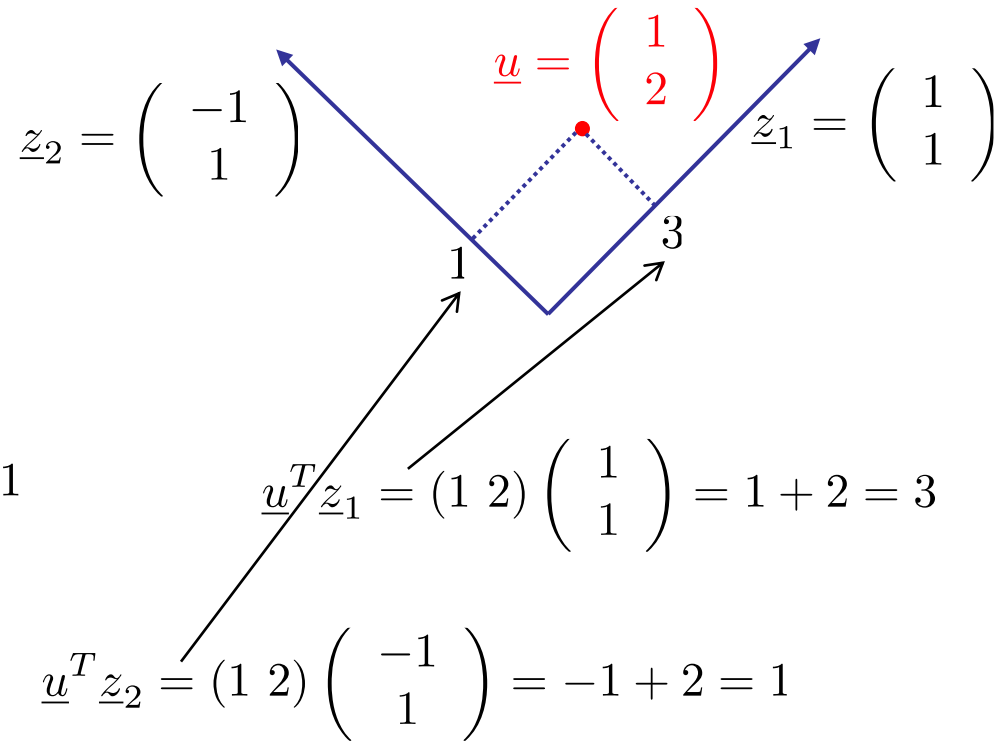
11.1 Principal Component Analysis

Projection of Vectors



$$|\underline{u}| = \sqrt{1^2 + 2^2} = \sqrt{5}$$

In order to keep the absolute value of \underline{u} constant, the vectors describing the coordinate axes have to be normalized to one, i.e.:



$$|\underline{u}_{\text{tansf}}| = \sqrt{3^2 + 1^2} = \sqrt{10}$$

$$\underline{z}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad \underline{z}_2 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

11.1 Principal Component Analysis

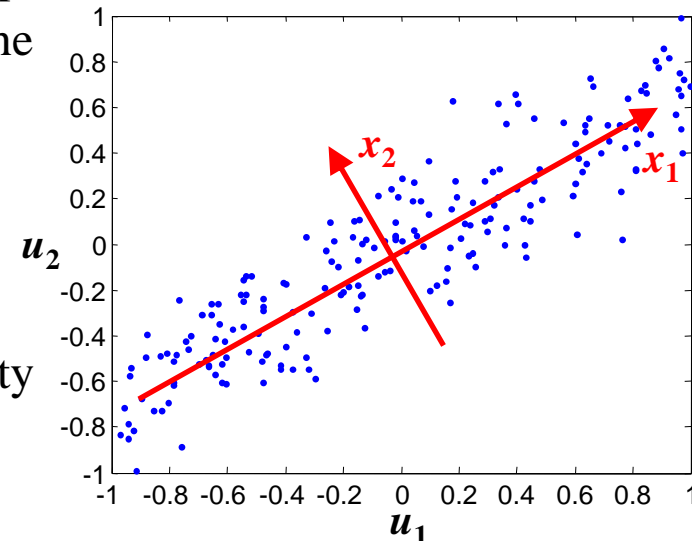
Transformation of the Coordinate System

With a principal component analysis (PCA) data is transformed from one coordinate system into a new one. The 1. new axis shall point in the direction of the highest variance of the data. The 2. new axis shall be orthogonal to the first and again in the direction of the highest data variance remaining, and so on. The idea behind this procedure is that data can often be described best in directions of high variance and often can be neglected in directions of low variance. The low variance directions typically represent just noise.

The example on the left illustrates this idea. The data distribution shows a strong correlation between u_1 and u_2 . It can be assumed that u_1 and u_2 may depend on each other,

e.g. $u_2 = au_1 + n$ with $a \approx 0.7$ and noise n . A PCA orients the 1. axis in direction of the highest variance, i.e., $x_1 = u_1 + au_2$ and the 2. axis orthogonally, i.e., $x_2 = u_2 - au_1$.

If the assumed relationship between u_1 and u_2 is indeed true then $x_2 = n$ and x_2 describes only noise and thus contains no information and can be removed (dimensionality reduction).



11.1 Principal Component Analysis

WARNING: The data needs to have zero mean!

Derivation of Principal Component Analysis (PCA)

Start with a p -dimensional space. The task of a PCA is to find new axes $\underline{x}_i = [x_{i1} \ x_{i2} \ \dots \ x_{ip}]^T$ for $i = 1, 2, \dots, p$, while the 1. axis point in the direction of the highest data variance, the 2. axis in the direction of the second highest, and so on. All axes shall be orthogonal to each other.

In the $N \times p$ data matrix \underline{U} all data is stored with respect to the original coordinate system:

$$\underline{U} = \begin{pmatrix} \underline{u}^T(1) \\ \underline{u}^T(2) \\ \vdots \\ \underline{u}^T(N) \end{pmatrix} = \begin{pmatrix} u_1(1) & u_2(1) & \cdots & u_p(1) \\ u_1(2) & u_2(2) & \cdots & u_p(2) \\ \vdots & \vdots & \ddots & \vdots \\ u_1(N) & u_2(N) & \cdots & u_p(N) \end{pmatrix}$$

← 2. data point

↑ 2. old axis

↑ N data points

↑ p dimensions

The scalar products $\underline{u}^T(k) \underline{x}$ are the projections of the $k = 1, 2, \dots, N$ data points onto an arbitrary axis $\underline{x} = \{x_1, x_2, \dots, x_p\}$. If the data has zero mean (if not then the mean has to be subtracted first) then the following expression corresponds to the squared distance to the mean (which is equal to 0): $(\underline{u}^T(k) \underline{x})^2$.

11.1 Principal Component Analysis

If we calculate this variance for each data point and sum them, we get the variance of the whole data along the new axis \underline{x} :

$$\begin{aligned}(\underline{U} \underline{x})^T (\underline{U} \underline{x}) &= (\underline{u}^T(1)\underline{x} \quad \underline{u}^T(2)\underline{x} \quad \cdots \quad \underline{u}^T(N)\underline{x})^T \begin{pmatrix} \underline{u}^T(1)\underline{x} \\ \underline{u}^T(2)\underline{x} \\ \vdots \\ \underline{u}^T(N)\underline{x} \end{pmatrix} \\ &= (\underline{u}^T(1)\underline{x})^2 + (\underline{u}^T(2)\underline{x})^2 + \dots + (\underline{u}^T(N)\underline{x})^2\end{aligned}$$

We want to maximize this expression. However, we must prevent that the variance becomes large just by shrinking the axis (and thereby generate large numbers). Thus the axes' scaling are restricted to a norm of 1:

$$\underline{x}^T \underline{x} = 1$$

This constraint is included in the optimization. With λ as Lagrange multiplier we achieve the following optimization problem:

$$(\underline{U} \underline{x})^T (\underline{U} \underline{x}) + \lambda (1 - \underline{x}^T \underline{x}) \longrightarrow \max_{\underline{x}}$$

11.1 Principal Component Analysis

The solution of this maximization yields the eigenvalue problem:

$$(\underline{U}^T \underline{U}) \underline{x} = \lambda \underline{x}$$

The eigenvector corresponding to the highest eigenvalue λ_1 is the 1. axis \underline{x}_1 , the eigenvector corresponding to the second highest eigenvalue λ_2 is the 2. axis \underline{x}_2 , and so on up to the smallest eigenvalue λ_p with the p . axis \underline{x}_p . The eigenvalues of $\underline{U}^T \underline{U}$ are the squared singular values of \underline{U} and thus can be computed with a *singular value decomposition (SVD)*. This can be done to a extremely high accuracy without explicitly squaring the matrix \underline{U} . These eigenvalues all are positive and the associated eigenvectors are orthogonal to each other.

Gene H. Golub, 1932-2007
(www.wikipedia.org)

Gene Golub's licence plate.

Photo of Professor Kroonenberg of the University Leiden.

For fun...

Gene Golub is computer scientist at Stanford University. He has contributed more than anyone else to make SVD the most powerful and common tool of modern linear Algebra (matrix computation).



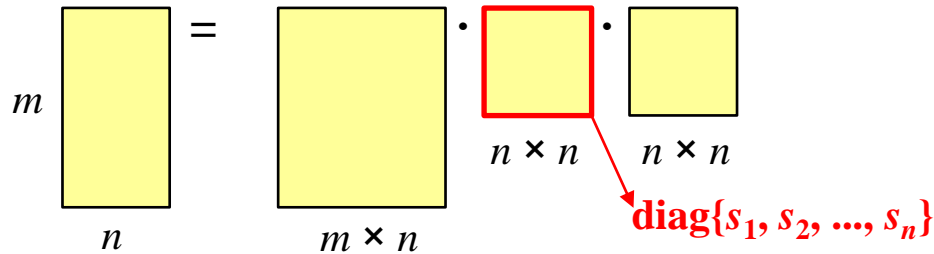
11.1 Principal Component Analysis

Singular Value Decomposition (SVD)

SVD computes the following matrix decomposition of an $m \times n$ matrix \underline{U} :

$$\underline{U} = \underline{W} \underline{S} \underline{V}^T$$

If \underline{U} has more rows than columns the following matrix dimensions arise:



The marked red quadratic matrix in \underline{S} contains the singular values of \underline{U} on its diagonal. They are identical to the square root of the eigenvalues of $\underline{U}^T \underline{U}$. They are sorted from large to small.

Therefore the matrix \underline{U} can be decomposed in a sum of n outer products (each has rank 1), whose influence becomes smaller through the decreasing singular values:

$$\underline{U} = s_1 \underline{w}_1 \underline{v}_1^T + s_2 \underline{w}_2 \underline{v}_2^T + \dots + s_n \underline{w}_n \underline{v}_n^T$$

mit

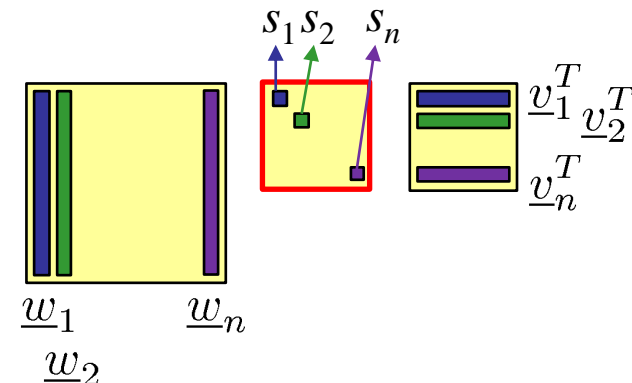
$$\underline{W} = (\underline{w}_1 \quad \underline{w}_2 \quad \dots \quad \underline{w}_n)$$

$$\underline{V} = (\underline{v}_1 \quad \underline{v}_2 \quad \dots \quad \underline{v}_n)$$

maximal rank = n

If the rank of \underline{U} is $r < n$ then

$$s_{r+1} = \dots = s_n = 0.$$



11.1 Principal Component Analysis

If \underline{U} is quadratic ($n = m$) then its eigenvalues λ_i and eigenvectors \underline{x}_i are given by:

$$\underline{U} \underline{x}_i = \lambda_i \underline{x}_i$$

If \underline{U} is rectangular ($n \times m$ with $m > n$ or $m < n$):

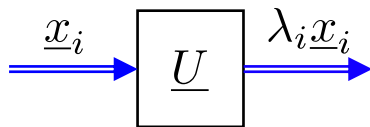
$$\underline{y}_i = \underline{U} \underline{x}_i$$

then \underline{x}_i is n -dim. but \underline{y}_i is m -dim., i.e., the mapping \underline{U} changes the dimension. No eigenvalues and eigenvectors can exist! But if one multiplies a second time with \underline{U}^T then one arrives in n -dim. space again and it is possible to calculate the “squares” of the eigenvalues:

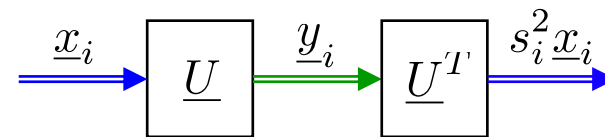
$$\underline{U}^T \underline{U} \underline{x}_i = s_i^2 \underline{x}_i$$

These singular values s_i correspond to the eigenvalues for rectangular matrices. They are the “gains” of matrix \underline{U} in its eigendirections. However, they are always positive.

\underline{U} quadratic



\underline{U} rectangular



11.1 Principal Component Analysis

Example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{bmatrix} = \begin{bmatrix} -0.1013 & 0.7679 & -0.0183 \\ -0.2486 & 0.4881 & 0.5367 \\ -0.3958 & 0.2082 & -0.8133 \\ -0.5430 & -0.0717 & 0.0896 \\ -0.6902 & -0.3515 & 0.2053 \end{bmatrix} \cdot \begin{bmatrix} 35.1826 & 0 & 0 \\ 0 & 1.4769 & 0 \\ 0 & 0 & 0.0000 \end{bmatrix} \cdot \begin{bmatrix} -0.5193 & -0.5755 & -0.6318 \\ -0.7508 & -0.0459 & 0.6589 \\ -0.4082 & 0.8165 & -0.4082 \end{bmatrix} \\
 = 35.1826 \begin{bmatrix} -0.1013 \\ -0.2486 \\ -0.3958 \\ -0.5430 \\ -0.6902 \end{bmatrix} \cdot \begin{bmatrix} -0.5193 & -0.5755 & -0.6318 \end{bmatrix} + 1.4769 \begin{bmatrix} 0.7679 \\ 0.4881 \\ 0.2082 \\ -0.0717 \\ -0.3515 \end{bmatrix} \cdot \begin{bmatrix} -0.7508 & -0.0459 & 0.6589 \end{bmatrix} + 0$$

U has only rank 2 since $s_3 = 0$ and thus the third singular value does not contribute to the rank.

Dimension Reduction by PCA

The PCA transforms data from one p -dimensional space into another p -dimensional space. This for itself can be an advantage because the new data distribution can be numerically better or easier to interpret. One step further is dimensionality reduction by PCA. Here all axes with low variance (below some threshold) are removed. The underlying (implicit) assumption is that these axes represent just noise. This is *especially* appropriate for extremely high-dimensional space where supervised technique would be too complicated.

11.1 Principal Component Analysis

Transformation

The columns of the matrix \underline{V} contain the eigenvectors of $\underline{U}^T \underline{U}$. They are also called the *right* singular vectors of \underline{U} . Correspondingly the *left* singular vectors of \underline{U} are in the columns of the matrix \underline{W} and are identical to the eigenvectors of $\underline{U} \underline{U}^T$. The data contained in the matrix \underline{U} can be transformed linearly into the new space by:

$$\underline{X} = \underline{U} \underline{V}$$

For the transformation back we have to calculate from \underline{X} to \underline{U} :

$$\underline{U} = \underline{X} \underline{V}^{-1} = \underline{X} \underline{V}^T$$

The last equality hold because \underline{V} is unitary, i.e., $\underline{V}^T \underline{V} = \underline{I}$ and $\underline{V} \underline{V}^T = \underline{I}$ and thus $\underline{V}^T = \underline{V}^{-1}$.

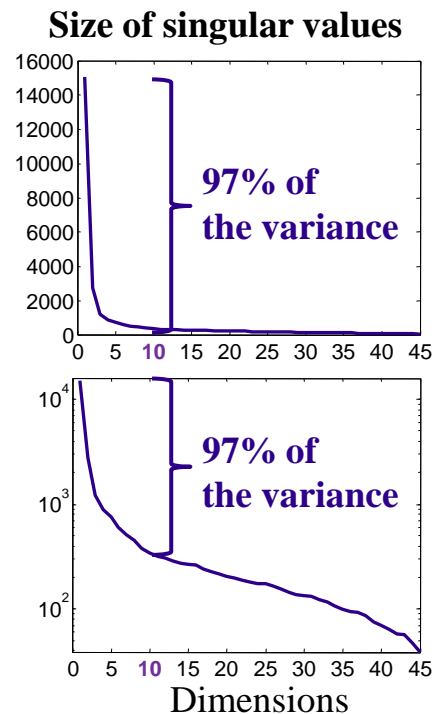
In the case of dimensionality reduction only the most important axes are selected. They belong to the largest eigenvalues of $\underline{U}^T \underline{U}$ or to the largest singular values of \underline{U} , respectively. Because a SVD sorts the eigenvalues according to their absolute values, this corresponds to the first singular values.

$$\underline{X}_{\text{red}} = \underline{U} \underline{V}_{\text{red}}$$

11.1 Principal Component Analysis

Example: Compression of a picture

- Picture with 128×45 pixels is represented as a 128×45 -dimensional matrix where “0” stands for “black” up to “255” for “white” and many grey shades in between.
- The most important 5-10 axes from a PCA already represent the picture quite well. The singular values quickly decline to 0.
- Computational effort is high. This method is not used in praxis.



Original

45

Dimensionality reduction to ? axes:

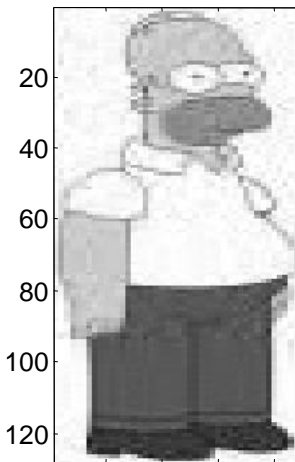
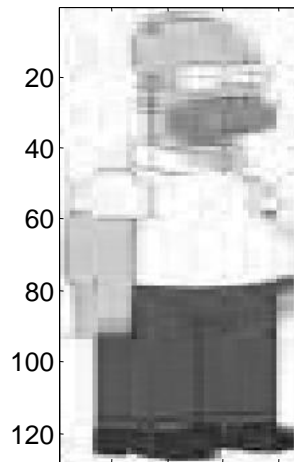
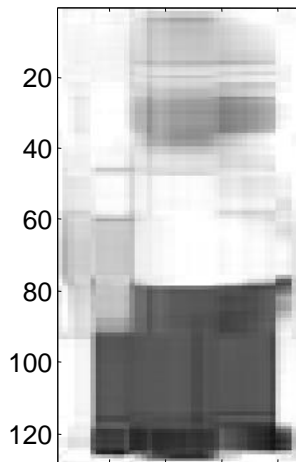
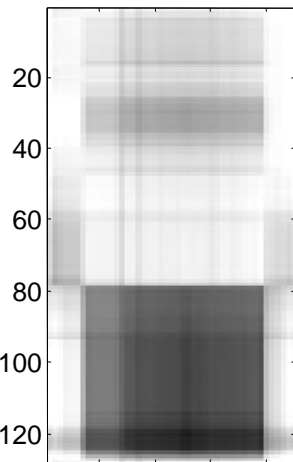
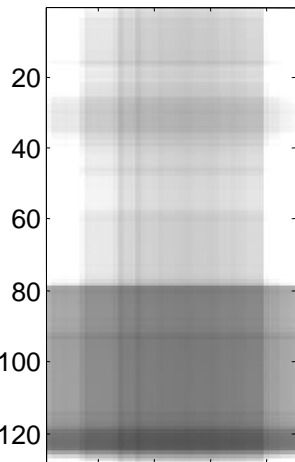
1

2

5

10

20

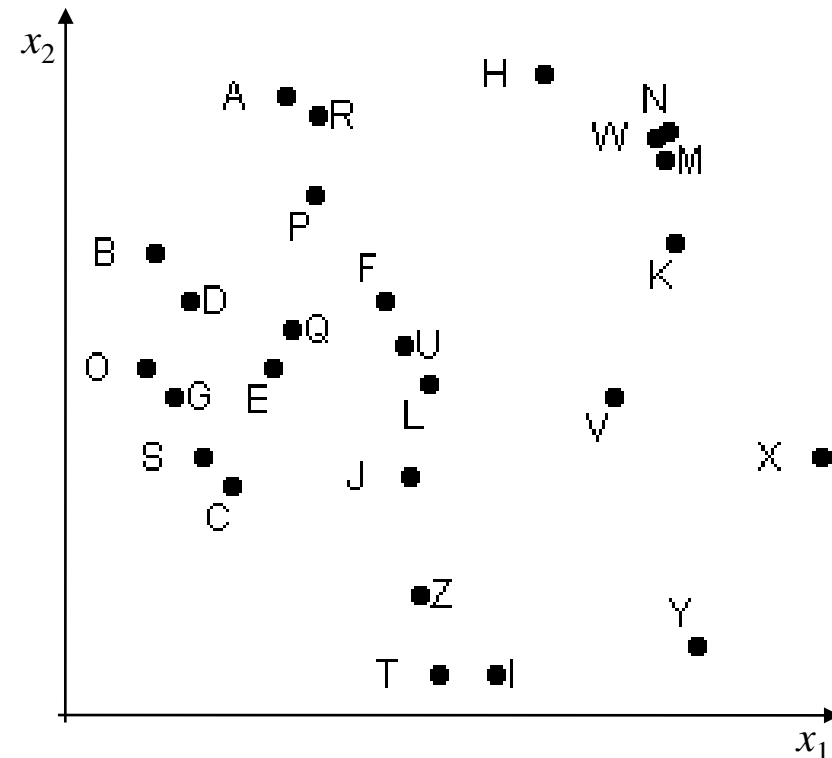
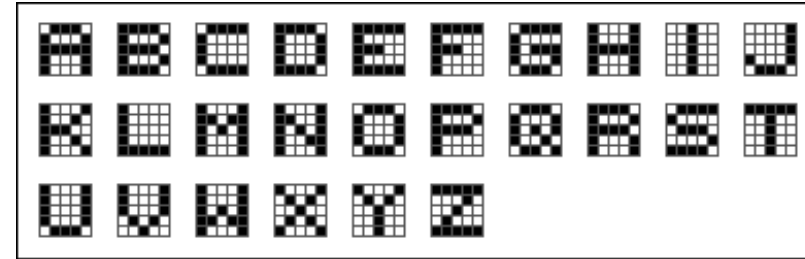


11.1 Principal Component Analysis

Example: Character Recognition

Source: <http://www.cs.mcgill.ca/~sqr/dimr/dimreduction.html>

- Characters A-Z with 5×5 pixels with “0” = “black” and “1” = “white”.
- Each pixel corresponds to one axis u_1, u_2, \dots, u_{25} .
- On each axis the pixel values (“0” or “1”) are entered, i.e., in each dimension only values at 0 and 1 appear.
- The 25-dimensional input space corresponds unit hyper-cube. Data only appears at the corners.
- PCA with dimensionality reduction to 2 axes x_1 and x_2 explains 44% of the data variance!
- “A” /”R” and “W”/”N”/”M” lie closely together. They are hard to distinguish from the 2 features alone. For “X”/”O” and “T”/”H” and “A”/”Y” the distinction is much easier!



11.1 Principal Component Analysis



Professor SVD

BY CLEVE MOIER

Stanford computer science professor Gene Golub has done more than anyone to make the singular value decomposition one of the most powerful and widely used tools in modern matrix computation.



Gene Golub's license plate, photographed by Professor P. M. Kroonenberg of Leiden University.

The SVD is a recent development. Pete Stewart, author of the 1993 paper "On the Early History of the Singular Value Decomposition", tells me that the term *valeurs singulieres* was first used by Emile Picard around 1910 in connection with integral equations. Picard used the adjective "singular" to mean something exceptional or out of the ordinary. At the time, it had

nothing to do with singular matrices. When I was a graduate student in the early 1960s, the SVD was still regarded as a fairly obscure theoretical concept. A book that George Forsythe and I wrote in 1964 described the SVD as a nonconstructive way of characterizing the norm and condition number of a matrix. We did not yet have a practical way to actually compute it. Gene Golub and W. Kahan published the first effective algorithm in 1965. A variant of that algorithm, published by Gene Golub and Christian Reinsch in 1970 is still the one we use today. By the time the first MATLAB appeared, around 1980, the SVD was one of its highlights.

We can generate a 2-by-2 example by working backwards, computing a matrix

The singular value decomposition (SVD), is a matrix factorization with a wide range of interesting applications.

from its SVD. Take $\sigma_1 = 2$, $\sigma_2 = 1/2$, $\theta = \pi/6$ and $\phi = \pi/4$. Let

$$U = \begin{pmatrix} -\cos \theta & \sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

$$V = \begin{pmatrix} -\cos \phi & \sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}$$

The matrices U and V are rotations through angles θ and ϕ , followed by reflections in the first dimension. The matrix Σ is a diagonal scaling transformation. Generate A by computing

$$A = U\Sigma V^T$$

You will find that

$$A = \begin{pmatrix} 1.4015 & -1.0480 \\ .4009 & 1.0133 \end{pmatrix}$$

This says that the matrix A can be generated by a rotation through 45° and a reflection, followed by independent scalings in each of the two coordinate directions by factors of 2 and $1/2$, respectively, followed by a rotation through 30° and another reflection.

The MATLAB function `eigshow` generates a figure that demonstrates the singular value decomposition of a 2-by-2 matrix. Enter the statements

```
A = [1.4015 -1.0480;
     -0.4009  1.0133];
eigshow(A)
```



11.1 Principal Component Analysis

a possibly different orthonormal basis for the range. The transformation becomes independent of scalings or dilatations in each coordinate direction.

The *rank* of a matrix is the number of linearly independent rows, which is the same as the number of linearly independent columns. The rank of a diagonal matrix is clearly the number of nonzero diagonal elements. Orthogonal transformations preserve linear independence. Thus, the rank of any matrix is the number of nonzero singular values. In MATLAB, enter the statement

`type rank`

to see how we choose a tolerance and count nonnegligible singular values.

Traditional courses in linear algebra make considerable use of the reduced row echelon form (RREF), but the RREF is an unreliable tool for computation in the face of inexact data and arithmetic. The SVD can be regarded as a modern, computationally powerful replacement for the RREF.

A square diagonal matrix is nonsingular if, and only if, its diagonal elements are nonzero. The SVD implies that any square matrix is nonsingular if, and only if, its singular values are nonzero. The most numerically reliable way to determine whether matrices are singular is to test their singular values. This is far better than trying to compute determinants, which have atrocious scaling properties.

With the singular value decomposition, the system of linear equations

$$Ax = b$$

becomes

$$UZV^T x = b$$

The solution is

$$x = VZ^{-1}U^T b$$

Multiply by an orthogonal matrix, divide by the singular values, then multiply by another orthogonal matrix. This is much more

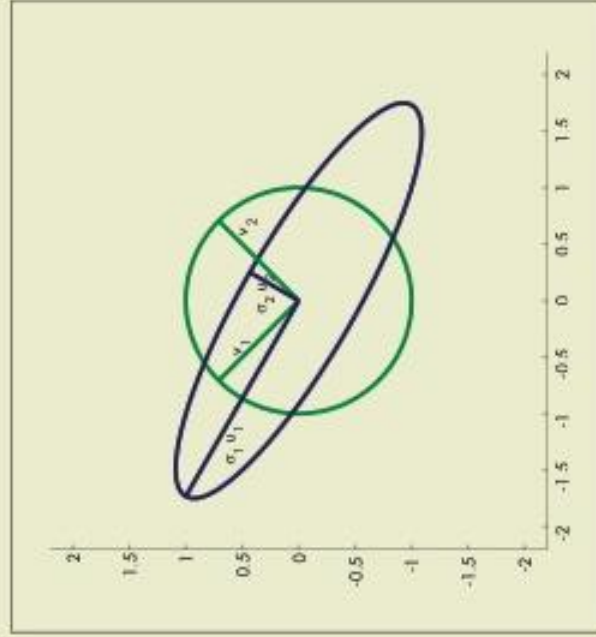


Figure 1. SVD figure produced by eLigthon.

Click the SVD button and move the mouse around. You will see Figure 1, but with different labels.

The green circle is the unit circle in the plane. The blue ellipse is the image of this circle under transformation by the matrix A . The green vectors, v_1 and v_2 , which are the columns of V , and the blue vectors, u_1 and u_2 , which are the columns of U , are two different orthogonal bases for two-dimensional space. The columns of V are rotated 45° from the axes of the figure, while the columns of U , which are the major and minor axes of the ellipse, are rotated 30° . The matrix A transforms v_1 into $\sigma_1 u_1$ and v_2 into $\sigma_2 u_2$.

Let's move on to m -by- n matrices. One of the most important features of the SVD is its use of orthogonal matrices. A real matrix U is orthogonal, or has *orthonormal* columns, if

$$U^T U = I$$

This says that the columns of U are perpendicular to each other and have unit length.

Geometrically, transformations by orthogonal matrices are generalizations of rotations and reflections; they preserve lengths and angles. Computationally, orthogonal matrices are very desirable because they do not magnify roundoff or other kinds of errors.

Any real matrix A , even a nonsquare one, can be written as the product of three matrices

$$A = UZV^T$$

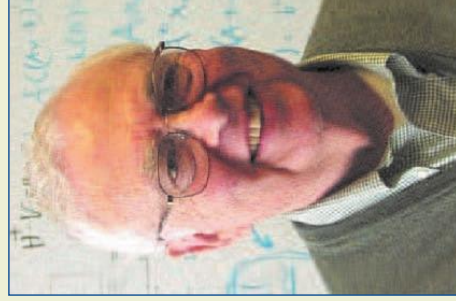
The matrix U is orthogonal and has as many rows as A . The matrix V is orthogonal and has as many columns as A . The matrix Z is the same size as A , but its only nonzero elements are on the main diagonal. The diagonal elements of Z are the *singular values*, and the columns of U and V are the left and right *singular vectors*.

In abstract linear algebra terms, a matrix represents a linear transformation from one vector space, the *domain*, to another, the *range*. The SVD says that for any linear transformation it is possible to choose an orthonormal basis for the domain and

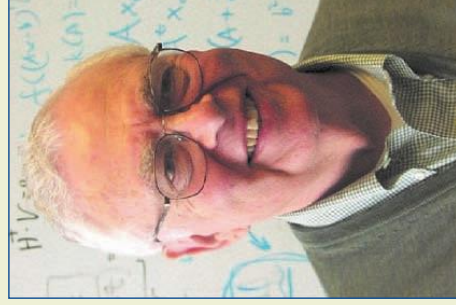
11.1 Principal Component Analysis



12



50



120

Figure 2. Rank 12, 50, and 120 approximations to a rank 598 color photo of Gene Golub.

computational work than Gaussian elimination, but it has impeccable numerical properties. You can judge whether the singular values are small enough to be regarded as negligible, and if they are, analyze the relevant singular system.

Let E_k denote the outer product of the k -th left and right singular vectors, that is

$$E_k = u_k v_k^T$$

Then A can be expressed as a sum of rank-1 matrices,

$$A = \sum_{k=1}^n \sigma_k E_k$$

If you order the singular values in decreasing order, $\sigma_1 > \sigma_2 > \dots > \sigma_n$, and truncate the sum after r terms, the result is a rank- r approximation to the original matrix. The error in the approximation depends upon the magnitude of the neglected singular values. When you do this with a matrix of data that has been centered, by subtracting the mean of each column from the entire column, the process is known as *principal component analysis* (PCA). The right singular vectors, v_k , are

the *components*, and the scaled left singular vectors, $\sigma_k u_k$, are the *scores*. PCAs are usually described in terms of the eigenvalues and eigenvectors of the covariance matrix, AA^T , but the SVD approach sometimes has better numerical properties.

SVD and matrix approximation are often illustrated by approximating images. Our example starts with the photo on Gene Golub's Web page (Figure 2). The image is 897-by-598 pixels. We stack the red, green, and blue JPEG components vertically to produce a 2691-by-598 matrix. We then do just one SVD computation. After computing a low-rank approximation, we repartition the matrix into RGB components. With just rank 12, the colors are accurately reproduced and Gene is recognizable, especially if you squint at the picture to allow your eyes to reconstruct the original image. With rank 50, you can begin to read the mathematics on the white board behind Gene. With rank 120, the image is almost indistinguishable from the full rank 598. (This is not a particularly effective image compression technique. In fact, my friends in image processing call it "image degradation.")

So far in this column I have hardly mentioned eigenvalues. I wanted to show that it is possible to discuss singular values without discussing eigenvalues—but, of course, the two are closely related. In fact, if A is square, symmetric, and positive definite, its singular values and eigenvalues are equal, and its left and right singular vectors are equal to each other and to its eigenvectors. More generally, the singular values of A are the square roots of the eigenvalues of AA^T or AA^T . Singular values are relevant when the matrix is regarded as a transformation from one space to a different space with possibly different dimensions. Eigenvalues are relevant when the matrix is regarded as a transformation from one space into itself—as, for example, in linear ordinary differential equations.

Google finds over 3,000,000 Web pages that mention "singular value decomposition" and almost 200,000 pages that mention "SVD MATLAB." I knew about a few of these pages before I started to write this column. I came across some other interesting ones as I surfed around.

Professor SVD made all of this, and much more, possible. Thanks, Gene. ◀◀



11.1 Principal Component Analysis

A Few Search Results for "Singular Value Decomposition"

- The Wikipedia pages on SVD and PCA are quite good and contain a number of useful links, although not to each other.
en.wikipedia.org/wiki/Singular_value_decomposition
en.wikipedia.org/wiki/Principal_component_analysis

- Rasmus Bro, a professor at the Royal Veterinary and Agricultural University in Denmark, and Barry Wise, head of Eigenvector Research in Wenatchee, Washington, both do chemometrics using SVD and PCA. One example involves the analysis of the absorption spectrum of water samples from a lake to identify upstream sources of pollution.

www.models.kvl.dk/users/rasmus
www.eigenvector.com

- Tammy Kolda and Brett Bader, at Sandia National Labs in Livermore, ca, developed the Tensor Toolbox for MATLAB, which provides generalizations of PCA to multidimensional data sets.
csni.ca.sandia.gov/~tgkolda/TensorToolbox

- In 2003, Lawrence Sirovich of the Mount Sinai School of Medicine published "A pattern analysis of the second Rehnquist U.S. Supreme Court" in the *Proceedings of the US National Academy of Sciences*. His paper led to articles in the *New York Times* and the *Washington Post* because it provides a nonpolitical, phenomenological model of court decisions. Between 1994 and 2002, the court heard 468 cases. Since there are nine justices, each of whom takes a majority or minority position on each case, the data is a 468-by-9 matrix of +1s and -1s. If the judges had made their decisions by flipping coins, this matrix would almost certainly have rank 9. But Sirovich found that the third singular value is an order of magnitude smaller than the first one, so the matrix is well approximated by a matrix of rank 2. In other words, most of the court's decisions are close to being in a two-dimensional subspace of all possible decisions.
www.pnas.org/cgi/rapidprint/100/13/7432

- Latent Semantic Indexing involves the use of SVD with term-document matrices to perform document retrieval. For example, should a search for "singular value" also look for "eigenvalue"? See a 1999 *SIAM Review* paper by Michael Berry, Zlatko Drmac, and Liz Jessup. "Matrices, Vector Spaces, and Information Retrieval!"
epubs.siam.org/SIREV/volume-41/art_34703.html

- The first Google hit on "protein svd" is "Protein Substate Modeling and Identification Using the SVD," by Tod Romo at Rice University. The site provides an electronic exposition of the use of SVD in the analysis of the structure and motion of proteins, and includes some gorgeous graphics.
bioc.rice.edu/~tromo/Sprez/loc.html

bioc.rice.edu/~tromo/Sprez/loc.html

- Los Alamos biophysicists Michael Wall, Andreas Rechsteiner, and Luis Rocha provide a good online reference about SVD and PCA, phrased in terms of applications to gene expression analysis.
public.lanl.gov/newall/kluwer2002.html

- "Representing cyclic human motion using functional analysis" (2005), by Dirk Ormonet, Michael Black, Trevor Hastie, and Hedvig Kjellstrom, describes techniques involving Fourier analysis and principal component analysis for analyzing and modeling motion-capture data from activities such as walking.
www.csc.kth.se/~hedvig/publications/ivc_05.pdf

- A related paper is "Decomposing biological motion: a framework for analysis and synthesis of human gait patterns" (2002), by Nicholas Troje. Troje's work is the basis for an "ei-genwalker" demo.
www.journalofvision.org/2/15/2
www.mathworks.com/moler/mcm/walker.m

www.journalofvision.org/2/15/2
www.mathworks.com/moler/mcm/walker.m

- A search at the US Patent and Trademark Office Web page lists 1,197 U.S. patents that mention "singular value decomposition." The oldest, issued in 1987, is for "A fiber optic inspection system for use in the inspection of sandwiched solder bonds in integrated circuit packages". Other titles include "Compression of surface light fields", "Method of seismic surveying", "Semantic querying of a peer-to-peer network", "Biochemical markers of brain function", and "Diabetes management".
www.uspto.gov/patft

RESOURCES

- On the Early History of the Singular Value Decomposition
locus.siam.org/SIREV/volume-35/art_1035134.html
- Cleve's Corner Collection
www.mathworks.com/es/cleve

11.1 Principal Component Analysis

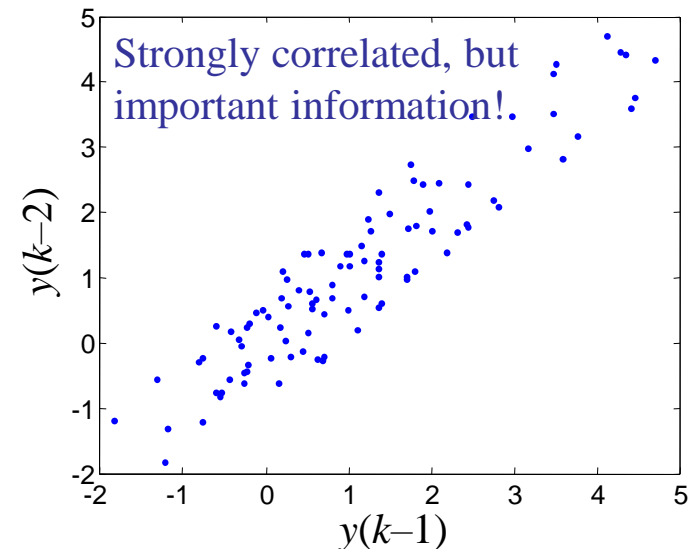
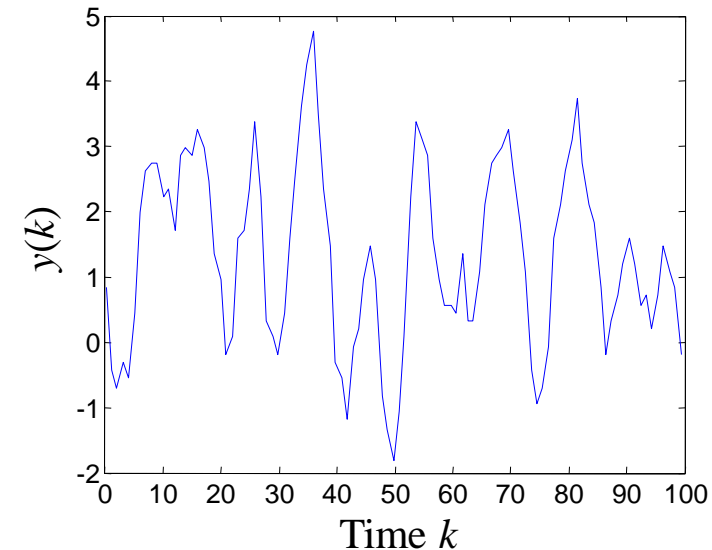
Difficulties with Dimensionality Reduction

The assumption that low variance axes are redundant and can be removed can be wrong! A small variance point towards a possible linear dependency but this is not necessarily the case. An analysis based on input space distributions only can never ensure this with certainty. The output has to be considered in order to be sure.

For example for dynamic processes a strong correlation of two subsequent outputs $y(k-1)$ and $y(k-2)$ occurs. However, they are *not* redundant if the process is of AR(2)-type as an example, that is it follows the equation:

$$y(k) = -a_1y(k-1) - a_2y(k-2) + v(k)$$

Although $y(k-1)$ and $y(k-2)$ are highly correlated (the higher, the smaller the sampling time is) both carry important information and are not redundant.

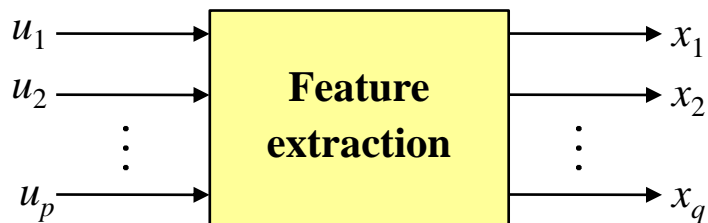


11.1 Principal Component Analysis

Feature Selection versus Feature Extraction

A dimensionality reduction with PCA yield a feature *extraction*. This means that from a many original inputs, say p , a smaller number of features, say q , are generated. However, they may depend on all original inputs. Therefore the next processing step requires are smaller number of inputs/features and is simpler to perform. But none of the original p measurements can be discarded.

A more radical approach is feature *selection*. Here the task is not only to reduce the dimensionality but also to remove inputs so that they don't have to be measured anymore. This simplifies not only the further processing but also the overall effort by requiring fewer sensors.



Each output x_i can depend on *all* inputs u_j !



Each output is identical to *one* input!

11.1 Principal Component Analysis

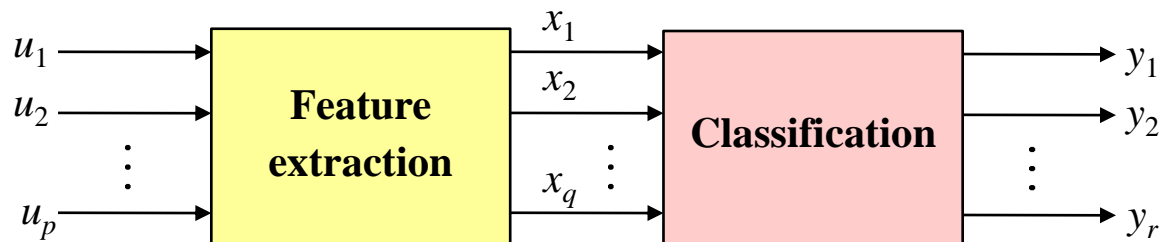
Application: Classification

A frequent application of PCA is data pre-processing, especially for dimensionality reduction in classification. The task is to correctly map measurements to r different classes. This can be done with the original measurements u_1, u_2, \dots, u_p or with features x_1, x_2, \dots, x_q extracted from these measurements. Usually $q \ll p$ which means that the classification problem becomes of much lower dimensionality.

In the *A-Z-character recognition* example we have $r = 26$ classes, $p = 5 \times 5 = 25$ original inputs and only $q = 2$ features (although for a higher classification accuracy than 44% we would require realistically 3-5 features).

For a *coin-operated machine* we would have to distinguish between $r = 9$ classes (1c, 2c, 5c, 10c, 20c, 50c, 1€, 2€, “no €-coin”). Possible inputs are

- Weight, color, diameter, thickness, reflectance, ...



11.2 Clustering

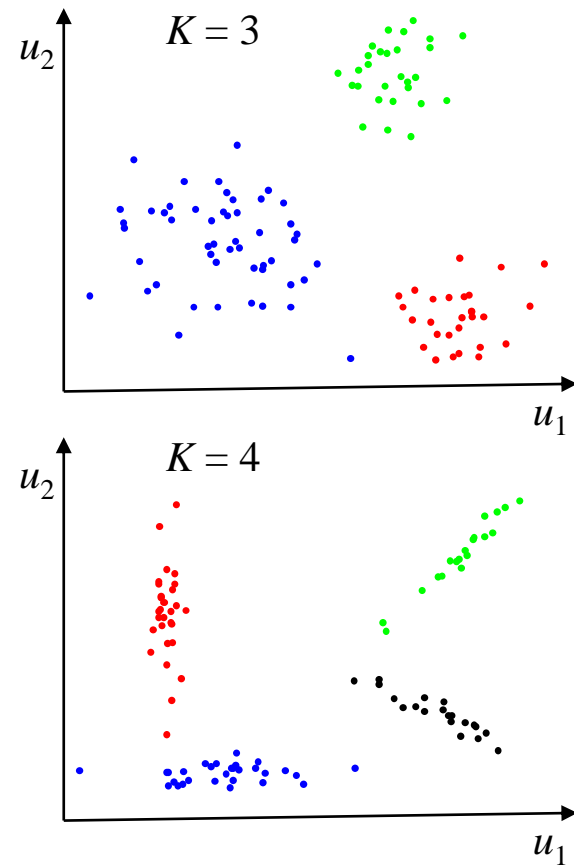
Basics of Clustering

Like PCA Clustering operates on the *input* data. The task is to find groups (clusters) of data points. These groups can be of different shapes and sizes. Depending on the method, a special *prototype* is defined that defines how a cluster should look like. In two dimensions examples are: hollow or filled circles or ellipsoids, lines, ...

A *similarity measure* is defined as a loss function. The similarity of each cluster is evaluated with this similarity measure. The famous *K-means clustering* for example utilizes the following type of loss function:

$$J = \sum_{j=1}^K \sum_{i \in S_j} \|\underline{u}(i) - \underline{c}_j\|^2 \longrightarrow \min_{\underline{c}_j}$$

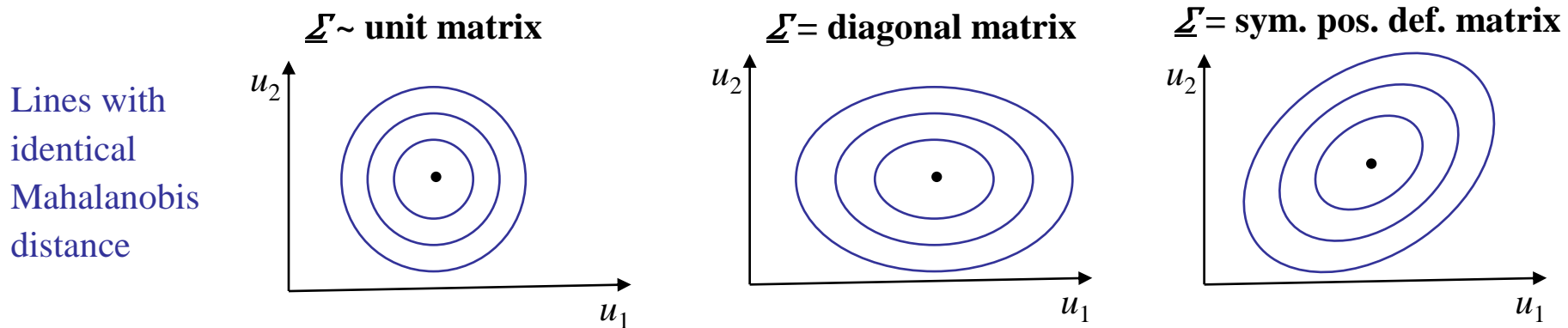
where K is the number of clusters and $i \in S_j$ runs over the data points belonging to the cluster j whose center of gravity is closest (in the Euclidian sense).



11.2 Clustering

K-means clustering tries to find K filled circles (or spheres) by minimizing the quadratic distances of all data points to the center of their associated cluster.

Instead of looking for circles (spheres) it can be easily extended to ellipses (ellipsoids) of a certain shape, i.e., a given covariance matrix $\underline{\Sigma}$. This can be done by replacing the Euclidian distance metric with the so-called *Mahalanobis* distance.



An extension to higher dimensions is easily possible.

It is possible as well to look for ellipse (ellipsoids) of variable covariance matrix (shape). However, this requires more complex algorithms as designed by *Gustafson* and *Kessel* or *Gath* and *Geva*.

11.2 Clustering

K-means Clustering

The K-means algorithm works as follows:

1. Choose the number of clusters K .
2. Initialize the cluster center with randomly selected data points.
3. Assign each data sample to the cluster with the closest center (according to the chosen distance metric).
4. Calculate the center of gravity for each cluster (averaging the associated data points).
5. Place the new cluster centers at those centers of gravity
6. If (at least) one cluster center has moved then go to step 3 otherwise STOP.

It can be shown that this algorithm minimizes the loss function (on the previous slide).

However, it can converge to a local optimum. Because the initialization is random, different initialization can be tried out and the best result can be selected.

A difficult “tuning factor” is the choice for the number of clusters K .

11.2 Clustering

Examples for K-means Clustering

Interpretation of the figures:

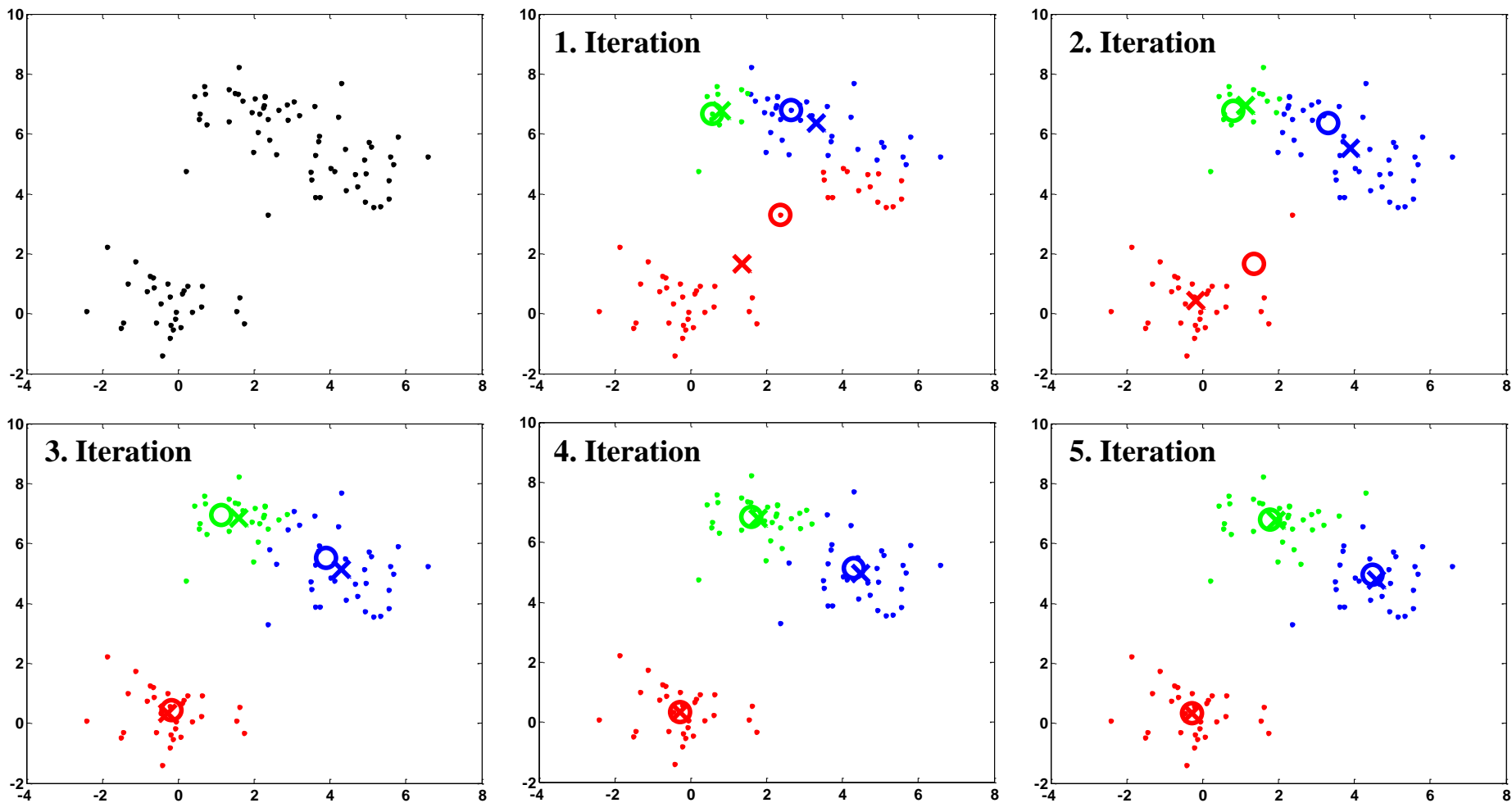
- Data points are marked by dots.
- The old cluster centers are marked by circles.
- The new cluster centers are marked by crosses.
- The color of the data points represents the association to the cluster of the same color.

Observations:

- Convergence is very fast; only a few iterations are needed.
- The global minimum of the loss function is reached in most cases.
- The sensitivity with respect to the initialization is low.
- For reasonable results the number of clusters has to be chosen in the right manner.
- Normalization of data is important because some dimensions can be dominant (and others almost irrelevant) if axes are scaled differently.

11.2 Clustering

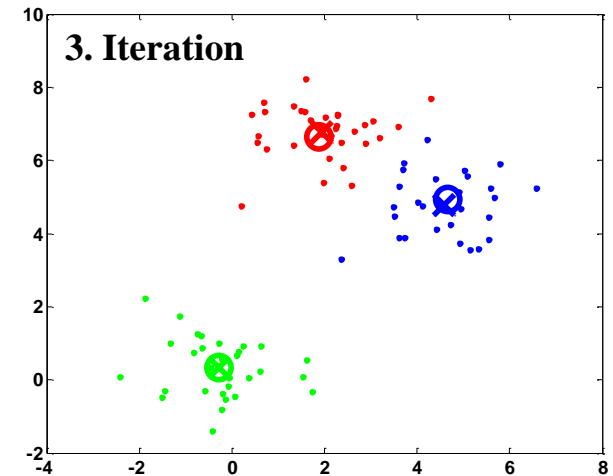
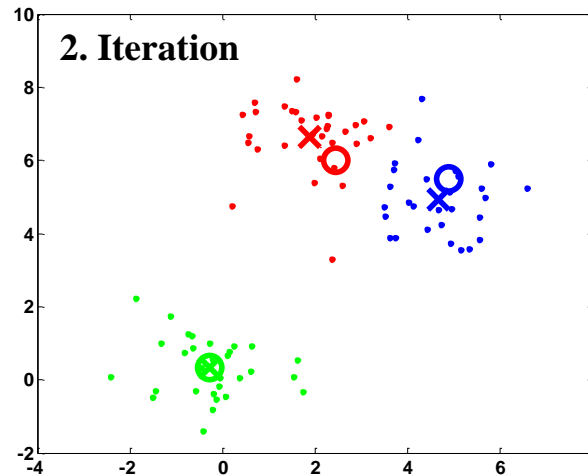
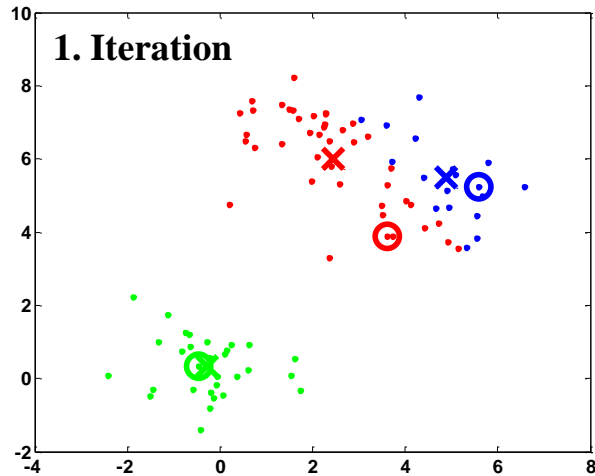
$K = 3$ 5 Iterations until convergence



11.2 Clustering

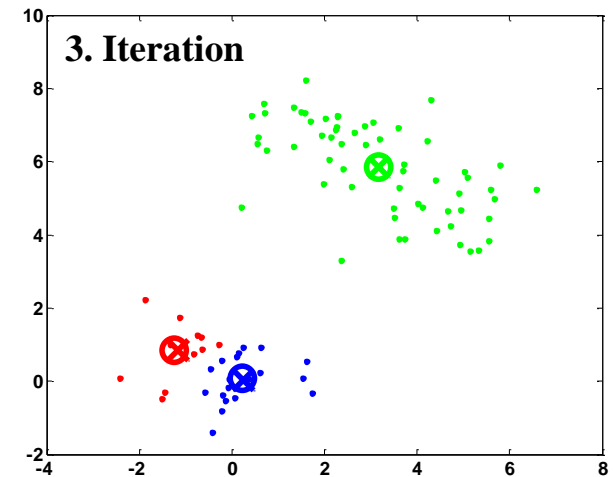
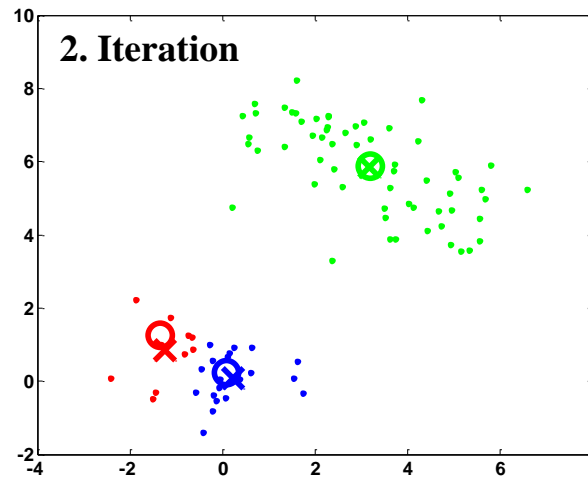
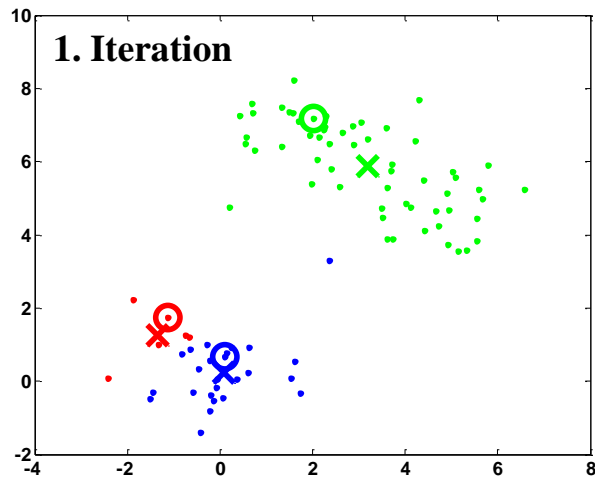
$K = 3$ 3 Iterations until convergence

Fast convergence due to lucky initialization



$K = 3$ 3 Iterations until convergence

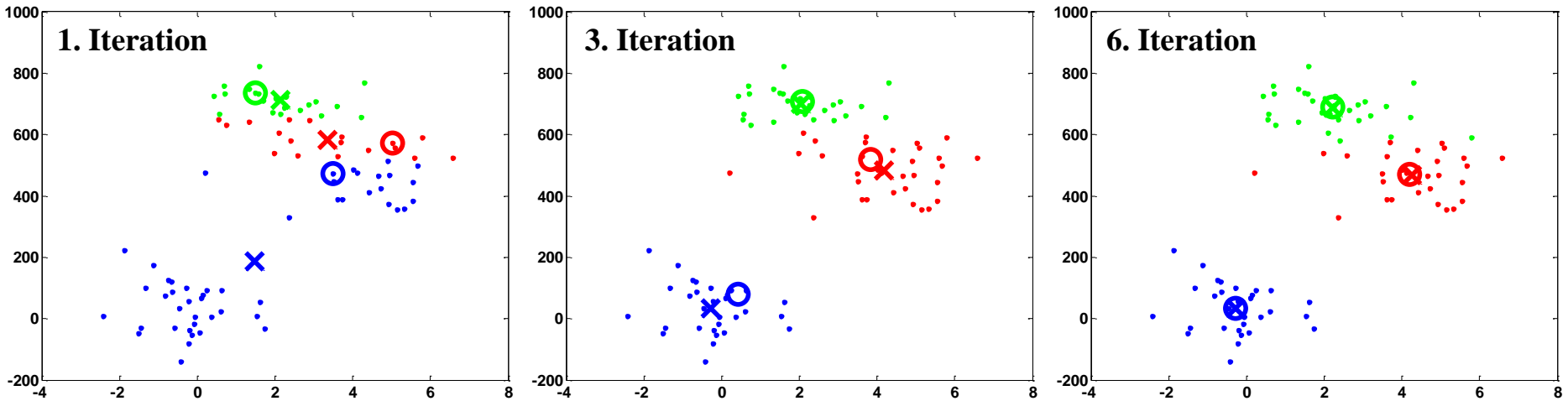
Bad result due to unlucky initialization



11.2 Clustering

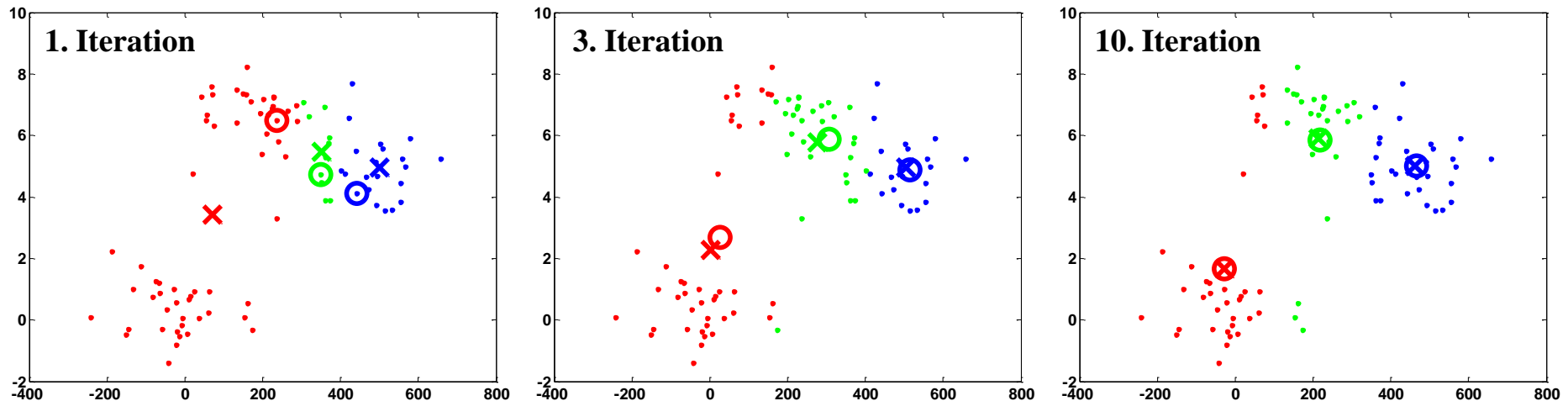
$K = 3$ 6 Iterations until convergence

Scaling of the y-axis is factor 100 larger



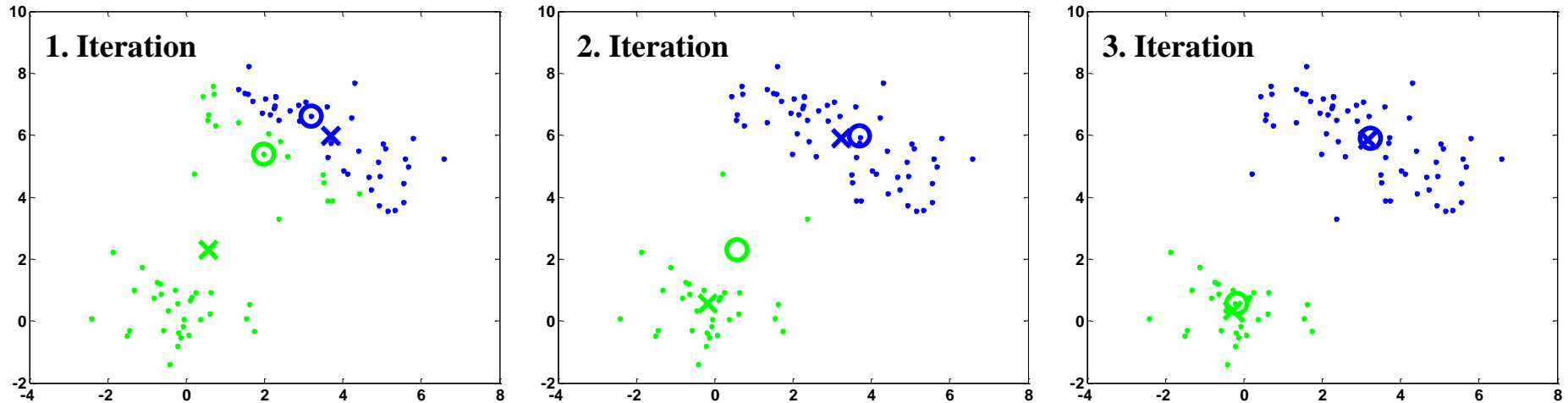
$K = 3$ 10 Iterations until convergence

Scaling of the x-axis is factor 100 larger

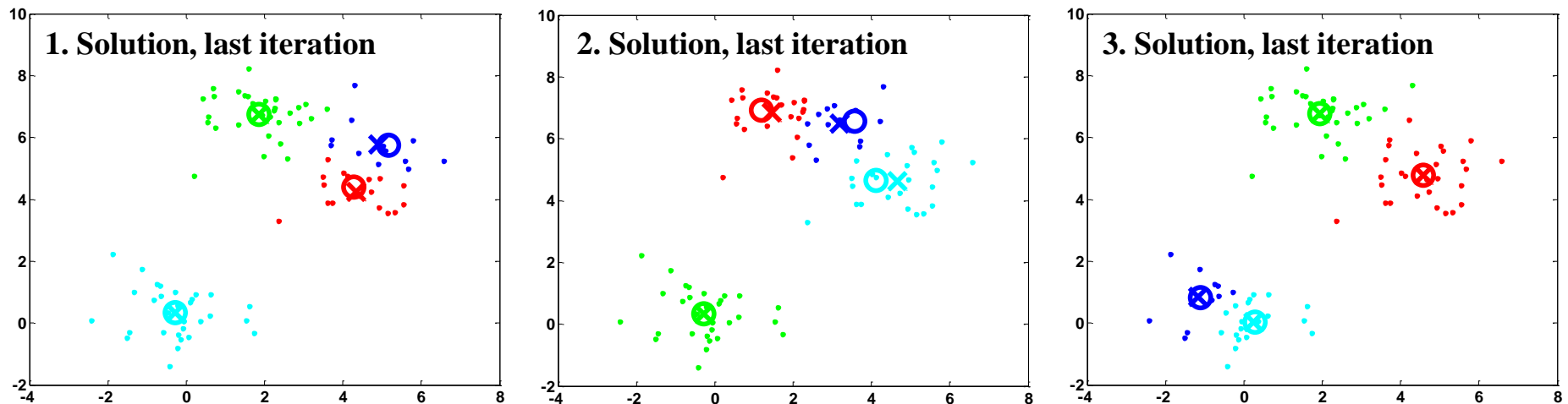


11.2 Clustering

$K = 2$ 3 Iterations until convergence Solution is stable, almost independent of initialization



$K = 4$ Many different solutions dependent on the initialization



11.2 Clustering

Fuzzy Clustering

The loss function known from K-means clustering can be re-written (extended):

$$J = \sum_{j=1}^K \sum_{i \in \mathcal{S}_j} \|u(i) - c_j\|^2 = \sum_{j=1}^K \sum_{i=1}^N \mu_{ij} \|u(i) - c_j\|^2 \longrightarrow \min_{c_j}$$

The second sum runs over all data points (not only those belonging to a single cluster j). K-means is a special case of fuzzy K-means with

$$\mu_{ij} = \begin{cases} 1 & \text{if data point belongs to cluster } j \\ 0 & \text{if data point does not belong to cluster } j \end{cases}$$

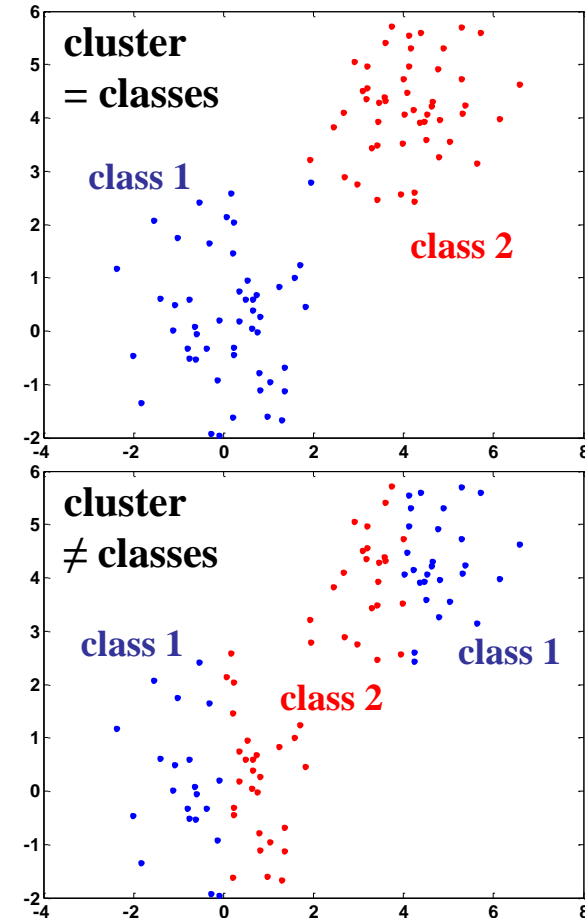
The variable μ_{ij} denotes the degree of membership to a cluster. A value of “1” means this point fully belongs to that cluster. A value of “0” means it doesn’t. The degree of membership μ_{ij} can be extended from a binary values to a real value between 0 and 1. Each point belongs to each cluster to a certain degree. They have to sum up to 1. A degree of membership of 0.51 to cluster A is similar to 0.49 to cluster B and would yield similar results. In the classical K-means it is binary and the point would fully be associated with cluster A und not at all with cluster B. Therefore, fuzzy clustering is less prone to bad initialization.

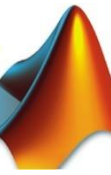
11.2 Clustering

Clustering for Classification

Like PCA clustering is suitable for data pre-processing. It is often utilized for solving classification problems. Instead of directly feeding the input features to the (supervised) classifier, they are clustered first. With the help of these cluster, the classifier has an easier task to perform the classification.

The underlying idea is that a certain distribution of the data reflects the associated classes. Often this is the case. However, this is not guaranteed. Therefore an unsupervised method can go astray.





PCA:

```
[COEFF,SCORE] = princomp(X);1
```

Singular Value Decomposition:

```
[U,S,V] = svd(X);
```

Fuzzy K-means Clustering:

```
[center,U,obj_fcn] = fcm(data,cluster_n);2
```

¹ : *Statistics Toolbox*

² : *Fuzzy Logic Toolbox*

5. Measurement Errors and Statistics

Contents of Chapter 5

5. Measurement Errors and Statistics

5.1 Measurement Errors

5.2 Accuracy Rating

5.3 Error Propagation

5.4 Histograms and Probability Density

5.5 Estimation of Mean and Variance

5.6 Confidence Intervals

5.1 Measurement Errors

Error Definitions

The **absolute error** e of some measurements is the difference between the displayed or outputted value y and the (typically unknown) true value Wert y_w :

$$e = y - y_w$$

The **relative error** e_r is absolute error divided by the true value y_w and commonly is given in percentage:

$$e_r = \frac{y - y_w}{y_w}$$

The true value y_w is unknown in practice (otherwise no measurement would be necessary). With additional effort it can be determined with high accuracy:

- Measurement with a precision instrument.
- Comparison with a measuring standard.

Often the **quadratic error** e^2 (absolute or relative) is utilized for optimization as an criterion. Many reasons for this exist. An important one is that the resulting optimization is particularly easy to solve and manage (*least squares*).

5.1 Measurement Errors

Systematic and Random Errors

Two error classes have to be distinguished:

- *Systematic errors*: Reason and kind of the error action are known. With a higher effort in the measurement system an improvement and/or compensation would be possible, at least in principle.

Examples: Temperature influence with strain gauges. Nonlinear characteristics.

- *Random or stochastic errors*: Repeated measurements under identical conditions yield different results. Typically the errors are different in size and sign (not necessarily, see quantization errors). The measurement values *scatter*! In contrast to systematic errors, random errors can not be predicted or compensated. With averaging (calculating the mean value), however, their influence can be reduced. The result will improve in quality typically with $1/\sqrt{N}$ where N is the number of trials that are averaged.

Examples: Brownian Motion. Fluctuations in material composition.

If we look very closely, most/all errors are of systematic nature. We have limited resources and cannot afford an arbitrary effort; we do not have infinite insights. Therefore we treat all errors that *seem* to be random as random! Typically many independent small systematic influences seem to be of random nature.

5.1 Measurement Errors

Error Causes

- *Disturbances*: It has to be distinguished between internal and external disturbances:
 - Internal disturbances affect the sensor itself, e.g. wear.
 - External disturbances come from the outside world, e.g. temperature influences.By accepting a high effort in the choice of a precision instrument and by changing the environment (e.g. climate chamber), disturbances can be kept to a minimum but they can never be annihilated.
- *Observation errors*: Error induced by the observer himself, e.g. by making a mistake during the measurement, wrongly reading the display, ... With care such errors can be avoided.
- *Feedback error*: Influence of the sensor on the object to be measured, e.g. the temperature of the thermometer changes the temperature of the body that shall be measured. The amount of such feedback depends on the measurement method. Radiation-based temperature measurement avoids such an unwanted feedback. Physics tells us some effect can never be completely eliminated (Heisenberg's uncertainty principle) but on a macroscopic level it can be negligible with the appropriate method.
- *Non-ideal characteristics*: The measurement system can possess **static** and **dynamic errors** and with a digital output it possesses **quantization errors** as well.

5.1 Measurement Errors

Non-ideal Sensor Characteristics

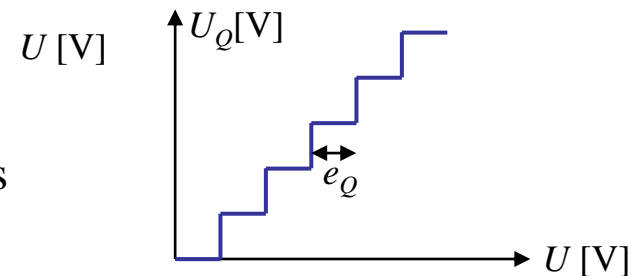
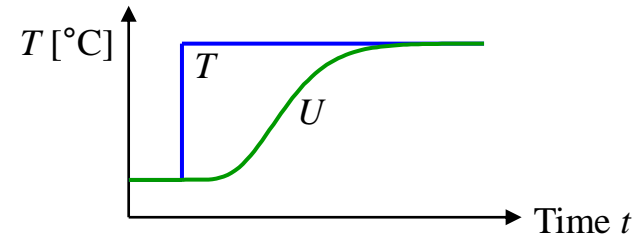
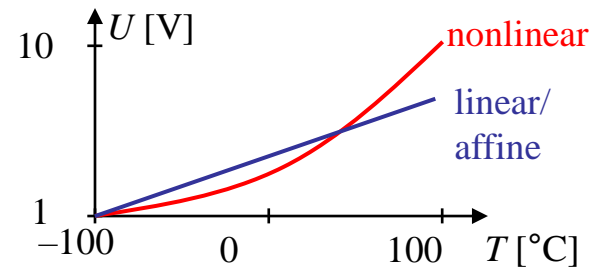
- *Static errors:* In the ideal case, the characteristics of the sensor is **linear/affine**.

In practice **nonlinearities** distort the result.

Example: quantity = temperature, output = voltage:

$T [^{\circ}\text{C}]$	-100	-50	0	50	100
$U [\text{V}]$	1	1.7	3	6	10

- *Dynamic errors:* If the measured quantity changes over time, the sensor follows with a time constant and delay. If we do not wait long enough until the measurement values reach steady state (settling time) a dynamic error occurs.
- *Quantization errors:* During the A/D conversion the discretization causes errors in time (through sampling) and in amplitude (through quantization). The latter corresponds to a stepwise characteristics. The maximum error is $e_Q/2$.



5.2 Accuracy Rating

The quality of measurement devices in practice is often characterized with their **accuracy rating** or **guaranteed minimum accuracy**. With this declaration a manufacturer guarantees that possible measurement errors within the specified conditions are limited to certain interval.

The **accuracy rating** declares the maximally to expect error in percentage of the instrument range.

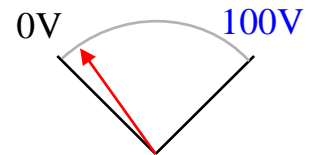
Typical accuracy ratings:
0,1; 0,2; 0,5; 1; 1,5; 2,5

Example: Voltage measurement, accuracy rating = 0,5

a) Range: 0V – 100V. Display: 7V.

$$\text{max. error} = 0,5\% \cdot 100\text{V} = 0,5\text{V}.$$

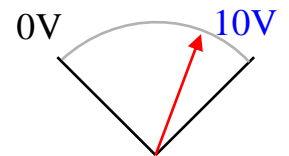
$$\text{guaranteed interval} = 7\text{V} \pm 0,5\text{V}.$$



b) Range: 0V – 10V. Display: 7V.

$$\text{max. error} = 0,5\% \cdot 10\text{V} = 0,05\text{V}.$$

$$\text{guaranteed interval} = 7\text{V} \pm 0,05\text{V}.$$



Recommendation: Always measure in the **upper third** of the instrument range!

5.3 Error Propagation

Problem

Commonly the quantity to be measured cannot be measured *directly* but has to be calculated from other measurements:

Examples:

- a) Determination of electrical power from voltage and current:

$$P = UI$$

- b) Determination of speed or velocity from distance and time interval:

$$v = \frac{s}{t}$$

- c) Determination of force via resistance change dependent on length, area, and specific conductivity:

$$R = \frac{l}{A}\rho \quad \text{with} \quad A = \pi r^2$$

How do errors in the measurement of U, I, s, t, l, A (or r), ρ affect the final results?

5.3 Error Propagation

Gaussian Error Propagation for *Systematic* Errors

The requested quantity y can be deduced from the measurement values x_i , $i = 1, \dots, n$, as follows:

$$y = f(x_1, x_2, \dots, x_n)$$

The errors of the single measurements x_i are denoted by Δx_i . This yields the following systematic error accumulation for the final output y :

$$\Delta y = \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 + \dots + \frac{\partial f}{\partial x_n} \Delta x_n$$

This equation directly is obtained from the Taylor series expansion of the function f , in which all higher than first order terms (linear) are neglected. Thus it is approximately correct if the errors are small, i.e., Δx_i is close to zero.

In the above equation, measurement errors can cancel or attenuate each other because they might be of opposite sign. Of course this requires knowledge about the right sign of Δx_i and the slope of $f()$ and therefore the systematic over- or underestimation.

5.3 Error Propagation

A different situation exists if just the maximal magnitude of errors can be assessed. The following worst case assessment is obtained.

Gaussian Error Propagation for *Maximal* Errors:

$$\Delta y = \left| \frac{\partial f}{\partial x_1} \Delta x_1 \right| + \left| \frac{\partial f}{\partial x_2} \Delta x_2 \right| + \dots + \left| \frac{\partial f}{\partial x_n} \Delta x_n \right|$$

Examples:

a) Power measurement: $P = UI$

$$\Delta P = \frac{\partial(UI)}{\partial U} \Delta U + \frac{\partial(UI)}{\partial I} \Delta I = I \Delta U + U \Delta I \qquad \frac{\Delta P}{P} = \frac{\Delta U}{U} + \frac{\Delta I}{I}$$

If for example the voltage is measured too small ($\Delta U < 0$) and the current too large ($\Delta I > 0$) (and $U > 0, I > 0$), then these error can (partly) compensate each other. If nothing is known about the sign of the errors and only their magnitude can be assessed, then a maximal error assessment has to be made in which the individual errors accumulate.

5.3 Error Propagation

Examples:

b) Speed measurement: $v = \frac{s}{t}$

$$\Delta v = \frac{\partial(s/t)}{\partial s} \Delta s + \frac{\partial(s/t)}{\partial t} \Delta t = \frac{1}{t} \Delta s - \frac{s}{t^2} \Delta t \qquad \frac{\Delta v}{v} = \frac{\Delta s}{s} - \frac{\Delta t}{t}$$

In this example a (partly) compensation happens if both, the distance and time interval, are over- or underestimated because of the “-” sign. Notice that the second term can become extremely large if the time interval t is chosen very small, i.e., then the speed measurement is very sensitive with respect to measurement errors in time.

c) Force measurement with strain gauges: $R = \frac{l}{A} \rho$ with $A = \pi r^2$

$$R = \frac{l\rho}{\pi r^2}$$

$$\Delta R = \frac{\rho}{\pi r^2} \Delta l - 2 \frac{l\rho}{\pi r^3} \Delta r + \frac{l}{\pi r^2} \Delta \rho \qquad \frac{\Delta R}{R} = \frac{\Delta l}{l} - 2 \frac{\Delta r}{r} + \frac{\Delta \rho}{\rho}$$

5.3 Error Propagation

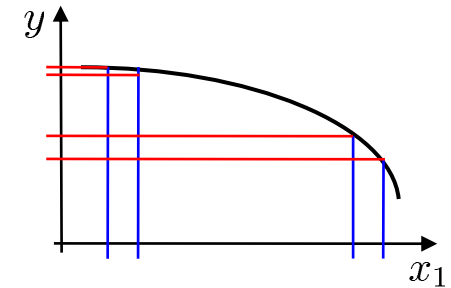
Gaussian Error Propagation for *Random* Errors

The quantity y to be measured depends on the input quantities x_i , $i = 1, \dots, n$, as follows

$$y = f(x_1, x_2, \dots, x_n)$$

The standard deviation of the individual input factors x_i shall be given by s_{xi} . Then the standard deviation of the output quantity y becomes:

$$s_y = \sqrt{\left(\frac{\partial f}{\partial x_1} s_{x1}\right)^2 + \left(\frac{\partial f}{\partial x_2} s_{x2}\right)^2 + \dots + \left(\frac{\partial f}{\partial x_n} s_{xn}\right)^2}$$



Example: Averaging of N measurements with equal standard deviations s_x

$$y = \frac{1}{N}(x_1 + x_2 + \dots + x_N) \quad \rightarrow \quad \frac{\partial f}{\partial x_1} = \frac{1}{N} \quad \dots \quad \frac{\partial f}{\partial x_N} = \frac{1}{N}$$

$$\rightarrow s_y = \sqrt{\left(\frac{s_x}{N}\right)^2 + \left(\frac{s_x}{N}\right)^2 + \dots + \left(\frac{s_x}{N}\right)^2} = \sqrt{N \frac{s_x^2}{N^2}} = \sqrt{\frac{s_x^2}{N}} \quad \rightarrow \quad s_y = \frac{s_x}{\sqrt{N}}$$

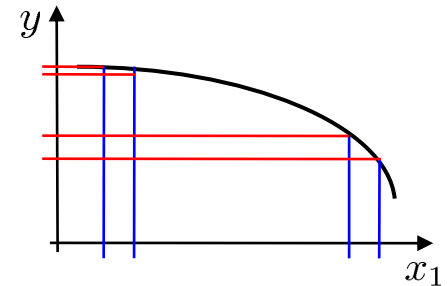
This is a universal statistical law! 100 times more measurement values improve the quality by a factor of 10 by reducing the standard deviation of the output y correspondingly.

5.3 Error Propagation

Approximation for *Random* Errors in Practice

Because it is difficult to estimate the standard deviations s_{x_i} for all quantities x_i , the following formula allows to assess the mean error of the output roughly (strictly speaking this formula is not exact):

$$\Delta y = \sqrt{\left(\frac{\delta f}{\delta x_1} \Delta x_1\right)^2 + \left(\frac{\delta f}{\delta x_2} \Delta x_2\right)^2 + \dots + \left(\frac{\delta f}{\delta x_n} \Delta x_n\right)^2}$$



The standard deviations s_{x_i} are approximated by $|\Delta x_i|$ roughly!

Difference of the effect of **systematic** and **random** errors

Systematic errors $\Delta x_i = \Delta x$, $i = 1, \dots, N$, add up:

$$y = x_1 + x_2 + \dots + x_N \rightarrow \Delta y = N \Delta x$$

Random errors $\Delta x_i = \Delta x$, $i = 1, \dots, N$, partly compensate each other:

$$y = x_1 + x_2 + \dots + x_N \rightarrow \Delta y = \sqrt{N} \Delta x$$

Therefore averaging yields benefits for **random** errors (smaller scattering)!

5.4 Histograms and Probability Density

Histograms

If we measure the same quantity N times under identical conditions, each outcome will be different due to random errors. In order to get an overview on the quality of the measurements and the size of the random errors, it makes sense to plot a **histogram**. This divides the measurements in intervals of size Δx . The number of measurement values that fall in the interval i are called **frequency** of the **observation** (German: **“absolute Häufigkeit”**) H_i . Each measurement falls in exactly one interval (with n_I intervals):

$$\sum_{i=1}^{n_I} H_i = N$$

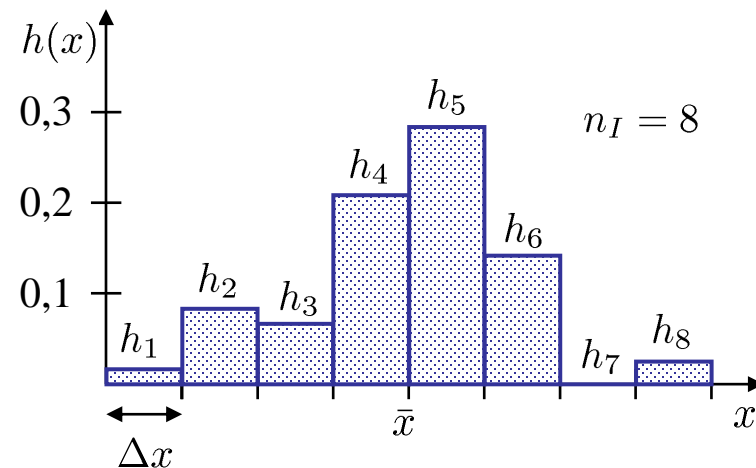
Recommendation for the number of intervals: $n_I \approx \sqrt{N}$

The relative value of H_i (Ger: **“relative Häufigkeit”**) h_i describes the fraction of H_i that falls into interval i :

$$h_i = \frac{H_i}{N}$$

The relative frequencies of observations sum up to 1:

$$\sum_{i=1}^{n_I} h_i = 1 = 100\%$$



5.4 Histograms and Probability Density

Probability Density Function (PDF)

With a histogram it is easy to see how the measurements are distributed, e.g. how strongly they scatter around their mean value \bar{x} . If we increase the number of measurements N and at the same time increase the resolution by making more intervals n_I smaller and smaller by decreasing Δx , then the histogram converges to the **probability density function (pdf)**:

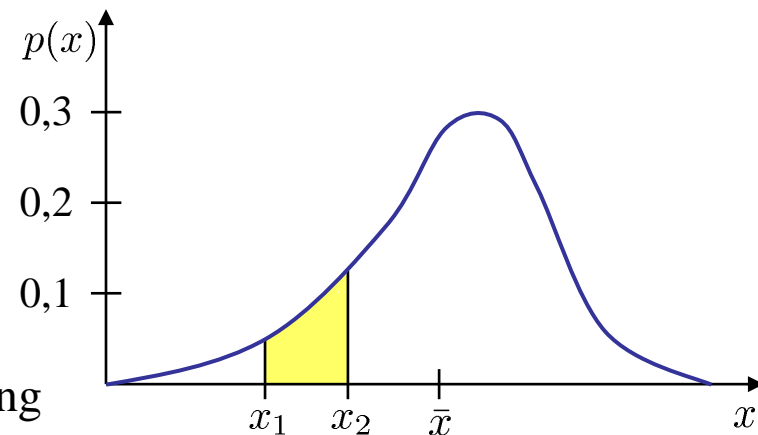
$$p(x) = \lim_{\Delta x \rightarrow 0} \left(\lim_{N \rightarrow \infty} h(x) \right)$$

$$\text{It is: } \int_{-\infty}^{\infty} p(x) dx = 1$$

The density $p(x)$ is a continuous and no stepwise function. We can calculate the **probability** of a measurement to fall into a certain interval $(x_1, x_2]$ by:

$$P(x_1 < x \leq x_2) = \int_{x_1}^{x_2} p(x) dx$$

The true density $p(x)$ according to which the measurements are distributed is usually unknown. Typically, realistic *assumptions* are made from insights in the first principles and a histogram. In most cases a **Gaussian** distribution is assumed if nothing contrary is known. Here is why... (see next slide)



5.4 Histograms and Probability Density

Normal Distribution (Gaussian)

A normal distribution with mean μ_x and variance σ_x^2 is defined as follows:

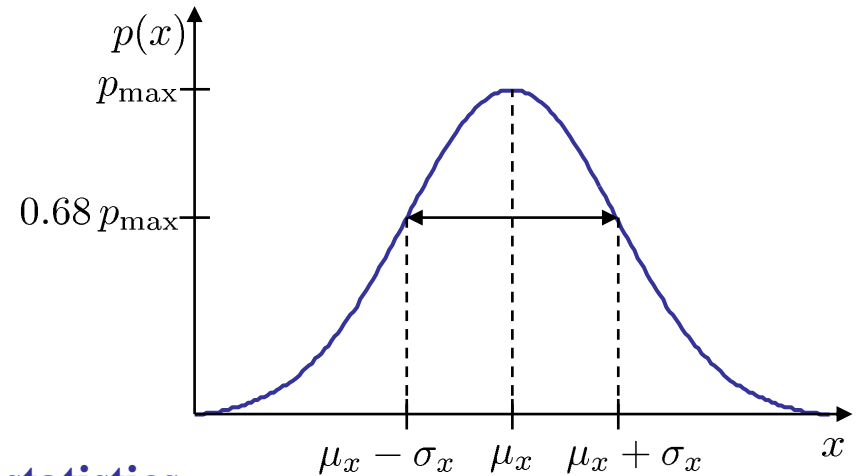
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2} \left(\frac{x-\mu_x}{\sigma_x} \right)^2}$$

It is of highest theoretical and practical importance. On the one hand, many other distributions can be approximated by the Gaussian (binomial-, t-/student distribution).

On the other hand, the **central limit theorem of statistics**

builds the key fundament for the essential normal distribution. It says that the sum of several independent random variables follows approximately a normal distribution. This is truly remarkable because it makes (almost, there are some minor exceptions) no restrictions on the distribution of each random variable!

In practice, most random errors are caused by many tiny effects that sum up. Therefore, almost all random errors are nearly Gaussian distributed. This explains why the Gaussian appears so often and is so well known.

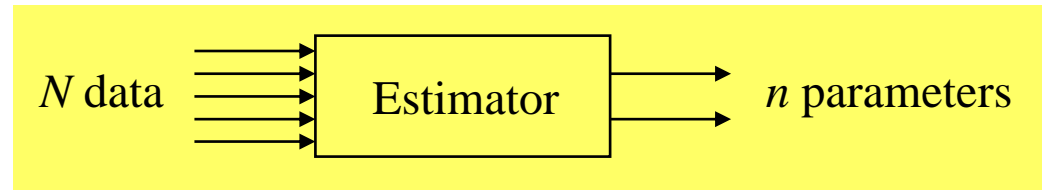


5.5 Estimation of Mean and Variance

Fundamentals of Estimation

An **estimation** in the statistical sense is the determination of one or many, in general n , quantities (parameters) by utilizing N measurement data. Typically the number of estimated parameters n is significantly smaller than the number of available data N :

$$n \ll N$$



Therefore an estimation can often be interpreted as a type of *data reduction* or *compression*. Common examples are the estimation of the:

- mean value of the measurement data ($n = 1$).
- standard deviation (scattering) of the measurement data ($n = 1$).
- auto- or cross-correlation function of a time signal ($n = \text{large}$).
- coefficients of a regression line ($n = 2$) or polynomial ($n = 3, \dots$).

The estimation results depend on the actual measurement data. If the same quantity is measured twice (even under identical conditions) we obtain different results and thus different *estimates*, because the random disturbances (noise) have different values.

5.5 Estimation of Mean and Variance

Properties of an Estimator: Variance

The estimation result depends on the random fluctuations of the disturbances which are modeled as random variables. Thus the estimation will yield different results for each data set. The estimation result is distributed according to an (unknown) probability density, e.g. an Gaussian normal distribution

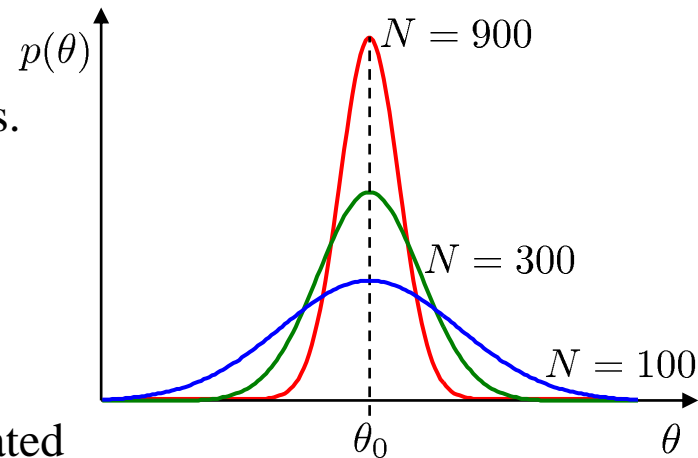
The quality of an estimation obviously is high if the estimated values are close to each other. This is the case, if the pdf is narrow, i.e. has a small variance. The smaller, the better.

A further demand on the properties of a good estimator is that the pdf becomes smaller the larger the amount of data N becomes. For many estimators indeed the variance follows the law:

$$\sigma_{\text{estimator}}^2 \sim \frac{1}{N}$$

$$\sigma_{\text{estimator}} \sim \frac{1}{\sqrt{N}}$$

A data set 4 times the size reduces the scattering by a factor of 2!



5.5 Estimation of Mean and Variance

$\hat{\theta}$: estimated parameter
 θ_0 : true parameter

Properties of an Estimator: Bias

In the previous slide it was assumed that the mean value of the pdf is identical to the true (but unknown) value θ_0 of the estimated parameters. If this is the case, the estimation is **without bias (unbiased)**:

$$E\{\hat{\theta}\} = \theta_0$$

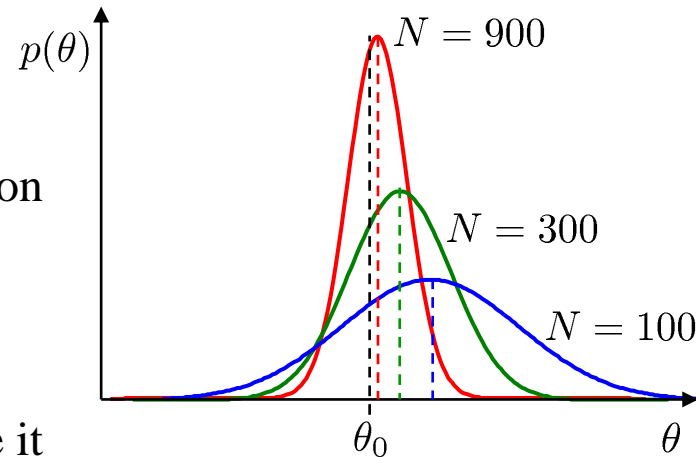
This is a desirable but not necessary property. Furthermore it is often traded for other advantages like a low variance!

If the estimation is not unbiased it possesses a **bias** (*systematic estimation error*) :

$$\text{Bias} = E\{\hat{\theta}\} - \theta_0$$

If the bias (and the variance) tend to 0 for $N \rightarrow \infty$, then we call this a **consistent estimation**:

$$\lim_{N \rightarrow \infty} \hat{\theta} = \theta_0$$



5.5 Estimation of Mean and Variance

Estimation of the Mean

Random errors can be reduced by averaging, i.e., calculating the mean value of several individual measurements. This is the simplest and most straightforward way to effectively lower scattering and noise influence. The estimation of the mean value thus plays an important role. We clearly distinguish between the true (but unknown) mean μ_x and the estimated mean value \bar{x} (also called **sample mean** or **empirical mean**).

$$\text{sample mean: } \bar{x} = \frac{1}{N} \sum_{i=1}^N x(i)$$

It can be shown that the sample mean approaches the true value (unbiased) if N becomes large

$$E\{\bar{x}\} = E\left\{\frac{1}{N} \sum_{i=1}^N x(i)\right\} = \frac{1}{N} \sum_{i=1}^N E\{x(i)\} = \frac{1}{N} \sum_{i=1}^N \mu_x = \frac{1}{N} N \mu_x = \mu_x$$

It can also be shown that for *statistically independent* data the variance of the sample mean estimation decreases for increasing data sets N , such as [4]:

$$\sigma_{\bar{x}}^2 = E\{(\bar{x} - \mu_x)^2\} = \sigma_x^2 / N \quad \text{bzw.} \quad \sigma_{\bar{x}} = \sigma_x / \sqrt{N}$$

5.5 Estimation of Mean and Variance

Estimation of the Variance

The variance σ_x^2 of the data is also an important quantity. It determines how widely the data is spread or scattered. The estimation of the data variance (**sample variance or empirical variance**) can be performed by:

$$\text{sample variance: } s_x^2 = \frac{1}{N-1} \sum_{i=1}^N [x(i) - \bar{x}]^2$$

μ_x is unknown!
 \bar{x} is its estimation!

The true mean μ_x is usually unknown and is replaced by its best estimate \bar{x} . Because of this the sum is divided by $N-1$ and not by N . One *degree of freedom (dof)* was already exploited or exhausted (figuratively speaking) for the estimation of this mean value and is not available anymore for the variance estimation. Only $N-1$ dof are remaining. It can also be shown theoretically that due to the denominator $N-1$ we have an **unbiased** estimation [4]:

$$E\{s_x^2\} = \frac{N}{N-1} (\sigma_x^2 - \sigma_{\bar{x}}^2) = \frac{N}{N-1} \left(\sigma_x^2 - \frac{\sigma_x^2}{N} \right) = \sigma_x^2 \rightarrow \text{unbiased!}$$

The *variance of an estimate* can be used for assessing the reliability of an estimate itself. It is required for example for determination of the **confidence intervals** that indicate the reliability of the estimate.

5.6 Confidence Intervals

Trust in a Measurement

A measurement or an estimated mean from many measurements is practically almost useless if its *reliability* is unknown. If its reliability is low then we cannot trust any information. Different information sources can be obtained with different reliabilities. A prerequisite for sensor fusion, for example, is some knowledge about their reliability. How can we quantify this?

Confidence Interval

The *trust* or *confidence* in an estimate can be quantified based on its probability density function (pdf). The pdf allows to calculate the probability that the true value lies within some interval. Typically a symmetric interval around the mean is considered. Most pdfs also have their maximal value at their mean. The probability that the deviation from the mean is smaller than $\pm\delta$ is:

$$P(\mu - \delta < x \leq \mu + \delta) = \int_{\mu - \delta}^{\mu + \delta} p(x) dx = ?\%$$

For any interval size (width) δ we can calculate the associated probability. It is called a **confidence interval**.

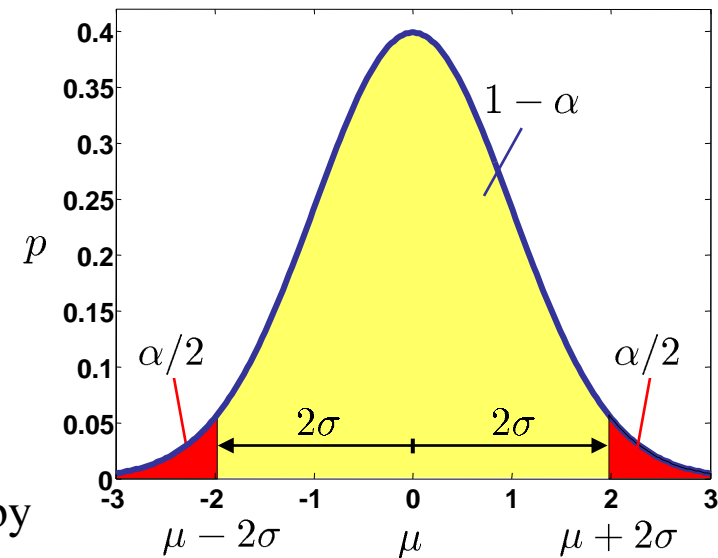
5.6 Confidence Intervals

Confidence Interval for Normal Distributions

The “width” of a pdf is determined by its standard deviation. Therefore it makes sense to measure the width of confidence intervals $\pm\delta$ in terms multiples of the standard deviation. For normal distributions the following confidence intervals are common:

Interval	Probability ($1-\alpha$)
$\mu_x - 1\sigma_x < x \leq \mu_x + 1\sigma_x$	68,27%
$\mu_x - 2\sigma_x < x \leq \mu_x + 2\sigma_x$	95,45%
$\mu_x - 3\sigma_x < x \leq \mu_x + 3\sigma_x$	99,73%
$\mu_x - 4\sigma_x < x \leq \mu_x + 4\sigma_x$	99,99%

The associated probability values $1-\alpha$ are called **confidence levels**. The **probability of error** is denoted by α and typically chosen as a small value like 5%, 1%, or even 0.1%. The less risk can be accepted the more multiples of the standard deviation must be accounted for. Such considerations are also part of any quality control system where error rates like 1 in 10.000 directly correspond to a multiple of σ .



5.6 Confidence Intervals

Decreasing the Standard Deviation

The quality of the estimator depends on the standard deviation that can be decreased by:

- *Improvement of the quality of the measurement:* Because we need to reduce *random* errors this is usually a complex and expensive task. Typical approaches are based on the isolation of environmental disturbances coming from temperature, air pressure, vibrations, radiation, etc.
- *Averaging over many measurements:* This is the typical approach to reduce random errors. The measurement is carried out several times and its average result is utilized. We know already that calculating the mean of N measurement values reduces the original standard deviation of the individual measurements σ_x as follows:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

This means it is possible, in principle, to decrease the standard deviation of the mean to an arbitrary accuracy. We just have to measure often enough! To double the accuracy we have to measure 4 times as many values. At the end, this is just a matter of cost and time.

5.6 Confidence Intervals

Confidence Intervals for Sample Mean With *Known* Standard Deviations

For random variables following a normal distribution, the confidence interval is

$$\bar{x} - c\sigma_x < x \leq \bar{x} + c\sigma_x$$

where the factor c corresponds to the requested *confidence level* $1-\alpha$ or *error probability* α , e.g. $c = 3$ for a confidence level of 99,73%.

Instead of measuring the value x a single time, the mean \bar{x} can be calculated from N measurements. Then we replace x with \bar{x} and its standard deviation decreases according to $1/\sqrt{N}$:

$$\bar{x} - c \frac{\sigma_x}{\sqrt{N}} < \bar{x} \leq \bar{x} + c \frac{\sigma_x}{\sqrt{N}}$$

$c = 1$: 68,27% confidence interval

$c = 2$: 95,45% confidence interval

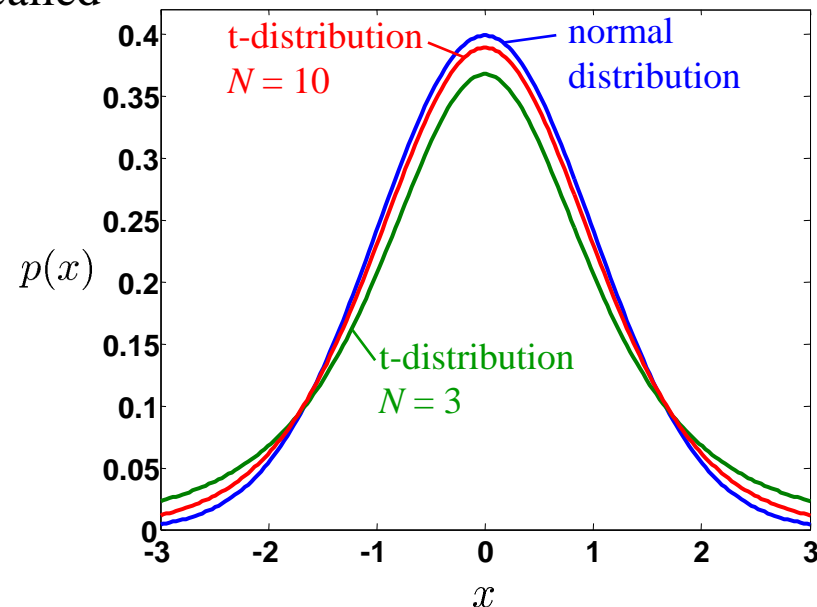
$c = 3$: 99,73% confidence interval

But this formula typically cannot be applied directly because the standard deviation σ_x is unknown. The next best thing to do, is to approximate it with the square root of the estimated sample variance s_x^2 . However, by using this approximation we make an (usually tiny) error.

5.6 Confidence Intervals

Confidence Intervals for Sample Mean With *Unknown* Standard Deviations

Because the estimated sample mean s_x is only an approximated value of the (unknown) true standard deviation σ_x the original confidence interval discussed above is not *exactly* accurate. In order to take this uncertainty into account the formula for the confidence interval has to be corrected. This can be done by replacing the normal distribution by the slightly wider **Student's t-distribution**. The t-distribution accounts for the additional uncertainty caused by the possible estimation error of the estimated instead of the true standard deviation. It thus depends on the number of measurements N , the so-called **degrees of freedom (dof)**. If the data set is huge ($N \rightarrow \infty$), the estimation error for s_x tends to zero, then Student's t-distribution converges to the normal distribution. However, for only a few measurements it becomes fatter at the outside making room for more uncertainty (*fat tail!*). This yields wider confidence intervals.



5.6 Confidence Intervals

Confidence Intervals for Sample Mean With *Unknown* Standard Deviations

For random variables that follow a t-distribution the formula for the confidence interval is basically unchanged:

$$\bar{x} - c \frac{s_x}{\sqrt{N}} < \bar{x} \leq \bar{x} + c \frac{s_x}{\sqrt{N}}$$

estimate for σ_x

but the factor c is larger than for a normal distribution (see table). For large N the factor c is hardly changed. But for small data sets (small N) it becomes significantly bigger.

The standard deviation is not known like for the normal distribution but estimated as follows:

$$s_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^N [x(i) - \bar{x}]^2}$$

= Gaussian distribution

Factor c for a t-distribution

N	$1-\alpha = 68,27\%$	$1-\alpha = 95,45\%$	$1-\alpha = 99,73\%$
5	1,11	2,65	5,51
10	1,05	2,28	3,96
20	1,03	2,13	3,42
50	1,01	2,05	3,16
100	1,00	2,03	3,08
200	1,00	2,01	3,04
∞	1,00	2,00	3,00

5.6 Confidence Intervals

Example: Confidence Intervals

A voltage meter yields measurement values that are corrupted by random errors. These errors come from an accumulation of many small disturbances which are not known in detail and whose sources are not studied. Therefore we can assume the overall error follows a normal distribution. From a long history of this voltage meter its behavior and accuracy are well known. The variance of the disturbance is determined to be $\sigma_x^2 = 0.01$ or $\sigma_x = \sqrt{0.01} = 0.1$.

a) The voltage meter displays: $U = 7$ V.

In which range will the true voltage be if we accept an error probability of maximal 0.3%?

→ Requested confidence level = 99.7%. For a normal distribution this corresponds to $c=3$.

$$\mu_x - c\sigma_x < x \leq \mu_x + c\sigma_x \quad \rightarrow \quad (7 - 3 \cdot 0.1) \text{ V} < x \leq (7 + 3 \cdot 0.1) \text{ V}$$

$$\rightarrow \boxed{6.7 \text{ V} < x \leq 7.3 \text{ V}}$$

The formula for *known* standard deviation is used, i.e., the confidence interval is calculated from the normal distribution because the standard deviation is well-known from a previous history of the instrument. (Or we assume $N \rightarrow \infty$ for the estimate).

5.6 Confidence Intervals

Example: Confidence Intervals

- b) The results in example a) does not fulfill our accuracy requirements. Therefore we decide to carry out 10 separate measurements and calculate its mean (average). This should get us closer to the true value than the above interval:

U [V]: 7.1 7.0 7.2 6.7 6.9 7.0 6.6 7.2 7.1 7.1

$$\text{Sample mean: } \bar{x} = \frac{1}{10} \sum_{i=1}^{10} x(i) = \frac{69.9 \text{ V}}{10} = 6.99 \text{ V}$$

$$\text{Standard deviation of the sample mean: } \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{10}} = \frac{0.1}{3.16} = 0.0316$$

$$\rightarrow (6.99 - 3 \cdot 0.0316) \text{ V} < x \leq (6.99 + 3 \cdot 0.0316) \text{ V}$$

$$\rightarrow \boxed{6.895 \text{ V} < x \leq 7.085 \text{ V}}$$

This result is more accurate by a factor of 3.16 for the same error probability of 0.3%. Even more measurement would improve the accuracy further.

5.6 Confidence Intervals

Example: Confidence Intervals

- c) We repeat the experimental setup from b) with a new instrument because the old one is broken. Thus a long history of the instrument's accuracy is not available. We do not know (as before) that the variance is 0.01. Therefore we have to estimate the instrument's accuracy by calculating the standard deviation of the 10 measurement values

$$\text{Sample standard deviation of the measurements: } s_x = \sqrt{\frac{1}{9} \sum_{i=1}^{10} [x(i) - 6.99]^2} = 0.2$$

$$\text{Sample standard deviation of the mean: } s_{\bar{x}} = \frac{s_x}{\sqrt{10}} = \frac{0.2}{3.16} = 0.0632$$

Factor c for the t-distribution with the confidence level of $1-\alpha = 99.7\%$: $c = 3.96$

$$\rightarrow (6.99 - 3.96 \cdot 0.0632) \text{ V} < x \leq (6.99 + 3.96 \cdot 0.0632) \text{ V} \quad \rightarrow \boxed{6.734 \text{ V} < x \leq 7.240 \text{ V}}$$

The larger interval range has two reasons:

- (i) factor 2 bigger standard deviations of the measurements (instrument is worse),
- (ii) factor 1.32 ($3.96/3$) bigger c -factor, because we need the t- not the normal distribution due to only estimated instrument quality.

5.6 Confidence Intervals

„Six Sigma (6σ)“ Quality Management System

This quality control management system was introduced in the mid 1980s by *Motorola* and since then has been adopted by many companies. It became particularly famous due to the introduction within *General Electric (GE)* by its CEO *Jack Welch* who made it a great success and the name “Six Sigma” became quite well-known.

The idea of Six Sigma is to reduce tolerances in a way, that the short term standard deviation becomes so small that the failure rate corresponds only to $6\sigma =$ quality of 1 ppb (parts per billion). According to expert knowledge, long term influences (mean changes slowly over time due to wear etc.) already cause approximately $\pm 1,5\sigma$. Thus the final quality will be in the range of $4,5\sigma =$ quality of 3,4 ppm (parts per million).

The implementation of “Six Sigma” is not only done in manufacturing. Rather *all* areas of a company are required to deliver a high quality level. An important feature of “Six Sigma” is an inherent **feedback control**. Quality is permanently measured and deviations from the required numbers cause *control actions*. The five main steps in “Six Sigma” are:

Define. Measure. Analyze. Improve. Control. (DMAIC).

The statistic evaluation plays an important role in “Six Sigma”.

6. Static and Dynamic Behavior of Sensors

Contents of Chapter 6

6. Static and Dynamic Behavior of Sensors

6.1 Overview

6.2 Static Behavior of Sensors

6.3 Dynamic Behavior of Sensors

6.1 Overview

Measurement errors have their reasons commonly in one or more of the following issues:

1. Nonlinear static characteristics of the instrument.
2. Dynamic transfer behavior of the instrument.
3. Noise superposes the desired signal.

Against these error sources counter measures can be taken that eliminate or at least reduce the error:

1. Compensation of the nonlinear distortion.
2. Compensation of the dynamic lag or waiting for the signal to settle (dynamics has faded).
3. Filtering to suppress noise.

Even if these counter measures are not completely successful or sufficient it is important to understand their effects. Only this allows one to assess the errors appropriately.

6.2 Static Behavior of Sensors

Linear Characteristics

The static characteristics between the input x and the output y can be described by a function:

$$y = f(x)$$

In sensorics we are primarily interested in the relationship between a measured quantity x , e.g. temperature, pressure, or displacement, and the yielded or displayed output y of the instrument, e.g. a voltage between 0V and 10V.

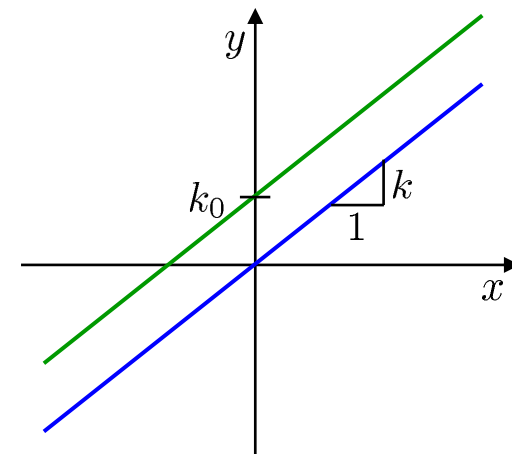
In the *ideal case*, this characteristics is **linear**, i.e., it exists a proportional relationship between input and output:

$$y = k x$$

For converting between input and output (or back) only the proportionality constant k is necessary. It is independent of the *operating point (OP)*. This is also true for the almost as simple **affine** relationship that includes an additional offset:

$$y = k x + k_0$$

By a simple transformation of the axis $\tilde{y} = y - k_0$ it can be transformed in the linear form $\tilde{y} = k x$



6.2 Static Behavior of Sensors

Advantages of a Linear (Affine) Characteristics

- Easy to understand and to handle.
- Described by one (two) parameters: k (and k_0).
- Identical sensitivities (slopes) in all operating points.

Life and Dead Zero

In measurement techniques the representation of the *origin* is practically important:

- **Dead Zero:** If the output $y = f(x) = 0$ for $x = 0$, i.e., the characteristics goes exactly through the origin of the coordinate system, as it is the case for *linear* systems.
- **Life zero:** If the output $y = f(x) \neq 0$ for $x = 0$, i.e., the characteristics does *not* go exactly through the origin of the coordinate system, as it is the case for *affine* systems.

A life zero offers an important practical advantage. It allows to distinguish between a zero measurement $x = 0$ with $y = k_0$ and a disconnection or other wire breakage ($y = 0$).

6.2 Static Behavior of Sensors

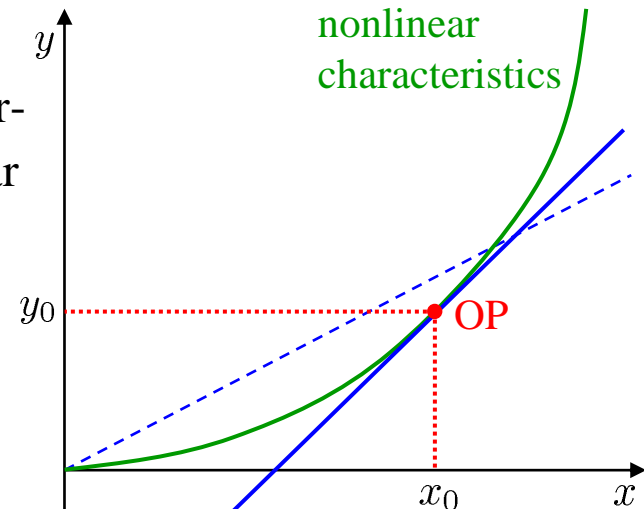
Linearization

In reality every instrument will possess a *nonlinear* characteristics. It is possible to *approximate* this relationship by linear or affine characteristics. Two alternative approaches exist:

1. *Global approximation*: The complete nonlinear characteristics in the *whole range* is approximated by a line (blue dashed).
2. *Linearization around an operating point (OP)*: The nonlinear characteristics in a *small range* around some *operating point (OP)* is approximated by a line (blue solid). Such an approximation is superior to the first approach as long the systems stays close to the OP (x_0, y_0) . Each OP requires an individual line since the slope and offset depends on the OP. The line follows the equation:

$$y = \left. \frac{dy}{dx} \right|_{x_0} \cdot (x - x_0) + y_0$$

Method 2 is better, if x changes slowly and it is possible to adjust the line as the OP changes. If the behavior is rapidly time-variant the 1. method might be better.



6.2 Static Behavior of Sensors

Sensitivity

The sensitivity S of an instrument is determined by the slope of its characteristics in the considered OP:

$$S = \left. \frac{dy}{dx} \right|_{x_0}$$

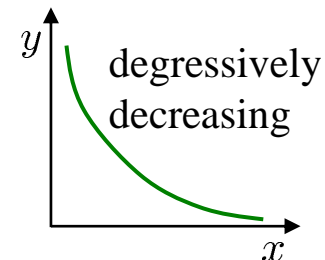
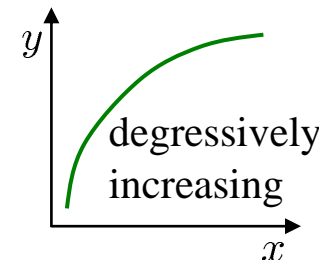
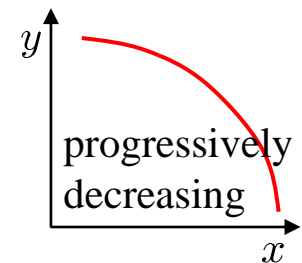
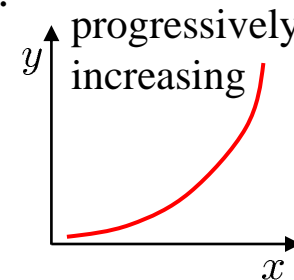
If the sensitivity is *low* a change in the measured value x hardly affects the output y of the instrument!

In general, the sensitivity of a nonlinear characteristics is operating point dependent, i.e., $S = S(x_0)$. For linear or affine characteristics the sensitivity is constant over the whole operating range because the slope never changes, i.e., $S = k$.

Common nonlinear characteristics possess a monotonically increasing or decreasing **sensitivity** (in absolute value).

The first is called **progressive**, the latter behavior is called **degressive**.

Of course, more complicated characteristics with inflection point(s) are possible as well. But the four main characteristics to the right cover at least 90% of all cases.



6.2 Static Behavior of Sensors

Compensation of Nonlinear Behavior

If the nonlinear characteristics of a sensor is known (from manufacturer's data or thorough measurements) it can be compensated at least partially. Two alternative exist:

- *Differential principle*: This is a popular approach for inductive and capacitive sensors and utilizes a bridge circuit. The nonlinearity often cannot fully be compensated but the approximation is commonly of high quality.
- *Inversion of characteristics*: By connecting the sensors and its inverted static characteristics in series theoretically both cancel each other. Theoretically, this is possible if the characteristics is strictly monotonous.

However, practical problems occur if the sensitivity is extremely small or large. The latter implies that the sensitivity of the inverted characteristics is extremely small.

This is also a standard method in control. Smart sensors commonly include such a compensation as well. Together with such a compensation they offer (almost) linear behavior which makes it very user friendly.

6.2 Static Behavior of Sensors

Compensation Via Difference Calculation

The key idea is to calculate the difference between two signals that are caused by counter-acting (e.g. opposite) effects. For inductive (or capacitive) displacement sensors e.g. one signal shows a positive and the other a negative influence. Calculating the difference yields:

$$y_1 = f(x) \quad y_2 = f(-x) \quad \rightarrow \quad y_d = y_1 - y_2 = f(x) - f(-x)$$

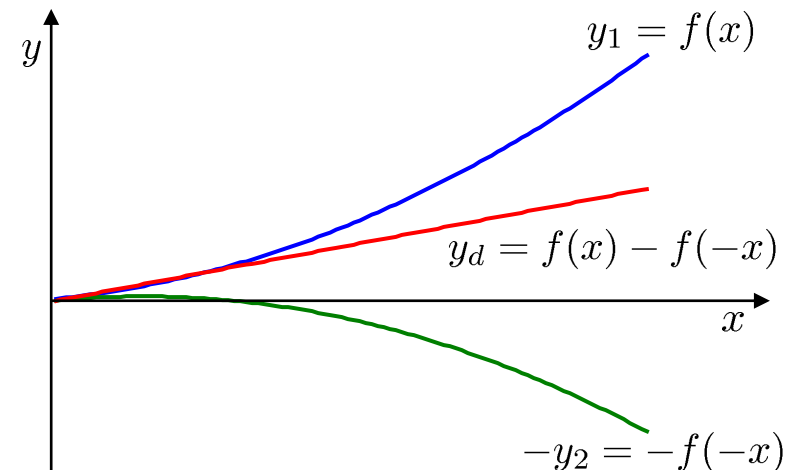
From a Taylor series expansion of the function f that gives

$$f(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + \dots,$$

we recognize the quadratic terms (and all terms of even powers) are eliminated in the difference calculation:

$$y_d = 2c_1x + 2c_3x^3 + \dots$$

By eliminating the quadratic terms the characteristics between x and y_d become more close to linear in a wider range. For all purely quadratic relationship the difference even yields an exact linear characteristics.



6.2 Static Behavior of Sensors

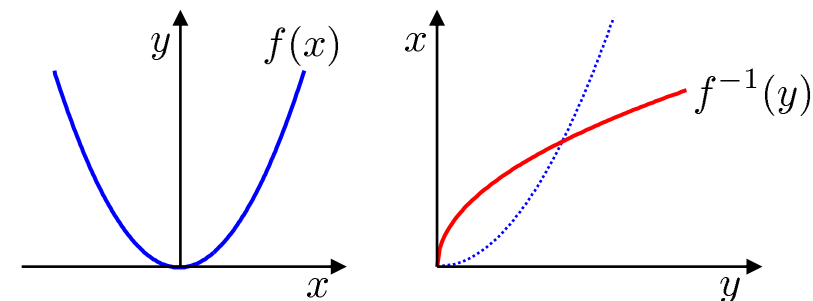
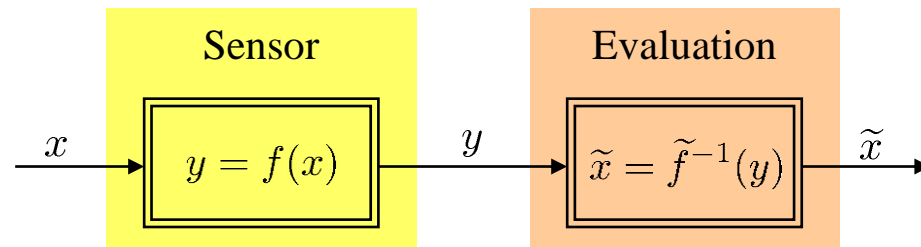
Compensation Via Inversion

The key idea is to isolated x as a function of y (inversion):

$$y = f(x) \quad \rightarrow \quad x = f^{-1}(y)$$

The inverse function only exists if $f(x)$ is *biuniquely*, i.e., if for every y from the physically reasonable range, *exactly one* x exists. If $f(x)$ does *not* fulfill this property (most will do) then the inversion can be carried out in intervals in which this property holds. By such an inversion, the electronics can compensate for all (at least most) nonlinearities in the sensor. The “~“ shall indicate that an exact inversion is never possible in practice.

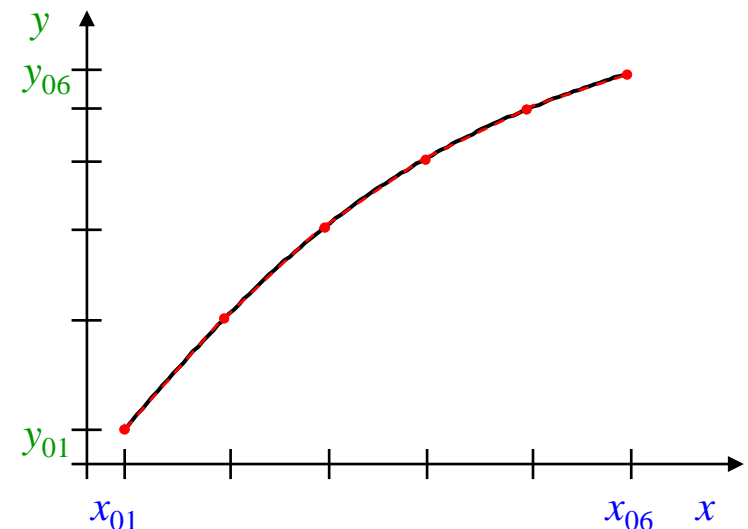
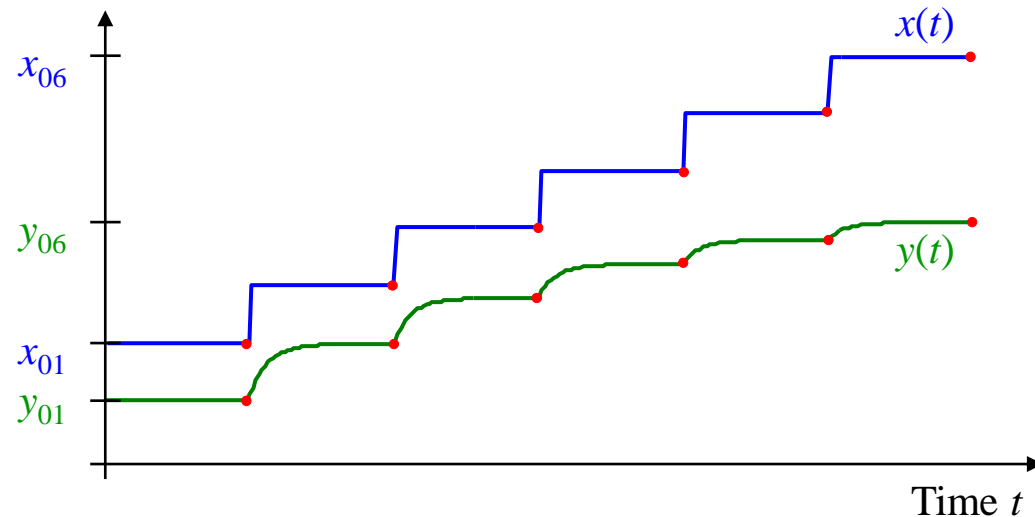
A prerequisite for an inversion is that the function $f(x)$ is known accurately. Special care is necessary for very small or large (where the inverse is very small) sensitivities because tiny errors cause huge deviations.



6.2 Static Behavior of Sensors

Determination of the Static Characteristics

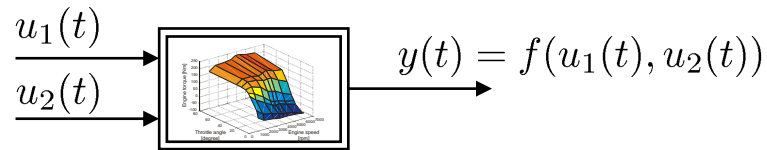
- The input signal must be held constant long enough that the output signal has time to settle. Then one point on the x - y -characteristics can be read out.
→ Time required for measuring through the entire characteristics is high!
- Characteristics typically are stored in a *look-up table* with linear interpolation (red dashed). Alternatives: Polynomials, neural networks, ...
- Characteristics for more than 1 input are called *characteristic maps*. They are commonly measured on a **grid**, e.g. 8×8 combinations for 2 inputs.



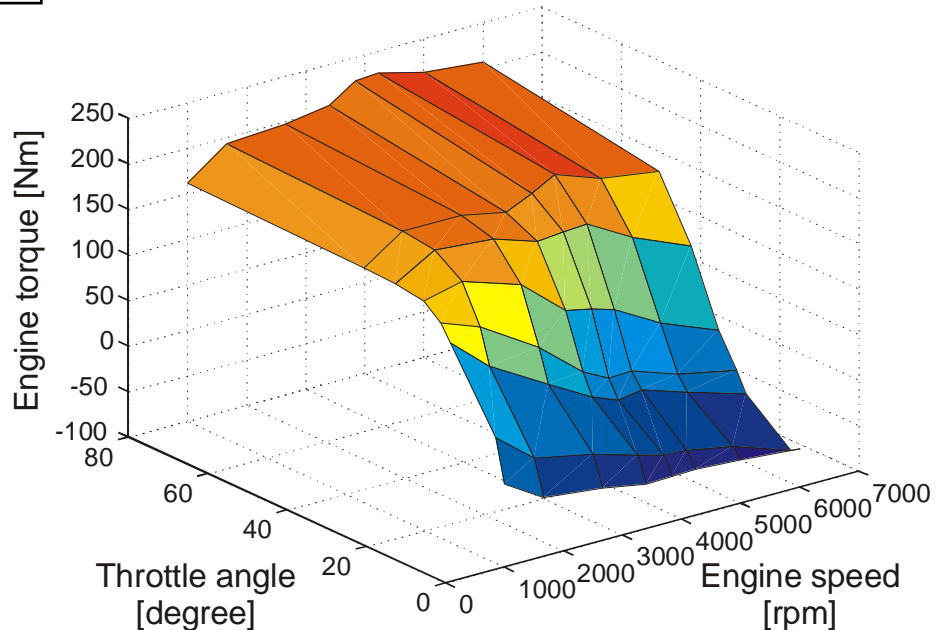
6.2 Static Behavior of Sensors

Characteristics in Lookup Tables

If a quantity depends in a nonlinear way on *several* other quantities, a **characteristic map** is required to describe such a behavior. For more than 2 input dimensions, however, only slices can be graphically illustrated. Therefore a 2-D example:



A typically characteristic map out of an automotive area: The control of combustion engines. The engine torque depends decisively on the engine speed and the throttle angle (for gasoline engines) or injection mass (for Diesel engines).



6.3 Dynamic Behavior of Sensors

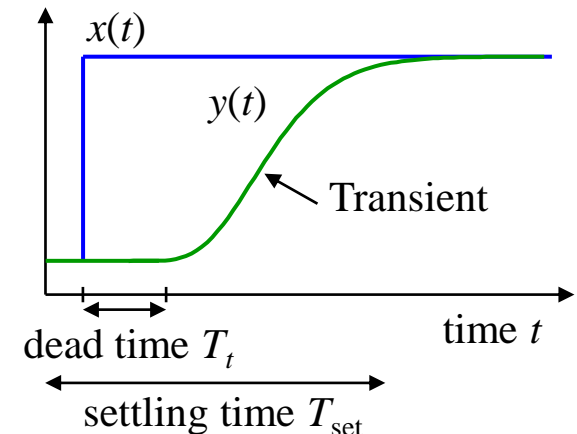
Dynamic Errors

The output y of an *ideal* instrument follows the input x instantaneously, i.e., without any time lag. In reality such an ideal behavior cannot be realized. Masses have to be accelerated, capacitors have to be charged, temperature must adjust, electric/magnetic fields have to build up, signals need to be processed. Such delays or lags cause a so-called **dynamic error**. Dynamic errors only show if the input signal *changes*. They are the higher, the faster these changes are. Examples for really fast input signals are impulses or steps.

To compare the dynamic behavior of sensors it makes sense to relate to a common scenario where the input changes step-wise and the deviation of the response y to a perfect step is measured. The response can be partitioned into 3 parts:

1. $0 \dots T_t$: $y(t)$ does not react at all.
2. $T_t \dots T_{set}$: $y(t)$ reacts.
3. $T_{set} \dots \infty$: $y(t)$ settles (almost) to its final value.

For Filters see Chapter 10



6.3 Dynamic Behavior of Sensors

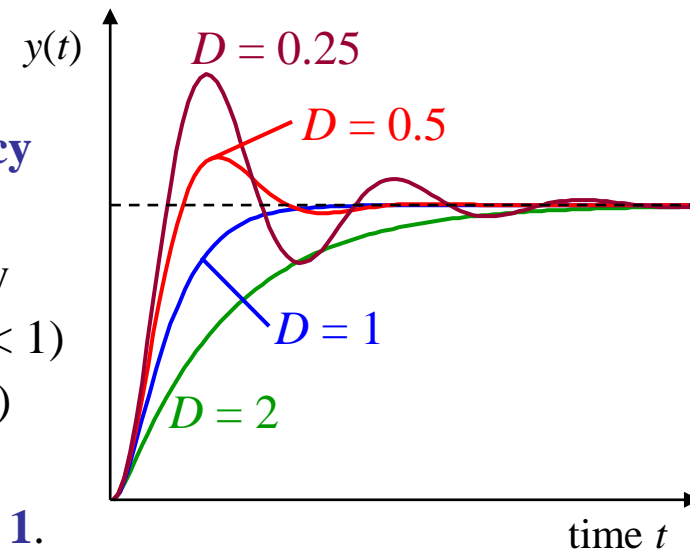
The smaller the dead time T_t and the settling time T_{set} are, the faster the sensor behaves and the smaller the dynamic error becomes. An ideal sensor has: $T_t = 0$ und $T_{\text{set}} = 0$ but of course this is not possible

Overshoot and Damping

Unfortunately the output is not always as nice with an asymptotic approach to its final value as shown in the last slide. Often the dynamic behavior (at least approximately) follows a differential equation of 2. order:

$$\ddot{y}(t) + 2D\omega_0\dot{y}(t) + \omega_0^2y(t) = x(t)$$

where D is the **damping** and ω_0 is the **resonance frequency** given by the physics of the sensor. The equation e.g. can describe a mass-spring-damper-system as it occurs in every instrument needle/pointer. If the damping D is too low ($D < 1$) oscillations will occur; if the damping D is too high ($D > 1$) the settling time will be too long. Therefore the best compromise is the so-called **aperiodic limit case** with $D = 1$.



6.3 Dynamic Behavior of Sensors

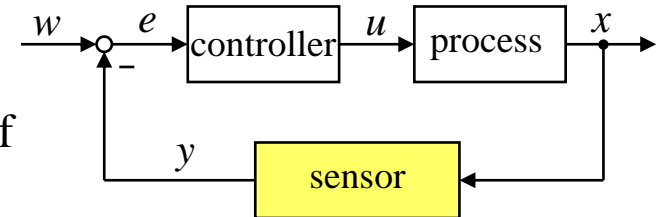
How to Avoid or Reduce Dynamic Errors?

1. Wait after a change in the measured quantity x until settling is reached after time period $T_t + T_{\text{set}}$ and then read output value y or process it further, respectively.
2. In a post-processing step the delayed and time-lagged output $y(t)$ is predicted into the future (non-causal system).
3. Reduce the time-lag in the dynamic error with dynamic filter with differential character. The price to be paid is a higher sensitivity to noise.

Method 1 and 2 can only work if the output y is not need at once! Method 1 additionally requires that the changes are step-wise and not continuous.

Method 1 and 2 thus cannot be used for feedback control systems! In feedback control it is crucial that the control variable x is fed back at once to the comparison with the desired value. The controller must act as quickly as possible with respect to deviations. Any additional delay will deteriorate the control performance.

That leaves us with method 3 where it is important to find a good trade-off between noise sensitivity and the reduction of dynamic errors.



2. Measurement of Electrical Quantities

Contents of Chapter 2

2. Measurement of Electrical Quantities

- 2.1 Moving Coil Mechanism
- 2.2 Measurement of Current
- 2.3 Measurement of Voltage
- 2.4 Measurement of Power and Energy
- 2.5 Measurement of AC Quantities
- 2.6 Measurement Methods and Amplifier Circuits

2.1 Moving Coil Mechanism

Why is the measurement of *electrical* quantities so important?

Electrical current possesses many advantages over alternative physical means to transport energy and information like with air pressure or hydraulics. Electricity is:

- Easy to measure with high efficiency.
- Easy and with high efficiency to transform to other quantities with motors (torque, speed), electric heating (heat) or air conditioning (coldness), lamp or LED (light).
- Well and easy to control.
- Efficiently to transport over long distances.
- Almost everywhere available.
- Standard means to transmit information.
- Easy to convert into digital signals and to process in a computer.

Because of these advantages electricity plays a dominant role in measuring things (**sensorics**) and manipulating things (**actuation**). At least the last part in sensorics and the first part in actuators is often of electrical nature to exploit the good controllability properties.

2.1 Moving Coil Mechanism

First Principles

A magnetic field of flux density B generates a force on a wire of length l that is orthogonal to the field and carries an electrical current I . The generated Lorentz force is calculated by

$$F = lBI$$

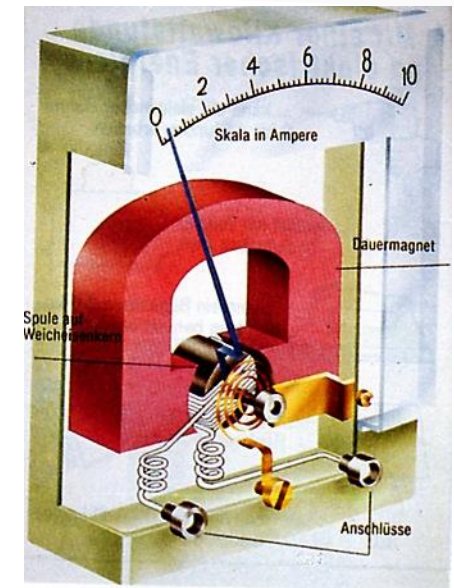
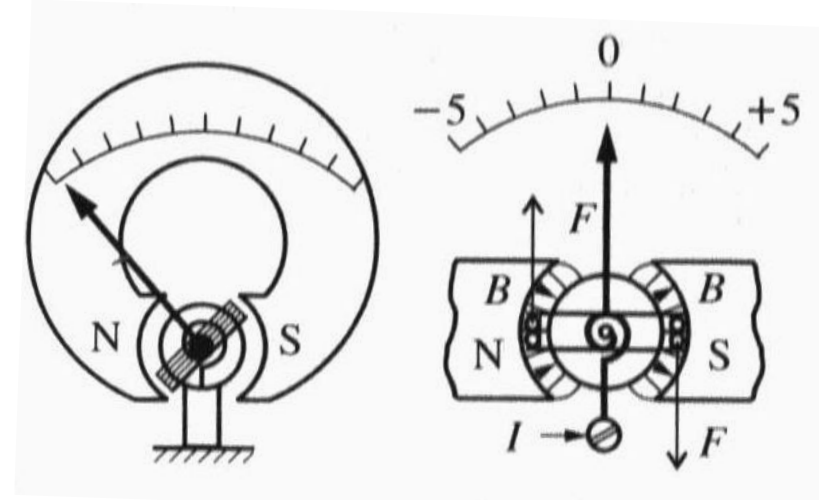
This force is proportional to the current and can be used to indicate its value. If this force is in balance with a spring, a pointer can display the size of the current.

More accurately, the force is generated in N windings of a coil. Because it acts on each side of the coil, the actual torque is twice this force times the distance r (diameter of the coil = $d = 2r$).

This gives the torque:

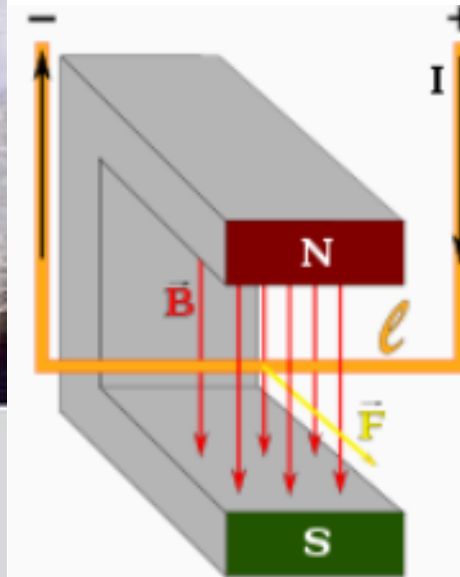
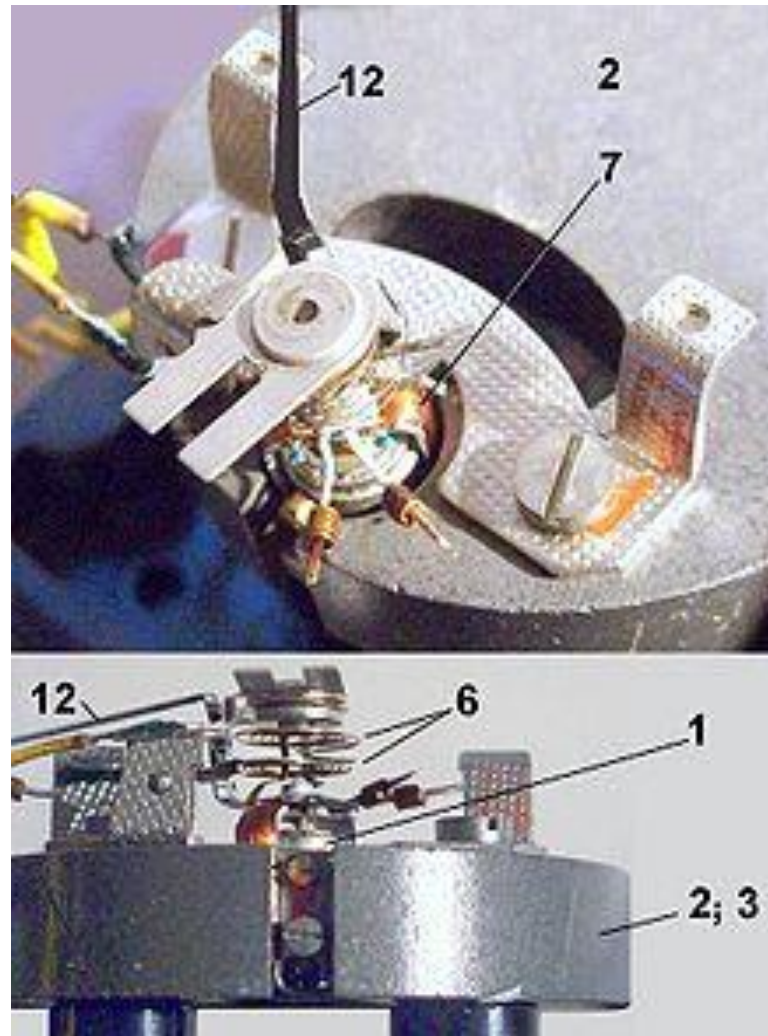
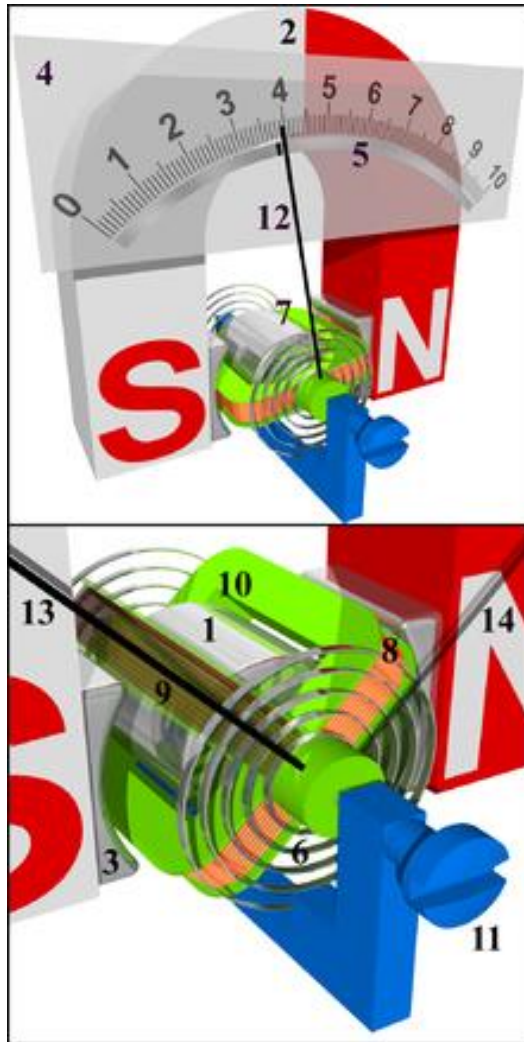
$$M = 2rNlBI = NdI$$

Acting on a torsion spring with torque $M = c \alpha$ results in a displayed angle α .



2.1 Moving Coil Mechanism

Source: <http://de.wikipedia.org/wiki/Drehspulmesswerk>



(1) Weicheisenkern, (2) Permanentmagnet, (3) Polschuhe, (4) Skale, (5) Spiegelskale, (6) Rückstellfeder, (7) Drehspule, (8) Ruhelage, (9) Maximalausschlag, (10) Spulenkörper, (11) Justierschraube, (12) Zeiger, (13) Südpol, (14) Nordpol

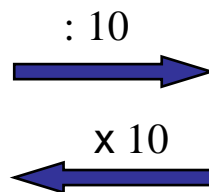
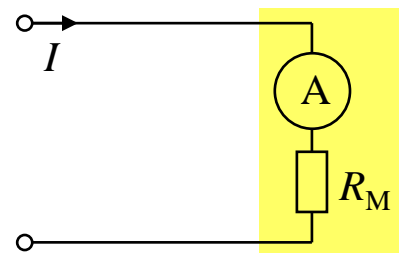
2.2 Measurement of Current

Some Facts on the Moving Coil Mechanism Meter

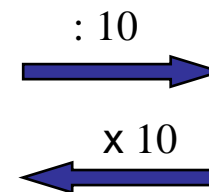
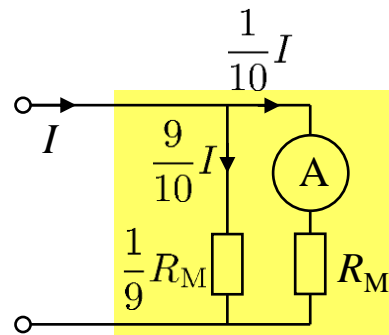
- Most *frequently* applied analog way to measure currents.
- Range: $10^{-6}\text{A} - 100\text{A}$. Accuracy: 0,1% – 1,5%. Settling Time: 0,5s – 1s.
- With resistors in parallel the *range* can be changed.
- By coupling it with an DC converter it can be used to measure an *AC* current.
- With an auxiliary resistor and Ohm's law, it can be used to measure *voltage*.
- Replacing the *permanent* magnet creating B by an electromagnet, the meter can be used for measurement of power.

Change of Range:

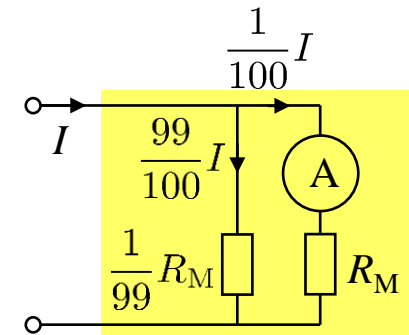
Internal Resistance: R_M



Internal Resistance: $10R_M$



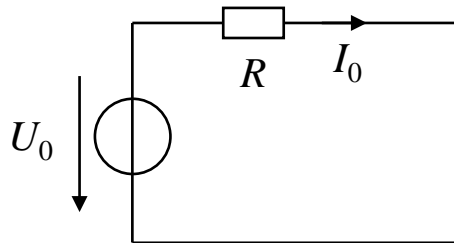
Internal Resistance: $100R_M$



2.2 Measurement of Current

Systematic Error of Current Measurement

Circuit without meter

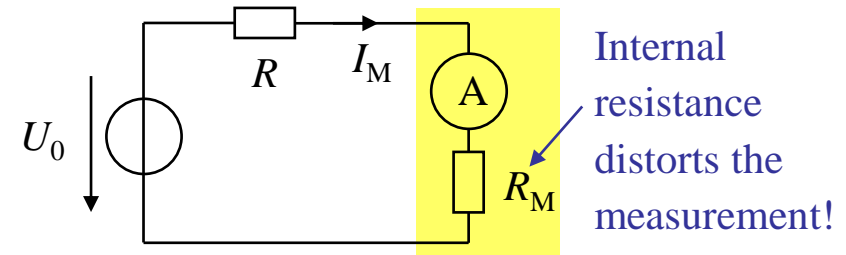


True current: $I_0 = \frac{U_0}{R}$

This leads to a relative error in the current measurement of:

$$\frac{I_0 - I_M}{I_0} = 1 - \frac{R}{R + R_M} = \frac{R_M}{R + R_M} \xrightarrow{\text{for } R_M \ll R} \approx \frac{R_M}{R} \xrightarrow{\text{for } R_M \rightarrow 0} 0$$

Circuit with meter



Measured current: $I_M = \frac{U_0}{R + R_M}$

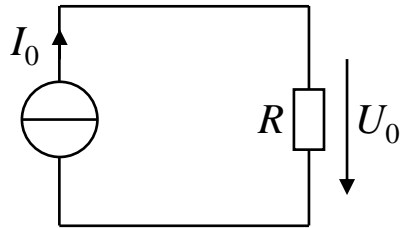
Current is measured too small!

Current meters should have an internal resistance as *small* as possible!

2.3 Measurement of Voltage

Using a Current Meter for Measuring a Voltage

Circuit without meter

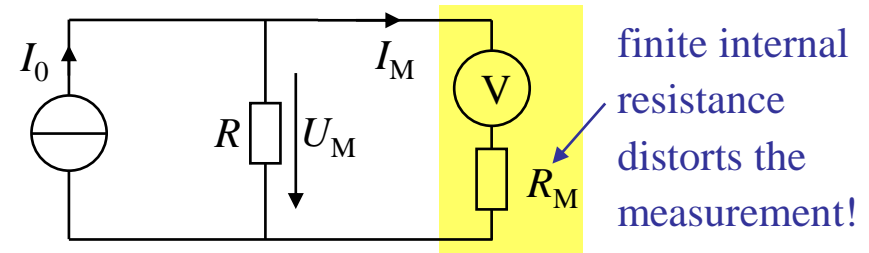


True voltage: $U_0 = RI_0$

This leads to a relative error in the voltage measurement of:

$$\frac{U_0 - U_M}{U_0} = 1 - \frac{R_M}{R + R_M} = \frac{R}{R + R_M} \xrightarrow{\text{for } R_M \gg R} \approx \frac{R}{R_M} \xrightarrow{\text{for } R_M \rightarrow \infty} 0$$

Circuit with meter



Measured voltage: $U_M = \frac{RR_M}{R + R_M} I_0$

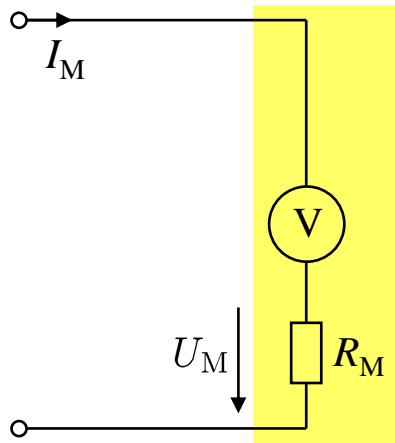
Voltage is measured too small!

Voltage meters should have an internal resistance as *large* as possible!

2.3 Measurement of Voltage

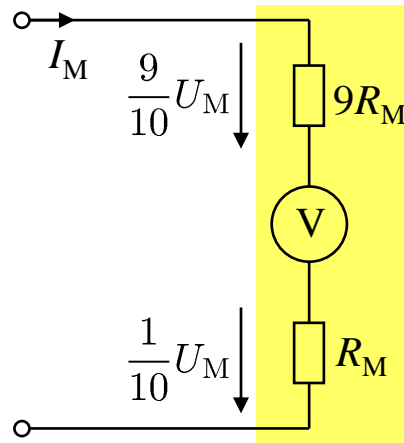
Change of Range

Internal Resistance: R_M



$: 10$

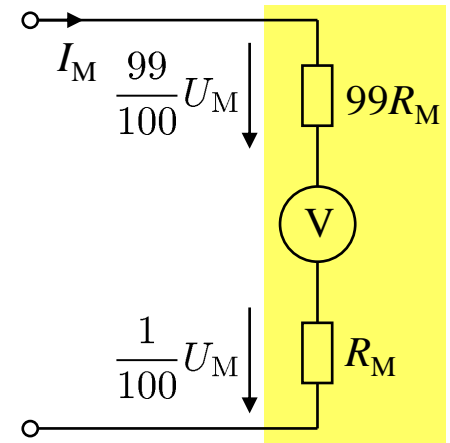
Internal Resistance: $10R_M$



$\times 10$

$: 10$

Internal Resistance: $100R_M$



$\times 10$

Nomenclature of Voltage Meters

The internal resistance is given in relation to the **upper range value** and in Ω / V .

E.g. “1 k Ω / V ” means:

- 100 k Ω internal resistance within the range 0...100 V.
- 10 k Ω internal resistance within the range 0...10 V, etc.

2.3 Measurement of Voltage

Considerations About the Systematic Errors in Current and Voltage Measurements:

- To reduce the deterioration in *current* measurements, we want to have a small internal resistance, in the ideal case $R_M = 0$.
- To reduce the deterioration in *voltage* measurements, we want to have a large internal resistance, in the ideal case $R_M = \infty$.
- The demand for a small internal resistance is much more difficult to fulfill than the demand for a large internal resistance, because
 - the coil of the moving coil mechanism naturally has a finite resistance, in particular if N is high,
 - also the connections/contacts where the meter is attached have a resistance,
 - amplifier circuits easily can generate a resistance close to $R_M = \infty$ (see Chapter 2.6).

These arguments show that a voltage measurement can be performed more accurately than a current measurement.

Therefore we can apply a trick to use voltage measurements for determining currents.

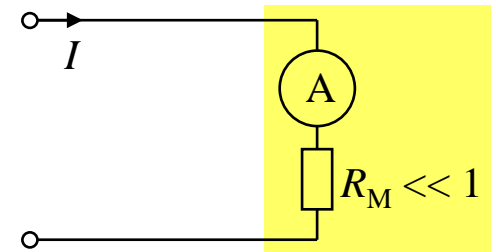
2.3 Measurement of Voltage

Indirect Current Measurement With a Shunt

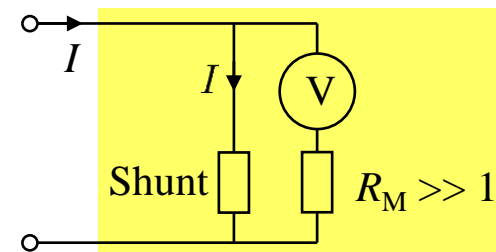
A **Shunt** is a measurement resistor that has been manufactured with care (expensive!) to ensure a low resistance with great accuracy almost independent of disturbing influences like temperature. The voltage drop over such a shunt is measured and by Ohm's law the flowing current is determined. Compared to a direct current measurement, which incorporates the meter in series within the circuit, the following advantages are obtained:

- The resistance of shunt is more accurate than the internal resistance of the meter.
→ Smaller measurement error.
- The resistance of the shunt can be chosen to be smaller than the internal resistance of the meter.
→ Smaller measurement error.
- The wires and connections to the meter lead to a voltage drop and are sources of measurement errors. Because the current through the voltage meter is tiny ($\ll I$), these are insignificant compared to the direct method.

Direct current measurement



Indirect current measurement via a shunt



2.4 Measurement of Power and Energy

First Principles

Electrical power is the product between voltage and current:

$$P = UI$$

Replacing the permanent magnet of the moving coil mechanism creating the magnetic field B by an electromagnet, constructs the **electrodynamic instrument**. It can measure power. If the electromagnet is fed with voltage U this creates a current and subsequently a magnetic field proportional to U :

$$B = kU$$

With the formula for the moving coil mechanism we obtain:

$$M = NdlBI = kNdlUI = kNdlP$$

The generated torque is proportional to the power P .

2.4 Measurement of Power and Energy

Measuring Electrical Energy

Measuring energy is based on the measurement of power. Energy is power integrated over time:

$$E = \int_0^t P(\tau) d\tau$$

If the power is constant over time, energy is simply power times time:

$$E = P t$$

Otherwise can be fed to an integration circuit (see Chapter 2.6) and be computed in an analog manner. Alternatively it can be measured (counted) by a **motor meter**. A motor meter basically is an induction measuring system (see Chapter 2.5) in which the electromagnets are replaced with an electromotor whose torque is proportional to the power. The number of revolutions of the disk is proportional to the energy.

2.5 Measurement of AC Quantities

Mean, Peak, Rectified, and Root Mean Square (RMS) Values

AC Quantities are periodic signals $x(t)$ with a period (cycle duration) of T . The following measures of “size” have to be distinguished:

Mean: $\bar{x} = \frac{1}{T} \int_0^T x(t) dt$ Peak: $\hat{x} = \max\{x(t)\}$

Rectified: $\overline{|x|} = \frac{1}{T} \int_0^T |x(t)| dt$ RMS: $X_{\text{eff}} = \sqrt{\frac{1}{T} \int_0^T x^2(t) dt}$

The by far most important periodic signal type is a sine or cosine signal. A sine oscillation with amplitude A has the following characteristic values:

Mean: $\bar{x} = 0$ Peak: $\hat{x} = A$
Rectified: $\overline{|x|} = \frac{2}{\pi} A = 0.637A$ RMS: $X_{\text{eff}} = \frac{1}{\sqrt{2}} A = 0.707A$

For a rectangular oscillation the mean, peak, rectified, and RMS values are all identical to its amplitude A . The rectified value is the mean of the absolute value. The RMS value is a measure for the signal power or energy.

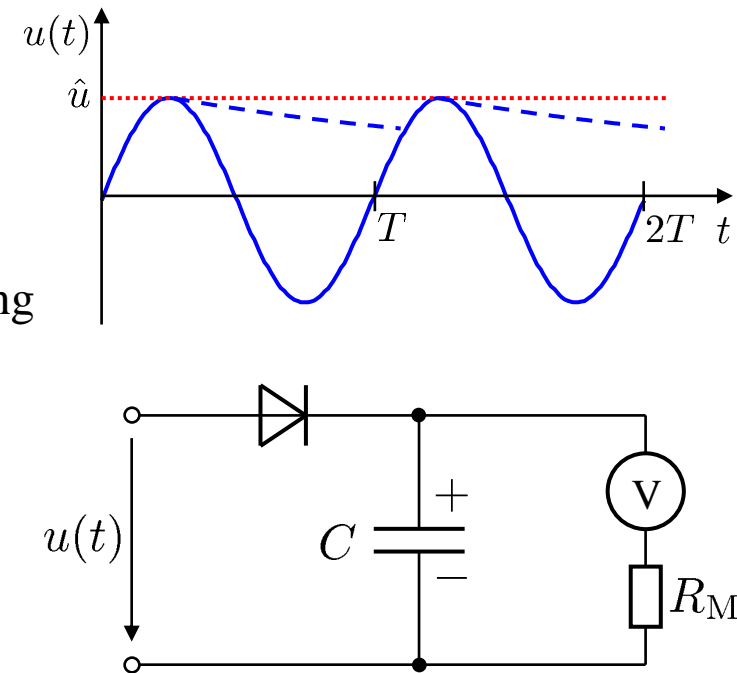
2.5 Measurement of AC Quantities

Measuring Mean Values

The mean value of an electrical AC quantity can be directly measured with a moving coil mechanism, if the frequency of the signal is high enough. Often occurring frequencies around 50 or 60 Hz (power net frequency) are so much higher than the bandwidth of the moving coil mechanism (around 1 Hz) that the instrument shows only the mean value. I.e., only the offset value of the AC signal is displayed.

Measuring Peak Values

A diode lets only the positive half part of an oscillation signal $u(t)$ pass. A capacitor C stores the highest occurring value of this voltage. Since the voltage meter has a very high internal resistance R_M , the capacitor will be hardly discharged (dashed line) before it is charged again at the next period T . A circuit manages to half the refresh times by an additional diode.



2.5 Measurement of AC Quantities

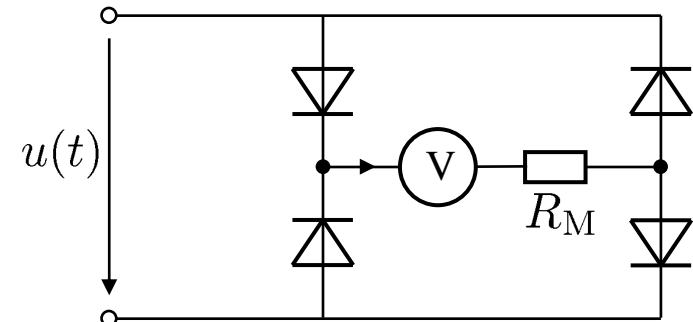
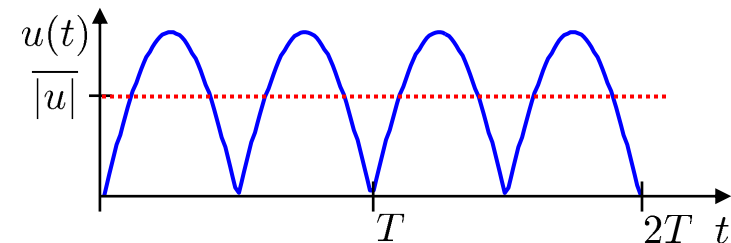
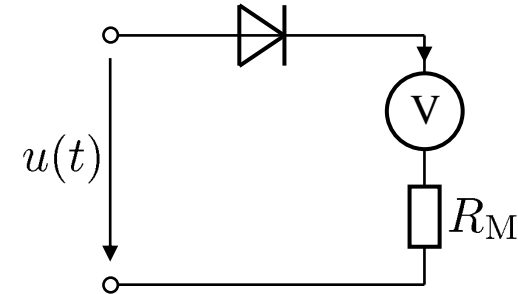
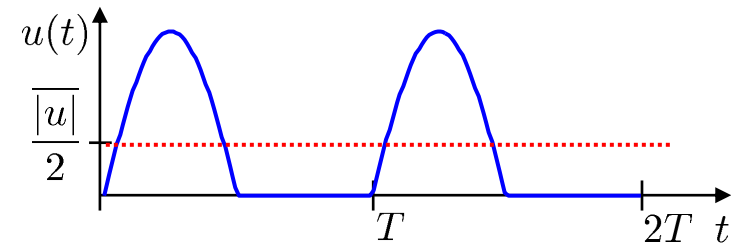
Measuring Rectified Values

The most straight forward way to rectify a signal is to let only the positive half of the oscillation pass by a diode. The negative halves are blocked. In contrast to its definition, this approach in the mean measures only $\frac{1}{2}$ of the rectified value. Therefore the result has to be multiplied by 2.

More advanced is the *Graetz* circuit which requires 4 diodes that manage to let the positive halves pass and let the negative halves pass in the other direction. Thus the full rectified value is determined.

Because for oscillations of sin type, the relation between the *rectified* value and both, the **peak** value and the **RMS** value are known, both values can be calculated from the rectified one:

$$\hat{u} = \frac{\pi}{2} \overline{|u|} = 1.571 \overline{|u|} \quad U_{\text{eff}} = \frac{\pi}{2\sqrt{2}} \overline{|u|} = 1.111 \overline{|u|}$$



2.5 Measurement of AC Quantities

Apparent, Active, and Reactive Power

In coils and capacitors where inductivity and capacity are the dominant factors, AC voltage and current are phase shifted by $+90^\circ$ and -90° , respectively. Thus, if not purely ohmic impedances are present, phase shifts φ between voltage and current have to be taken into account in any AC circuit in general. The **apparent power** P_S in such a impedance is simply the product between the RMS values (called “effective” in German) of voltage and current:

$$P_S = U_{\text{eff}} I_{\text{eff}}$$

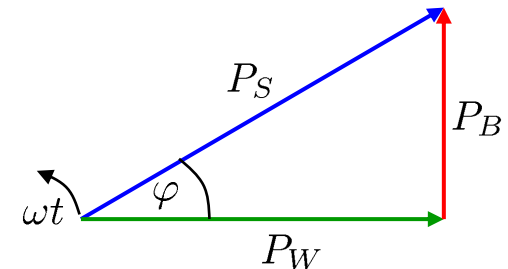
But the entire apparent power cannot perform work. One part of it just oscillates around the mean value 0. The really useful part of it is called **active power** (“Wirkleistung” in German). This part can perform work and is calculated by:

$$P_W = P_S \cos\varphi$$

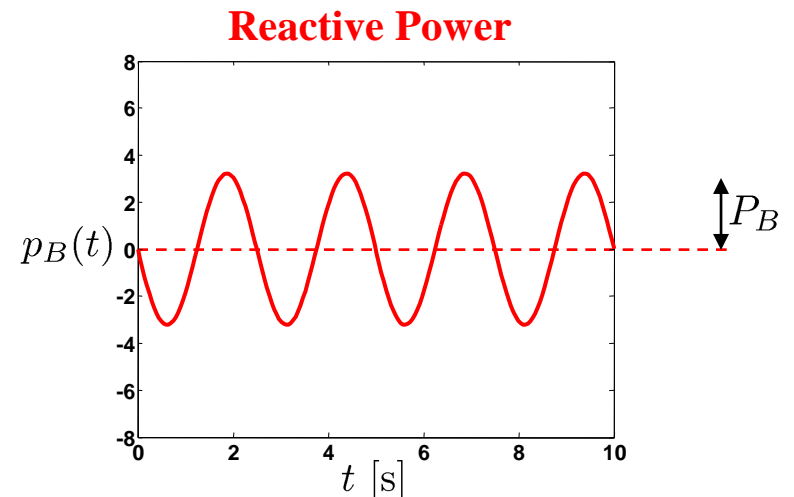
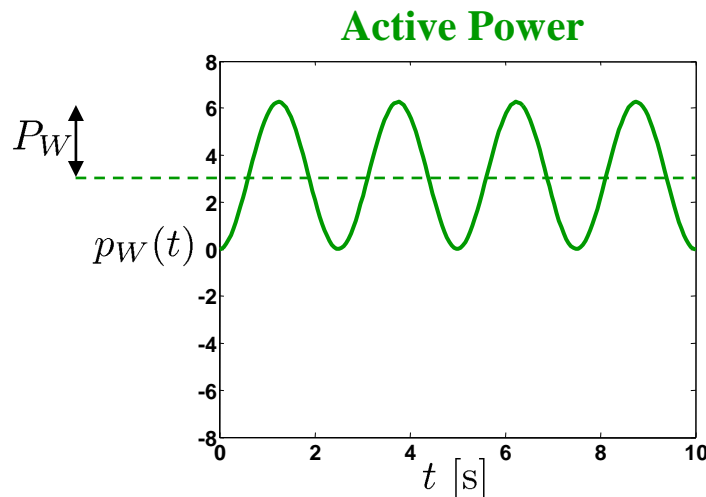
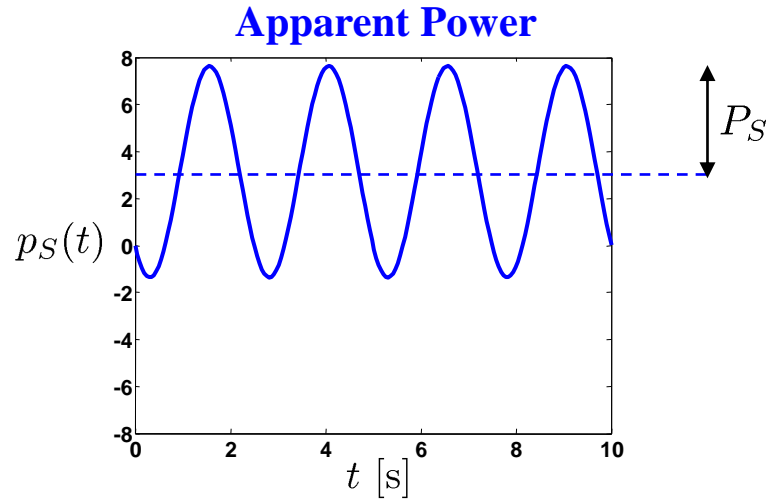
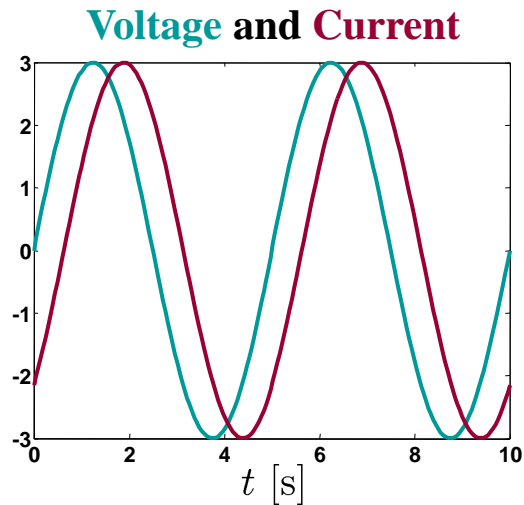
The part that cannot perform any work is called **reactive power** (“Blindleistung” in German) and calculated by:

$$P_B = P_S \sin\varphi$$

If voltage and current are not phase-shifted ($\varphi = 0$), then the reactive power = 0 and apparent power = active power.



2.5 Measurement of AC Quantities



2.5 Measurement of AC Quantities

Power Measurement

What happens if we measure an AC current with a moving coil mechanism instrument like a DC current?

$$u(t) = \hat{u} \sin(\omega t) \quad i(t) = \hat{i} \sin(\omega t + \varphi)$$

The displayed deflection is proportional to the product between voltage and current

$$p_S(t) = u(t)i(t) = \hat{u} \hat{i} \sin(\omega t) \sin(\omega t + \varphi) = \frac{1}{2} \hat{u} \hat{i} [\cos\varphi - \cos(2\omega t + \varphi)]$$

The 2. cos term is averaged out to 0, because we can assume a high frequency of AC quantities (e.g. 50 Hz) compared to the bandwidth of the instrument (around 1 Hz). This gives the mean value of the *apparent power* $p_S(t)$ which is identical to the mean of the amplitude of the **active power**:

$$\frac{1}{2} \hat{u} \hat{i} \cos\varphi = U_{\text{eff}} I_{\text{eff}} \cos\varphi = P_W$$

The reactive power can be measured by shifting the voltage by -90° before feeding it to the instrument. The displayed value is proportional to the **reactive power**:

$$\frac{1}{2} \hat{u} \hat{i} \cos(\varphi - \pi/2) = U_{\text{eff}} I_{\text{eff}} \sin\varphi = P_B$$

2.5 Measurement of AC Quantities

Measuring the Apparent Power

One way to measure apparent power is to measure the RMS of voltage and current separately and subsequently multiply them:

$$P_S = U_{\text{eff}} I_{\text{eff}}$$

An alternative is to let this multiplication happen in a moving coil mechanism instrument by physical law. To do this, the instrument has to be fed with the rectified values of voltage and current. The scale must then consider the quadratic nature of the result and the conversion factor between rectified and RMS values.

Measuring the Phase Shift

There are instruments to measure the phase shift between voltage and current. If this is determined, the active and reactive powers can be calculated from the apparent power.

Besides these possibilities there are some tricky measurement circuits for three-phase systems that are beyond the scope of this chapter.

2.5 Measurement of AC Quantities

Energy Measurement

Because only active power can perform work, the energy (work) can be *calculated* by integration:

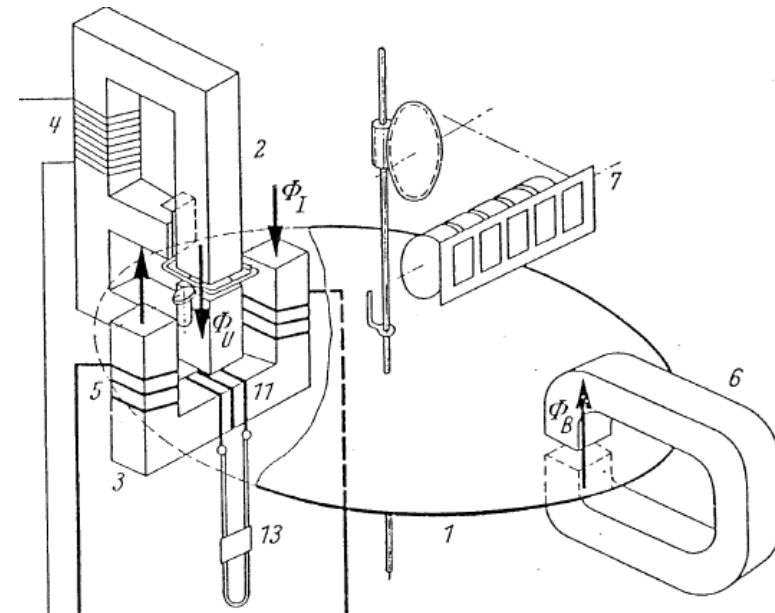
$$E = \int_0^t P_W(\tau) d\tau$$

If the power is constant over time this gives:

$$E = U_{\text{eff}} I_{\text{eff}} \cos \varphi t$$

To really *measure* the energy, can be done by an **induction-based system**. Such a reliable measurement system is very common, e.g. in any household for measurement of the consumed electricity (“Stromzähler”).

An electromagnet generates a field that creates eddy currents in the revolving disk. These cause a torque which is proportional to the product of voltage and current, i.e., the active power.



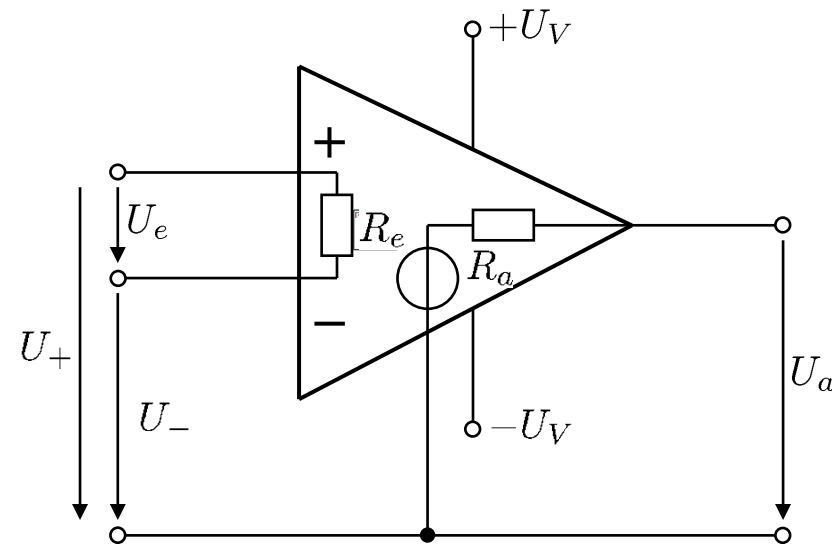
2.6 Measurement Methods and Amplifier Circuits

Operational Amplifier

An operational amplifier (OpAmp) is an *active* component. This means that it needs an external energy source which is given by a supply voltage U_V . An OpAmp is a multi-stage amplifier circuit that incorporates many transistors. Since 1962 it is available as an integrated circuit on a chip. Practically all measurement circuits are realized with the help of OpAmps. It is easy to build filter, integrator, differentiator and many more kind of circuits. **Analog computers** are based on OpAmp circuits and allow to simulate differential equations in a straight forward manner. They can be seen as the predecessor of **Simulink**.

A real OpAmp has the following properties:

- 2 inputs U_+ und U_- , whose difference U_e is amplified and generates the output $U_a = V u_e$.
- Input resistance R_e is in the mega ohm range.
- output resistance R_a is only a few ohm.
- Gain V is in the range 10.000 – 100.000.

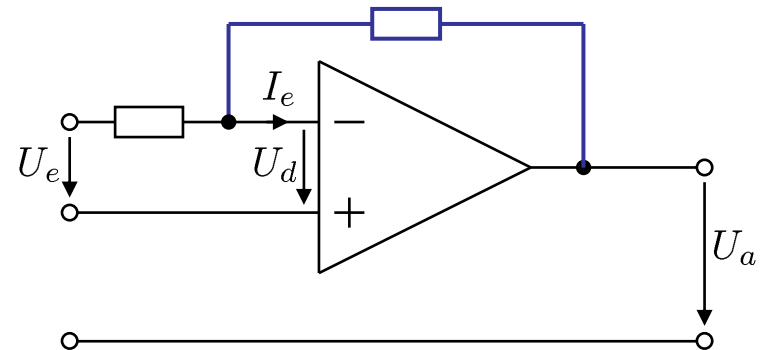


2.6 Measurement Methods and Amplifier Circuits

Ideal Operational Amplifier

Idealized an OpAmp can be described by the following approximations:

- Input resistance $R_e = \infty$.
- Output resistance $R_a = 0$.
- Gain $V = \infty$.



Amplifier with Feedback

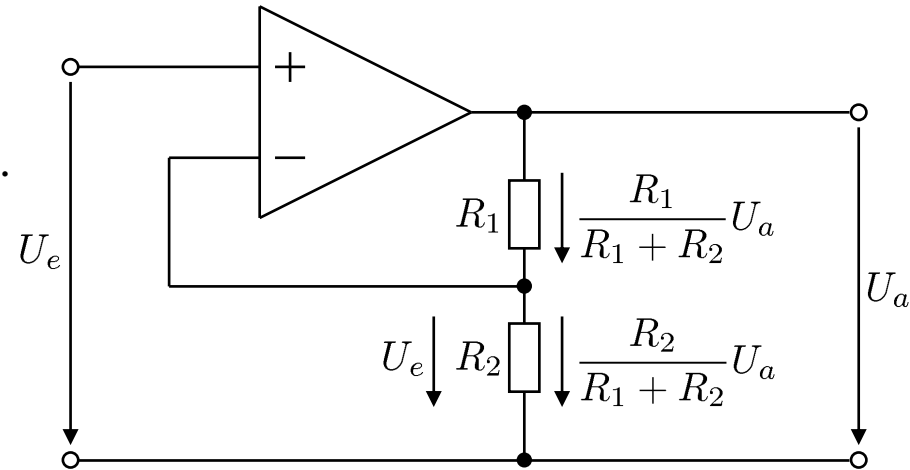
An OpAmp is either used as a switch (comparator) or most frequently applied with feedback that typically is used with negative sign (like in feedback control). I.e., the output is fed back to the “-”-input. This ensures that the input voltage U_d becomes very small since $U_d = U_a/V$ with $V = \infty$. Furthermore, the current into the OpAmp is insignificant since the input resistance is huge ($R_e = \infty$). Therefore, all fed back OpAmps are assumed to follow the important simplifications:

- OpAmp input voltage $U_d = 0$.
- OpAmp input current $I_e = 0$.

2.6 Measurement Methods and Amplifier Circuits

Voltage Amplification (Non Inverting)

A voltage amplifier has the task to convert an input voltage U_e in an output voltage $U_a = KU_e$. Moreover the load on the input voltage should be as small as possible, i.e., only a tiny current should be drawn from the circuit at the input. On the other side, the output should be capable to drive significant currents.



The gain of the voltage amplification has to be adjusted by the components within the circuit easily.

U_e can be measured over the resistor R_2 , because between the “+” and “-” inputs of the OpAmp almost no voltage drops. U_e splits according to the standard voltage divider rules onto both resistors, since almost no current goes into the OpAmp. Therefore the transfer function becomes:

$$U_e = \frac{R_2}{R_1 + R_2} U_a \quad \rightarrow \quad U_a = \frac{R_1 + R_2}{R_2} U_e = \left(1 + \frac{R_1}{R_2} \right) U_e$$

2.6 Measurement Methods and Amplifier Circuits

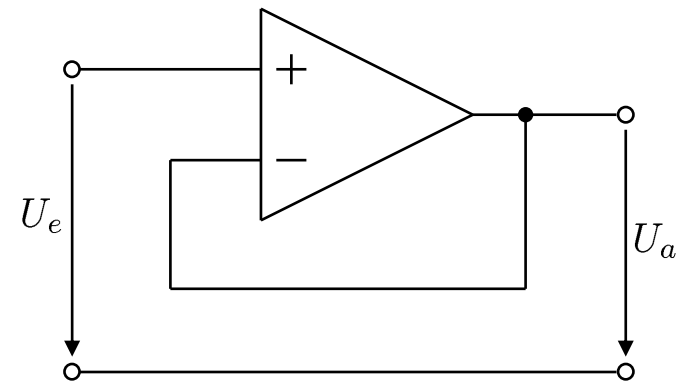
Application of a Voltage Amplifier

- *Voltage Measurement:* The voltage that shall be measured is connected to the input. At the output any circuit can draw a high current without influencing the measurement circuit. The evaluation circuit itself does not need to possess a very high resistance.
- *Constant Voltage Source:* If a voltage source is connected to the input, the OpAmp output can draw big currents without putting any load on the input. The voltage source is then in no danger to break down.
- *Voltage Amplification:* With an appropriate choice of R_1 and R_2 almost any desired gain $K > 1$ can be created.

Voltage Follower / Impedance Converter

Interesting is the special case $R_1 = 0$ (short circuit) and $R_2 = \infty$ (wire open). Such a circuit just converts the resistance/impedance. The transfer function is unity:

$$U_a = U_e$$



2.6 Measurement Methods and Amplifier Circuits

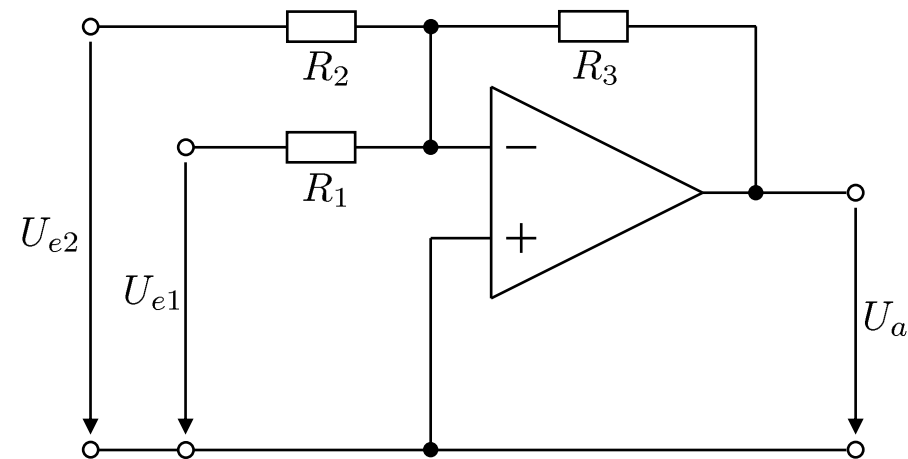
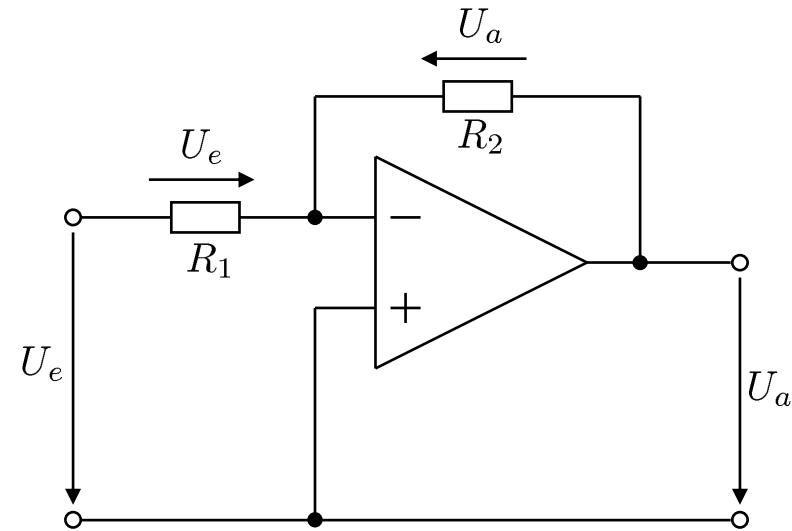
Voltage Amplification (Inverting)

The voltage amplification circuit has a small input resistance. Furthermore, it changes sign (inverting). U_e also drops at the resistor R_1 , because between the “+” and “-” inputs almost no voltage drops. According to the same argument, the output voltage U_a drops over R_2 . No current flows into the OpAmp. This means:

$$\frac{U_e}{R_1} + \frac{U_a}{R_2} = 0 \rightarrow U_a = -\frac{R_2}{R_1}U_e$$

It is also possible to add additional input in parallel. It can be used to build more complex addition or subtraction circuits., e.g.:

$$U_a = -\frac{R_3}{R_1}U_{e1} - \frac{R_3}{R_2}U_{e2}$$



2.6 Measurement Methods and Amplifier Circuits

Creation of Desired Dynamic Behavior

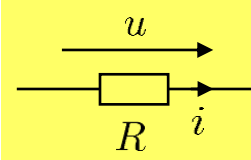
With the OpAmps any dynamic behavior can be achieved by using not only ohmic impedances, but also applying frequency-dependent components like capacitors and coils.

With a current of sin-type we get:

$$i(t) = \hat{i} \sin(\omega t)$$

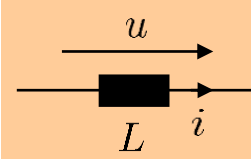
At a **resistor** with resistance R the voltage becomes:

$$u(t) = R \hat{i} \sin(\omega t)$$


$$u(t) = R i(t)$$
$$i(t) = \frac{1}{R} u(t)$$

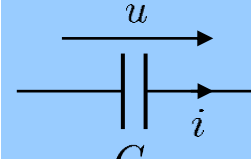
At a **coil** with inductivity L :

$$u(t) = L \hat{i} \cos(\omega t) \cdot \omega = \omega L \hat{i} \sin(\omega t + \pi/2)$$


$$u(t) = L \dot{i}(t)$$
$$i(t) = \frac{1}{L} \int_0^t u(\tau) d\tau$$

At a **capacitor** with capacity C :

$$u(t) = -\frac{1}{C} \hat{i} \cos(\omega t) \cdot \frac{1}{\omega} = \frac{1}{\omega C} \hat{i} \sin(\omega t - \pi/2)$$

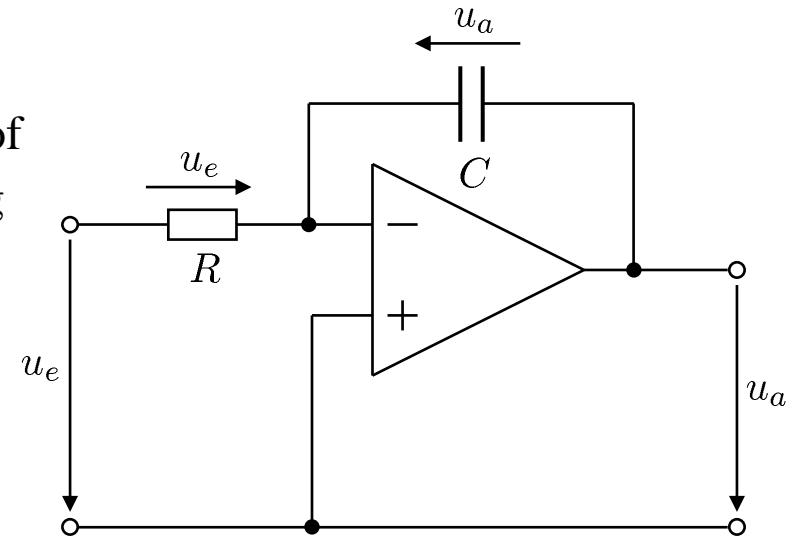

$$u(t) = \frac{1}{C} \int_0^t i(\tau) d\tau$$
$$i(t) = C \dot{u}(t)$$

2.6 Measurement Methods and Amplifier Circuits

Integrator

An integrator circuit is needed e.g. for the simulation of differential equations. It is also required for computing energy from power, speed from acceleration, distance from speed, electrical charge from current etc.

$$\frac{u_e(t)}{R} + C \dot{u}_a = 0 \quad \rightarrow \quad u_a(t) = -\frac{1}{RC} \int_0^t u_e(\tau) d\tau$$

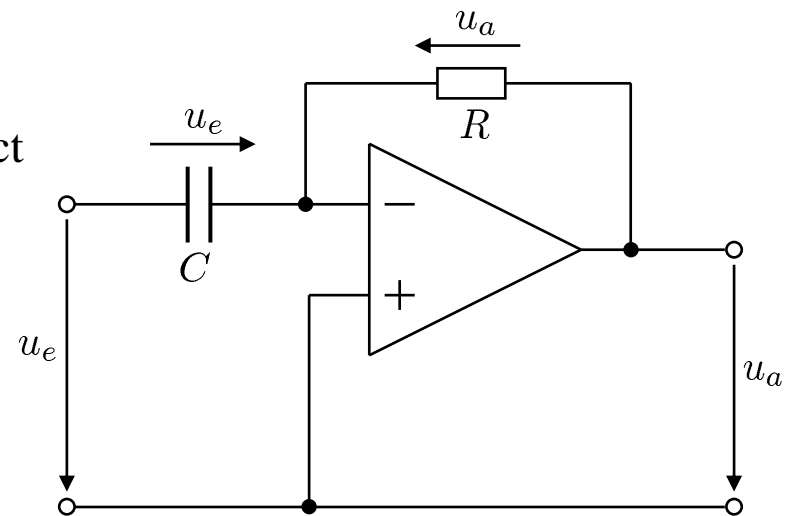


Differentiator

At the OpAmp circuit it is obvious, that this is the exact opposite of the integrator shown above.

$$C \dot{u}_e(t) + \frac{u_a}{R} = 0 \quad \rightarrow \quad u_a(t) = -RC \dot{u}_e(t)$$

With R and C the proportionality (time) constant can be adjusted.



2.6 Measurement Methods and Amplifier Circuits

Low Pass Filter

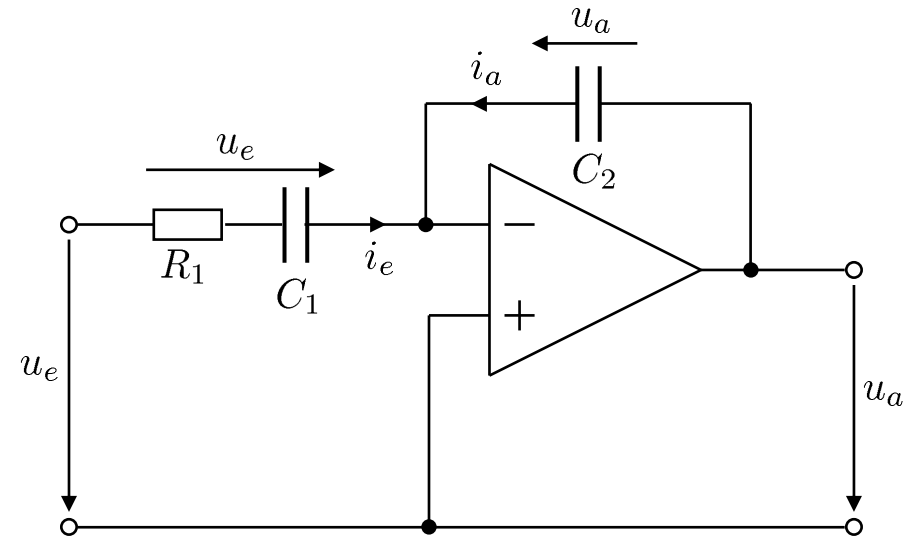
This circuit simulates a first order differential equation. It is a simple low pass filter (PT₁) to suppress high frequency disturbances like noise.

$$u_e(t) = R_1 i_e(t) + \frac{1}{C_1} \int_0^t i_e(\tau) d\tau$$

$$i_e(t) = -i_a(t) = -C_2 \dot{u}_a(t)$$

$$\rightarrow u_e(t) = -R_1 C_2 \dot{u}_a(t) - \frac{C_2}{C_1} u_a(t)$$

The factor $-C_1/C_2$ is a gain factor, i.e., it determines the static gain of the transfer function. For a filter it is thus reasonable to choose $C_1 = C_2$. A subsequent inverter should be used to get rid of the “-“ sign. $R_1 C_1$ is the time constant and $1/R_1 C_1$ is called the corner frequency which determines the filter bandwidth.



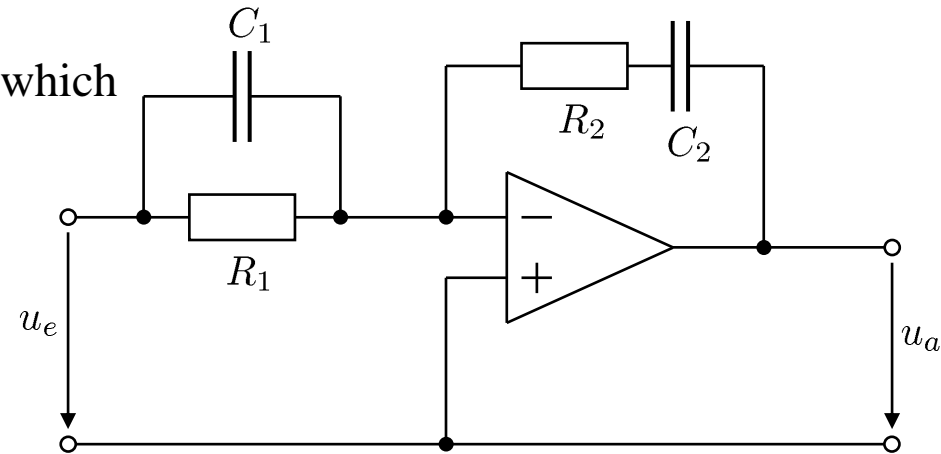
$$\rightarrow -\frac{C_1}{C_2} u_e(t) = u_a(t) + R_1 C_1 \dot{u}_a(t)$$

$$G(s) = -\frac{\frac{1}{sC_2}}{\frac{1}{sC_1} + R_1} = -\frac{\frac{C_1}{C_2}}{1 + sR_1 C_1}$$

2.6 Measurement Methods and Amplifier Circuits

PID Control

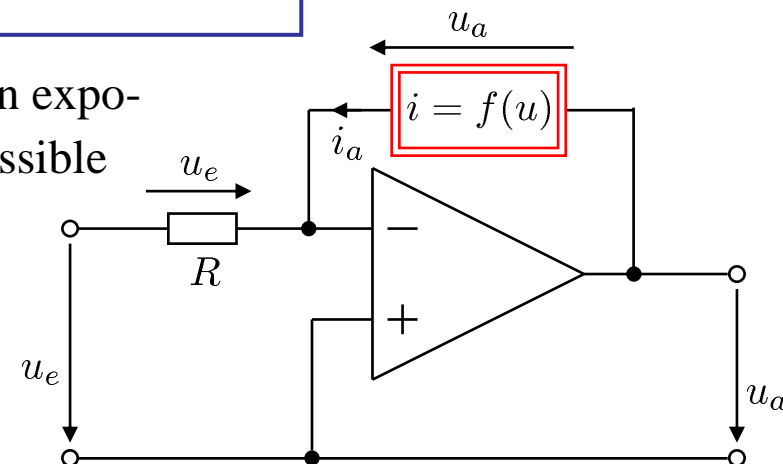
This OpAmp circuit realizes a PID controller, which is the most widely used controller type. The (P) part realizes the proportional, the (I) part realizes the integrative, and the (D) part realizes the derivative action. The respective values can be adjusted by the corresponding resistors and capacitors.



$$u_a(t) = -\frac{R_1 C_1 + R_2 C_2}{R_1 C_2} u_e(t) - \frac{1}{R_1 C_2} \int_0^t u_e(\tau) d\tau - R_2 C_1 \dot{u}_e(t)$$

With help of **nonlinear** components like diodes, e.g. an exponential characteristics can be constructed. It is even possible to construct circuit that calculate the logarithm. Based on these, multiplication and division are easy to build.

$$\frac{u_e(t)}{R} + f(u_a(t)) = 0 \rightarrow u_a(t) = f^{-1}\left(-\frac{u_e(t)}{R}\right)$$



2.6 Measurement Methods and Amplifier Circuits

Bridge Circuit

Measurement of impedances (purely ohmic or frequency-dependent) can be reduced to a simple voltage and current measurement and a subsequent division. But very powerful and widely used are direct measurements via a **bridge circuit**. For simplicity, the procedure shall be explained for resistances but an extension to any kind of impedance is straight forward.

There are 2 alternative approaches:

1. The unknown resistance is compared to an adjustable resistance.
The adjustable resistance will be tuned as long the bridge circuit is balanced.
2. The unknown resistance deviates only insignificantly from its (known) nominal value.
In this case, it is possible to calculate the resistance from the diagonal bridge voltage.

Method 1 has the advantage that the diagonal bridge voltage has to be measured only for very small (positive or negative) values around 0. It is not necessary to have an instrument that can handle large amplitudes. It is possible to achieve with a high accuracy with simple instruments. On the other hand the tuning can be tedious.

Method 2 is fast and effective but works only around an operating point, i.e., if the resistance is close to its nominal value.

2.6 Measurement Methods and Amplifier Circuits

Balance the Bridge

This bridge circuit was invented and first applied by *Wheatstone* in 1843. Under the following condition this bridge is balanced, i.e., the diagonal voltage is zero ($U_d = 0$):

$$U_1 = U_3$$

According to the voltage divider rule this means:

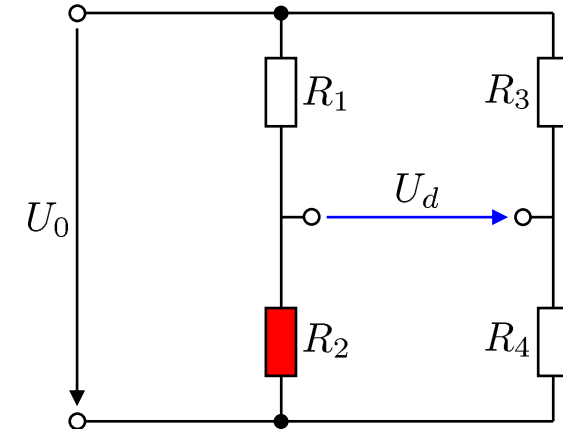
$$\frac{R_1}{R_1 + R_2} = \frac{R_3}{R_3 + R_4} \rightarrow R_1 R_3 + R_1 R_4 = R_1 R_3 + R_2 R_3 \rightarrow \boxed{R_1 R_4 = R_2 R_3}$$

If the resistance R_2 is unknown, we can tune one resistor (in principle, any one or more than one) until the diagonal voltage is zero: $U_d = 0$. The **bridge** then is **balanced**. The unknown resistance thus can be calculated from:

$$\boxed{R_2 = \frac{R_1 R_4}{R_3}}$$

Advantage: Independent of quality of the voltage source U_0 . Only measurement of U_d around zero is necessary.

Drawback: Tedious tuning of the comparing resistance.



2.6 Measurement Methods and Amplifier Circuits

Bridge Voltage

If the unknown resistance deviated only slightly from its nominal value, the diagonal voltage can be used as a measure of this resistance:

$$U_d = \frac{R}{2R} U_0 - \frac{R}{2R + \Delta R} U_0 = \frac{\Delta R}{2R + \Delta R} \frac{U_0}{2}$$

If the resistance deviation ΔR is small compared to R , in approximation we have:

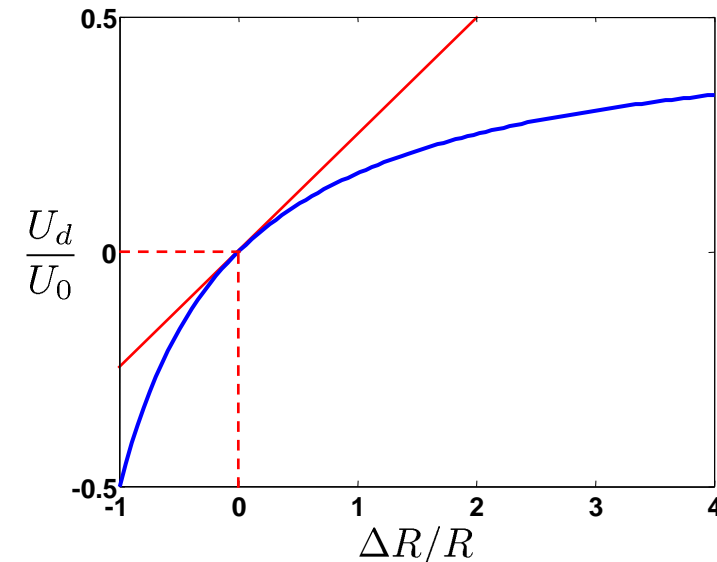
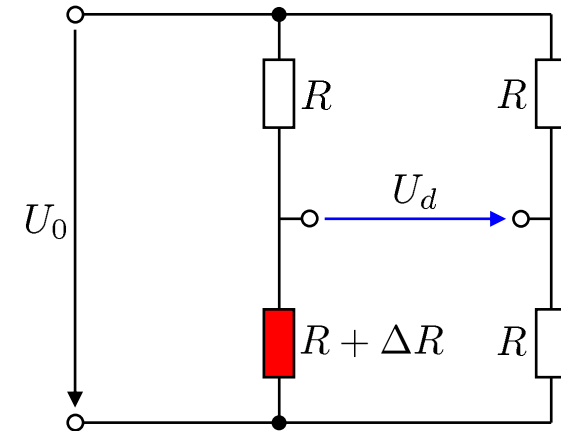
$$U_d \approx \frac{\Delta R}{R} \frac{U_0}{4}$$

However, the relation between ΔR and U_d is only approximately linear:

$$\Delta R = 0 \rightarrow U_d = 0$$

$$\Delta R = R \rightarrow U_d = \frac{U_0}{6}$$

$$\Delta R = -R \rightarrow U_d = -\frac{U_0}{2}$$



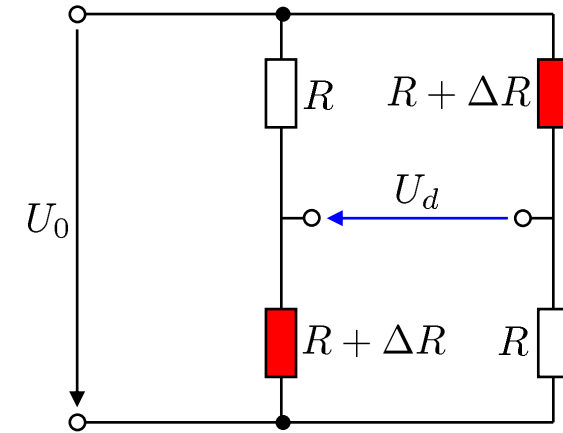
2.6 Measurement Methods and Amplifier Circuits

Increase of Sensitivity

Half Bridge

The sensitivity of the measurement can be doubled by utilizing 2 measurement resistors (red) instead of 1:

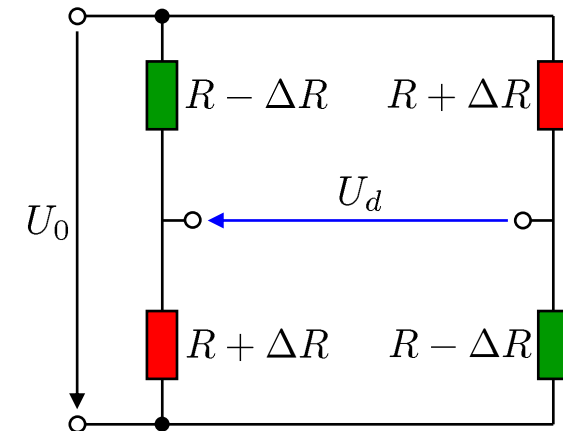
$$U_d \approx \frac{\Delta R}{R} \frac{U_0}{2}$$



Full Bridge

A further increase of sensitivity can be achieved by utilizing 2 positively (red, $R + \Delta R$) and negatively (green, $R - \Delta R$) changed resistances. This is e.g. a common approach for resistance strain gauges. Typically the strains are attached on opposite sides of a bar.

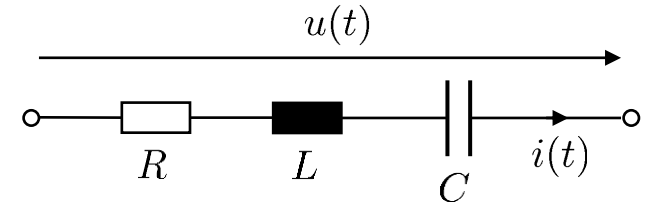
$$U_d = \frac{\Delta R}{R} U_0$$



2.6 Measurement Methods and Amplifier Circuits

Oscillators

Electrical oscillators consist of a capacitor with capacity C and a coil with inductivity L and a resistor with (relatively small) resistance R .



Such an oscillator is the equivalent to a mass-damper-spring system in mechanics. Only in the resistor or the damper, respectively, energy is lost (more strictly speaking converted to heat). Without these dissipative elements, they would oscillate forever with their **resonance frequency** ω_0 . This resonance frequency depends on C and L (or the spring constant c and the mass m , respectively). Therefore, it can be utilized to measure capacities and/or inductivities in an indirect manner.

Electrical oscillators follow the relationship between voltage and current given by:

$$u(t) = R i(t) + L \dot{i}(t) + \frac{1}{C} \int_0^t i(\tau) d\tau \quad \rightarrow \quad \dot{u}(t) = R \dot{i}(t) + L \ddot{i}(t) + \frac{1}{C} i(t)$$

With a current of sin-type: $i(t) = i_0 \sin(\omega t)$

$$\dot{u}(t) = R i_0 \omega \cos(\omega t) - L i_0 \omega^2 \sin(\omega t) + \frac{1}{C} i_0 \sin(\omega t)$$

2.6 Measurement Methods and Amplifier Circuits

Resonance in Oscillators

In the case of resonance, the change of voltage at the capacitor and the coil cancel each other exactly. Resonance happens for:

$$L \omega^2 = \frac{1}{C}$$

Then, the impedance of the oscillator is purely ohmic. In the ideal case of no energy loss ($R \rightarrow 0$ or in the mechanical case damper constant $d \rightarrow 0$, respectively) the current would be of infinite amplitude and oscillating at the resonance frequency of:

$$\boxed{\omega_0 = \frac{1}{\sqrt{LC}}} \quad \text{or for the mechanical counter part: } \omega_0 = \sqrt{\frac{c}{m}}$$

The resonance frequency ω_0 can be used to determine:

- the inductivity L if C is known,
- the capacity C if L is known.

Coils, capacitors, and whole oscillators can naturally be build in OpAmp circuits.

3. Measurement of Non-Electrical Quantities

Contents of Chapter 3

3. Measurement of Non-Electrical Quantities

- 3.1 Sensors and Sensor Systems
- 3.2 Displacement and Angles
- 3.3 Speed
- 3.4 Acceleration
- 3.5 Force, Torque, Pressure, and Mass
- 3.6 Temperature
- 3.7 Flow
- 3.8 Miscellaneous

3.1 Sensors and Sensor Systems

Desired Properties for Sensors

- Conversion of a physical measurement quantity into a signal that is suitable for further processing. Typically, this is an electrical signal because it is especially well suited for this task.
- *Sensitivity*: High as possible reaction with respect to the quantity that shall be measured.
- *Selectivity*: Low as possible reaction with respect to everything else.
- *Stability*: Constant as possible behavior with respect to all environmental changes like temperature and aging.

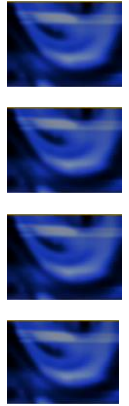
Sensor Systems

- Sensors integrated with intelligent components such as micro-controllers with software (also called **smart sensor**).
- Combination of many identical or different sensors.
- Integration of sensors, actuators, and appropriate control equipment.

3.1 Sensors and Sensor Systems

Sensor Fusion

- Information of many sensors is combined in a clever way to achieve advantages.
- Stochastic measurement errors can be reduced by averaging.
- Different principles can be combined to reduce their weaknesses and gain strengths from synergy effects.



Examples for Sensor Fusion:

- *Stereo Vision:* 2 cameras build up a 3D picture or video.
- *Navigation System:* Modern such systems for planes, ships, and cars make use of the satellite-based GPS and combine it with local sensing of speed, steering angle, etc.
- *Driver Assistance:* Adaptive cruise control (ACC), lane detection, night vision, lane changing assistant (blind spot detection), etc. are based on a variety of different sensors like radar, laser, CCD camera, ultrasonic, navigation maps, ...
- *Smart Dust:* Next page.

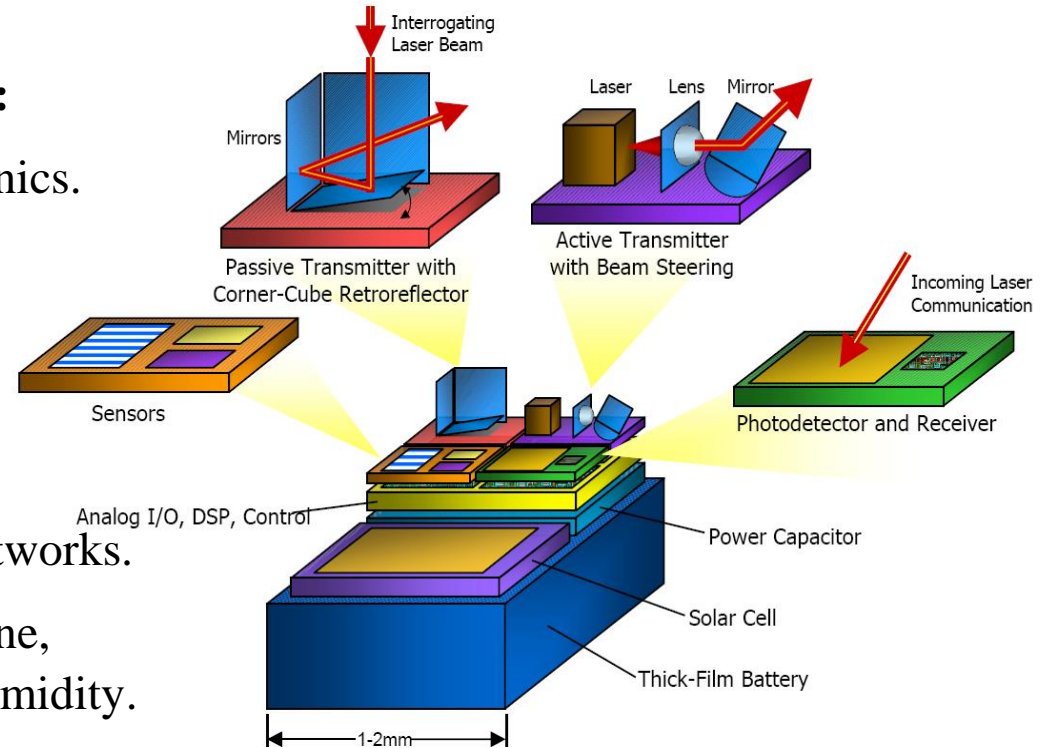
3.1 Sensors and Sensor Systems

Smart Dust

A few cm^3 small, intelligent sensor systems communicate over a wireless network with a base station and possibly with each other. This is performed by the means of laser beams. These concepts are currently developed at UC Berkeley by *Pister* and some ideas and problems are known from the novel “Prey” by *Crichton*. Maybe it becomes reality!

Integration of Different Technologies:

- Ultra energy efficient micro-electronics.
- MEMS: micro-electro-mechanical systems.
- Wireless laser-based communication (1 kB/s).
- Management of huge distributed networks.
- Possible sensors: camera, microphone, acceleration sensor, temperature, humidity.
- Extremely cheap.



3.2 Displacement and Angles

Resistive Measurement Methods

Many of the techniques to measure displacement and angles can also be used for the determination of force, torque, and pressure. It is just necessary to have a spring whose displacement is proportional to these quantities.

Principle of Resistive Displacement Measurement

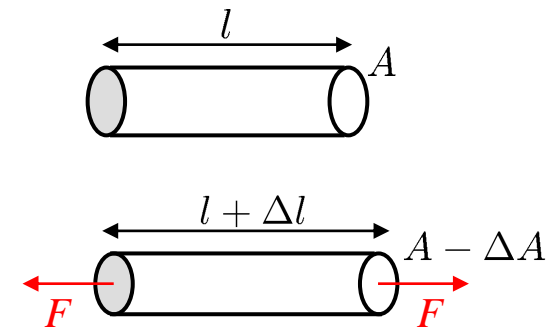
The ohmic resistance of a electric wire depends on its length l , its cross-section area and its specific resistance ρ which in turn depends on the material:

$$R = \frac{l}{A} \rho$$

If the wire is pulled apart with a force F this is influencing the relative resistance:

$$\frac{\Delta R}{R} = \frac{\Delta l}{l} - \frac{\Delta A}{A} + \frac{\Delta \rho}{\rho} = K \frac{\Delta l}{l} = K \varepsilon \quad \text{mit } \varepsilon = \frac{\Delta l}{l}$$

The factor K summarizes the influence of length and area change and the variation of the specific resistance.



3.2 Displacement and Angles

Resistive Measurement Method: Strain Gauge

Resistive strain gauges utilize the resistance change caused by a length change ε . They are commonly manufactured as an elastic foil and glued on the body to be measured. It can be distinguished between different material types:

- **Metal:** Typical sensitivity is around $K = 2$. The resistance change is mainly based on the length and area change. Specific resistance changes only insignificantly.
- **Semiconductor:** Typical sensitivity is very high in absolute values, either around $K = -100$ or around $K = 100$ for n- or p-doped semiconductors. The **piezoresistive effect** is utilized, i.e., the internal generation of electrical charge resulting from an applied mechanical force. It changes the specific resistance significantly. This extremely high sensitivity must be paid for by an undesirable high temperature dependency.

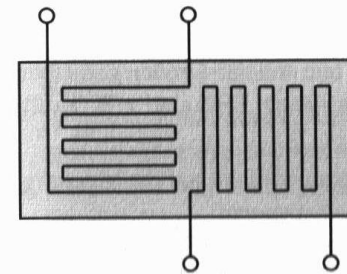


Bild 3.91 Zwei Foliendehnmessstreifen mit um 90° versetzten Beanspruchungsrichtungen

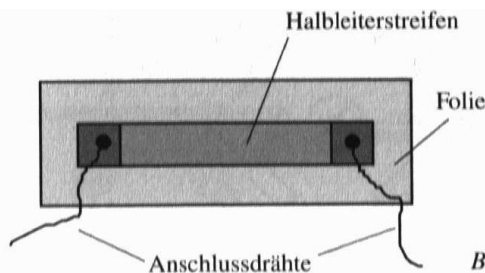


Bild 3.92 Halbleiterdehnmessstreifen

3.2 Displacement and Angles

Resistive Method: Strain Gauge Embodiments [3]

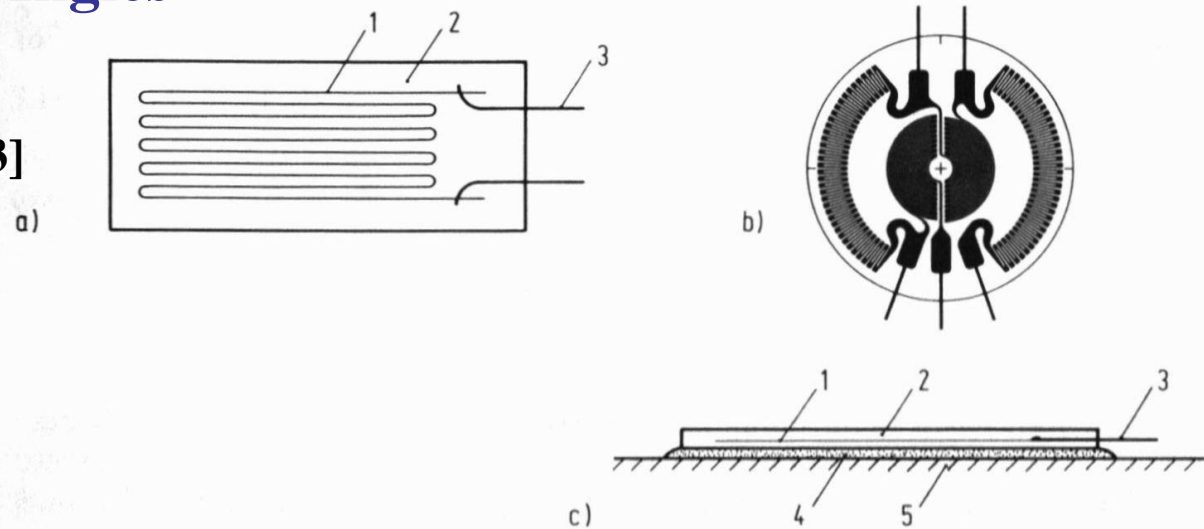


Bild 3.47: Dehnungsmeßstreifen

a) Drahtmeßstreifen

b) Folienmeßstreifen als Membranrosette (Hottinger Baldwin Meßtechnik)

c) Querschnitt durch einen aufgeklebten Meßstreifen; 1 Meßgitter, 2 Abdeckung, 3 Streifenanschluß, 4 Kleber, 5 zu untersuchendes Werkstück

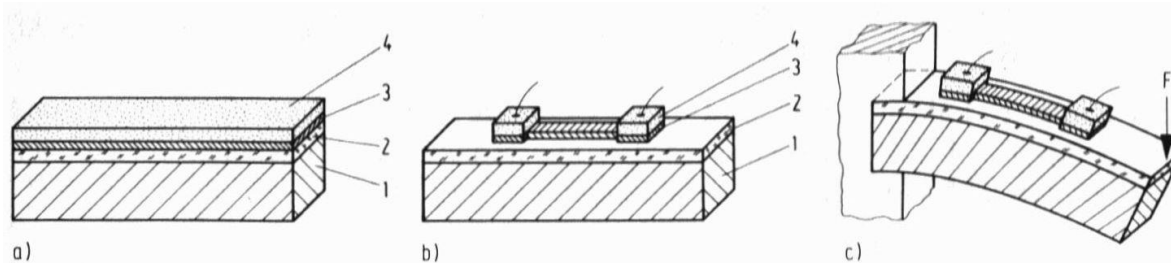


Bild 3.48: Biegebalken-Kraft-Meßaufnehmer mit Dünnschicht-DMS [3.18]

a) nach einseitig-vollflächiger Beschichtung; 1 Federkörper aus Bronze oder Stahl, 2 Isolierschicht, 3 dehnungsempfindliche Widerstandsschicht, 4 niederohmige Leiterschicht

b) nach photolithographischer Strukturierung

c) Aufnehmer unter Belastung

3.2 Displacement and Angles

Resistive Measurement Method: Placement of Strain Gauges

Applying multiple strain gauges can improve the sensitivity of the measurement. Like shown below, in a bridge circuit the sensitivity can be quadrupled (4x). The higher selectivity of such an approach is desirable. However, most important is the robustness against temperature changes because the temperature effects (and others) cancel each other. If the resistances are all changed relatively in the same manner, the bridge voltage is not affected at all.

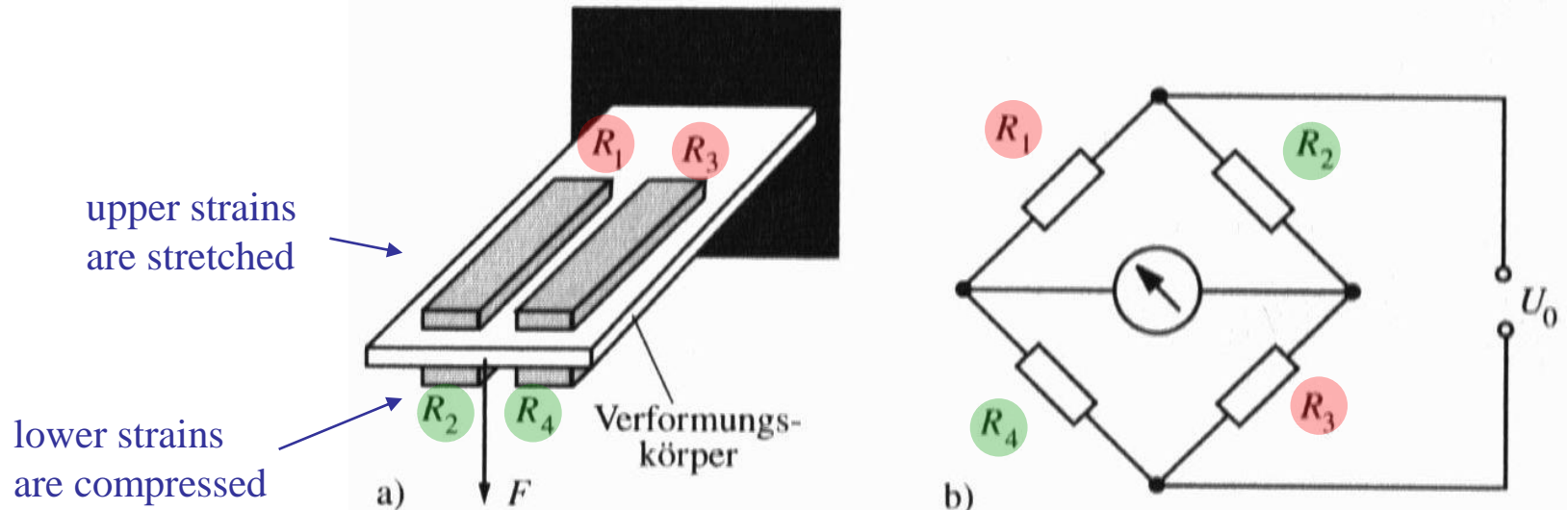


Bild 3.93 Dehnmessstreifen in einer Vollbrücke
a) Anordnung, b) Schaltung

3.2 Displacement and Angles

Resistive Measurement Method: Effect of Magnetic Field

- *Hall sensors:* A magnetic field orthogonal to an electrical current leads to an Lorentz force on the electrons. This causes a **Hall voltage** orthogonal to magnetic field and current. For currents around 100...500 mA the voltage is typically around 50...400 mV with reasonable field strength. Such Hall sensors are commonly used as limit switches.
- *Field-plates:* The Hall effect deflects the current and enforces it not go the direct way but to take a detour. As a consequence, the **resistance** increases (magneto resistive effect). A quadratic characteristic results. It can be compensated by a differential bridge circuit.

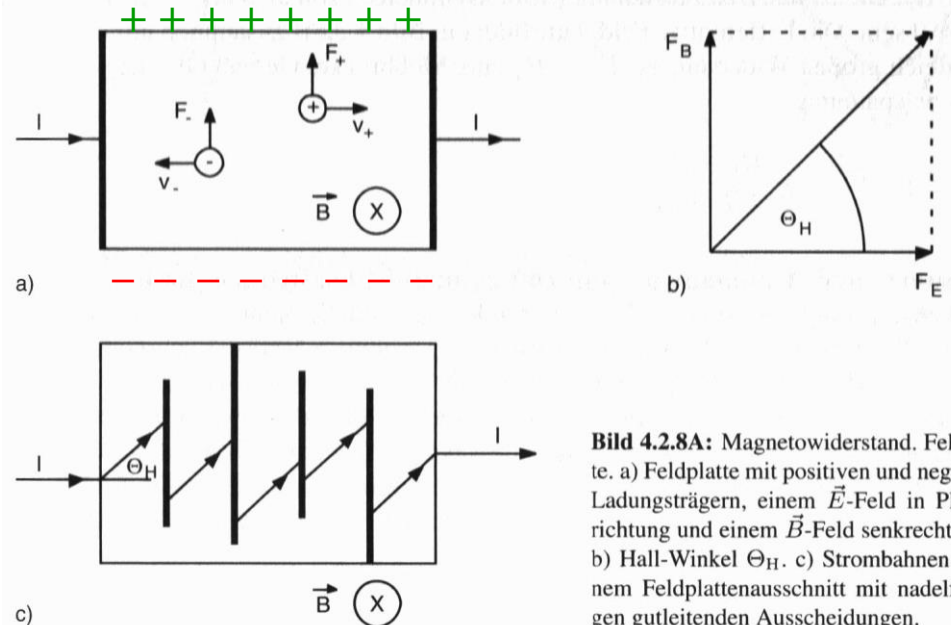
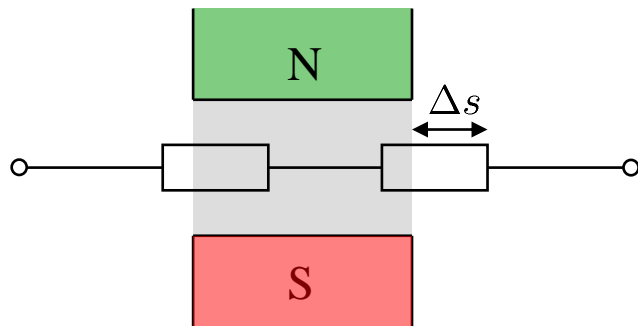


Bild 4.2.8A: Magnetowiderstand. Feldplatte. a) Feldplatte mit positiven und negativen Ladungsträgern, einem \vec{E} -Feld in Plattenrichtung und einem \vec{B} -Feld senkrecht dazu. b) Hall-Winkel Θ_H . c) Strombahnen in einem Feldplattenausschnitt mit nadelförmigen gutleitenden Ausscheidungen.

3.2 Displacement and Angles

Inductive Measurement Method: Inductivity of a Coil [3]

The inductivity of a coil can be calculated from its number of windings N and its magnetic resistance R_m :

$$L = \frac{N^2}{R_m} \quad \text{with} \quad R_m = \frac{s}{\mu_0 \mu_r A}$$

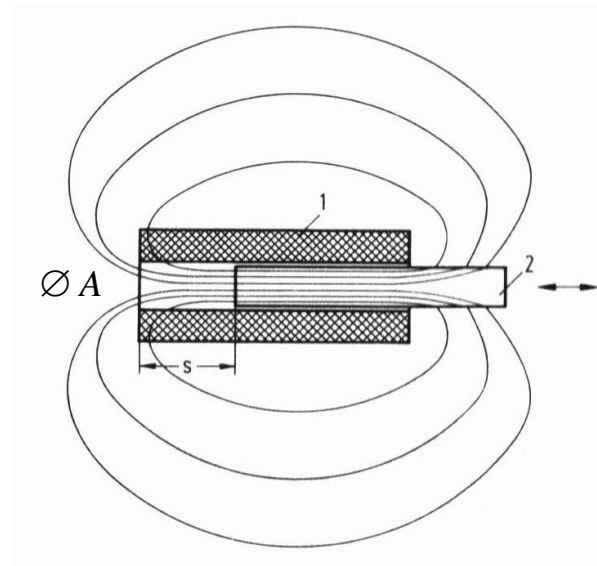
where s is the length of the flux lines, A is the area where the flux lines pass through, and μ_r is the relative magnetic

permeability of the material. For the coil three such parts add up: a) inside the coil in a part that is filled with iron ($\mu_r \gg 1$), b) inside the coil in a part that is filled with air or nothing ($\mu_r \approx 1$) and c) outside the coil that usually also consists of air or nothing ($\mu_r \approx 1$):

$$R_m = R_{m,a} + R_{m,b} + R_{m,c} = \frac{s_{\text{iron}}}{\mu_0 \mu_r A} + \frac{s}{\mu_0 A} + \frac{s_{\text{outside}}}{\mu_0 A_{\text{outside}}} \approx \frac{s}{\mu_0 A}$$

The 1. term can be neglected due to the very high value for μ_r . The 3. term can be neglected due to the large area A_{outside} outside. This leaves us the 2. term. Therefore the inductivity is inverse proportional to the length of the part inside the coil which is *not* filled:

$$L = \frac{N^2 \mu_0 A}{s} = \frac{k}{s}$$



3.2 Displacement and Angles

Inductive Measurement Method: Plunger and Differential Plunger

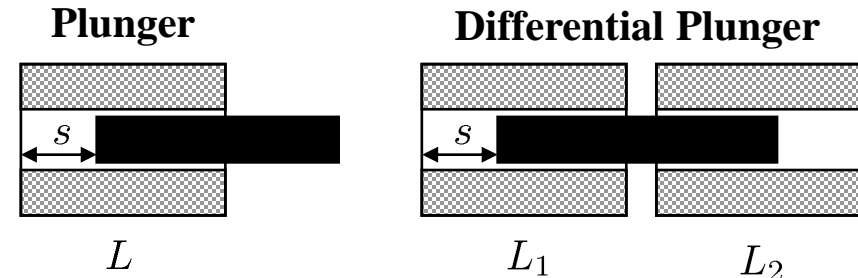
A small displacement Δs of the armature from the operating point s influences the inductivity in a *nonlinear* way as follows:

$$L = \frac{k}{s + \Delta s} = \frac{\tilde{k}}{1 + \frac{\Delta s}{s}}$$

This means that only for tiny displacements the inductivity L is roughly proportional to the displacement Δs (with negative sign, i.e., $\Delta s > 0 \rightarrow \Delta L < 0$). To enlarge the roughly linear range, the differential approach was developed. The idea is to introduce a second coil whose inductivity operates in the other direction. The displacement drives the armature opposite to the first coil and a displacement Δs leads to a decrease in the first but increase in the second coil, or the other way round:

$$L_1 = \frac{k}{s + \Delta s} \quad L_2 = \frac{k}{s - \Delta s}$$

A clever combinations of both inductivities by a circuit creates a *linear* measurement characteristics. The linear behavior is achieved in an exact way, not only by approximation or linearization, which would be valid only for small displacements Δs .



3.2 Displacement and Angles

Inductive Measurement Method: Differential Measurement Principle

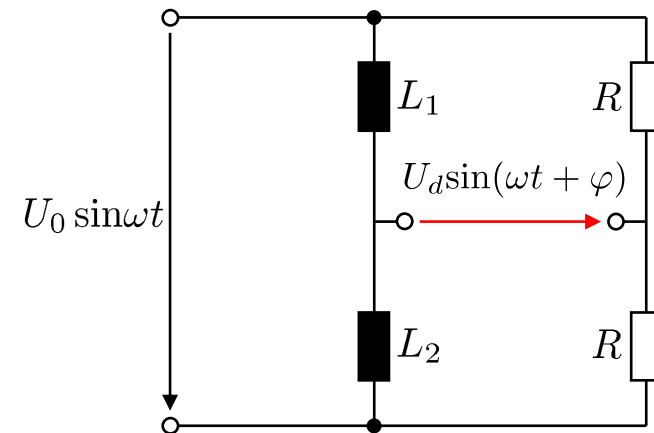
Such a bridge circuit can be used to create a linear characteristics. The diagonal bridge voltage U_d is equal to the difference between the voltage drop along the upper resistance and inductivity:

$$U_d = \frac{R}{R + R} U_0 - \frac{L_1}{L_1 + L_2} U_0 = \frac{1}{2} U_0 - \frac{L_1}{L_1 + L_2} U_0 = \frac{L_1 + L_2 - 2L_1}{2(L_1 + L_2)} U_0$$

$$U_d = \frac{L_2 - L_1}{2(L_1 + L_2)} U_0 = \frac{\frac{1}{L_1} - \frac{1}{L_2}}{\frac{1}{L_1} + \frac{1}{L_2}} \frac{U_0}{2}$$

Introduction of the dependency on the displacement gives:

$$U_d = \frac{s + \Delta s - (s - \Delta s)}{s - \Delta s + s + \Delta s} \frac{U_0}{2} = \frac{2\Delta s}{2s} \frac{U_0}{2} = \frac{\Delta s}{s} \frac{U_0}{2}$$



The differential principle together with the bridge circuit results in an *exact* proportionality between displacement and diagonal voltage. This type of “physical linearization” is widely applied in many circumstances (also with capacitor, etc.).

3.2 Displacement and Angles

Inductive Sensors [3]

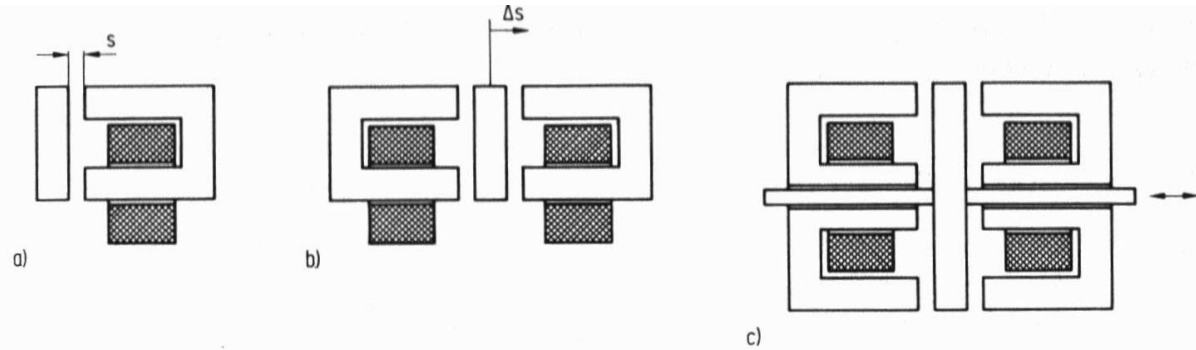
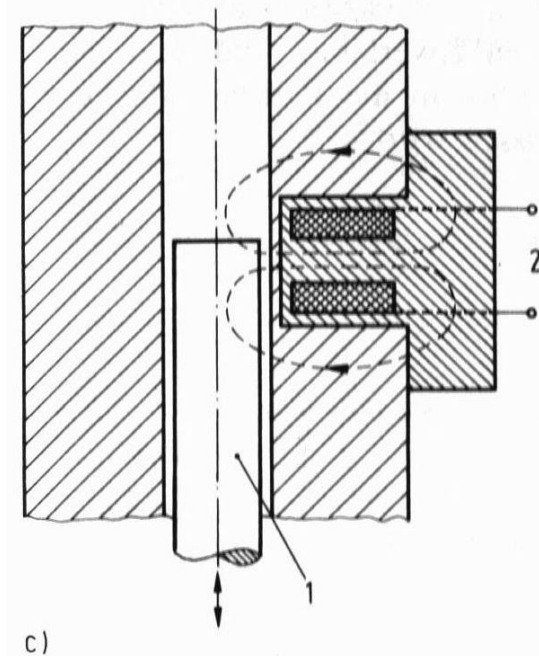
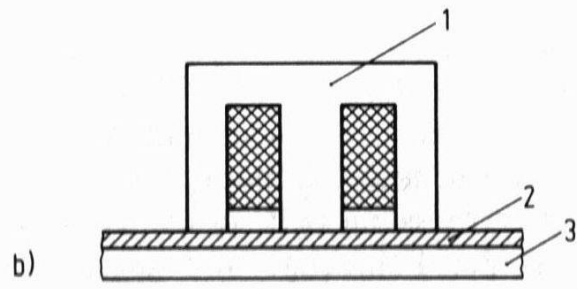
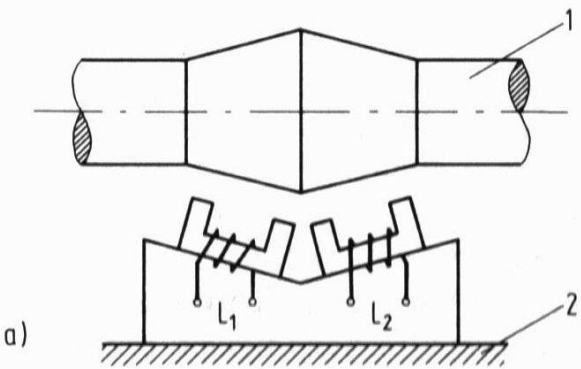


Bild 4.18: Querankeraufnehmer; einfache Ausführung (a), Differential-Querankeraufnehmer (b) und Differential-Querankeraufnehmer mit Topfkern (c)

Bild 4.20: Anwendung von induktiven Aufnehmern

- a) Messung der Relativdehnung zwischen Turbinenwelle 1 und Gehäuse 2 [0.17]
- b) Messung der Dicke von nichtmagnetischen Schichten; 1 Drossel, 2 nichtmagnetische Komponente (Folie, Lack-schicht), 3 Eisenkern
- c) Messung der Ventilstellung in einer Hochdruck-Dampfleitung [4.4]; 1 Ventilstange, 2 Anschlüsse der Spule



3.2 Displacement and Angles

Capacitive Measurement Method

The capacity C of a plate capacitor depends on the distance between the plates d , the area of the plates A and the permittivity ε_r , determined by the material between the plates:

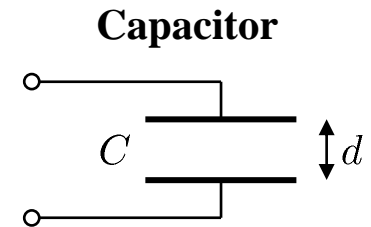
$$C = \frac{\varepsilon_0 \varepsilon_r A}{d}$$

Change of Capacitor Plate

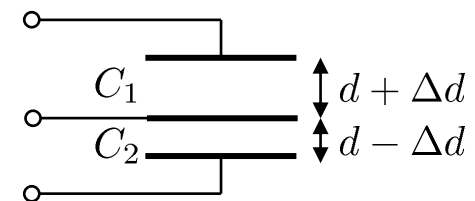
A change in the distance between both plates has the same *nonlinear* effect as the just discussed displacement in the inductivity:

$$C = \frac{k}{d + \Delta d} = \frac{\tilde{k}}{1 + \frac{\Delta d}{d}}$$

Similar to the inductivity change, the capacitor can be built according to the differential principle. Again, together with a bridge circuit a *linear* characteristics can be created.



Differential Capacitor

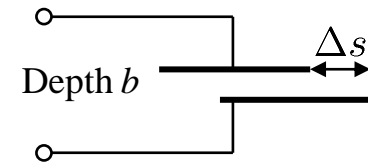


3.2 Displacement and Angles

Capacitive Measurement Method: Change of Capacitor Plate

A change of the plate area directly (without any tricks) effects the capacity in a *linear* way:

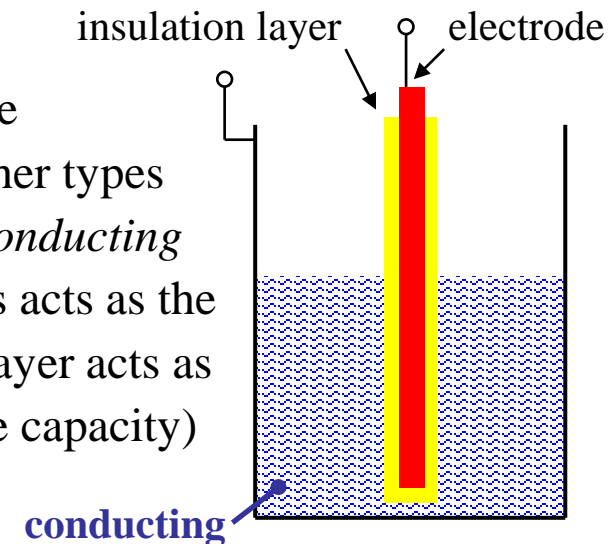
$$C = k(A + \Delta A) = \tilde{k} \left(1 + \frac{\Delta A}{A} \right)$$



With an original plate area of $A = bs$ this yields a change of that area of $\Delta A = b \Delta s$. Thus, the capacity changes linearly with the displacement of the plates against each other:

$$C = kb(s + \Delta s) = c(s + \Delta s) = \tilde{c} \left(1 + \frac{\Delta s}{s} \right)$$

This approach is commonly applied for displacement and angle measurement as well as **fill level** measurement in tanks and other types of reservoirs. It is important to notice that the liquid must be *conducting* electricity. The reservoir together with the conducting contents acts as the one capacitor plate, the electrode as the other. The insulation layer acts as dielectric medium. The effective plate area (proportional to the capacity) is proportional to the fill level.



3.2 Displacement and Angles

Capacitive Measurement Method: Change of Dielectric Medium

With the shown approach the thickness of layers can be measured if their permittivity ϵ_{r2} is known. On one capacitor plate the material layer is applied; the remaining part is typically filled just with air, i.e., $\epsilon_{r1} = 1$.

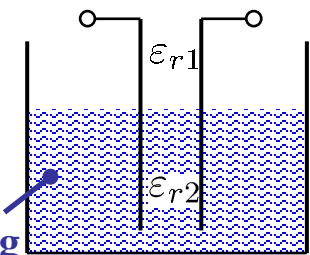
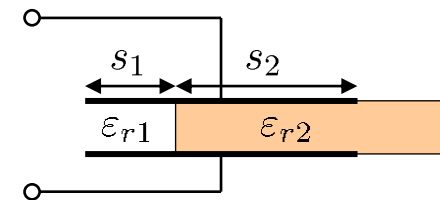
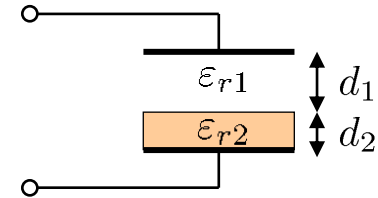
The capacity of the capacitor is influenced in a *nonlinear* way by the thickness of the material layer d_2 . According to the rule of a series connection of two capacitors, we get the following overall capacity ($d_1 = d - d_2$):

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} = \frac{d_1}{\epsilon_0 \epsilon_{r1} A} + \frac{d_2}{\epsilon_0 \epsilon_{r2} A} \rightarrow C = \frac{\epsilon_0 A}{d_1/\epsilon_{r1} + d_2/\epsilon_{r2}}$$

Displacement, angles, and fill levels even of *non-conducting* materials (as long as ϵ_{r2} is significantly different from ϵ_{r1}) can be measured with the approach shown to the right. Here, the relationship between the displacement s_2 and the overall capacity follows the rule of two capacitors in parallel which yields a linear relationship ($s_1 = s - s_2$):

$$C = C_1 + C_2 = \frac{\epsilon_0 \epsilon_{r1} b s_1}{d} + \frac{\epsilon_0 \epsilon_{r2} b s_2}{d} = \frac{\epsilon_0 b}{d} (\epsilon_{r1} s_1 + \epsilon_{r2} s_2)$$

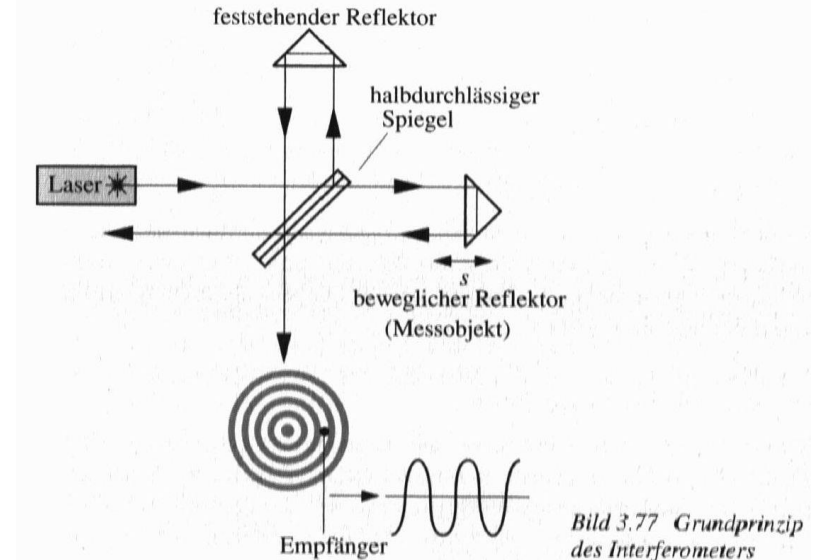
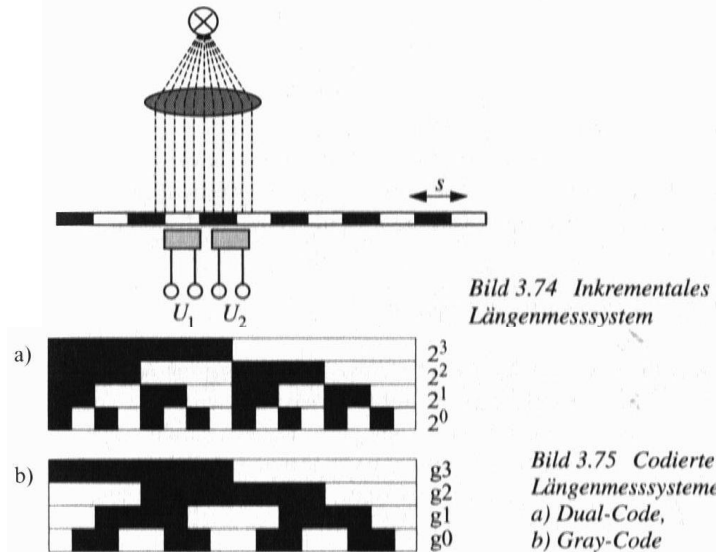
non-conducting



3.2 Displacement and Angles

Optical Measurement Techniques

- *Incremental Displacement Measurement:* The distance is divided into equidistant intervals whose width determines the resolution of the measurement. The intervals are counted and the measurement is always *relative* to a starting point.
- *Coded Displacement Measurement:* Coding of the position allows to determine the *absolute*, not only the relative, position.
- *Interferometric Displacement Measurement:* Highly accurate measurement based on interference of laser beams. Displacement around $\lambda/8$ can be determined ($\lambda \sim 600 \text{ nm}$).



3.2 Displacement and Angles

Miscellaneous

Displacement measurement techniques applied in modern driver assistant systems:

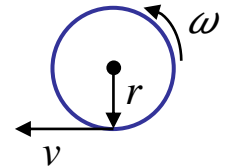
- *Infrared*: Based on the emission and reflection of laser impulses and the measurement of their time delay (ns range!). Can be used to measure the distance to the ahead driving car for *adaptive cruise control* systems. Good visibility is required, but then good signal quality can be expected. Quite low price.
- *Radar*: An alternative to infrared technology. Typically, realized with 77 GHz radar frequency. According to the *Doppler* principle besides the distance, the *relative velocity* to the next car can be measured. Bad visibility is no handicap. Relatively expensive. In the short range 24 GHz radar is used for *parking* sensors.
- *Ultrasound*: Used for parking sensors (only short distances!). High importance for nondestructive material testing.
- *CCD camera*: Together with powerful but expensive and complex image data processing, this can support the other sensors. It is necessary for lane and blind spot detection. Very flexible but complicated. Not very robust.

3.3 Speed

Possibilities for Speed Measurement

Three main alternatives are available for speed measurement:

1. Measurement of a time interval Δt , in which a certain distance Δs is covered. Subsequently the speed can be calculated by $v = \Delta s / \Delta t$. Speed measurement is done by measuring distance and time.
2. Measurement of a rotational speed ω and conversion into the translational speed with $v = \omega r$.
3. Direct measurement of speed by the use of:
 - Doppler effect of acoustic waves.
 - Doppler effect of electromagnetic waves with radar or light.
 - Combination of 2 cameras and correlation analysis (strictly speaking based on method 1, but only used for speeds).



3.3 Speed

Doppler Effect for Acoustic Waves

The Doppler effect describes the *relative* velocity between the object that emits the waves and the object that reflects the waves. The acoustic Doppler effect is typically used in the **ultrasonic range**.

For departing objects the frequency shift becomes ($c =$ speed of sound):

$$f_- = f \frac{c - v}{c + v} \quad \rightarrow \quad v = c \frac{f - f_-}{f + f_-}$$

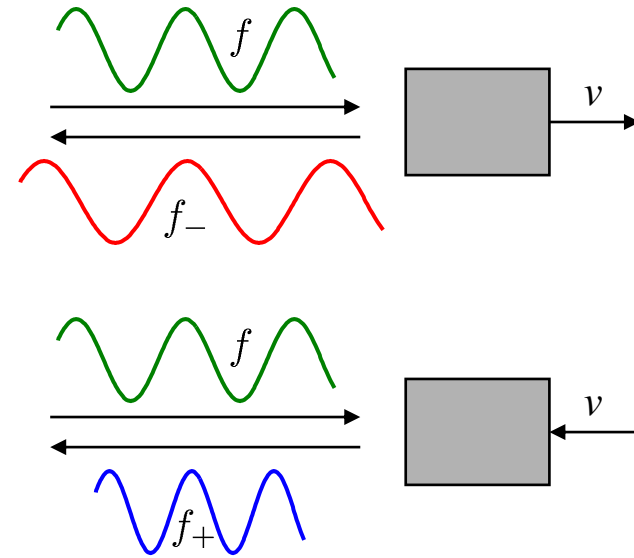
For approaching objects:

$$f_+ = f \frac{c + v}{c - v} \quad \rightarrow \quad v = c \frac{f_+ - f}{f_+ + f}$$

Doppler Effect for Electromagnetic Waves (Radar, Light)

Due to the theory of special relativity ($c =$ speed of light):

$$f_- = f \sqrt{\frac{c - v}{c + v}} \quad \rightarrow \quad v = c \frac{f^2 - f_-^2}{f^2 + f_-^2} \quad f_+ = f \sqrt{\frac{c + v}{c - v}} \quad \rightarrow \quad v = c \frac{f_+^2 - f^2}{f_+^2 + f^2}$$



3.3 Speed

Speed Measurement with 2 Cameras and Correlation Analysis

With well-structured surfaces like bulk on a conveyor belt or a street below a car, these patterns can be recorded with 2 distant cameras.

Comparing both camera signals with the help of correlation analysis, yields the time interval between both signals. With known camera distance d , the speed can be calculated from $v = d / \Delta t$.

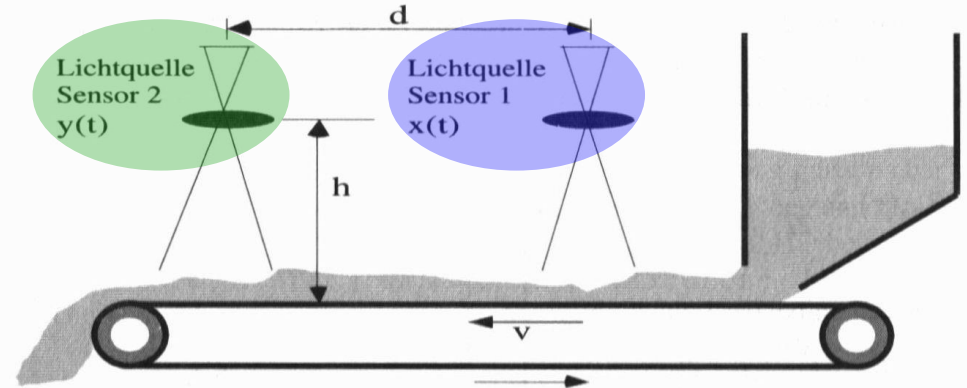
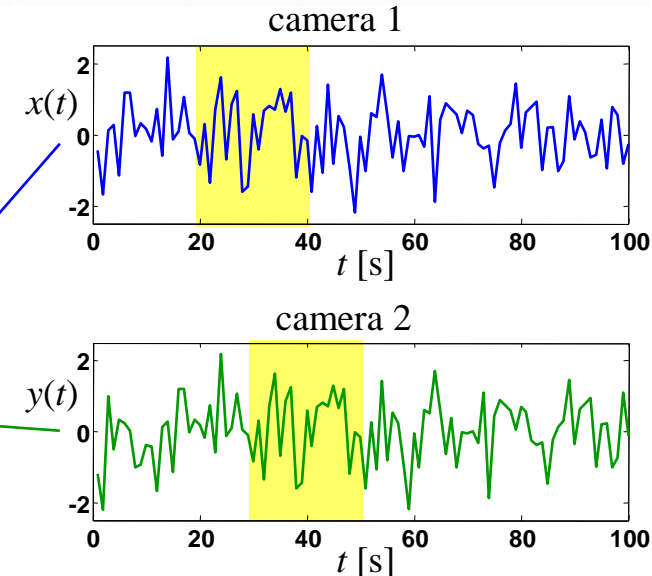
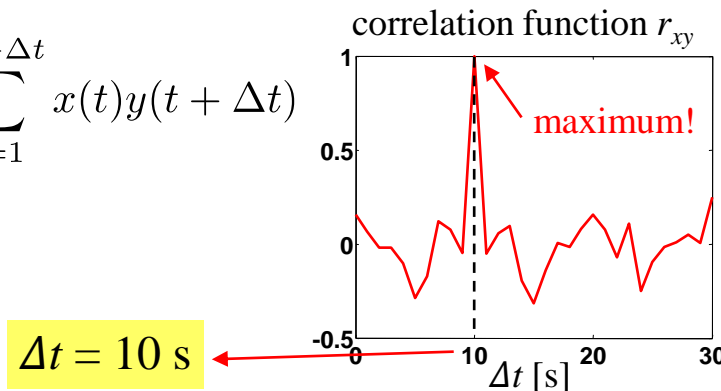


Abbildung 6.4. Modell eines Förderbandes zur berührungslosen Geschwindigkeitsmessung über das Laufzeitkorrelationsverfahren

$$r_{xy}(\Delta t) = \frac{1}{N} \sum_{t=1}^{N-\Delta t} x(t)y(t + \Delta t)$$

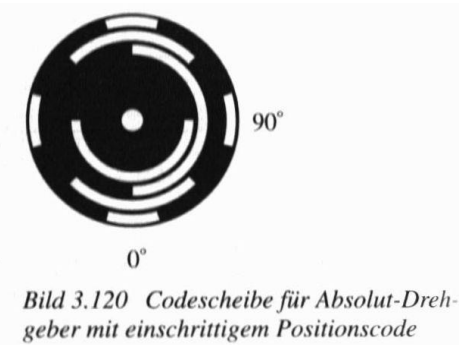
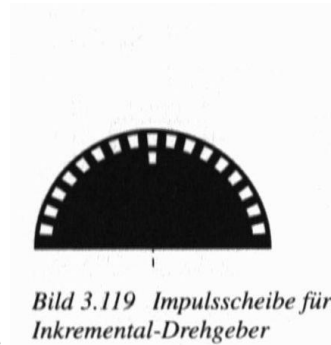


3.3 Speed

Speed Measurement: Optical Methods

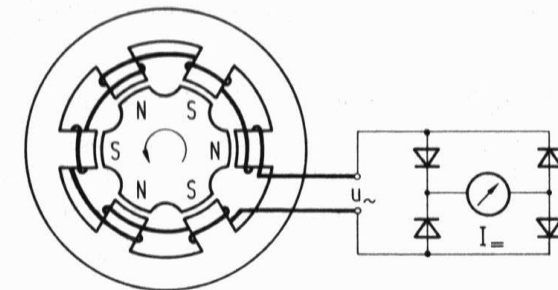
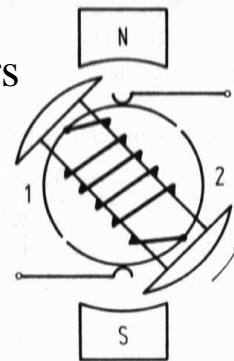
A disc as shown to the right can be mounted on an axle and illuminated by a light source. The reflected light can be accepted from a photo diode.

The discs can be marked incrementally or coded. Typically they have a marking of the initial point, that give an absolute reference for the incremental disc. The speed range that can be covered by this kind of approach is typically around 0 – 12000 min⁻¹.



Speed Measurement: Tachogenerators

A generator can be used for speed measurement. DC motors/generators yield a DC voltage proportional to the speed. AC motors/generators yield an AC voltage that has to be rectified before its amplitude is proportional to the speed. However the direction information (sign) is lost by this procedure.



3.3 Speed

Speed Measurement: Inductive Method

The inductivity of a coil depends on the relative magnetic permeability μ_r of the material through which the field line pass. Therefore, teeth and gaps can be detected, if the cog wheel is built of ferromagnetic material. In contrast to optical speed sensors, this approach is very robust against dirt and other environmental disturbances.

Thus, they are commonly used in automotive industry.

A double gap marks the initial point.

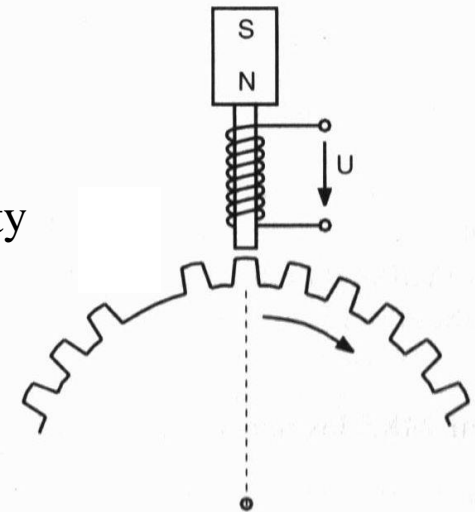


Bild 8.3.2: Induktiver Aufnehmer für Drehzahl und Drehwinkel, Ferromagnetische Zahnscheibe mit Zahnflücke A, Dauermagnet mit Weicheisenkern in einer Induktionsspule.

Speed Measurement: Magneto Resistive Method

It is similar to the inductive method. With a field plate the dependency of the electrical resistance of a resistor on the strength of a magnetic field is utilized (see Hall effect).

By using nonsymmetrical teeth-gap sequences, even the speed direction can be recovered.

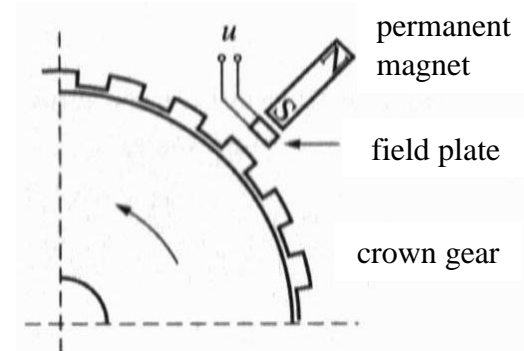


Bild 3.123 Feldplatte in Drehzahlmessanordnung

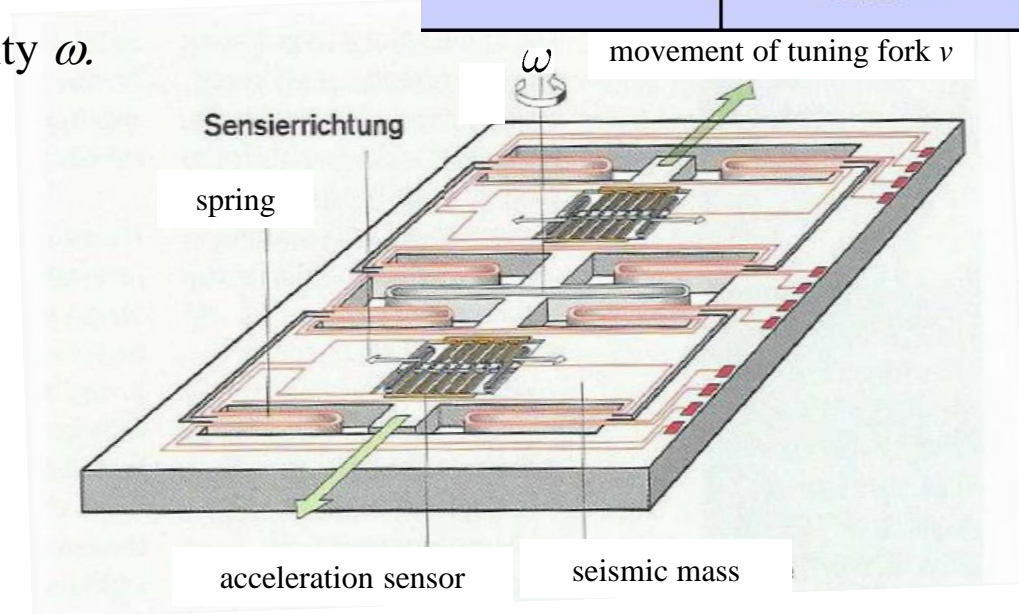
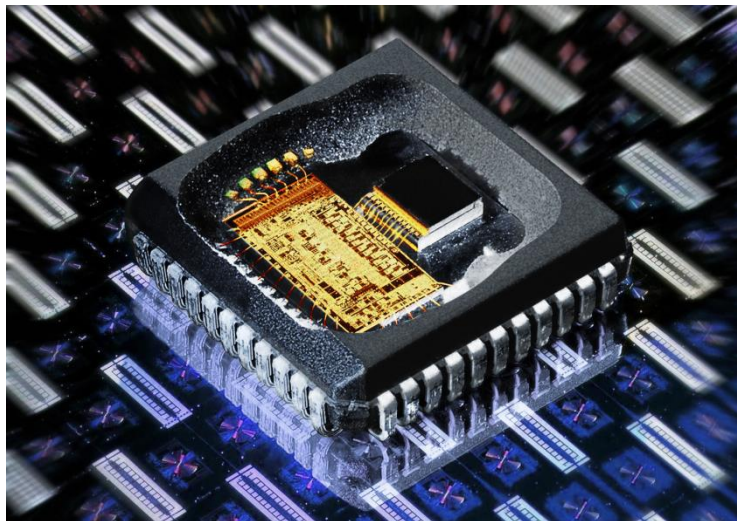
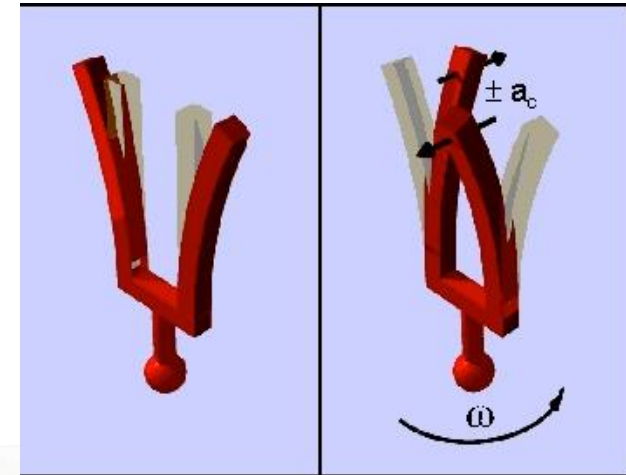
3.3 Speed

Yaw/Angular Velocity Measurement: Coriolis Principle (e.g., for ESP)

With micromechanics it is possible to realize an equivalent of a tuning fork that can be excited by permanent oscillations (in direction left/right). Due to these oscillations, the endings of the fork move with speed v . An angular velocity ω (from outside) with the oscillation orthogonal with respect to the movement creates the Coriolis force orthogonally

$$F_C \sim \omega \times v$$

which is proportional to the angular velocity ω .



3.4 Acceleration

Measurement Principles

For measurement of accelerations (translationally or rotationally) the following two approaches are important:

- The derivative of speed signals (attention: Derivatives enhance the noise!).
- Measurement of the force F or torque M at a body with mass m or a moment of inertia Θ and determination of acceleration via:

$$F = m a \quad \text{or} \quad M = \Theta \dot{\omega}$$

The first approach leads to the two previous sections. Therefore only the second approach is pursued here. Hereby the inertia of a **mechanical resonator** acts on a **seismic mass**. The equations of motion are those of a standard spring-damper-mass system:

$$m \ddot{r}(t) + d \dot{r}(t) + c r(t) = -m \ddot{x}(t) \quad \rightarrow \quad \ddot{r}(t) + \underbrace{\frac{d}{m}}_{2D\omega_0} \dot{r}(t) + \underbrace{\frac{c}{m}}_{\omega_0^2} r(t) = -\ddot{x}(t)$$

3.4 Acceleration

$$\omega_0 = \sqrt{\frac{c}{m}}$$

$$D = \frac{d}{2\sqrt{mc}}$$

Measurement of Acceleration with a Seismic Mass

With the usual notations for the damping D and the resonance frequency ω_0 , a seismic mass follows the equation:

$$\ddot{r}(t) + 2D\omega_0\dot{r}(t) + \omega_0^2 r(t) = -\ddot{x}(t) = -a(t)$$

If ω_0 is chosen to be big (via a stiff spring and a small mass) then the 3. term dominates the left part of this equation which yields approximately:

$$a(t) = -\omega_0^2 r(t) = -\frac{c}{m} r(t) = -\frac{F_c(t)}{m}$$

Acceleration measurement:

$$c \gg 1, m \ll 1, D \ll 1, \omega_0 \gg 1$$

Velocity measurement:

$$c \ll 1, m \ll 1, D \gg 1$$

Displacement measurement:

$$c \ll 1, m \gg 1, D \ll 1, \omega_0 \ll 1$$

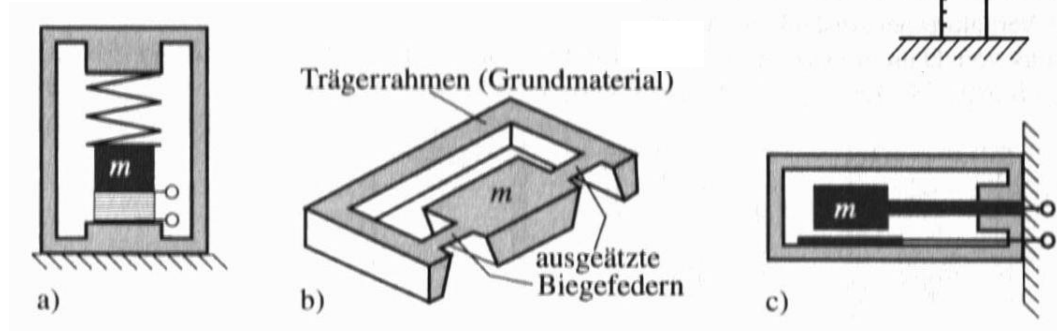
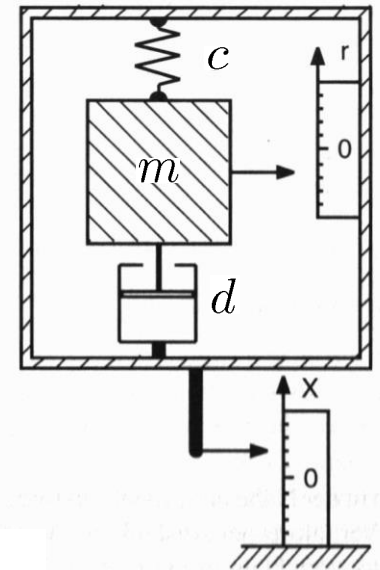


Bild 3.101 Beschleunigungssensoren
 a) piezoelektrischer Beschleunigungssensor, b) monolithischer Si-Beschleunigungssensor, c) kapazitiver Beschleunigungssensor

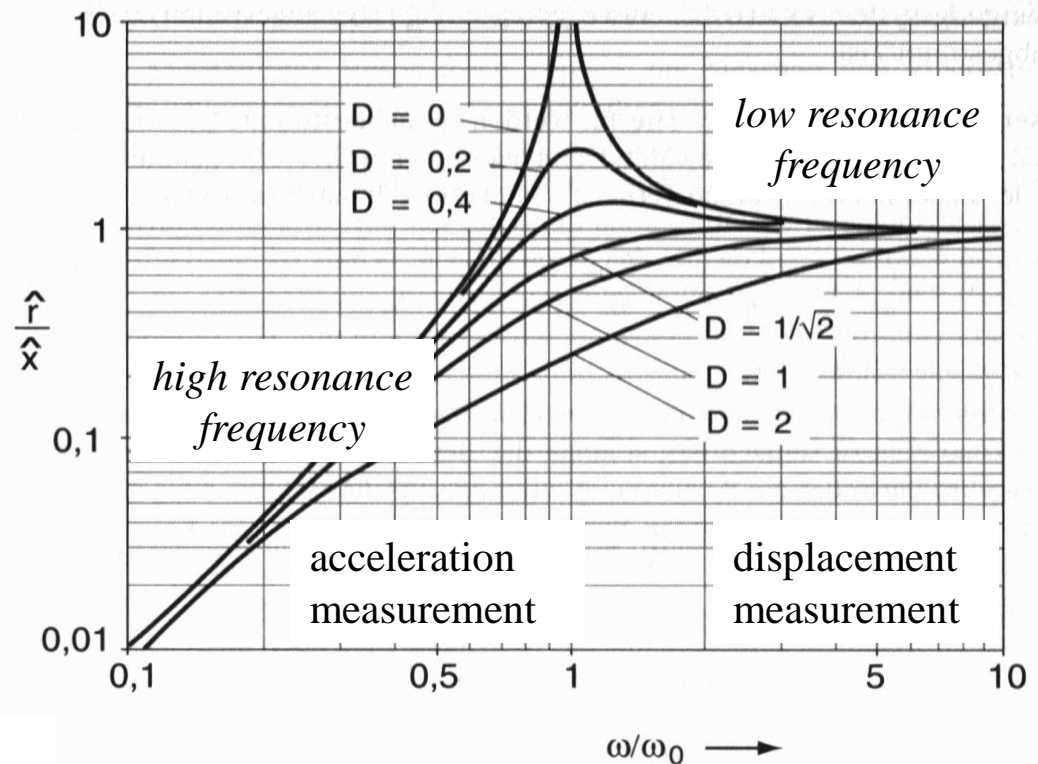
3.4 Acceleration

Frequency Response of a Seismic Mass

The frequency response is very dependent on the damping D around $\omega \approx \omega_0$. Therefore, either the low frequency range $\omega \ll \omega_0$ (tuned to high resonance frequency) or the high frequency range $\omega \gg \omega_0$ (tuned to low resonance frequency) is utilized. The high frequency range is used for measuring accelerations (slide before), the low frequency range is used for measuring displacement of oscillations.

The frequency response shown on the right is given by the relationship:

$$\frac{\hat{r}}{\hat{x}} = \frac{\omega^2}{\sqrt{(\omega_0^2 - \omega^2)^2 + (2D\omega_0\omega)^2}}$$



3.5 Force, Torque, Pressure, and Mass

Force Measurement

Measurement of force is typically achieved via measurement of displacements. The following principles are the most common ones:

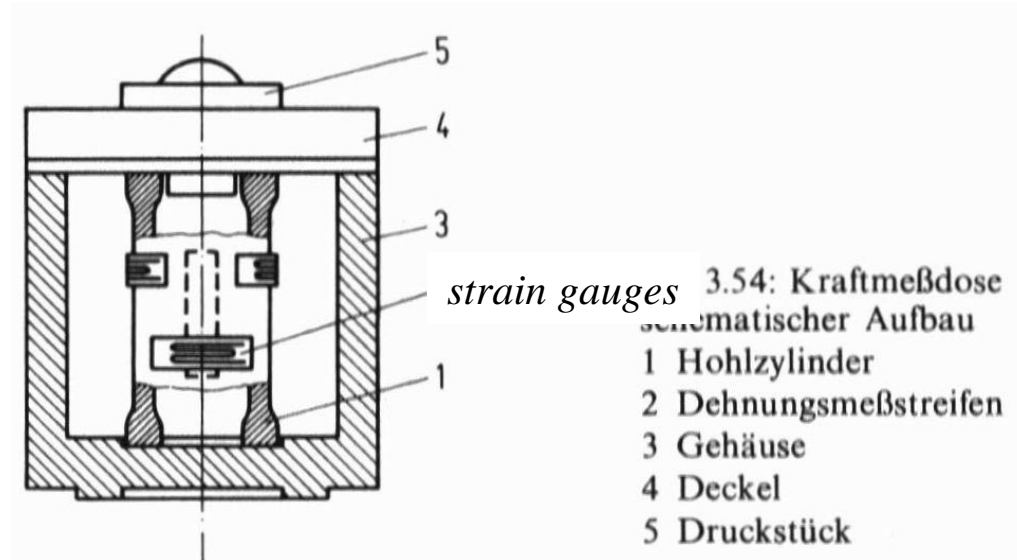
- *Strain Gauges:* In elastic deformation the force is proportional to the change of length which in turn results in a change of electric resistance (see Chapter 3.2).
- *Piezoelectric Effect:* A force or stress applied to a crystal generates an electric charge (“piezo” means “squeeze” or “press” in Greek). This principle is well-suited to measure highly dynamic (fast and/or oscillating) forces.
This effect is a reversible process, i.e., a mechanical force is generated if an electrical field is applied to the crystal. The force \rightarrow field effect can be used for sensors; the field \rightarrow force effect can be used to build actuators. The latter is e.g. used to generate ultrasound or for injection valve control of modern Diesel engines.
- *Magnetoelastic Effect:* The dependency of the magnetic properties of certain alloys with respect to an external force can be used to measure this force. The caused displacement is minimal.

3.5 Force, Torque, Pressure, and Mass

Load Cell

A load cell consists of an elastic, cylindrical body that is compressed or elongated by an external force. Strain gauges are glued on this body which measure the resulting stress.

- Range: 50 N ... 5 MN.
- Uncertainty ~ 0,05%.
- Applications, e.g. electromechanical scales (balances):
 - Commercial balances.
 - Horizontal containers.
 - Weighbridges.
 - Rail scales.
 - Belt scales.



3.5 Force, Torque, Pressure, and Mass

Piezoelectric Force Measurement

Certain types of crystal, e.g. SiO_2 , generate an electric field in response to mechanical force or stress. Dependent on the polarization direction, an electric charge gathers on the stressed areas (longitudinal effect, “Längseffekt”) or in the orthogonal direction (transverse effect, “Quereffekt”) or from a shear force (shear effect, “Schereffekt”)

The amount of electric charge Q is proportional to the causing force F :

$$Q = kF \quad \text{with } k = 2,3 \cdot 10^{-12} \text{ As/N}$$

In order to increase this tiny amount of charge, those crystals are typically build as stacks, i.e., many crystals are placed in series.

Shortly after their generation the charges try to balance each other. Thus, the effect is only temporarily. The electric charge has to be stored somehow after its generation.

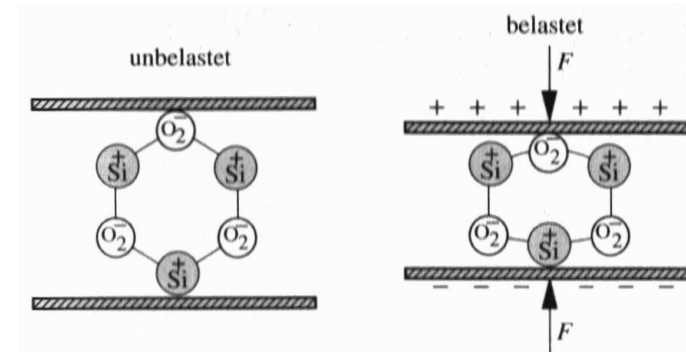


Bild 3.94 Prinzip des piezoelektrischen Effektes

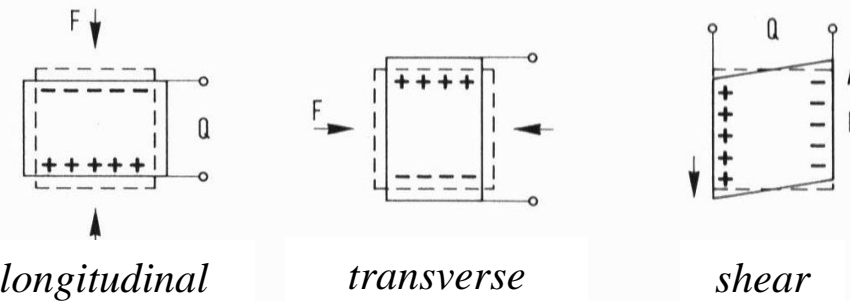


Bild 2.130: Wirkungsrichtungen des piezoelektrischen Effekts

3.5 Force, Torque, Pressure, and Mass

Piezoelectric Force Measurement: Dynamic Behavior

It is possible to describe the crystal as a current source with an internal resistance R_q and a capacity of C_q (see figure b) below). If a force appears suddenly (step input), then quickly a charge Q_0 is generated. With a time constant of $R_q C_q$ this charge exponentially fades away although the force continues to act. Via the internal resistance the capacitor discharges. If it is required to measure static forces, it is therefore necessary to feed the voltage to an integrator OpAmp circuit.

This transient behavior of the piezoelectric effect is a drawback for *stationary* measurements, but is well-suited for *fast dynamic* measurement because it possesses a high bandwidth.

The voltage generated as a result of the electric charge can be calculated as:

$$U_q = Q / C_q$$

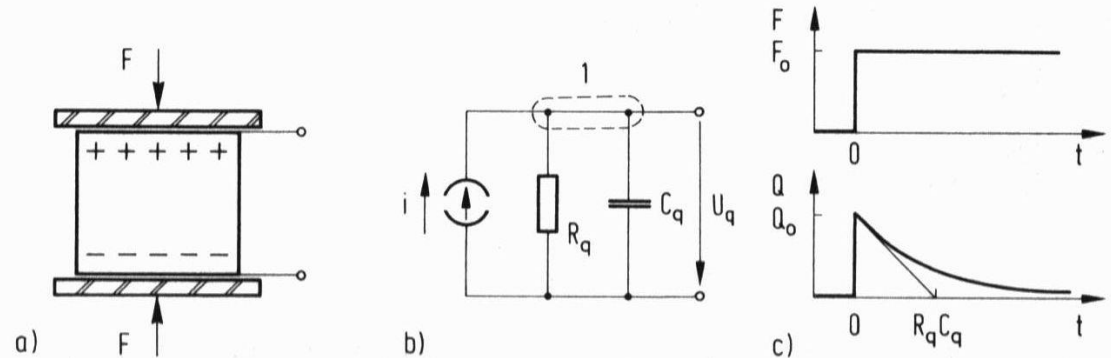


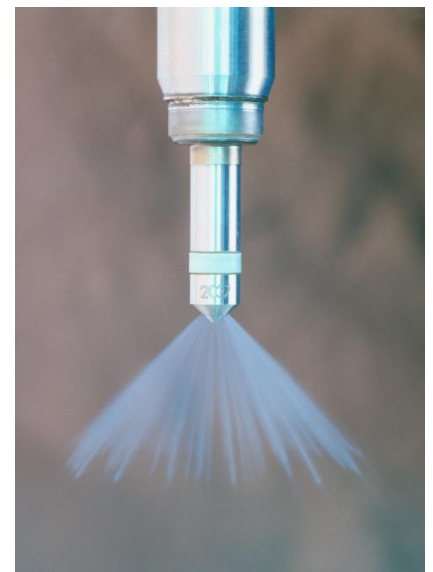
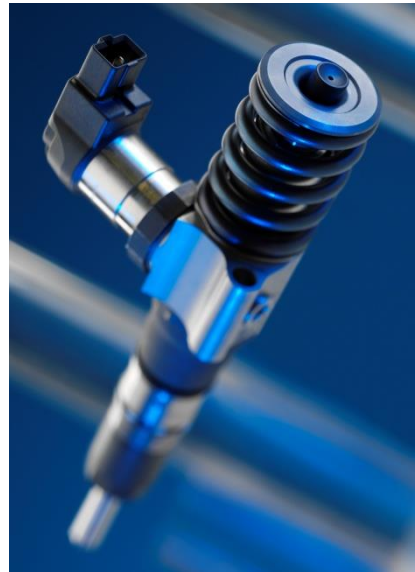
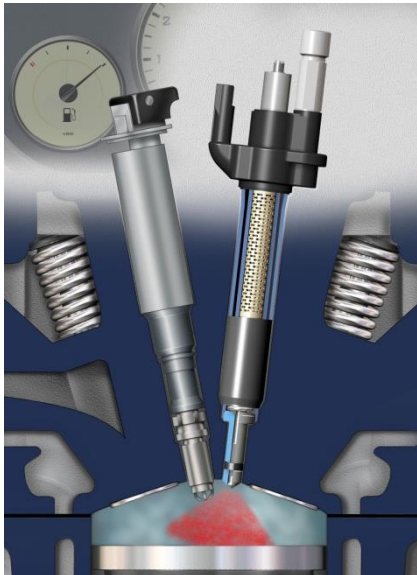
Bild 2.132: Aufbau (a), Ersatzschaltbild (b) und Sprungantwort (c) eines piezo-elektrischen Aufnehmers

3.5 Force, Torque, Pressure, and Mass

Reversed Piezoelectric Effect: Principle of Actuators

The piezoelectric effect offers new possibilities in actuation because of its high bandwidth. High injection pressures of 2000 bar spray the Diesel fuel very accurately and smoothly into the cylinder. This allows to partition the injection into several small injections to shape the combustion profile. Thereby, it is possible to make the explosion more efficient and at the same time optimize its other properties like decent acoustics.

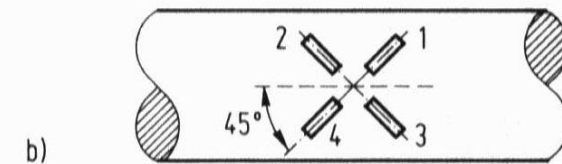
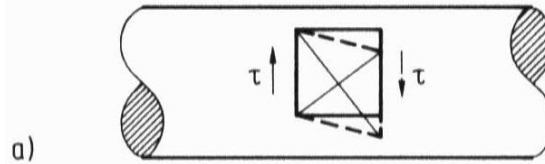
Piezoelectric Injector for Diesel Engines [Siemens VDO]



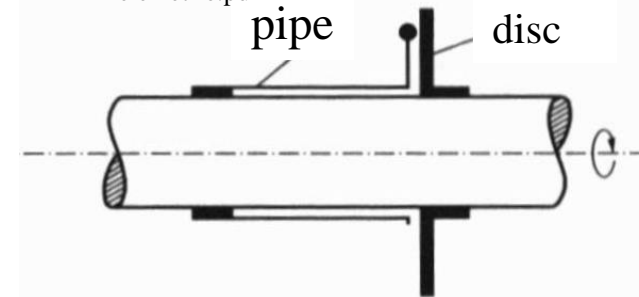
3.5 Force, Torque, Pressure, and Mass

Torque Measurement

For the measurement of torque the technique discussed for force measurement can be applied. Strain gauges can be applied to an axle to measure torsional stress. The change in resistance can be evaluated in a bridge circuit. A different possibility is to measure the torsional displacement between a flange-mounted disc and a pipe mounted in further distance. The displacement measurement can be performed inductively or capacitively.



Source: http://www.telemetrie-world.de/fachartikel/7._Drehmomentmessung_mit_Telemetrie.pdf



Signal Processing

One difficulty with measuring torques is the transmission of the measurement signals outside of the rotating axle to a fixed system around. This can be solved via slip rings. A more robust technique is via a transformer. Modern systems are based on infrared or radio systems.

3.5 Force, Torque, Pressure, and Mass

Pressure Measurement

Pressure measurement is typically based on the measurement of force. The force acts on a defined area, normally a membrane. Actually, **pressure differences** are measured, i.e., the deviation between a pressure and some reference pressure:

- If the reference is equal to the atmosphere pressure, measurement value is called *excess (over) pressure* or *under pressure*. Example: tire of a car.
- Sometimes the reference pressure is zero (vacuum). Then, the measurement value is called *absolute pressure*.

The difference pressure lifts or lowers the membrane.

By this, the pressure difference is converted into a displacement. This can either be displayed directly (see figure) or it can be further converted with the principles discussed in Chapter 3.2 (resistive, inductive, capacitive) into an electric signal.

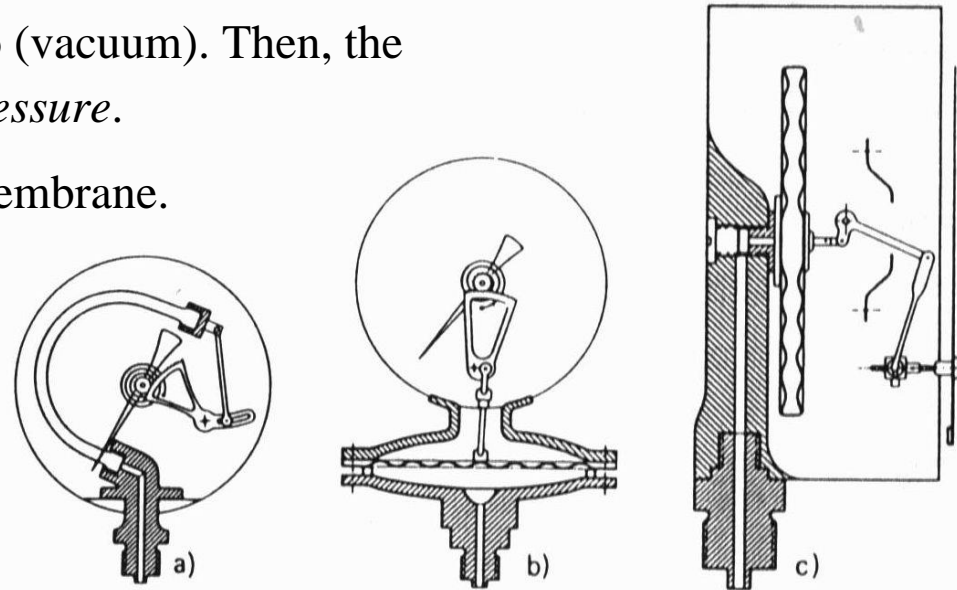


Bild 7.2: Federdruckmesser [0.17]

a) Rohrfedermeßwerk, b) Plattenfedermeßwerk, c) Kapselfedermeßwerk

3.5 Force, Torque, Pressure, and Mass

Mass Measurement

Masses m can be determined via their proportional weight force F . The proportional constant is the acceleration due to gravity g :

$$F = m g$$

A counter force is created that balances the weight force. If the counter force is also generated by masses, the acceleration g cancels out. If, on the other hand, the counter force is generated by springs, magnetic or electric fields, or similar, the scale has to be calibrated dependent on the location because g is influenced by the location on earth (not a perfect, homogeneous sphere!), even so in higher heights.

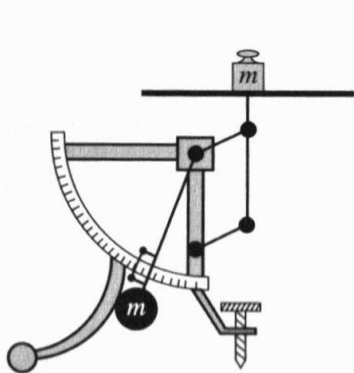


Bild 3.102 Neigungswaage mit Parallelogrammführung

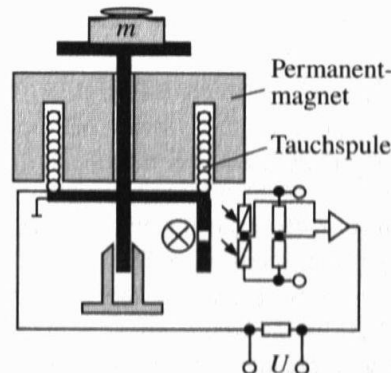


Bild 3.103 Elektrodynamische Kraftkompensationswägezelle

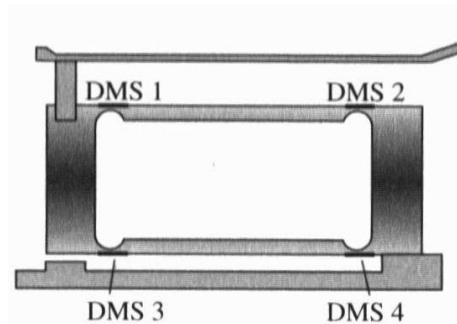


Bild 3.104 DMS-Wägezelle

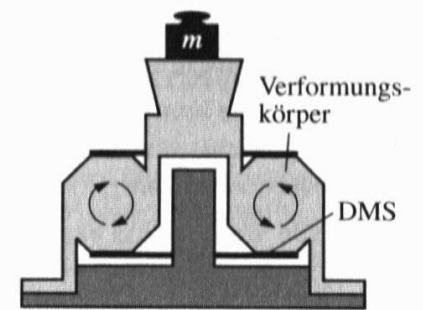


Bild 3.105 Ringtorsionswägezelle

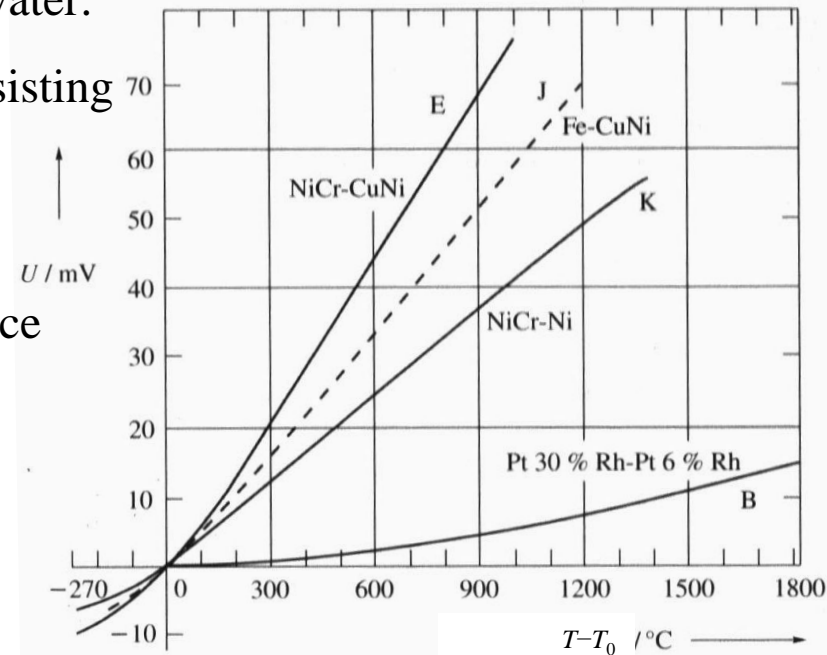
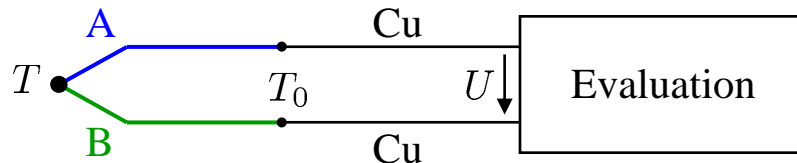
3.6 Temperature

A) Thermocouples

- 2 wires consisting different materials (usually metal alloys) A and B that produce a voltage proportional to a temperature difference between either end of the pair of conductors. Thermocouples are a widely used type of temperature sensor for measurement and control of temperature T .
- At the ends of the wires a circuit is connected. These connection have temperature T_0 . For reference, these connection can be put into ice water.
- The voltage generated by the thermocouple consisting of wire A and B is given by:

$$U = k_{AB}(T - T_0)$$

The proportionality constant k_{AB} and the reference temperature T_0 have to be known a priori!



3.6 Temperature

B) Resistance Thermometer (PTC, Positive Temperature Coefficient) – Metal

The ohmic resistance of a metal wire depends on the temperature T approximately as follows:

$$R = R_0 [1 + \alpha(T - T_0) + \beta(T - T_0)^2]$$

The coefficients α and β are material dependent, R_0 denotes the resistance at a reference temperature T_0 (as well material dependent). Because β is much smaller than α , the quadratic term can be neglected – at least for small and moderate temperature changes.

Typically the reference temperature is chosen as $T_0 = 0^\circ\text{C}$:

$$R = R_0(1 + \alpha\vartheta)$$

where ϑ denotes the measured temperature in $^\circ\text{C}$.

The temperature coefficient α describes the relative change of the resistance with the temperature:

$$\alpha = \frac{1}{R_0} \frac{dR}{d\vartheta}$$

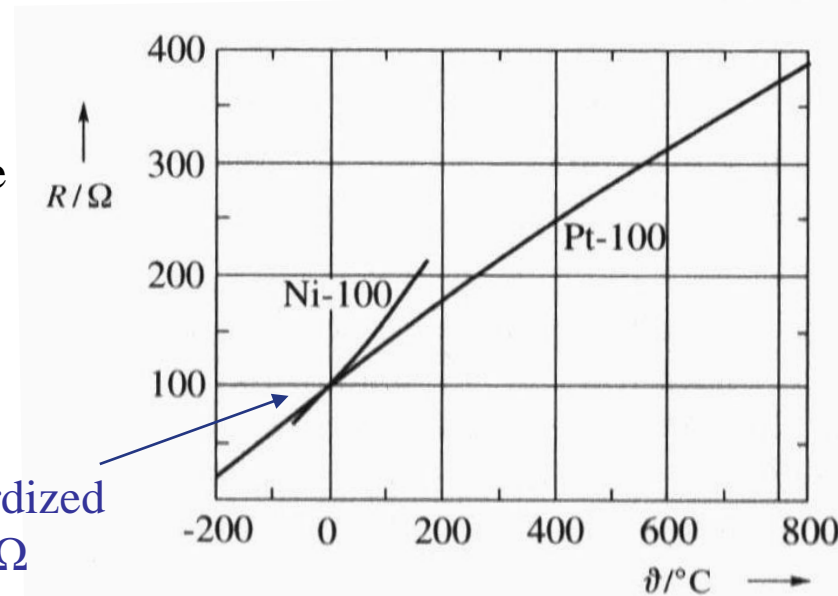
$\alpha > 0$ for PTC

$\alpha < 0$ for NTC

From the measured value we obtain:

$$\vartheta = \frac{R - R_0}{\alpha R_0}$$

standardized
at 100 Ω



3.6 Temperature

Signal Processing for Resistance Thermometer

- Processed with bridge circuits
- Direct voltage measurement possible if current is forced by a constant current source.
- CAUTION: The current through the measurement resistor must be small enough that the power loss (dissipation) is negligible. Otherwise, the heat can distort the temperature measurement.
- For the Pt-100 resistance thermometer two accuracy classes are standardized:

Class A: $\pm (0.15 + 0.002|t|)^\circ\text{C}$

Class B: $\pm (0.30 + 0.005|t|)^\circ\text{C}$

PTC Resistance Thermometer (Metal)

more accurate

up to max. 850°C

slower (large time constant)

no point-wise measurement

Thermocouples

less accurate

even for higher temperatures

faster (small time constant)

point-wise measurement

3.6 Temperature

C) Resistance Thermometer (NTC, Negative Temperature Coefficient) – Semiconductor

In semiconductors the number of free electrons grows with the temperature significantly. The intrinsic conductivity increases, the resistance decreases. With the material constant b and the resistance R_0 at temperature T_0 the following relationship holds:

$$R = R_0 e^{b\left(\frac{1}{T} - \frac{1}{T_0}\right)}$$

With the constant $K_0 = R_0 e^{-b/T_0}$ this yields:

$$R = K_0 e^{b/T}$$

Thus, the sensitivity becomes:

$$\frac{dR}{dT} = -\frac{b}{T^2} R$$

The temperature coefficient is:

$$\alpha = \frac{1}{R} \frac{dR}{dT} = -\frac{b}{T^2}$$

$\alpha < 0$: negative temperature coefficient!

Applications: car, appliances.

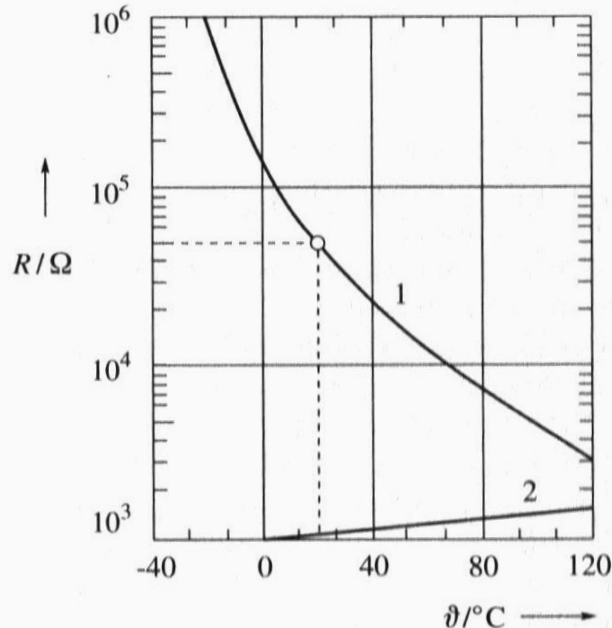


Bild 3.7 Widerstand eines Halbleiters 1 (Nennwiderstand $50 \text{ k}\Omega$) und eines Platinwiderstandsthermometers 2 (Nennwiderstand 1000Ω) in Abhängigkeit von der Temperatur ϑ .

3.6 Temperature

PTC Resistance Thermometer

α is positive and small

α is almost constant
(~ linear characteristics)

resistance is small; Calibration
of the wires is necessary

extensive in space
no point-wise measurement

slow

high accuracy

high long-term-stability

NTC Resistance Thermometer

α is negative and has large absolute value

α is strongly temperature dependent
(strongly nonlinear characteristics)

resistance is so large that no calibration of the
wires is necessary

manufactured in tiny sizes
point-wise measurement possible

fast

medium accuracy

little long-term-stability

3.6 Temperature

D) PTC Resistance Thermometer – Semiconductor

PTC thermometers consisting of semiconducting and *ferromagnetic* material and not of metal. In the low temperature range it has a small resistance with negative temperature coefficient. Above a material dependent critical temperature T_A , the *Curie temperature*, the unified orientation dissolves. This leads to an exponential increase of the resistance in a small temperature band ($T_N - T_E$). In this range the approximate relationship holds:

$$R = R_0 e^{b(T-T_0)}$$

Sensitivity and temperature coefficients are:

$$\frac{dR}{dT} = bR \quad \alpha = \frac{1}{R} \frac{dR}{dT} = b$$

The temperature coefficient is 5 x higher as with NTC. Drawbacks are the extremely dispersive material properties and volatile stability. Thus, only a low precision is possible.

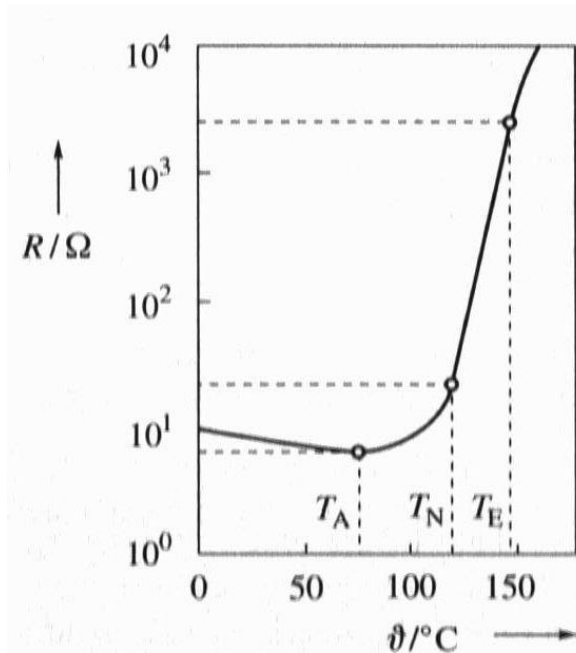


Bild 3.8
Widerstand eines Kaltleiters in
Abhängigkeit von der Temperatur
 T_A Temperatur, bei der der Temperaturkoeffizient positiv wird
 T_N Nenntemperatur, Beginn des steilen Widerstandsanstiegs
 T_E Endtemperatur, Ende des steilen Widerstandsanstiegs

3.6 Temperature

E) Miscellaneous

Besides the discussed temperature measurement approaches, there exist many alternatives that also work according to a **contact principle**. The following things have to be considered:

- First, the sensors measure their *own* temperature.
- The instrumentation engineer has to ensure that the sensor adopts the temperature of the medium which shall be measured.
- The sensors affect the medium which shall be measured. Thus, the sensors can introduce or draw heat from the medium. This means, the measurement is *interacting*!

Alternatively, there exist sensors which work according to the **radiation principle**.

Especially for high temperatures this is a common approach. The sensors do not have any contact to the measured medium. They evaluate its radiation, e.g.:

- Thermopile: Series connection of thermocouples that are sensitive to heat radiation.
- Pyroelectric temperature sensor (see picture): Based on the change of polarization of certain dielectric materials whose charge density on their surface is measured.
- Radiation pyrometer: Based on the measurement of the radiation power density $\sim \sigma T^4$.



3.7 Flow

Volume Flow and Mass Flow

The volume flow is defined as:

$$Q_V = \dot{V} = \frac{dV}{dt}$$

The mass flow is defined as:

$$Q_m = \dot{m} = \frac{dm}{dt}$$

Both quantities are related via the **density** ρ of the fluid:

$$Q_m = \rho Q_V$$

If the density is known theoretically (commonly the case for incompressible fluids) or can be measured, then it is possible to convert volume flow in mass flow and vice versa.

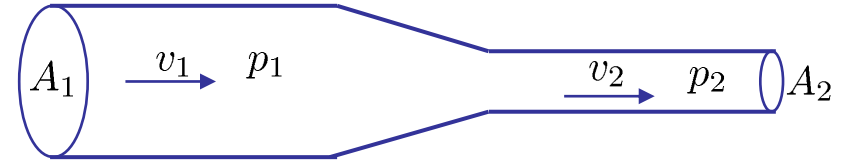
Mass flow as a quantity has the advantage that it is constant in closed systems, while volume flow of compressible fluids depends on their density and thus also on pressure and temperature. On the other hand, the measurement of volume flows is cheaper, simpler, and more widely used.

3.7 Flow

A) Differential Pressure Method

The flow measurement is indirectly performed by measuring pressures. A narrowing pipe increases the flow velocity due to a decreasing cross-section. Following Bernoulli, the flow velocity increases accordingly:

$$p_1 + \frac{\rho}{2}v_1^2 = p_2 + \frac{\rho}{2}v_2^2$$



The pressure drop therefore becomes:

$$\Delta p = p_1 - p_2 = \frac{\rho}{2}(v_2^2 - v_1^2) = \frac{\rho}{2}v_1^2 \left(\frac{v_2^2}{v_1^2} - 1 \right) = \frac{\rho}{2}v_1^2 \left(\frac{A_1^2}{A_2^2} - 1 \right) = \frac{\rho}{2}Q_V^2 \left(\frac{1}{A_2^2} - \frac{1}{A_1^2} \right)$$

The volume flow can be calculated from the square root of the difference pressure:

$$Q_V = k \sqrt{\frac{\Delta p}{\rho}}$$

Dependent on the kind of narrowing (orifice, nozzle, venturi), an additional pressure drop of 9% – 60% has to be considered due to turbulence (energy loss). That has to be taken into account with a proportionality factor k .

With a Pitot tube well-known through *Prandtl* such difference pressures can be measured.

3.7 Flow

Different Kinds of Narrowing

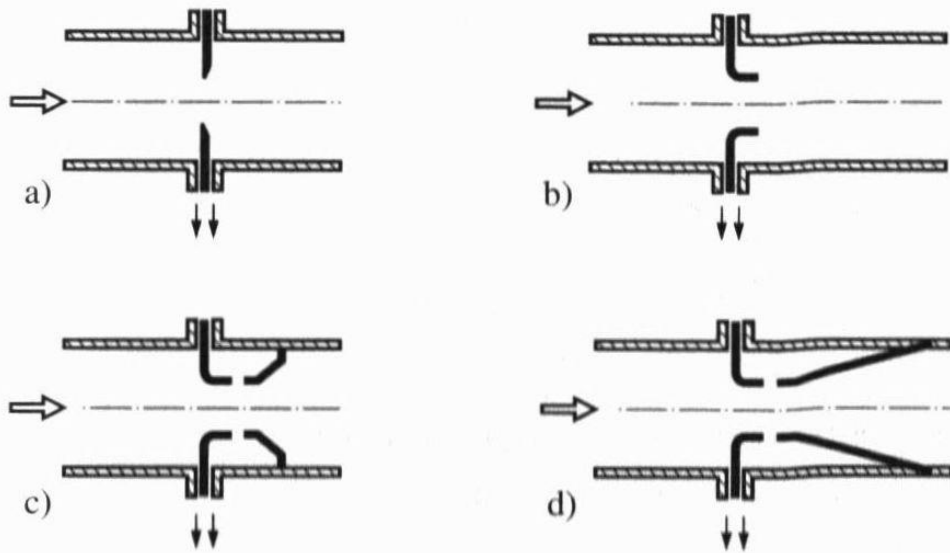
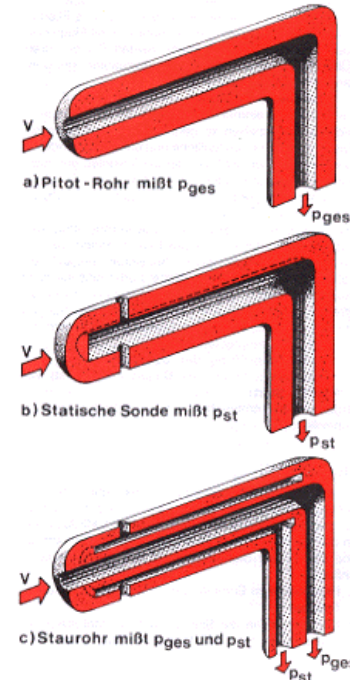


Bild 3.33 Übersicht über die in der Bundesrepublik Deutschland genormten Drosselgeräte

a) Blende, b) Düse, c) Venturidüse, kurz, d) Venturidüse, lang

Pitot Tube



Mounted on Airbus A380

Source:

http://en.wikipedia.org/wiki/Pitot_tube

Properties of Flow Measurement with the Differential Pressure Method

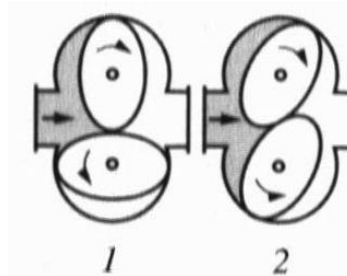
- Robust, simple and resistant (endure hard environmental conditions).
- No moving parts. Limited measurement range due to quadratic pressure dependency.
- Most commonly used and standardized approach.

3.7 Flow

B) Volume Counter Measurement

Volume counter with metering chamber

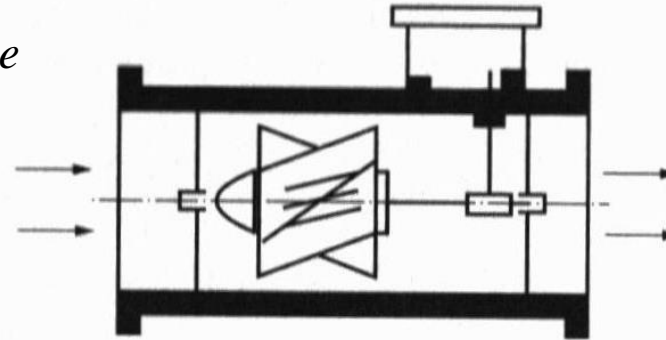
Transports fluid in chambers and thus counts its amount and therefore flow.



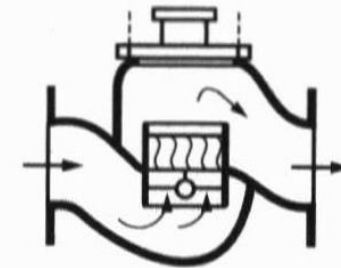
oval
gear meter

Volume counter with hydrometric vane

A wheel with vanes (or blades) is turned by the fluid flow. Actually, the flow velocity is measured but a multiplication with the cross-section yields the volume flow.



meter with axial wings



meter with
vertical wings

Modern method: The energy for turning the wheel is not taken from the fluid flow. Rather it is supplied from outside. The pressure drop is feedback controlled to zero.

Properties: Large measurement range, independent of viscosity, sensitive with respect to contamination of the fluid because of moving parts.

3.7 Flow

C) Float Measurement

A floating body with large cross-section A_K is placed inside the fluid flow. It is lifted to a height where the force caused by the flow balances exactly the force caused by its weight:

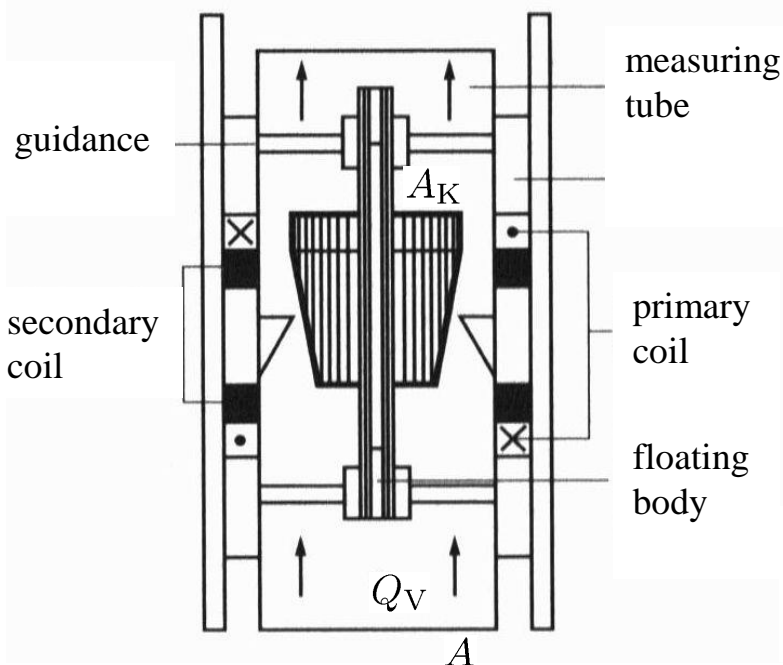
$$F = c_W A_K \frac{\rho v^2}{2}$$

Here v is the flow velocity inside the ring-formed opening $A - A_K$ between the tube and the floating body. According to the balance of continuity the flow is proportional to the square

of the height h (\sim diameter):

$$Q_V = v(A - A_K) \sim h^2$$

In order to not only display the height but transmit the signal to the outside world, it is reasonable to convert it into an electrical signal. An effective way to realize that, is to use a *ferromagnetic floating body* as coupling between two coils works like in a transformer.



3.7 Flow

D) Magnetic Inductive Measurement

For all conducting fluids, the flow can be measured based on **Faraday's law** in a *contact-free* manner. Orthogonally to the flow a magnetic field with density B is generated. Thus, in a moving conductor (as such the fluid can be interpreted) orthogonal to the field, a voltage is induced. This voltage is generated orthogonal to the magnetic field and to the flow direction and amounts to:

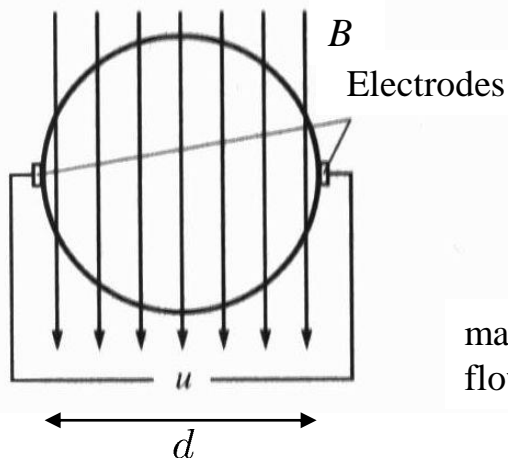
$$u = -\frac{d\Phi}{dt} = -\frac{B dA}{dt} = -Bvd$$

$$\rightarrow v = -\frac{u}{Bd}$$

The flow can be calculated by multiplication of velocity v with cross-section A .

Properties:

- Very good linearity, big measurement range.
- Independent of density, viscosity, pressure, temperature.
- Also suitable for corrosive fluids and fluids that contain solids.
- No internal constructions necessary.
- Minimum conductivity is necessary.



3.7 Flow

D) Remark: Reversal of the Sensor Principle as Actuator

In the movie “The Hunt for Red October” [Sean Connery, Alec Baldwin] a new and silent drive system plays an important role. This is no science fiction! The movie refers to a so-called **magneto-hydrodynamic drive**, which is constructed without any moving parts. However, it works only in salt water because it is based on Faraday’s law and requires a conducting medium.

The magnetic field is generated by a superconductive generator. Orthogonal to the field an electric current is send through the water. Together with the current the magnetic field results in a force on the water that is accelerated orthogonal to field and current. This causes the water to shoot outside the ship without any propeller!

The picture shows the first ship of this type with superconductive magneto-hydrodynamic drive [Mitsubishi, 1998].



3.7 Flow

E) Coriolis-based Measurement

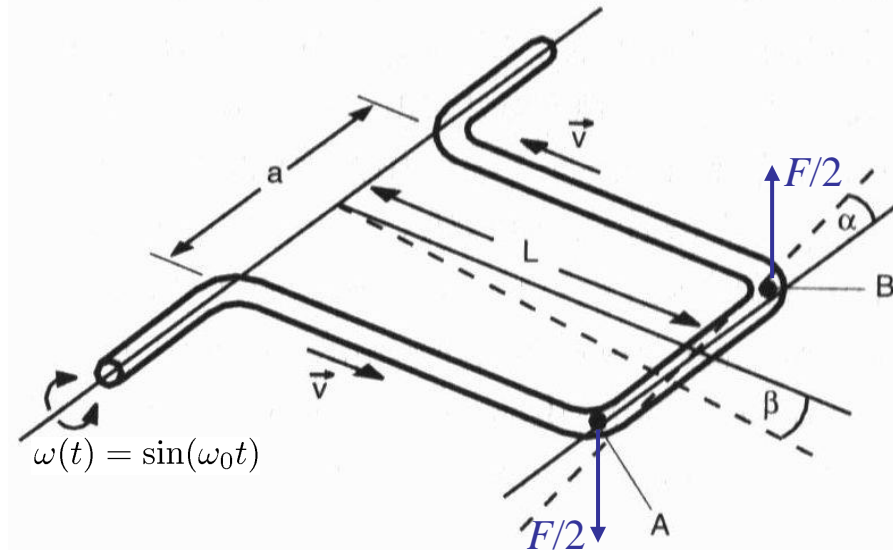
A body that rotates with the angular velocity ω that moves with a speed of v orthogonal to the axis of rotation experiences a **Coriolis force** orthogonal to this axis and the speed direction of

$$F = m v \omega(t) = 2\rho A L v \omega(t)$$

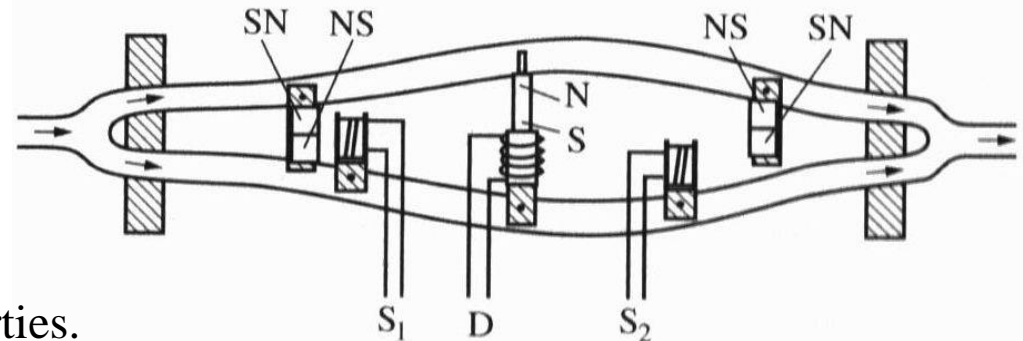
This force bends the U-pipe to an angle α .
With a sin-type excitation, a phase lag exists between point A and B. This phase lag is proportional to the mass flow.

Properties:

- No constructions inside necessary.
- Robust with respect to all fluid properties.
- Suitable for liquids and gases.



Coriolis flow meter in U-pipe configuration



Coriolis flow meter in straight configuration

3.7 Flow

F) Hot Wire Measurement

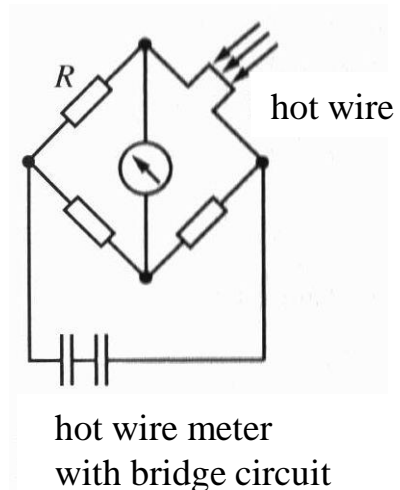
A **hot wire** or a **hot foil** are heated by an electric current via a constant voltage or current source.

The flow that flows around the wire or foil decreases its temperature. This temperature drop causes a change in the electric resistance that is measured (typically by a bridge circuit).

Here the *mass* flow is directly measured because the cooling is proportional to the temperature difference between wire/foil and fluid and proportional to the number of molecules that impact. Corrections with respect to density or pressure changes are superfluous.

Properties:

- Especially well suited for low velocities.
- Sensitive with respect to dirt and burn-out.
- Because of aging frequent calibrations are necessary.



3.7 Flow

G) Miscellaneous

Beside the approaches discussed more detailed above, many alternatives are worth at least to mention briefly:

- *Vortex method*: The **frequency** of a *vortex shedding* (Karman vortex street) behind a body where a fluid flows around is proportional to the velocity of the fluid.
- *Transit time method*: Within a short interval a short injection is carried out into a pipe. The velocity of the fluid is determined by measuring the time interval and the distances between 2 points of the solution clouds.
- *Laser Doppler flow measurement*: The frequency shift of laser light that is scattered on particles inside the fluid yields a point-wise velocity measurement.
- *Ultrasound flow measurement*: a) Transition time method: A sound wave runs inside the medium, i.e., the speeds add up (wave + medium). This speed minus the wave speed in the resting medium yields the medium (fluid) speed. b) Doppler method: The frequency shift of reflecting sound waves is used. It is dependent on the speed of the medium.

3.8 Miscellaneous

Many other quantities can be measured which are not discussed here. The most prominent certainly are:

- *Density*: Weighting methods determine the mass and the volume via suppression. The density can be calculated by division. For solid materials, the uplift in liquids or gases can be used. For liquids, the hydrostatic pressure difference can be used. For gases, Bunsen's law describing the relationship between volume flow and density for exhausting gas through a hole can be used.
- *Concentration*: A huge number of special methods exist dependent on the kind aggregate state of the studied material. Frequently these methods are based on *absorption*, *emission*, or *reflection* of **radiation**. For **Chromatography** different delays of different components inside an intermixture are used. For **Spectroscopy** different properties of atoms or molecules (mass, spin, ...) are used for their division. The **Refractometry** uses changes in the optical refractive index and **Polarimetry** uses changes in the polarization level.

3.8 Miscellaneous

- *Concentration*: Changes in the thermal conductivity can be utilized. Of particular importance are the measurement of:
 1. *Humidity*: Many approaches exist based on changes in the evaporation rate, conductivity, permittivity inside a capacitor ϵ_r .
 2. *pH Value*: Between electrodes within different liquids a voltage occurs, an effect known from a galvanic cell (battery). A diaphragm enables the exchange of ions but prevents the mixing of the liquids.
 3. *Particle*: E.g, the particle-induced *cloudiness* or *scattering* of light is measured.
- *Light*: **Photoresistors** are resistors whose resistance depends on the amount of light they measure. **Photodiodes** and **CCDs** (*charge coupled devices*) convert light (point-, row-, or matrix-wise) into electrical current. The sensitivity depends strongly on the wave length of the light.
- *Sound*: A **dynamic microphone** works according to *Faraday's law*. This means a membrane is coupled with a wire that moves through an magnetic field. The induced voltage in this wire is proportional to membrane movement. However, **Capacitor microphones** are based on a capacity change dependent on the membrane movement.